Teacher's Reference Handbook

PHYSICS

AN ROINN
OIDEACHAIS
AGUS EOLAÍOCHTA

DEPARTMENT OF
EDUCATION
AND SCIENCE

Department of Education and Science: *Intervention Projects in Physics and Chemistry*
With assistance from the European Social Fund

CONTENTS

ACKNOWLEDGEMENTS

This handbook has been produced as part of the Department of Education and Science's scheme of ***Intervention Projects in Physics and Chemistry***. These projects were commissioned by the Department's Equality Committee, chaired by Mr Denis Healy, Assistant Secretary General. The projects have been funded by the Department with assistance from the European Social Fund.

Authors of Modules

Dr Anne Breslin, University College, Dublin

Mr Noel Brett, Coláiste Chríost Rí, Cork

Mr Peter Burke, Ballyfermot Senior College, Dublin

Ms Sheelagh Clowry, Presentation/De La Salle College, Bagenalstown, Co. Carlow

Mr Dan Condren, St Fergal's Vocational School, Templemore, Co. Tipperary

Ms Frances Feeley, formerly of Dominican College, Taylor's Hill, Galway

Ms Catherine McDonagh, St Michael's Holy Faith Secondary School, Dublin

Mr Pat O'Leary, Nagle Community College, Cork and formerly of the Vocational School, Athlone, Co. Westmeath

The section on 'Gender and Science' was written by Dr Sheelagh Drudy, Education Department, NUI Maynooth, in association with Ms Marion Palmer, Dun Laoghaire Institute of Art, Design and Technology and formerly of Mount Temple School, Dublin. The section on 'Some Irish Contributions to Physics, Mathematics and Technology' was written by Dr Charles Mollan, Samton Ltd, Newtownpark Avenue, Blackrock, Co. Dublin.

Editorial Board

Mr Séamas Ó Míocháin, Gormanston College, Co. Meath

Mr George Porter, Beneavin College, Dublin, on secondment to the Department of Education and Science

Mr Seán Ó Donnabháin, Senior Inspector, Department of Education and Science

The Department of Education and Science is grateful to the following for the provision of additional material and for their expert advice and support at various stages of the preparation of this handbook.

Mr Brendan Campbell, formerly of Clongowes Wood College, Co. Kildare

Mr Owen Clancy, St Muredach's College, Ballina, Co. Mayo

Dr Ian Elliott, Dublin Institute of Advanced Studies, Dunsink Observatory, Dublin

Dr Eric Finch, Trinity College, Dublin

Mr Randal Henly, formerly of Mount Temple School, Dublin

Dr Des Hickey, Dublin Institute of Technology, Kevin Street, Dublin

Mr Larry Keating, Coláiste Éinde, Galway

Dr Paula Kilfeather, St Patrick's College, Drumcondra, Dublin

Dr Joe Lennon, University College, Cork

Dr James Lunney, Trinity College, Dublin

Dr Eithne McCabe, Trinity College, Dublin

Dr Colette McDonagh, Dublin City University, Dublin

Dr Sara McMurry, Trinity College, Dublin

Sr Elizabeth McNamee, Mercy Convent, Longford

Dr Peter Mitchell, University College, Dublin

Mr Enda O'Flaherty, Dublin Institute of Technology, Kevin Street, Dublin

Mr Art O'Hare, Dublin Institute of Technology, Kevin Street, Dublin

Mr Diarmuid O'Leary, Scoil Pól, Kilfinane, Co. Limerick

Dr Colm T. O'Sullivan, University College, Cork

Mr Tim Regan, St Brendan's College, Killarney, Co. Kerry

Mr Bob Richardson, The Grammar School, Sligo

Dr Adrian Sommerfield, formerly of St Columba's College, Dublin

The Department of Education and Science also wishes to acknowledge the provision of material from the following sources.

Fisher Scientific UK Ltd; CERN; The Irish Times, Dr Brendan McWilliams; The Walton Family.

INTRODUCTION

This handbook has been produced as part of the Department of Education and Science's Equality of Opportunity Programme. The project developed out of the Department's scheme of ***Intervention Projects in Physics and Chemistry*** which was implemented from 1985 with a view to increasing the participation of girls in the study of the physical sciences. It is hoped that the material contained in this book will assist teachers in presenting physics in a manner which will give due cognisance to gender differences in relation to interests and attitudes. It is also hoped that the material will help teachers in their continuing quest to develop new approaches to their teaching which will make physics more interesting and exciting for all their students.

Physics can be taught purely as a set of laws and definitions; as a series of equations which are derived from first principles and which are applied in solving mathematical problems. The minimum number of experiments can be done to verify laws and to investigate relationships.

Alternatively *physics can be made interesting, relevant and challenging.* Every teacher knows the sense of satisfaction to be gained from giving a class that is well prepared and well presented and that holds the students' attention from start to finish.

The problem is that good preparation involves more than knowing the course backwards. If you can prove the equation for the period of a satellite in circular motion, while blindfolded, standing on one leg, with one hand tied behind your back and whistling 'Molly Malone', your students may be impressed by the physical dexterity involved but not necessarily by the physics. You need to know more than the basic physics. You need to know a little about the people involved, the common applications, the implications for society and popular conceptions or misconceptions. In short you need some background information. This is what brings the subject alive and makes it relevant to the students' experience.

In preparing for a student experiment or demonstration it is not enough to know the account that is given in the text book. It is important to know the limitations of the experiment and of the apparatus and to be aware of the things that can go wrong and how to deal with them. It is said that, 'if it's green or wriggles it's biology, if it stinks it's chemistry and if it doesn't work it's physics'. Physics experiments **do** work provided they have been set up properly, sources of error have been identified and proper precautions have been taken.

All of this is easier said than done. Where do you get this background information? Text books tend to be fairly concise in their treatment of the course and are written with the student, not the teacher, in mind. Background information tends to be gleaned from general reading, from magazine and newspaper articles, from television programmes and from lectures. It is generally accumulated over many years of teaching.

These modules were written by teachers, for teachers, to provide the kind of background information that makes the teaching of physics easier and better. The authors are all experienced teachers of physics who have pooled their knowledge and experience to produce a reference book that contains facts, explanations, and anecdotes that will help to make the subject more interesting to student and teacher alike.

Each module follows the relevant section of the syllabus fairly closely but this is of no particular significance since there is a fairly logical order in which the various topics might be introduced which is obvious to both planners and authors alike. Each module contains background information which may be of use to the teacher in the class room, suggestions on the teaching methods that the authors have found beneficial over the years and recommendations regarding student experiments and teacher demonstrations. Worked examples are included which the teacher may find useful in the class room or for homework. It is important to realise that these modules do not define the syllabus. They do not determine the scope of the syllabus nor the depth of treatment that is required or recommended.

Physics is an experimental subject. General principles and concepts are more easily understood if they are demonstrated in the laboratory. Laws and relationships are more fully appreciated if the student investigates and verifies them at the laboratory bench. While experiments should be carried out as rigorously as the available apparatus will allow, the principles should be illustrated as simply as possible. For example, an acceptable value for the acceleration due to gravity can be obtained in the class room using a bunch of keys suspended from a shoe lace. The laboratory experiment involves precautions which are taken to ensure a more accurate result.

Students should be aware that physics is involved in most of the technology on which we rely in industry, medicine, entertainment and in the home. Nuclear physicists may have been involved in the development and manufacture of weapons of mass destruction but they have also been involved in the development of diagnostic and therapeutic techniques in nuclear medicine. The theory of resonance can be applied in the design and construction of earthquake resistant buildings. The importance of science in general, and physics in particular, in a technological society should be considered when relevant. Many useful examples will be found in this book.

It is not envisaged that the individual modules in the book will be read line by line from cover to cover although that is not to be discouraged. Rather it is intended that the teacher consult the book to find interesting questions or snippets of information that might help to enliven a particular topic or challenge a particular group of students.

While details of individual experiments are not given, advice is given on how to deal with common problems that arise and on trouble shooting techniques when experiments do not work as expected. Therefore it might be advisable to consult the relevant section of the book as part of the preparation for a student experiment or teacher demonstration. The material in this handbook is arranged into ten modules plus a section on Gender and Science (see Contents). Each module is paginated independently. One index is provided for all the material, with page numbers preceded by a number to indicate the module. Thus, for example, '**2**: 47' refers to Module 2, page 47.

All of the content is also provided on the attached CD, along with the material from the Chemistry Handbook. It is intended that this will facilitate teachers in finding specific items of interest and in maximising on the use of the material in their classes. To this end teachers may print selected sections from the CD for class handouts or overhead transparencies. They may also incorporate selected topics into on-screen presentations and into class materials prepared in other software packages.

GENDER AND SCIENCE

DR SHEELAGH DRUDY

Introduction

The Department of Education introduced the scheme of Intervention Projects in Physics and Chemistry in the 1980s. It formed part of the Department's programme for Equality of Opportunity for Girls in Education, and it arose from the observation that, though half the population is female, the majority of scientists and engineers are male. In particular, it was a response to a concern in many quarters that females were very under-represented among Leaving Certificate Physics and Chemistry candidates. The lack of representation and participation of females in the physical sciences, engineering and technology is not confined to Ireland. For the past two decades it has been a major concern throughout most of the industrialised world.

Although girls and women are well represented in biological science, their participation and achievement at all levels of education in Physics and Chemistry have been the focus of much research. This research has sought not only to describe and analyse female participation in Physics and Chemistry but to identify ways of improving it.

It is essential that girls and women orient themselves to the physical sciences and related areas for two principal reasons. Firstly there is the question of women's relationship to the natural world. It is vital that women's and girls' participation in the physical sciences is improved in order to increase women's comprehension of the natural world and their appreciation of the contribution of Physics and Chemistry to human well-being. It is also important to improve women's capacity to control nature and to appreciate the beauty of Physics and Chemistry.

Just as important is the issue of employment. There is no doubt that, in the future, the potential for job opportunities and careers in the scientific and technological areas will increase in significance for women, in comparison to the 'traditional' areas of female employment. In so far as women are under-represented in scientific and technological areas, they are disadvantaged in a labour market increasingly characterised by this form of employment.

Achievement and Participation

International Comparisons

In order to put the issue of female participation in Physics and Chemistry into context, it is useful to consider the international trends. Since the 1970s a number of international comparisons of girls' and boys' achievements in science have been conducted. The earliest of these indicated that boys achieved better than girls in all branches of science at ages ten and fourteen, and at pre-university level. Studies also showed that while girls were more likely to take Biology as a subject, they were a lot less likely than boys to take Physics, Chemistry or Higher Mathematics.

A recent international comparison focused on the performance of girls and boys in science at age thirteen. This study indicated that in most of the 20 countries participating, 13 year old boys performed significantly better than girls of that age. This significant gender difference in performance was observed in Ireland, as well as in most of the other participating countries. This difference was observed in spite of the fact that in Ireland, as in the majority of countries, most students had positive attitudes to science and agreed with the statement that 'science is important for boys and girls about equally'.

Let us now consider the Irish context. Firstly, we should remember that relatively little science is taught in schools before this age, so it could be suggested that boys have greater socialisation into scientific 'culture' by early adolescence (for example, through very gender-differentiated toys, comics and television programmes). In addition, the international tests mentioned above relied heavily on the use of multiple-choice. There is evidence that the multiple-choice mode of assessment disadvantages girls.

Junior and Leaving Certificates: Achievement[1]

The important relationship between mode of assessment and performance in the sciences becomes evident when we examine the performance of girls in *public examinations* in Ireland. Analysis of recent results from the Junior Certificate and the Leaving Certificate examinations reveals some interesting patterns. For example, at Junior Certificate level, of the candidates taking science, a higher proportion of girls are entered at Higher Level than is the case with boys. A higher proportion of girls than boys receive grades A, B or C. This is also the case at Ordinary Level.

At Leaving Certificate level, while fewer girls than boys are entered for Physics, a higher proportion of the girls who do take the subject are entered at Higher Level. Traditionally, Chemistry has also been a male-dominated subject, though to a lesser extent than Physics. More recently, the numbers of girls taking the subject has equalled, or even slightly exceeded, the numbers of boys. However, as for Physics, a higher proportion of the girls who take the subject do so at Higher Level. As regards grades awarded, boys are generally somewhat more likely to receive an award in the A categories on the Higher Level papers. However, a higher proportion of girls than boys receive awards at the B and C levels, so overall a higher proportion of girls are awarded the three top grades in Physics and Chemistry at Higher Level. The same is true at Ordinary Level. Grade Point Average for girls in Physics and Chemistry is higher for girls than for boys at both Higher and Ordinary levels. Thus, in public examinations in Ireland, girls outperform boys in Physics, Chemistry and Junior Certificate Science. There is, therefore, no support for the notion that girls underachieve in the physical sciences in Ireland, when results are based on performance in public examinations.

This should not be taken to suggest that there is no longer a problem in relation to gender and science among Irish school-children. There is still a very serious problem in relation to differential take-up rates in science.

Junior and Leaving Certificates: Participation

At Junior Certificate level a lower proportion of girls than boys are entered for science. At Leaving Certificate level there are very marked variations in participation in science by gender. Indeed, overall, more girls than boys sit for the Leaving Certificate. Biology is the science subject most frequently taken by both boys and girls at Leaving Certificate. In terms of participation *rates*, though, it is predominantly a 'female' subject since two-thirds of the candidates are girls.

By contrast, in terms of participation rates, Physics is still very much a 'male' subject. Just under three-quarters of the Physics candidates are male. It is worthwhile noting that, although there is a great disparity in male and female take-up rates in Physics, there has been a marked increase in the proportion of females taking Physics since the early 1980s. This increase (albeit from a very low base) has been the result of a number of factors - one of these is the response of second-level schools (especially girls' schools) to the findings of a major survey by the ESRI (Sex Roles and Schooling[2]) in the early 1980s. This study highlighted the extremely low proportion of girls taking Physics. Another factor is the growing awareness among girls of the importance of science for future careers. A further important component in the improvement in the take-up of Physics by girls is the impact of the various phases of the Intervention Projects in Physics and Chemistry. Evaluations of this programme have shown that it has had an important impact on the participation levels in Physics and Chemistry among girls in the target schools. This present handbook is the most recent example of the work of these worthwhile Intervention Projects. However, while it is important to note the increase in participation by girls, it is also a matter of serious concern that the imbalance in take-up rates between girls and boys in Physics is still so very great.

In summary, then, as regards achievement and participation in the physical sciences, it would appear that girls are capable of the highest levels of achievement. Indeed in terms of overall performance rates at Junior Certificate Science and in Leaving Certificate Physics and Chemistry girls now outperform boys. However, major problems still remain in relation to participation rates. While some of the variation in these rates is no doubt due to the attitudes and choices of girls, there is equally no doubt that they are significantly affected by school policy, particularly as it relates to the provision of the subjects and the allocation of pupils to them within the school.

School Policy

Provision

Whether or not a subject is provided in a school is clearly a matter of policy for that particular school. Obviously, there are constraining factors such as the availability of teachers with appropriate qualifications. This problem was the principal focus of the Intervention Projects in Physics and Chemistry. The efforts made in these Projects have met with some considerable success. Nevertheless, the variability in provision in Ireland, according to school type, indicates a very strong element of policy decisions in the provision of Physics and Chemistry.

Variations in provision according to school type

A survey in the early 1980s indicated a very considerable discrepancy in the provision of Physics and Chemistry to girls and boys. Since then provision of these subjects has improved, especially in girls' schools. However, the most recent analysis of provision indicates the persistence of the problem. Although provision for girls is now best in single-sex schools, girls' secondary schools are less likely to provide Physics to their pupils than are boys' secondary schools. In co-educational schools girls are proportionately less likely to be provided with the subject.

Given the differential provision in the various school types, the provision of Physics and Chemistry for girls has been linked with the debate on co-education. This is an important debate in an Irish context, in the light of the overall decline in pupil numbers, the resulting school amalgamations and the decline in the single-sex sector. As indicated above, for girls, the best provision in these two subjects is in girls' single-sex secondary schools. Provision for girls is less favourable in all other types of school. However, it is very important to realise that research has pointed to the close relationship of take-up of science and social class. Thus the weaker provision in vocational schools and community/comprehensive schools probably reflects their higher intake of working-class children rather than their co-educational structure. Nevertheless, this explanation, on its own, would not account for the variation between co-educational and single-sex secondary schools.

Concern with the effects of co-education on girls, especially with regard to take-up and performance in maths and science, is not confined to Ireland. For example, in Britain, it has been the focus of heated debate. Some have suggested that girls have more favourable attitudes to physical science in single-sex schools than in co-educational schools. It must be noted that the results on attitudes in Ireland, from a major Irish survey, are directly contrary to this. In the United States major controversies have arisen with the introduction of men to formerly all-women's colleges. In Australia also the issue is a major policy one. There also it has been suggested that the somewhat contradictory evidence must be assessed bearing in mind the higher ability intake in the majority of single-sex schools which are academically selective, and also the different social class intakes between types of school. It is not possible to reach a conclusion here on the relative merits of co-educational or single-sex schools. However, anywhere there is a lack of provision of key subjects such as Physics and Chemistry it should be a matter of concern to school authorities.

Allocation

Closely linked to the matter of *provision* of Physics and Chemistry is that of the *allocation policy* of the school. Research has shown that even where schools provide these subjects there tend to be considerable variations in access to them within the school. Allocation relates to a number of factors. In particular, it involves the academic prerequisites that the school demands before allowing pupils to choose particular subjects. At Senior Cycle it also connects to the timetabling and availability of different options.

This may be of particular relevance to girls. They may be less inclined to choose a 'non-traditional' option, such as Physics, if it is timetabled against a subject perceived as more 'feminine' such as Biology or Home Economics. The differential achievement levels apparent in Physics and Chemistry mentioned earlier, i.e. the greater likelihood of girls obtaining grades A - C and of boys obtaining Grades E - NG at Higher Level, and the proportionately greater number of awards to boys than to girls in the D, E, F, and NG grades at Ordinary level in Physics and Chemistry, suggest that schools tend to encourage only their 'star' female pupils to take these subjects, and thus reflect a more selective access of female pupils to these subjects at Leaving Certificate. Boys, on the other hand, appear to be allowed greater access irrespective of their ability levels.

Choice

As we have seen, schools need to critically evaluate their policies on the provision of Physics and Chemistry, and their internal practices of allocation within the school, in order to improve the access of girls. However, such improvements do not necessarily mean that girls will choose these subjects unless other barriers to participation are also addressed. For example, the ESRI study mentioned earlier found that even where schools offered Physics to pupils, over four times the proportion of boys to girls chose to do it. Analysis in the 1990s shows that, in spite of a rise in girls' take-up rates in Physics, a marked differential is still apparent.

Before we turn to an assessment of why comparatively few girls choose to do Physics, even when it is available to them, let us consider briefly the concept of 'choice' itself. A great deal of educational research has shown that so-called 'choices', made by school-children in relation to subject options and careers, are themselves highly structured by the social situations within which they are made. When choosing options young people take account of the existing social context and the structure of opportunity available to their social class or their sex. Thus, the perceptions that young people have of the appropriateness of a particular option to their social class and their sex become an important element in their decision-making. Much of the international debate about girls and science has pointed to the very masculine image of the physical sciences. This masculine image, it has been suggested, contributes in an important way to the formulation of girls' attitudes to the physical sciences.

Attitudes

'Masculine' Image of Science

The formation of attitudes is a highly complex process. When attitudes relate to issues as fundamental as gender roles the processes become even more complex. Such attitudes are deeply rooted in early socialisation within the family, in the school and via the media. The debate about girls and science suggests that there may be a mismatch between the 'masculine' image of science (especially Physics and Chemistry), and girls' identification with the 'feminine' role at a critical period of adolescent development.

The first sense in which science can be regarded as 'masculine' is in the sense that men are numerically predominant. Perusal of any science textbook will suggest to the reader that the vast majority of scientific discoveries of any importance were made by men. To a very large degree this does represent a reality. However, this cannot be fully understood without the realisation that scientific discovery is itself as much a social process as a scientific one.

For example, at a time when the foundations of contemporary scientific enquiry were being laid, especially in the late nineteenth and early part of the twentieth century, women were not permitted to take degrees in many of the European universities at the forefront of scientific research, nor could they be members of the principal scientific societies. Even where women were involved in scientific discoveries it appears that they were allowed to carry out practical work but the named authors were mostly male.

Recent years have seen the publication of a number of accounts of the ways in which women's scientific contributions have been rendered invisible through the operation of male power structures in scientific institutions and universities. One of the best documented cases is that of Rosalind Franklin whose crucial work on the analysis of DNA received scant recognition as a result of the difficulties experienced by her in a male-dominated institution. Cases such as these no doubt have had an impact on women's self-image in the scientific community.

Kathleen Lonsdale[3] pinpointed the causes for the relative dearth of women scientists as the relative lack of women school teachers in 'hard' scientific subjects, girls opting for 'general science' rather than mathematics, pure Physics or Chemistry, more men than women continuing research after graduating and the small percentage of women appointed to responsible or creative academic posts. The 'masculine' image of science reflects unequal access to, and unequal relationships in, the scientific community. However, there are other ways in which this masculine image can be interpreted.

It has been argued that because science develops in relation to specific social and historical circumstances, in order to understand why girls choose Biology rather than Physics we need to consider the social role of science in general and the particular roles of the different sciences. It is suggested that, in advanced industrial societies, science has developed in relation to two major objectives: increasing the efficiency of production (economic) and the development of means of social control (military). The status of the different sciences will vary according to their perceived economic and military significance at any particular time. The higher their status, then the greater the exclusion of women in male-dominated societies, and the less relevant the subject matter is to issues of female concern. Therefore the variation in male domination between Physics and Biology follows from their different historical, economic and military significance.

Female and Male Roles and Behaviour

Let us turn now to examine the link between the image of the physical sciences, on the one hand, and male and female roles and behaviour, on the other. Physical science not only has a masculine image but also an impersonal image. Research among school-children shows that they think that science has to do with things rather than people. This image may be more off-putting to girls than to boys. Caring for people, both physically and emotionally, is an important part of the female role. A subject which appears to ignore people can seem irrelevant to girls' concerns. Indeed, transnational research on the attitudes and work of women scientists indicates that the scientific interests of women are interrelated with a profound commitment to humanity and society.

Women have also been found to define themselves in terms of a network of relationships. Research has suggested that their positive view of themselves is linked to their judgement of their capacity for caring. Women tend to see the world as networks of relationships, men to see it in terms of hierarchies. Boys, research indicates, are concerned with making sense of the world through rules, and achieving an individual identity; they aspire to a position in a hierarchy. Girls' concerns centre on developing a network of close relationships with other people. These differences in self-image and interest have an impact on subject choice and career orientation.

Another factor of relevance in the formation of attitudes during adolescence is the difference in maturity between girls and boys. It has frequently been observed that girls tend to mature physically earlier than boys. At a period when important subject choices are being made in schools, girls - especially working class girls - are also becoming involved with a 'culture of femininity' which involves an ideology of romance, marriage, family life, fashion and beauty. Involvement with this culture, which places heavy emphasis on traditional female roles, may well make girls reluctant to select subjects which have a strong masculine image, such as Physics.

On the other hand Biology is very much a female dominated science. It can be argued that part of the reason for this is because of its image as a 'life' science - and because it is the only science subject which deals at all with the topic of sexuality which is so fascinating for adolescent girls. Sexuality is also fascinating for adolescent boys, of course. As we have already seen, more boys also select Biology than any other science subject, even if they are in a minority compared to girls. A further reason that girls may orient themselves to Biology in such disproportionate numbers is because of its perceived function as a prerequisite for certain 'traditional' female careers, such as nursing.

Research has also suggested that, apart from being perceived as masculine, Physics and Chemistry have another 'image problem' which affects attitudes. Young people, it has been found, perceive Physics and Chemistry as being very difficult, very mathematical, heavily content-loaded, very dull, and demanding passive reception rather than active involvement with the learning process. This perception of Physics and Chemistry as difficult may be more off-putting to girls than to boys, although girls are more successful when they enter for these subjects (linked, perhaps, to girls' more selective entry). Boys' higher entry to Physics may be linked to higher levels of self-confidence at that age. They may be more likely to 'have a go', irrespective of the level of difficulty. Again, it may be that traditional 'male' careers, such as engineering, exercise a greater attraction for boys thus making it easier, psychologically, for them to take on subjects perceived as difficult.

The twin images of Physics and Chemistry as both masculine and difficult, and the impact of these on the attitudes of girls, have significant implications. These revolve around the operation of school policy, the selection and presentation of the curriculum in both subjects, and teaching methodologies in the classroom and laboratory.

Strategies For Action I: School Policy

Provision

If the proportion of girls taking Physics and Chemistry is to be improved, one of the most obvious strategies for action for school authorities is to improve the provision of these subjects in schools. Although marked improvements in provision have taken place over the last decade, there still are considerable variations, especially in some of the co-educational schools. This issue needs to be addressed as a matter of priority.

Allocation

In addition to increasing the provision of these subjects to both sexes, but especially to girls, the policy with relation to allocation within the school must be addressed, both at junior cycle and at senior cycle. At junior cycle there is a strong argument for making science a 'core' subject. Although this is, in effect, the practice in the majority of schools, the evidence shows that girls are still less likely than boys to be entered for Junior

Certificate Science. This can only serve to perpetuate the notion that boys are somehow more 'suitable' for science.

At senior cycle the two key allocation factors appear to be a) the combination in which various subject options are offered and the way they are timetabled; and b) the degree of selectivity practised by the school for entry to Leaving Certificate Physics and Chemistry classes (e.g. the requirement of higher level Maths (Syllabus A) at Junior Certificate, or a high grade in higher level Junior Certificate Science). There is little hope of improving girls' participation in Physics, for example, if taking Physics eliminates the possibility of 'traditional' female options such as Biology or Home Economics, to which girls are likely to be very well predisposed. Furthermore, there is evidence to suggest that a high proportion of girls have less confidence in their abilities to do Physics and Chemistry than boys do. Girls require considerable encouragement from their teachers to take up these subjects, or related ones such as Higher Mathematics. These factors are important for all girls but more especially for working class girls who are the group least likely to orient themselves to Physics and Chemistry.

The Allocation of Teachers to Science Classes

The allocation of teachers to science classes can reflect schools' policies (and teachers' attitudes to science). Many science teachers, particularly in girls' schools, are Biology and/or Chemistry graduates. They may not feel very comfortable with Physics, and possibly Chemistry, at junior cycle. The Physics teachers may teach higher level mathematics, as well as Physics, and may not be allocated to Junior Science classes. This may lead to the perception among students that to do Physics you should be doing higher level mathematics, that subjects such as Physics and Chemistry are difficult. Secondly, it means that senior cycle Physics and Chemistry teachers may not know the junior cycle science pupils. This may affect take-up, given that girls may be more likely to link their choices in some way to their relationship with their teachers.

Whole School Approach

The allocation issues have implications for teachers of other subjects as well as for science subjects. This requires a holistic approach to science, as well as staff development in relation to general equality issues. The need for a whole school approach has been emphasised in evaluation of earlier phases of the Intervention Projects in Physics and Chemistry.

Careers Guidance

The whole school approach has implications not only for option choices and timetabling across the whole curriculum but also, very particularly, for careers guidance. Careers advice can be influential in encouraging girls in science. There is evidence that while both boys and girls benefit from careers advice on science, girls notice it and respond to it more than boys. Research indicates that where boys choose Chemistry, for instance, they say that they do so because they will need it in their careers when in fact the careers chosen have no such requirement (e.g. accountancy, banking, law). Conversely, the same research shows that girls often choose careers, such as nursing or catering, where Chemistry would be a useful qualification but do not link the choice to Chemistry. It seems that while careers guidance can encourage more girls to orient themselves to Physics and Chemistry, it can also discourage them if the guidance counsellor is not fully aware of the importance of these subjects for girls.

Parents

The school can also play a very positive role in raising parental awareness of the importance of subjects such as Physics and Chemistry for their daughters. In international comparisons, there is a statistically significant relationship between parental interest in science and pupil performance in science in many countries. Work on the earlier phases of the Intervention Projects in Physics and Chemistry has also indicated the importance of involving parents.

In sum, then, school policy has an important role to play in increasing the participation of girls in Physics and Chemistry through the provision and allocation of the subjects, through a whole school approach to staff development and through the involvement of parents.

Strategies For Action II: The Curriculum

Humanising Physics and Chemistry

We have seen that Physics and Chemistry developed as subject areas within a context of male domination in universities and research institutions. This, combined with the strong influence of economic and military interests in the generation of research in the physical sciences has given rise to their largely ‘masculine’ image. This, it has been argued, has been one barrier to the recruitment of girls and young women to science.

Many experts have suggested that, in order to attract more girls to Physics and Chemistry, it is essential to tackle the curriculum itself. To make science meaningful, science teachers must personalise and carefully contextualise science itself, whilst invoking and accepting the previous experiences and prior knowledge that girls bring to lessons. Students, especially girls, can be motivated by being helped to see the physical sciences (and mathematics) as a human creation, developed in a particular cultural and historical context, by individuals who were influenced by the needs and values of the society in which they lived. Although science has been dominated by men, it is also important to emphasise the areas (and there are many) where women have made significant discoveries and achievements.

As well as presenting Physics and Chemistry within historical and social contexts, research has shown that girls are more encouraged when examples are used which relate developments and materials to contexts in which females have dominated. Females have dominated in the domestic sphere, but of course in many other domains as well. This approach is particularly easy in Chemistry, but can be adopted within Physics. There is also evidence that girls’ interest in physical science is increased by stressing its relevance to human biology. If, for example, girls are uninterested in finding out how machines work, they may be introduced to moments and forces by studying how muscles work.

Transition Year

The introduction of Transition Year into schools provides an opportunity to encourage more girls into the physical sciences. There are issues both of school policy and of curriculum in the Transition Year that can affect the take-up of Physics and Chemistry in the senior cycle.

The structure of the science courses within the year, the allocation of teachers, the question of choice, can all be considered when setting up the Transition Year. The Transition Year science course(s) may be optional modules, or part of a core curriculum. It may be the separate sciences, or general science with no explicit Physics and/or Chemistry components. The course may be taught by one teacher, or by a number of teachers with a variety of science specialisms. Students who did not take science for the Junior Certificate may or may not be permitted to take up any of the science subjects at this stage. The effect that policy decisions on the structure and organisation of the Transition Year may have on the take-up of Physics and Chemistry in the senior cycle needs to be considered.

School policy decides and defines the structure of the courses taught in Transition Year. The courses taught in Transition Year are developed by the teachers in the school and often reflect their interests and ideas. It can be an ideal opportunity to develop girls’ confidence and competence in the Physical sciences before they begin the two year Leaving Certificate syllabuses. The contact with an interested and enthusiastic Physics/Chemistry teacher and the development of confidence, particularly in the ability to cope with the mathematics, may be what is needed to encourage some students to take the physical sciences. It is an opportunity for teachers both to humanise Physics and Chemistry and to develop ‘girl-friendly’ science.

'Girl-Friendly' Science

'Girl-friendly' science is an approach that attempts to place science in a context that appeals to both girls and boys, rather than to boys only. Traditionally, many areas in the physical sciences have focused on issues which have been predominantly of interest to boys rather than girls. Guidelines for a 'girl-friendly science curriculum', arising from an English intervention project, the 'Girls Into Science and Technology' (GIST) project, include the following.

(a) Set experiments in context by providing background information about the possible uses and applications of scientific principles. Do this, if possible, before the ideas are derived by experiment - tell the pupils where they are going and why.

(b) Link physical science principles to the human body.

(c) Stress safety precautions rather than dangers.

(d) Discuss scientific issues, e.g. the microprocessor revolution and unemployment, energy and the bomb, aiming at a balanced view of the benefits and disadvantages of scientific developments.

(e) Make aesthetically appealing exhibitions.

(f) Use imaginative writing as an aid to assimilating scientific principles and ideas.

Girl-friendly teaching is good teaching.

Strategies For Action III: Teaching Methods[4]

Teachers' Perceptions of Pupils

There is evidence that, when it comes to girls participation in Physics and Chemistry, teachers are particularly important. Educational research shows that, across a range of areas, teachers' expectations have an impact on pupil performance. Some studies have shown a clear tendency on the part of teachers to overrate the work of a boy compared with that of a girl. These results suggest that, even if there are differences between boys' and girls' interest in and attitudes towards science, teachers may be further magnifying these differences. There is some evidence that teachers believe science education to be of greater importance to boys than to girls. If teachers believe that boys are naturally better and more interested in scientific and technical subjects, this may become a 'self-fulfilling prophecy'.

It is known also from various studies that girls, especially those in their early teenage years, have less confidence in their own ability than boys do. This is particularly the case in relation to traditionally male-dominated subjects such as Physics and Chemistry, and is especially manifested in the context of mixed classes. Since lack of self-confidence militates against achievement - and enjoyment - it is vitally important that girls' confidence in their own ability be raised as much as possible.

The analysis of the Junior and Leaving Certificate examinations mentioned earlier shows that, while girls are participating less in Junior Certificate Science and in Physics, their level of achievement in these subjects, and in Chemistry, is, on average, better than that of boys. However, their level of confidence in these subjects may not be commensurate. Consequently, teachers can do much to improve participation by support, encouragement and high expectations. Girls need to be told they are capable, even if to the teacher this seems obvious. Teachers should strive to create a relaxed, supportive, non-competitive environment where pupils can gain and maintain confidence.

Encouraging Pupils in Single Sex and Mixed Classrooms

Pupils in the classroom interact in a variety of different ways with teachers. Different strategies are needed for girls-only classrooms, for boys-only classrooms and for mixed classrooms if pupils are to experience Physics and Chemistry in an inclusive way.

In girls-only classrooms, girls can be encouraged by teaching styles that develop their confidence and enable them to cope with the different ideas and skills that Physics and Chemistry demand. Girls often prefer to work in a co-operative way, helping each other. The team skills that girls prefer can be useful in a career that develops out of their studies of Physics or Chemistry. The use of social, personal and biological applications of Physics and Chemistry should help in developing a warm, non-threatening classroom atmosphere. It is important to praise girls, not just for their neatness and hard work, but for the excellence of their work. Practical work is essential and girls should develop their skills and competence in this area.

In boys-only classrooms, it is equally important that boys realise that Physics and Chemistry include men and women. The image of Physics and Chemistry presented should be inclusive of girls and women. Comments below about the layout and presentation of laboratories and of posters apply equally to boys' classrooms. Women and men should be shown working in Physics and Chemistry in non-stereotypical ways. Boys should be encouraged to work co-operatively and to share equipment and other resources. A range of applications should be taught including the social, personal and biological, not just the mechanical and technical. Physics and Chemistry are about people and boys need to realise this just as much as girls.

In mixed classrooms, there has been a considerable amount of research which shows that teachers interact differently with girls and boys in all subject areas. This is as much the case in science classrooms as in any other. Boys both demand and get more teacher attention. A significant amount of the attention boys get is in the form of disciplinary interventions. However, they also receive more praise from teachers. In some science classes teaching may be mainly directed to the boys with girls 'listening in'. However, if teachers find that for various reasons (perhaps disciplinary) they are unavoidably spending more time interacting with boys in the lesson, to compensate they can spend more time with the girls after class, or during practical sessions. Teachers should ensure that practical work is organised fairly for girls and boys. Among the considerations should be the organisation of working groups (mixed or single-sex), the allocation and issuing of equipment to the groups and the attention paid by the teacher during the practical sessions. Boys can tend to both hog the equipment and take the active roles in practical work. It is important that girls are as much expected to do the work as the boys. It may be possible to involve more girls in the lesson by gearing the work to their interests and aptitudes. A more humanistic approach to Physics and Chemistry could be tried which might also improve girls' attitudes to the social implications of these subject areas.

It has been suggested by research that the introduction of co-operative learning activities benefits female students. Teachers who use student-centred, 'problem-driven', enquiry methods in science have been found to be more effective in maintaining girls' liking for the subject. These methods involve co-operation among students and allow less opportunity for a competitive spirit to develop. Collaborative, problem-solving work allows girls to make effective use of their verbal abilities which can help to clarify their ideas and boost self-confidence.

Laboratory Organisation

The layout and presentation of the laboratory/classroom is an important consideration in encouraging students to take up Physics and Chemistry. A clean, bright room can be friendly and welcoming. Posters that reflect a range of applications, not just mechanical and technical, and that show that people (both women and men) are involved in Physics and Chemistry can encourage girls and give them the confidence to feel that they can belong in the world of the physical sciences.

Teaching Behaviour and Techniques for Encouraging Girls into Physics and Chemistry and Retaining Them

Research in the United States by Jane Kahle has summarised successful strategies for encouraging girls to pursue science. The teaching behaviour and techniques that are effective for retaining girls and women in science are as follows:

DO

- use laboratory and discussion activities
- provide career information
- directly involve girls in science activities
- provide informal academic counselling
- treat both sexes equally in the science classroom
- directly involve girls in science activities
- allocate and issue equipment fairly to all students

DON'T

- use sexist humour
- use stereotyped examples
- distribute sexist classroom material
- allow boys to dominate discussions or activities
- allow girls to passively resist

Conclusion

In conclusion, teachers can play a key role in increasing the participation of girls in Physics and Chemistry. They can do this through having high expectations of their female students, by giving them strong support and encouragement, by ensuring a fair, relaxed and hassle-free atmosphere in mixed classrooms, by using a good variety of teaching methods which are student-centred and problem-solving and by using group as well as individual work. Excellent science teaching must also be innovative and exciting. 'Science must be presented as not only basic but beautiful.'

Notes

1. The commentary on gender differences in uptake and performance in the sciences in public examinations is based on analysis of the 1993 Junior and Leaving Certificate Examination results. This detailed analysis is provided in Drudy, 1996.

2. In order to assist the flow of the text, the material is not referenced in the usual academic format. The bibliography gives a list of sources. Detailed referencing is to be found in Drudy, 1996.

3. Kathleen Lonsdale was born in Newbridge, Co. Kildare in 1903 – see module on Some Irish Contributions to Chemistry, p. 25.

4. I would like to gratefully acknowledge the contribution and assistance of Marian Palmer with this section and with the section on the Transition Year.

Bibliography

BOURDIEU, P. and PASSERON, J. P. 1977. *Reproduction in Education, Society and Culture,* London: Sage.

BRUSH, S. G. 1991. 'Women in Science and Engineering', *American Scientist*, Vol. 79, September-October, pp. 404-419.

BURTON, L. 1990. *Gender and Mathematics: an International Perspective*, London: Cassell.

BYRNE, E. 1993. *Women and Science: The Snark Syndrome*, London: Falmer.

CLARKE, B. 1989. *Women and Science*, Hove, East Sussex: Wayland.

COLDRON, J. and BOULTON, P. 1990. *Integrating Equal Opportunities in the Curriculum of Teacher Education: Phase II (1989-90) Report*, Tenet European Community Action-Research Project, Sheffield City Polytechnic: School of Education.

DEPARTMENT OF EDUCATION, 1994. *Leaving Certificate and Junior Certificate Examination Statistics 1993*, Dublin: Department of Education, Circular S17/94.

DRUDY, S. 1981. *School Leavers in Transition*, Cambridge: University of Cambridge, Unpublished Ph. D. Dissertation.

DRUDY, S. 1996. 'Gender Differences in Participation and Achievement in the Physical Sciences and Mathematics' in *Teacher Education for Equality*, E. Befring (Ed.), Oslo: Oslo College.

DRUDY, S. and LYNCH, K. 1993. *Schools and Society in Ireland*, Dublin: Gill and Macmillan.

HANNAN, D. et al. 1983. *Schooling and Sex Roles: Sex Differences in Subject Provision and Student Choice in Irish Post-Primary Schools*, Dublin: Economic and Social Research Institute.

INTERNATIONAL ASSOCIATION OF EDUCATIONAL PROGRESS, 1992. *Learning Science*, New Jersey: Educational Testing Service.

INSTITUTE OF PHYSICS EDUCATION. 1986. *Girls and Science: a Better Deal. A Resource Pack for Science Teachers*, London: Institute of Physics.

KAHLE, J. B. (ed.), 1985. *Women in Science: a Report from the Field*, Philadelphia and London: Falmer.

KELLY, A. 1978. *Girls and Science*, Stockholm: Almqvist and Wiksell International.

KELLY, A. (ed.). 1981. *The Missing Half: Girls and Science Education*, Manchester: Manchester University Press.

KELLY, A. (ed.), 1987. *Science for Girls?*, Milton Keynes: Open University Press.

MARTIN, M. O. and HICKEY, B. L. 1992. *The 1991 Leaving Certificate Examination: A Review of the Results*, Dublin: NCCA.

MARTIN, M. O. and HICKEY, B. L. 1993. *The 1992 Junior Certificate Examination: A Review of the Results*, Dublin: NCCA.

McEWEN, A. and CURRY, C. 1987. 'Girls' Access to Science: Single Sex Versus Co-Educational Schools', in R. Osborne et al. (eds.), *Education and Policy in Northern Ireland*, Belfast: Queen's University (Policy Research Institute).

McROBBIE, A. 1978. 'Working Class Girls and the Culture of Femininity', in Women's Studies Group, Centre for Contemporary Cultural Studies (eds.), *Women Take Issue*, London: Hutchinson.

NÍ CHÁRTHAIGH, D. and O'SULLIVAN, C. 1988. *Pilot Intervention Projects in Physics and Chemistry: Report of External Evaluators*, Dublin: Department of Education.

Ó CONAILL, N. 1991. 'Girls and Science: Equality in School or Society?' *Irish Educational Studies*, Vol. 10, pp.82-95.

PHILIPS, P. S. and MCKAY, R. 1994. 'Women in Science: a Brief History Within Chemistry', *SSR*, Vol. 76, No. 274, pp. 132-137.

PORTER, G. 1993. *A Comparative Study of the Career Path Choices of Girls From Intervention Project Schools and From Other Schools*, Dublin: Department of Education, Intervention Projects in Physics and Chemistry.

ROBERTS, K. 1968. 'The Entry into Employment: an Approach Towards a General Theory', *Sociological Review*, Vol. 16, pp. 165-184.

SAYRE, A. 1978. *Rosalind Franklin and DNA*, London: W. W. Norton and Company.

STOLTE-HEISKANEN, V. (ed.). 1991. *Women in Science: Token Women or Gender Equality*? Oxford/New York: Berg Publishers.

WATTS, M. and BENTLEY, D. 1994. 'Humanising and Feminizing School Science: Reviving Anthropomorphic and Animistic Thinking in constructivist Science Education', *International Journal of Science Education*, Vol. 16, No. 1 pp. 83-97.

WOOLNOUGH, B. E. 1994. 'Why Students Choose Physics, or Reject It', *Physics Education*, Vol. 29, pp. 368-374.

YOUNG, D. 1994. 'Single-sex Schools and Physics Achievement: Are Girls Really Disadvantaged?', *International Journal of Science Education*, Vol. 16, pp. 315-325.

MODULE 1

Light

CONTENTS

CHAPTER 1

REFLECTION AND REFRACTION

1.1 Introduction

This chapter covers the basic principles of reflection and refraction and related topics, i.e. it deals with the geometrical optics section of the Leaving Certificate syllabus.

Under reflection the topics covered include the laws of reflection, images formed by plane and spherical mirrors, and use of the formulae $1/f = 1/u + 1/v$ and $m = v/u$.

The topics covered under refraction include the laws of refraction, refractive index, total internal reflection, transmission of light through optical fibres, and images formed by single thin lenses, including use of the relevant formulae. The power of a lens and of two lenses in contact is also dealt with.

The optical structure of the eye is described, along with a number of defects of the eye and their correction.

Suggested Teaching Approach

Order of topics

As no prior knowledge is needed for this section of the syllabus it can be studied at any time. It is a good starting point in the syllabus.

The order of topics as given in the syllabus is an ideal working order.

For each topic

Choose a few questions from the Do You Know section to arouse the interest of the pupils.

Use demonstrations to provide the answers to these questions if possible.

Discuss the relevant theory involved.

Carry out any mandatory student experiments.

Co-ordinate the theory to the experiment.

Demonstrate how problems are solved by worked examples.

Discuss the applications of the theory in every day life. Use more demonstrations to reinforce the theory.

Use the Do You Know section to check the understanding of the topic.

1.2 Background

Light fills the universe. Even on the darkest night light still fills the universe. Light is a form of energy. When we watch television every picture sends about one millionth of a joule of energy into our eyes. Twelve hours' viewing would give about one joule which is roughly the energy needed to put on a pair of jeans.

Our main source of light energy is the sun. The nuclear fusion reactions occurring in the core of the sun give off this energy. Hydrogen nuclei combine to produce helium nuclei with the release of energy according to the equation, $E = mc^2$, where m is the change in mass and c is the speed of light in vacuum.

The sun produces a vast amount of energy as about 6×10^8 tonnes of hydrogen are converted every second. The earth, because of its size and distance from the sun, receives about 0.000 000 05% of this energy as heat and light.

The visible part of the sun, the intensely bright sphere of white light, is called the photosphere.

The temperature of the surface of the sun is about 6000 °C. Sun spots (dark patches) are areas on the sun's surface at a lower temperature due to variations in the sun's magnetic field.

The sun, like most celestial bodies, spins on its axis. As the sun is not a solid body all parts of it do not rotate in unison; the equator rotates once every 25 days while points near the poles take 34 days for one rotation.

The earliest speculations about the nature of light are attributed to the Greek philosopher, Pythagoras (c.582 BC–c.497 BC), who proposed that light consists of tiny particles which are sent out by the object being viewed to the eyes of the viewer. An alternative theory, attributed to the Greek mathematician Euclid (325 BC–?), is that the eyes send out rays which strike objects and give the sensation of sight. Euclid also studied mirrors and discovered the laws of reflection.

The Greek astronomer, Ptolemy (c.75 AD–?) studied the refraction of light. He recognised the need to make adjustments to the apparent positions of planets in the sky to arrive at their actual positions. This change in location is due to the light bending on entering the earth's atmosphere. He studied air/water, air/glass and water/glass interfaces and concluded that the angle of incidence was proportional to the angle of refraction which we now know is incorrect.

The Arabian physicist, Alhazen (c.965–1038), applied mathematics to his study of plane and curved glasses and mirrors. He showed that spherical mirrors could not bring parallel rays to a sharp focus – spherical aberration – and that light travels more slowly through a more dense medium, causing the bending at interfaces between media of different density.

It was initially thought that the transmission of light was instantaneous as all attempts to measure the speed of light failed.

The first measurement of the speed of light was made by the Danish astronomer, Ole Rømer (1644–1710) about the year 1676. He observed the planet Jupiter and its satellites. Each of these satellites is eclipsed when it moves behind the planet. The time between successive eclipses of a particular satellite should be the same. Rømer found that when the earth was approaching Jupiter the eclipses became progressively earlier and that when the earth was receding from Jupiter the eclipses became progressively later.

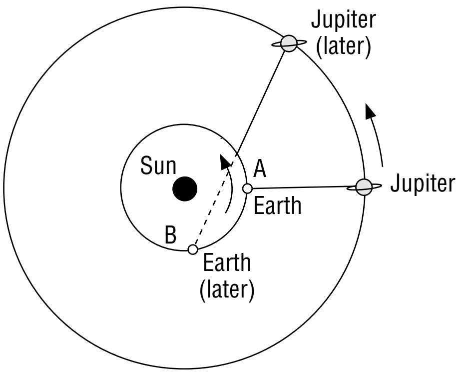


Fig. 1.1

He correctly attributed this variation to the variation in the time it took light to come from Jupiter to Earth as the distance between the two planets changed, Fig. 1.1. (The variation in the distance between Earth and Jupiter is due to the fact that Jupiter, being farther from the sun, takes much longer – approximately twelve times longer – to complete an orbit of the sun than the earth does.) Based on the information available at that time Rømer calculated a value of 2.28×10^8 m s^{-1} for the speed of light.

In 1849 the French physicist, Armand Fizeau (1819–1896) carried out the first terrestrial measurement of the speed of light using mirrors to make light travel a round trip of some 17 km. He used a rotating toothed wheel W, Fig. 1.2, to control the emerging and reflected light.

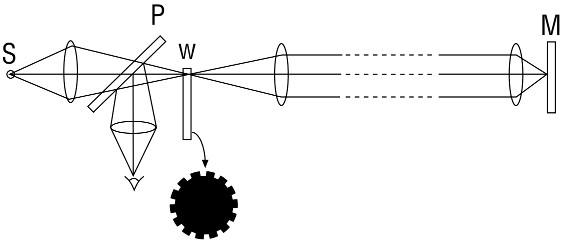


Fig 1.2 *Fizeau's experiment*

Light from the source S was focused by a converging lens through a half-silvered plate P onto the edge of the toothed wheel. If the light was not blocked by a tooth then it passed through and travelled a distance of 8.6 kilometres to the mirror M and back to the wheel's edge. He adjusted the speed of rotation of the wheel until he got no reflected light. He now knew that the light had travelled the round trip in the time it took the gap in the wheel to be replaced by a neighbouring tooth, blocking the light.

This method was adapted by Michelson (1852–1931) who used a rotating mirror to replace the toothed wheel.

Today, as part of the 1983 redefinition of the metre, the speed of light in vacuum, *c*, has been assigned an exact value. By definition the value of *c* is now 299 792 458 m s^{-1} exactly. Experiments which formerly measured the speed of light in vacuum are now used to calibrate length standards.

The daylight we use to see this print left the sun 8.5 minutes ago and travelled in a straight line until it encountered the earth's atmosphere, where it was continuously refracted due to the density gradient of the atmosphere, until it reached the page. From the time man discovered fire he has used light. He extended his day and defended his cave family using bundles of burning twigs. Oil from seals and other animals was used in stone or shell lamps. Roman and Greek lamps had spouts to hold the wick in the oil while the people of the Shetland Islands pushed wicks into the bodies of the very oily stormy petrel birds and used them as lamps. As the quality of lamps improved their uses increased, especially in the field of communications.

The ancient Romans used light to send messages from one army post to another. They used large wooden flags and codes to convey the message but they were limited by the fact that these stations depended on direct line of vision. The discovery of electric current gave us the electric telegraph and the telephone but, as these required wire connections, armies still used light – flashing lamps and mirrors – to transfer information.

Marconi invented the wireless telegraph and this gave us radio, radar and television. As a result the use of light for transferring information decreased. However, as the amount of information being sent from place to place increased, the air waves became crowded and light is now being used again.

Light has a much higher frequency (c. 10^{15} Hz) than radio waves (c. 10^5–10^7 Hz) or microwaves (10^8–10^{10} Hz). A system that operates at the frequency of light can transmit much larger quantities of information than radio or microwaves. In this system the information is encoded in pulses of light which are transmitted over large distances using optical fibres, at the end of which the information is retrieved. Light wave telecommunication systems using optical fibres already span the world and are being used to carry voice, video and computer data. Very fast computers of the future will probably use light.

The microscope, which consists of two converging lenses, was invented by Zacharius Janssen about 1590. In the microscope the first lens is called the objective lens and the object to be viewed is placed just outside its focus, to form a magnified inverted real image. This real image is the object for the second lens, called the eyepiece lens, and is placed just inside the focus of the eyepiece lens. The final image formed is magnified, virtual, still inverted and formed far away from the eyepiece lens so that the it can be viewed with a relaxed eye.

The telescope was invented by a Dutch optician, Hans Lippershey (1587–1619) in 1608. See the Appendix to this chapter, p. 22, for more details of the telescope.

1.3 Do You Know

A periscope with a rotating top mirror may be used to view either in front or behind. What problems would be encountered by a large vehicle using a periscope to view the traffic behind? What would be seen if the top mirror were rotated through 90° so that it pointed to the side?

When a periscope is used to view in front, the image is erect but when viewing behind the image is inverted. The traffic behind the vehicle would appear inverted. If the top mirror is rotated through 90°, the image will be half way between erect and inverted, that is, on its side.

Glass from the scene of a crime is of vital importance to forensic science. How can it be identified?

The glass is identified by its refractive index. A small piece of the glass is placed in oil and is observed with a microscope. The oil is heated gently until the glass disappears. At this temperature the refractive index of the glass is equal to the refractive index of the oil and, as no optical boundary exists, the glass is invisible. This enables a very accurate measurement of the refractive index of the glass and allows it to be identified.

Is the sun exactly where we see it in the sky?

The sun always appears higher in the sky than it actually is due to refraction. This gives us five to ten minutes of extra daylight.

Why can we not focus clearly under water yet swimming goggles will restore clear focus?

The focal length of a lens depends on the relative refractive indices of the two media. Light entering our eyes from air will be focused on the retina. In water, which has nearly the same refractive index as the cornea, the focus will be far behind the retina, resulting in a blurred image. The wearing of goggles introduces an air/cornea interface so the eyes focus normally.

What does a fish see when it looks up from the pond?

The fish sees a circle of light directly above and can see the pond surroundings (the bank, the trees, etc.) in this circle. Beyond this circle the surface of the water acts as a mirror, due to total internal reflection, and other fish in the water will be reflected in it. The radius of the circle depends on the refractive index of the water and the depth of the fish below the surface.

How do diving ducks and birds catch fish since their eyes, which give a clear image in air, lose focus when they enter the water?

Certain diving birds have a nictitating membrane, a third eyelid, which has a central region of high refractive index, providing clear focus under water.

Why are mirrors used instead of lenses in large telescopes?

It is very difficult to manufacture large glass lenses which have exactly the same refractive index throughout the lens. Also, as glass is not rigid the particles can move under their own weight and change the shape, and therefore the focal length, of the lens.

If a small circular beam of light is passed through a prism, will the image formed on a screen be circular?

The image will be elongated due to dispersion and therefore no longer circular. Colours are spread out along a line perpendicular to the vertex of the prism, red at one end and violet at the other.

Why are prisms used as reflectors in some optical instruments?

Prisms are used because they give almost 100% reflection (due to total internal reflection), no double images and do not deteriorate. Normal silvered mirrors reflect only about 70% of the light (the rest is absorbed), they give multiple images and become tarnished with time.

What causes red eyes in photographs?

Red eyes result when a flash is fired directly into the eyes of the subject, especially in dim light when the pupil is wide open.

The light enters the eye and is focused on the retina where it is reflected. The retina, with its network of blood vessels, reflects mainly the red light, absorbing the other colours. The camera interprets this red light as coming from the pupil.

A zoom lens can change focal length from 30 mm to 80 mm. How is the focal length changed?

A zoom lens is not a single lens, it consists of a combination of lenses. It is the relative positions of these lenses that give the overall focal length. The focal length is varied by changing the distance between the lenses.

In William Golding's novel Lord of the Flies the character Piggy uses his glasses to focus the sun's rays and light a fire. Later the boys abuse Piggy and break his glasses. He cannot identify them as he is short-sighted. How valid is the story?

If Piggy can focus the sun's rays with his spectacles then they are converging lenses. Converging lenses are used to correct long-sightedness.

Why is it an advantage for carnivores to have their eyes in the front of their head?

With their two eyes on the front of the head the carnivore has binocular vision. This gives a complete view of the environment and enables them to judge distance accurately. This is of extreme importance when the dinner has to be caught.

Why have you spots before your eyes when you wake up in the morning in a bright room?

The network of blood capillaries in the retina causes a definite shadow on the retina. The brain normally ignores these shadows as they are constant but on the first opening of the eyes in the morning the shadows are a change so they are recorded by the brain as spots before the eyes. They gradually disappear as the brain ignores them.

Why are you advised not to water plants on a bright sunny day?

The water forms droplets on the leaves. These droplets act as converging lenses and focus the sun onto the leaves, burning them. As a result the leaves will have brown spots.

How does leaving an empty glass bottle in a forest create a potential environmental disaster?

An empty glass bottle can act as a converging lens and focus the sun's rays onto dry vegetation and set it alight. This can result in a forest fire.

1.4 Conceptual Approach

How the Eye Sees

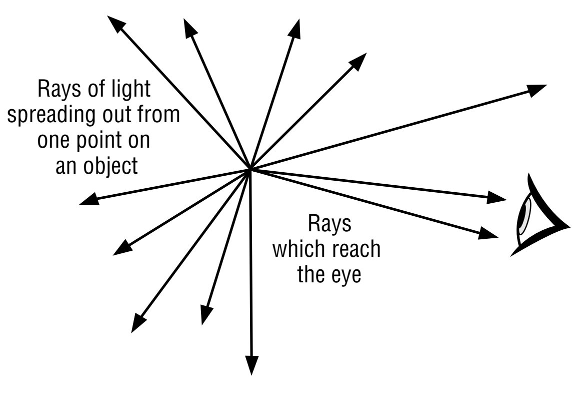


Fig. 1.3

An object is visible if it is luminous, that is gives out light, or if rays from a source of light bounce off the object. The light given off from luminous objects or reflected by non-luminous objects normally spreads out from each point on the object in every direction. If some of this diverging light enters the eye the object can be seen.

As the object moves further from the eye the rays of light become less diverging and eventually, if the object is a large distance away, the rays will be nearly parallel.

This can be demonstrated using a length of elastic to represent the rays, Fig. 1.4. Fasten the ends of the elastic to a board about 1 cm apart and draw an eye on the board where the elastic enters the top and bottom of the pupil as shown in the diagram, Fig. 1.4.

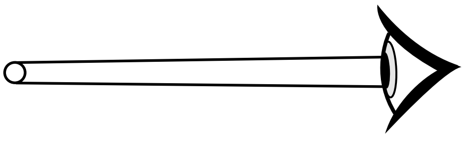


Fig. 1.4

Insert a pencil into the loop of elastic to represent a point on the object. As the object moves away from the eye the rays will become more and more parallel.

When the eye muscles are fully relaxed parallel rays entering the eye will be focused exactly on the retina in a person with perfect sight; in a short-sighted person they focus in front of the retina, while in a long-sighted person they focus behind the retina.

As the object moves closer to the eye, the rays from the object become more divergent and eventually, a point is reached where the muscles of the eye cannot change the lens sufficiently to give clear focus on the retina. This near point is taken to be 25 cm for the normal eye and is known as the least distance of distinct vision.

If for any reason the light rays are bent from their direct path the eye may still see the object but not in its true position. The rays *appear* to be spreading out from some point beyond the object, Fig. 1.5. The eye sees the point from which the rays *appear* to have come, as the eye and the brain do not realise that the rays have been bent. The eye is looking at an image of the object. The eye cannot distinguish objects from images.

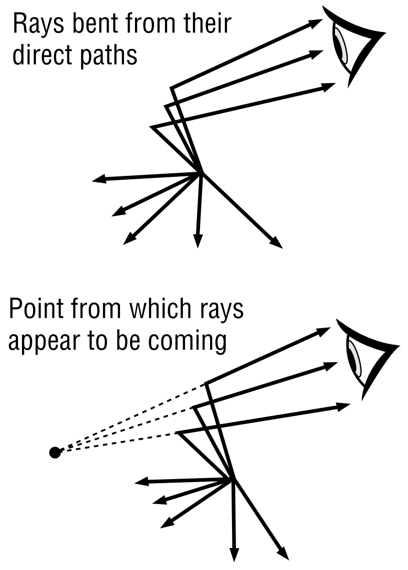


Fig. 1.5

Formation of Images

Rays can be bent by plane mirrors, concave mirrors, convex mirrors, and converging and diverging lenses, so these all form images.

If the rays reflected off a plane mirror to the eye are traced back they appear to come from a point behind the mirror, Fig. 1.6, but no light passes through this point, so the image is said to be virtual. If a piece of paper is placed at the point from which the rays appear to come nothing is seen on the paper.

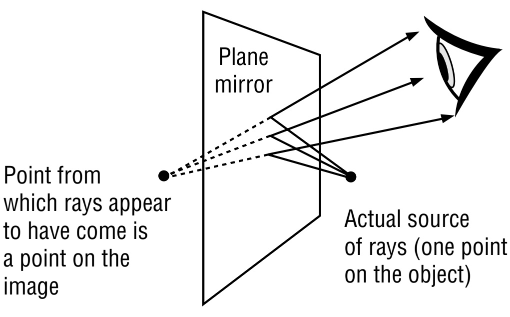


Fig. 1.6 *Formation of an image in a plane mirror*

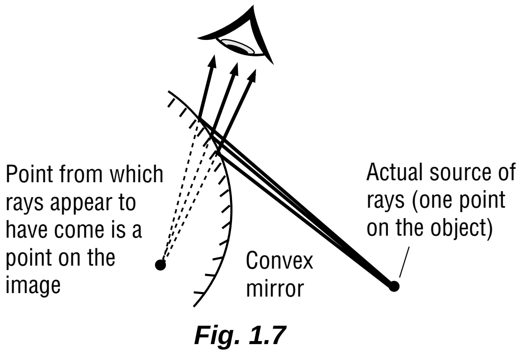


Fig. 1.7

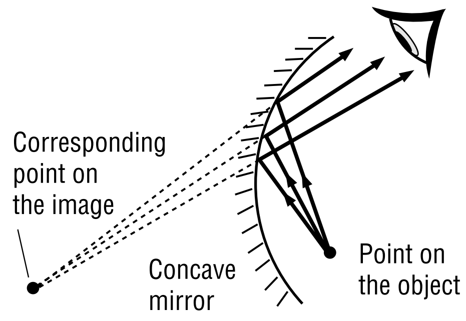


Fig. 1.8

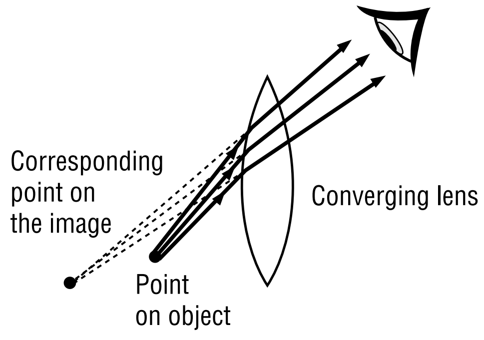


Fig. 1.9

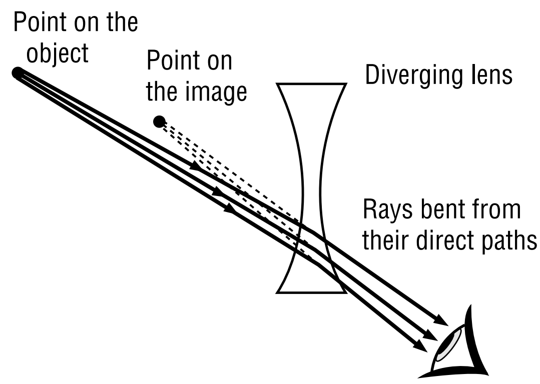


Fig. 1.10

Converging lenses and concave mirrors are capable of forming images that can be caught on a piece of paper just as the image is formed on the screen in the cinema. These images are called real images and are formed when the object is placed outside the focus of the lens or mirror, Figs. 1.11 and 1.12.

Rays spreading out from a distant object which meet a converging lens are bent to such an extent that they converge at a point and then diverge, Fig. 1.11. If these diverging rays reach an eye, the eye sees the image at the point from which they have diverged. The image is real because real rays of light spread out from it but it is inverted because of the crossing-over of the light rays.

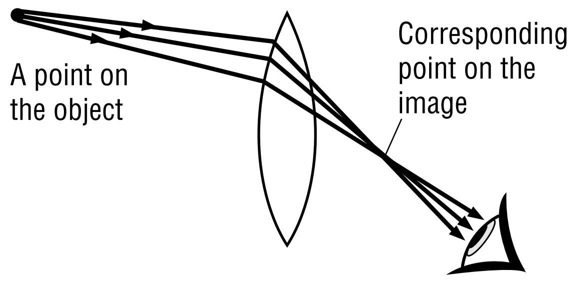


Fig. 1.11 *Formation of a real image in a converging lens*

A concave mirror produces real images of distant objects in a similar fashion but in this case the paths of the rays are changed by reflection, Fig. 1.12.

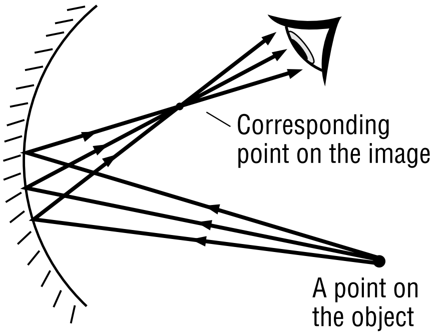


Fig. 1.12 *Formation of a real image in a concave mirror*

Size and the Eye

The light rays from the top and the bottom of an object or image enter the eye at a certain angle and the brain uses this angle to determine the apparent size.

It is therefore the angular size of an object that determines the apparent size of the object in the brain; objects with the same angular size will have the same apparent size. The sun and the moon look the same size in the sky because, although the diameter of the sun is about 400 times larger, it is also 400 times farther away.

There are two types of optical instrument. The first type forms real images and includes cameras and projectors. The second type forms virtual images and is designed to assist the eye. This type includes telescopes and microscopes. These instruments increase the angular size of the image and so increase the apparent size.

No Parallax

Parallax is defined as the apparent displacement of an object caused by an actual change in the position of the observer.

Hold your index finger at arm's length and view its position relative to a point on the wall with your left eye only. Now view it with your right eye only; its relative position has changed.

Hold up both index fingers, one in front of the other, and repeat the experiment. The relative positions will change except in the one case where the two fingers are together.

If the position of no parallax is found between two objects, two images, or an object and an image then they must be in exactly the same position.

1.5 Experimental Approach

Plane Mirror

NB. Refer to the instructions and precautions given in Chapter 3 on the use of the ray box and the laser.

Use a ray box or a laser to show the following.

(i) Light travels in a straight line. Chalk dust scattered in the path of the light will make the path visible.

(ii) Light shining on an object enables it to be seen from all directions due to the light being reflected through various angles.

(iii) Light shining on a plane mirror from a particular direction is reflected in one direction only.

Curved Mirrors

Use a ray box and comb filter to show the effect of a concave mirror and of a convex mirror on a parallel beam of light. Define real focus and virtual focus.

Use a single beam of light and a concave mirror of known focal length to show the following.

(i) A ray of light entering the mirror parallel to the principal axis is reflected through the principal focus.

(ii) A ray of light entering the mirror through the principal focus is reflected parallel to the principal axis.

(iii) A ray of light entering the mirror through the centre of curvature is reflected through the centre of curvature.

Refraction

Place a small object, e.g. a coin, in a plastic basin. Mark its position and let the students move back until they just cannot see the object over the edge of the basin, Fig. 1.13. Carefully add water to the basin. The object again becomes visible to the students due to refraction at the surface of the water.

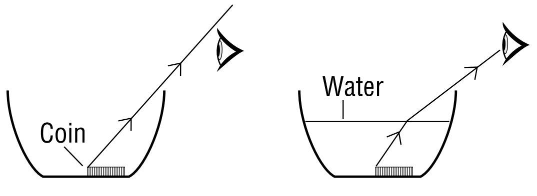


Fig. 1.13 *Refraction at water surface*

Place a glass block on print. Observe the print. The print appears to rise up into the glass due to refraction at the surface of the glass.

Use a ray box or a laser to produce a ray of light. Mark the position of the ray. Place a glass block in the path of the ray, firstly perpendicular to the ray and then at an angle to it, Fig. 1.14. This shows the bending of light as it passes from one medium to another of different optical density, i.e. refraction.

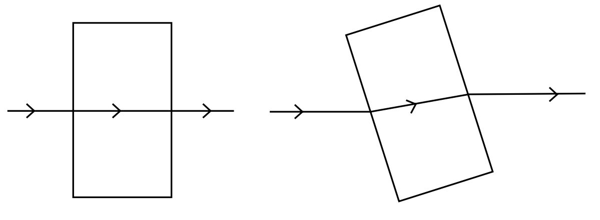


Fig. 1.14 *Demonstrating refraction*

Total Internal Reflection

Use a ray box or a laser with a semicircular block to show the bending of light as it passes from glass or perspex to air. To ensure good results the semicircular block must be centred accurately. This is done as follows.

Measure the diameter of the block circle. Calculate the radius and draw a circle with this radius. Draw a horizontal and a vertical diameter in the circle. Place the plane surface of the semicircular block along the vertical diameter. The horizontal diameter is then the normal to the plane surface at the centre of the block. All lines passing through this point are normal to the curved surface of the block. Therefore, any incident ray passing through the block to this point will be undeviated at the air/perspex interface since the angle of incidence is zero.

When the angle of incidence at the plane surface is zero the light will pass straight through the perspex/air interface, Fig. 1.15(a). As the angle of incidence is increased some of the light will be reflected at the plane surface and some will be refracted, Fig. 1.15(b). As the angle of incidence increases further the proportion of reflected light increases and the angle of refraction increases to 90°, with the refracted ray emerging along the surface, Fig. 1.15(c). The angle of incidence which gives an angle of refraction of 90° is called the critical angle, *C*. Further increase of the angle of incidence gives total internal reflection and the laws of reflection are obeyed.

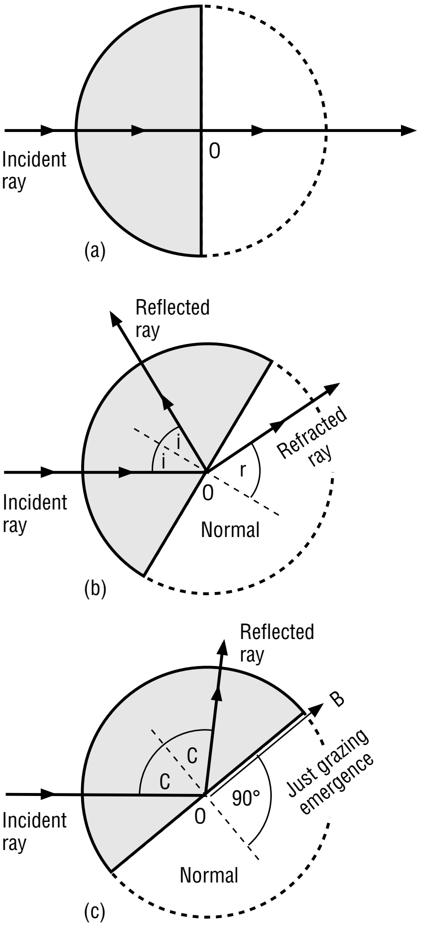


Fig. 1.15 *Total internal reflection*

Snell's law can be applied to the critical angle, C. The refractive index of the perspex is equal to the sine of the angle between ray and normal in air divided by the sine of the angle between ray and normal in the perspex. Thus, refractive index, $n = \sin 90°/\sin C = 1/\sin C$.

Demonstrations of Internal Reflection

1. Place a test-tube with a little water in it into a beaker of water and observe the sides of the test-tube from above, Fig. 1.16. The test-tube will appear silvered like a mirror except at the bottom where there is water. The silvery appearance is due to total internal reflection at the glass/air interface.

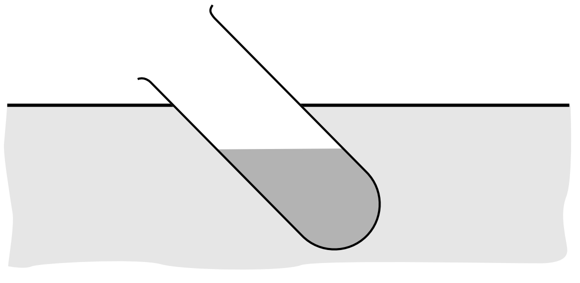


Fig. 1.16 *Demonstrating total internal reflection*

2. Completely cover the brass ball from a ball-and-ring (expansion of a metal by heating) apparatus with soot by holding it in the flame from burning turpentine or a candle. Lower it gently into a beaker of water, where it will appear silvery. The soot traps a layer of air around the ball and the silvery appearance is due to total internal reflection at the water/air interface. Remove the ball carefully and it should still be dry due to the layer of air trapped in the soot.

3. The laser is a very useful tool for demonstrating total internal reflection. Details are given in Chapter 3.

4. A torch where the bulb cover is replaced by optical fibres is available in some toy shops. This easily demonstrates the use of the optical fibre endoscope in medicine and of optical fibres in general. In the endoscope one set of fibres carries light into the body and another carries the reflected light back. Both sets of fibres are small enough to pass down a vein and bring back the picture of the organ being examined.

5. Use a ray box and a 45° prism to show how the prism can be used to bend light through 90° and 180°, Fig. 1.17.

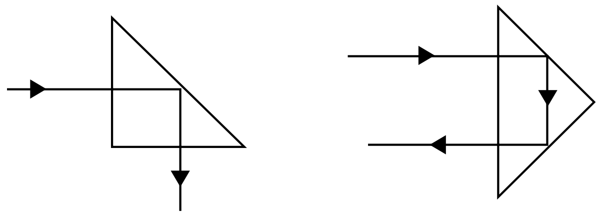


Fig. 1.17 *Total internal reflection in a prism*

Lenses

Use two 60° prisms, one on top of the other, with a ray box and comb to show that the top prism bends light down while the bottom prism bends light up.

Use a converging lens with a parallel beam of light to show the convergence of the light to a focus. Emphasise that the light does not stop at the focus. Repeat the experiment with a diverging lens to give meaning to the concept of a virtual focus.

Use a single ray to show the path taken by

(i) a ray parallel to the principal axis,

(ii) a ray passing through the focus,

(iii) a ray through the centre of the lens.

1. Use a lamp box, a converging lens and a screen (e.g. greaseproof paper on a card former) to show that when the object, i.e. the lamp box, is far away, the image is inverted and formed at the focus, Fig. 1.18. (The light from a point on a distant object is nearly parallel when it reaches the lens.)

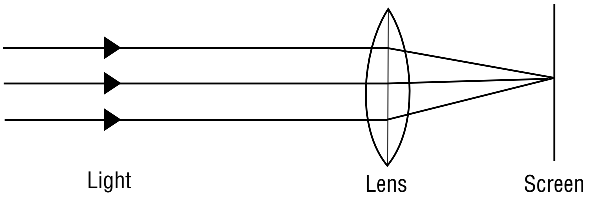


Fig. 1.18

2. Move the lamp box closer to the lens. The image moves away from the lens and the screen must be adjusted to obtain a sharp image; the image increases in size. As the object moves closer, the light from it is more

divergent when it reaches the lens; the lens bends it by the same amount – consequently it takes a greater distance for it to converge and form the image.

3. Move the lamp box closer again, almost to the focus. The image is now formed very far off in the distance (at infinity), so it cannot be caught on the screen. The light is very divergent when it reaches the lens; again the lens bends it by the same amount so it is still parallel when it leaves the lens, i.e. the image is at infinity.
4. Move the lamp closer again, the light is so divergent that the lens cannot bend it sufficiently to make it converge. The light leaving the lens is still divergent so it will never come to a focus so no real image will be formed. To the eye the divergent light appears to come from a larger source on the same side of the lens as the object, i.e. a virtual, magnified image is formed, Fig. 1.19. The lens now acts as a magnifying glass.

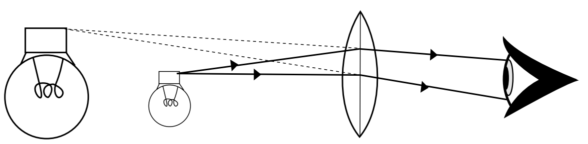


Fig. 1.19

The Eye

A full description of the eye is given in the Appendix, p. 20.

1. Use a ray box with a parallel beam of light to represent light from a distant object reaching the eye. Use a converging lens to represent the optical system of the eye. Show that the lens brings the light to a focus at a given point. This point represents a point on the retina. Draw the retina in this position. Adjust the ray box to give a diverging beam, representing an object close to the eye. The light will no longer be focused on the retina. This demonstrates the need for a lens of variable focal length in the eye.
2. Using the above arrangement show how long-sightedness occurs and how it can be corrected using a converging lens.
3. Discuss how short-sightedness occurs and how a diverging lens can be used to correct same.
4. Dissect a bull's eye.

The Power of a Lens

The power of a lens is a measure of the extent to which it bends light passing through it. The power of a lens depends on its focal length; the shorter the focal length the more powerful the lens. The focal length of a lens depends on both the refractive index of the glass and the radii of curvature of the surfaces. The shorter the radii of curvature, the more the light will be bent and so the shorter the focal length.

The power of a lens is defined as the reciprocal of the focal length in metres. The unit of power is the m^{-1}. Opticians use the dioptre as the name of the unit of power but this is not used in the SI system. The power is taken as positive for converging lenses and negative for diverging lenses.

1. Use a parallel beam of light from a ray box to determine the focal length of a thin lens. Calculate the power of the lens using the formula $P = 1/f$.
2. Use a second thin lens and find its power in the same manner.
3. Fix the two thin lenses together and find the power of the combination. Hence show that the power of the combination is equal to the sum of the powers of the individual lenses, i.e. $P = P_1 + P_2$.
4. Use a combination of a converging lens and a less powerful diverging lens to find the power of the diverging lens. This will show that diverging lenses have negative powers.

1.6 Applications

Mirrors

In home decoration, clever use of mirrors will give the impression of space. Safety must be considered as the illusion of space created by mirrors may lead to accidents.

Placing a mirror behind a light will increase the brightness of the room.

The stealth F-117A bomber is almost invisible to radar as the flat panels are arranged to reflect the radar signal up or down – not back to the radar station.

A kaleidoscope is a toy which produces symmetrical patterns by the use of mirrors. Inexpensive kaleidoscopes and periscopes are available from toy shops.

Make-up mirrors or shaving mirrors are concave mirrors. The face must be held close to the mirror, that is, inside the focus, for the mirror to give an erect magnified image.

Concave mirrors are used to focus light in telescopes.

Parabolic mirrors are used in spot or search lights: when the bulb is placed at the focus of the parabola the reflected light is in a parallel beam.

Security mirrors and driving mirrors are convex mirrors as they give a much wider field of vision and the image is always erect.

The sparkle of a diamond is due to its high refractive index. Most of the light entering the diamond is reflected due to total internal reflection at the cut faces.

A mirage is the result of successive refractions until the angle of incidence exceeds the critical angle and total internal reflection takes place. A tar or concrete road absorbs energy from the sun and becomes much hotter than the atmosphere. This creates a temperature gradient in the layer of air above the road, and this in turn causes a density gradient. Successive refractions and eventual total internal reflection in this layer of the 'bluish' light from the sky give the illusion of a puddle of water on the road.

The multiple images which can be seen in a mirror are due to light which was reflected at the silvered back of the mirror, being internally reflected at the front of the mirror and then again reflected at the back and so on.

Prisms are often used in optical instruments as reflectors because they give almost 100% reflection (due to total internal reflection), no double images and do not deteriorate. Normal silvered mirrors reflect only about 70% of the light; the rest is absorbed, they give multiple images and become tarnished with time.

Lenses

Converging lenses are used in spectacles to correct long-sightedness, in the simple camera to focus the light on the film etc., while combinations of converging lenses are used in microscopes, telescopes, binoculars, spectrometers and other optical instruments. Diverging lenses are used in spectacles to correct short-sightedness.

Lenses always reflect some of the light which falls on them and this reduces the quality of the images they produce. To reduce this reflection the lenses of some optical instruments, including cameras, are covered with an antireflection coating consisting of a thin layer of transparent material (the lenses are sometimes said to be 'bloomed'). The refractive index of the material is mid-way between that of air and than of the lens. The thickness of the layer is equal to one quarter of the mean wavelength of visible light in the layer. Light reflected from the bottom surface of the antireflection coating interferes destructively with light reflected from the top surface since the path difference is approximately half a wavelength. The intensity of the reflected light is zero for only the mean wavelength of the visible spectrum but the intensities of the other wavelengths are also significantly reduced. The effect is least with the longest and the shortest wavelengths, i.e. the red and blue ends of the spectrum. As a result, surfaces with an antireflection coating appear dark with a pale purplish hue.

Optical Fibres

Optical fibres enable information to be carried using light. The optical fibre acts as a conductor for light. Optical fibres can be thinner than human hairs. They work by total internal reflection, so no light escapes. Optical fibres are much thinner than wires yet they can carry much more information than radio waves or electrical signals, so they can provide many more telephone lines and TV channels than conventional cables. The standard copper telephone cable can carry up to a thousand conversations at the same time but the much, much smaller optical fibre can carry at least eleven times that number. Optical fibre telephone lines cannot be bugged or tapped.

The Endoscope

This is an instrument used in hospitals to examine the internal organs of patients. It consists of two very fine bundles of optical fibres in a protective casing. It is small enough to slip down blood vessels to the organ under investigation. One set of fibres carries the light down to the organ while the second bundle carries the reflected light, which is a picture of the organ, back to the operator.

1.7 Worked Examples

Rules for ray diagrams

1. A ray of light which is parallel to the principal axis will be reflected (refracted) through the principal focus.
2. A ray of light which passes in through the principal focus will be reflected (refracted) parallel to the principal axis.
3. A ray of light which passes in through the centre of curvature will be reflected back along its own path. (A ray of light which strikes the centre of a lens passes straight through.)

   Draw a diagram to scale and work out the magnification, i.e. the ratio of the image size to object size.

Examples

1. An object 4 cm high is placed at right angles to the axis of a concave mirror and at a distance of 30 cm from the mirror. If the focal length of the mirror is 10 cm find the position, size and nature of the image.

By ray tracing

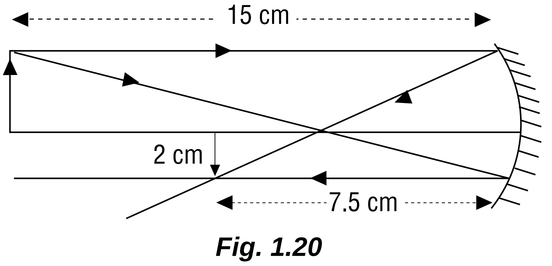


Fig. 1.20

Using graph paper, choose a suitable scale, e.g. 2 cm = 1 cm. Use the information given to draw the diagram, Fig. 1.20.

Draw a ray from the top of the object parallel to the principal axis. This ray will be reflected through the focus, and so can be drawn. Draw a second ray from the top of the object in through the focus. This ray will be reflected parallel to the principal axis, and so can be drawn. Where these two rays intersect, the image of the top of the object will be formed. Measure the distance of the image from the mirror and use the scale to find the actual distance.

As the rays actually meet at a point the image is real. Measure the height of the image and use the scale to find the actual size.

By calculation (using real-is-positive sign convention)

Since u = 30 cm and f = 10 cm

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

$$\frac{1}{30} + \frac{1}{v} = \frac{1}{10}$$

$$\frac{1}{v} = \frac{1}{10} - \frac{1}{30}$$

$$= \frac{(3-1)}{30}$$

$$= \frac{2}{30}$$

$$= \frac{1}{15}$$

Therefore v = 15 cm and the image is real as v is positive.

Size of image.

Magnification formula:

$$\frac{\text{size of image}}{\text{size of object}} = m$$

$$= \frac{v}{u}$$

$$= \frac{15}{30}$$

$$= \frac{1}{2}$$

$$\frac{\text{size of image}}{4} = \frac{1}{2}$$

Therefore image is 2 cm high.

2. Where must an object be placed in front of a concave mirror of focal length 20 cm so that a twofold magnified image will be produced?

A concave mirror can form both a real image and a virtual image so two answers are required.

Real image

Magnification = 2

$$\Rightarrow \frac{v}{u} = 2$$

$$\Rightarrow v = 2u.$$

Also, f = 20 cm.

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

$$\frac{1}{u} + \frac{1}{2u} = \frac{1}{20}$$

$$\frac{3}{2u} = \frac{1}{20}$$

$$2u = 60$$

$$u = 30 \text{ cm}$$

First answer: 30 cm from the mirror.

Virtual image

Magnification = –2

$$\Rightarrow \frac{v}{u} = -2$$

$$\Rightarrow v = -2u.$$

Also, f = 20 cm.

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

$$\frac{1}{u} + \frac{1}{-2u} = \frac{1}{20}$$

$$\frac{1}{u} - \frac{1}{2u} = \frac{1}{20}$$

$$\frac{1}{2u} = \frac{1}{20}$$

$$2u = 20$$

$$u = 10 \text{ cm}$$

Second answer: 10 cm from the mirror.

3. An object 3 cm high is placed at right angles to the principal axis of a convex mirror and at a distance of 10 cm from it. If the radius of curvature of the mirror is 20 cm, find the nature, size and position of the image formed.

As the radius of curvature is twice the focal length, the focal length is 10 cm. As the mirror is convex the focus is virtual, thus f = –10 cm. The object distance is 10 cm, i.e. u = 10 cm.

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

$$\frac{1}{10} + \frac{1}{v} = \frac{1}{-10}$$

$$\frac{1}{v} = \frac{1}{-10} - \frac{1}{10}$$

$$\frac{1}{v} = \frac{-2}{10}$$

$$= \frac{-1}{5}$$

$$v = -5 \text{ cm}$$

The image is virtual and 5 cm behind the mirror.

$$\text{Magnification} = \frac{\text{size of image}}{\text{size of object}}$$

$$= \frac{v}{u}$$

$$\frac{\text{size of image}}{3} = \frac{5}{10}$$

$$\text{size of image} = \frac{15}{10}$$

$$= 1.5 \text{ cm}$$

The image is 1.5 cm high.

4. A ray of light strikes an air/glass interface at an angle of incidence of 60°. If the refractive index of the glass is 1.5 calculate the angle of refraction.

$$\text{Refractive index, } n = \frac{\sin i}{\sin r}$$

$$1.5 = \frac{\sin 60°}{\sin r}$$

$$\sin r = \frac{\sin 60°}{1.5}$$

$$\sin r = 0.5774$$

$$r = 35.3°$$

Angle of refraction = 35°.

5. A microscope is focused on an ink spot through a glass block 4.5 cm thick. When the glass block is removed the microscope must be lowered 1.5 cm to regain focus. Find the refractive index of the glass.

$$\text{Refractive index} = \frac{\text{real depth}}{\text{apparent depth}}$$

Real depth = 4.5 cm and

apparent depth = (4.5 – 1.5) cm

= 3.0 cm

$$\text{Refractive index} = \frac{4.5}{3.0}$$

$$= 1.5$$

Refractive index of the glass is 1.5.

6. Find the speed of light in water of refractive index 1.33. ($c = 3.00 \times 10^8$ m s^{-1}.)

$$\text{Refractive index} = \frac{\text{speed of light in medium 1}}{\text{speed of light in medium 2}}$$

$$1.33 = \frac{3 \times 10^8}{\text{speed in water}}.$$

$$\text{Speed in water} = \frac{3 \times 10^8}{1.33}$$

$$\text{Speed in water} = 2.25 \times 10^8 \text{ m s}^{-1}.$$

7. Given that the refractive index of water is 1.33 find the critical angle for water.

$$\text{Refractive index} = \frac{1}{\sin C}$$

$$1.33 = \frac{1}{\sin C}$$

$$\sin C = \frac{1}{1.33}$$

$$= 0.7519$$

$$C = 48.76°$$

Critical angle for water is 48.8°.

8. What is the minimum refractive index of glass prisms which can be used as reflectors in optical instruments?

When prisms are used as reflectors the angle of incidence is 45°.

For total internal reflection to occur the critical angle must therefore be less than 45°.

$$\text{Refractive index, } n = \frac{1}{\sin C}$$

$$\text{Minimum refractive index} = \frac{1}{\sin 45°}$$

$$= \frac{1}{0.7071}$$

$$= 1.41$$

Minimum refractive index = 1.4.

9. A converging lens of focal length 10 cm produces a twofold magnified real image of an object placed perpendicular to the principal axis. Find the distance of the object from the lens.

By ray tracing

Choose a suitable scale, e.g. 4 cm = 1 cm.

Draw the lens and mark in the focus, Fig. 1.21.

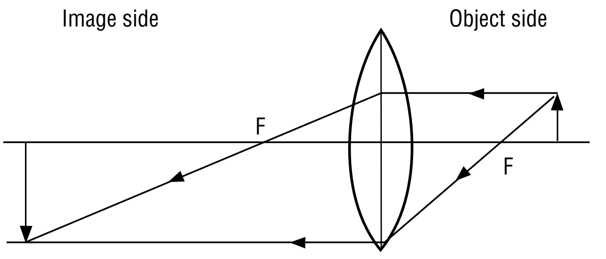


Fig. 1.21

Let the object be 4 cm high and draw in the parallel ray (1 cm above the principal axis) from the top of the object which, on passing through the lens, goes through the focus.

The image will be 8 cm high, which is 2 cm to scale, so draw in a ray parallel to the principal axis and 2 cm below it. This ray must have come in through the focus so its position can be drawn. Where these two rays intersect the object and image are located. Measure the distances and convert to actual distances using the scale.

By calculation

$$\text{Magnification, } m = \frac{\text{image distance}}{\text{object distance}}$$

$$= \frac{v}{u}$$

$$= 2$$

$$\Rightarrow v = 2u$$

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

$$\frac{1}{u} + \frac{1}{2u} = \frac{1}{10}$$

$$\frac{3}{2u} = \frac{1}{10}$$

$$2u = 30$$

$$u = 15 \text{ cm}$$

The object is 15 cm from the lens.

10. The near point of a long-sighted person is 100 cm from the eye. What spectacle lenses would she require to enable her to read a book at 25 cm from the eye?

The lens must have a focal length such that the virtual image of the book is formed at the near point 100 cm from the eye, Fig. 1.22.

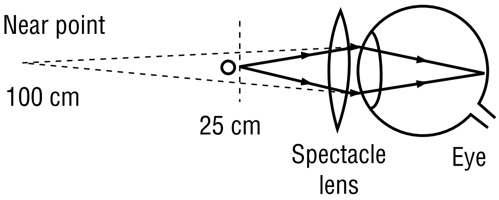


Fig. 1.22

Therefore u = 25 cm and v = –100 cm (real-is-positive).

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

$$\frac{1}{25}+\frac{1}{-100}=\frac{1}{f}$$

$$\frac{3}{100}=\frac{1}{f}$$

$$f = 33.3 \text{ cm}$$

Therefore converging lenses of focal length 33 cm are required.

The power of the lens is $1/f$ where f is expressed in metres.

$$\text{Power} = \frac{1}{0.333}\ \text{m}^{-1}$$

$$= 3\ \text{m}^{-1}$$

$$= 3 \text{ dioptres.}$$

The power of the lenses is 3 m^{-1}.

11. The furthermost point of distinct vision of a short-sighted person is 200 cm. What spectacles would she require to enable her to see distant objects clearly?

The virtual image of distant objects must be formed 200 cm from the eye. Therefore v = –200 cm and u is infinite.

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

$$\frac{1}{\infty}+\frac{1}{-200}=\frac{1}{f}$$

$$0-\frac{1}{200}=\frac{1}{f}$$

$$f = -200$$

Therefore diverging lenses of focal length 200 cm are required.

The power of the lens = $\frac{1}{f}$ where f is expressed in metres.

$$\text{Power} = \frac{1}{-2}$$

$$= -\,0.5\ \text{m}^{-1}$$

The power of the lenses is –0.5 m^{-1}.

1.8 Student Experiments

Measurement of the Focal Length of a Concave mirror

Find the focal length of a concave mirror by focusing a distant object on a screen, Fig. 1.23. screen

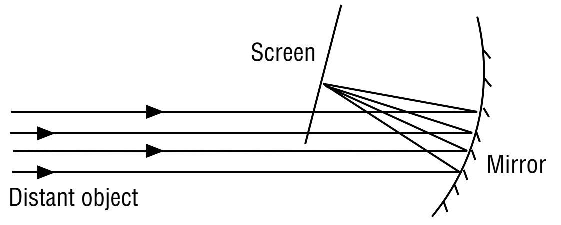


Fig. 1.23

The light from a point on a distant object will be nearly parallel so the image will be formed at the focus.

Find the focal length of the same mirror more accurately by using a lamp box as the object and catching the image formed on a screen, (greaseproof paper on a card former is a convenient screen). The position of the screen is adjusted until a sharp image is formed, Fig. 1.24.

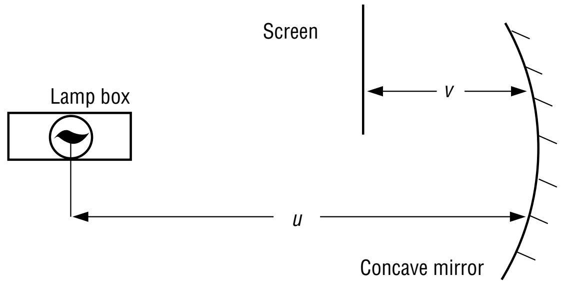


Fig. 1.24

Measure accurately the distance from the bulb to the mirror – the object distance, u – and the distance from the screen to the mirror – the image distance, v. Then use the mirror formula, $1/u + 1/v = 1/f$, to calculate f, the focal length of the mirror. Remember, it is the filament of the bulb which is acting as the object, so measurements should be made from the position of the filament.

Verification of Snell's Law of Refraction

This experiment can be carried out either by using a ray box or by representing the ray of light by pins. A combination of both methods will reinforce the fact that the pins are visible through the glass due to the light travelling from them.

Set up the apparatus using two pins, as far apart as possible for greater accuracy, to represent the incident ray, Fig. 1.25.

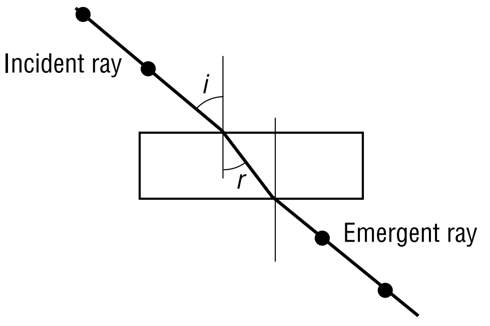


Fig. 1.25

View the pins through the glass block and place two more pins in line with them, again as far apart as possible for accuracy. These last two pins represent the emergent ray – the refracted ray is inside the glass block and can only be drawn after the glass block is removed. Mark the positions of the four pins and draw the incident ray and the emergent ray.

Now use the ray box to pass a ray of light along the marked incident ray; it should emerge along the marked emergent ray. This is a good check on the accuracy of the experiment using the pins.

Now draw the outline of the glass block, remove it, and draw in the refracted ray. Measure the angle of incidence and the angle of refraction. Repeat for various angles of incidence and plot a graph of sin *i* against sin *r*, Fig. 1.26.

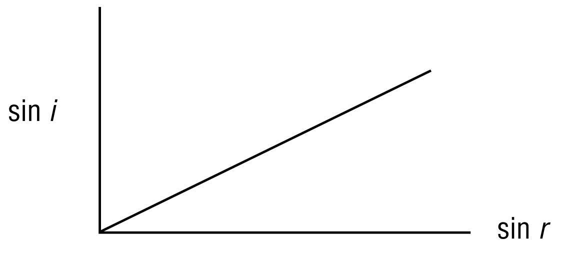


Fig. 1.26

As the graph is a straight line through the origin the equation of the graph must be of the form $y = mx$, where m, the slope, is a constant.

In this graph, y represents sin i, and x represents sin r.

Therefore, $\sin i = \text{constant} \times \sin r$,

or $\dfrac{\sin i}{\sin r}$ = constant, i.e. Snell's law.

But, $\sin i / \sin r$ = refractive index. Therefore the slope of the graph is equal to the refractive index.

Measurement of Refractive Index of a Liquid or a Solid

The previous experiment can be used to find the refractive index of glass or perspex. A rectangular container filled with liquid can be used in place of the block to find the refractive index of the liquid provided the sides of the container are reasonably thin.

Plot the graph of sin *i* against sin *r*. Draw the best straight line through the points.

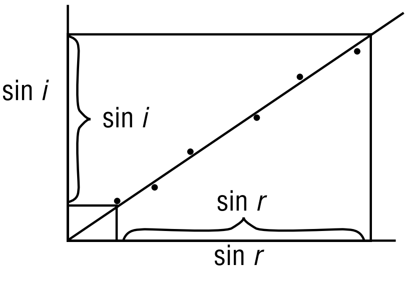


Fig. 1.27

Take two points on the line, again as far apart as possible for accuracy. Mark in the value of sin *i* and sin *r* as calculated between the points on the graph; this makes it easier for students who have problems with graphs.

Calculate the refractive index in the usual manner using the values calculated from the graph.

$$\text{Refractive index} = \frac{\sin i}{\sin r}.$$

Alternative method: real and apparent depths

Place a glass block on print so that the concept of apparent depth is grasped.

Stick a tall pin to one edge of a glass block using Sellotape. Stick a small plane mirror to the other edge of the glass block as shown in Fig. 1.28.

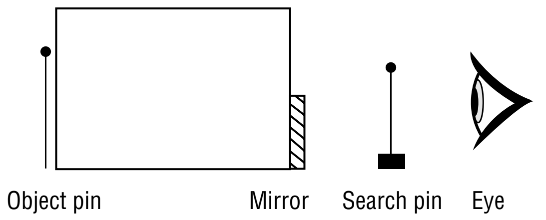


Fig. 1.28

Stress the fact that the apparent position of the pin when viewed through the glass block is somewhere inside the block and so it can only be found by getting it to coincide with the virtual image of the search pin in the mirror.

Move the search pin until the position of no parallax between the image of the object pin as seen through the block and the image of the search pin in the plane mirror is found. (See p. 7 for a discussion of parallax.) Pins with different coloured heads will assist recognition of the images. Each student must find the position of no parallax for themselves to understand the concept.

Stress the fact that the image of the search pin is the same distance behind the mirror as the search pin is in front of it.

The apparent depth can be found by measuring the distance of the search pin from the mirror.

$$\text{Refractive index} = \frac{\text{real depth}}{\text{apparent depth}}.$$

This experiment can be used to find the refractive index of a liquid by using a square box filled with the liquid in place of the glass block, again provided the sides of the box are reasonably thin.

It is easier to perform the experiment by this method than by using the beaker of liquid; measurements are easier and there is less chance of breaking the mirror. However, ideally, a number of boxes of different size are required and, as already noted, the sides of the boxes must be thin.

Measurement of Focal Length of a Converging Lens

A. Using a distant object

Find the focal length of the converging lens by focusing a distant object such as a tree or house on a screen, Fig. 1.29.

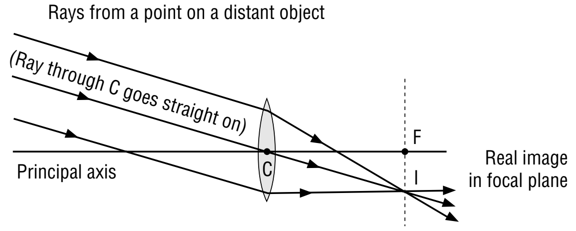


Fig. 1.29 *Rays from a distant point are parallel to each other*

The light from a point on the distant object is almost parallel so it will converge to the focus and form an image in the focal plane of the lens. The distance between the lens and the screen is therefore approximately equal to the focal length.

B. Using a lamp box

A more accurate method is to use a lamp box, a lens and a screen set up as shown in Fig. 1.30.

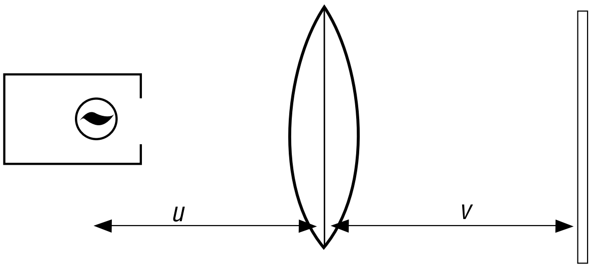


Fig. 1.30

The position of the lens is adjusted until a clear image of the bulb filament is formed on the screen. In measuring the distance of the object from the lens, u, you must measure from the bulb filament to the centre of the lens. The distance from the screen to the centre of the lens is v. The focal length is found using the formula:

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

C. Using pins

Place one pin in front of the lens to act as an object. This object pin must be outside the focus of the lens to form a real image. A second pin, the search pin, is placed on the other side of the lens, Fig. 1.31.

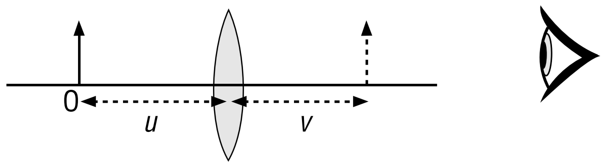


Fig. 1.31

The search pin is moved until the position of no parallax is found between the search pin and the image of the object pin seen through the lens. This means that the image of the object pin and the search pin are in exactly the same position. The distance of the object pin from the centre of the lens is u while the distance of the search pin from the centre of the lens is v. The focal length is calculated using the lens formula.

D. Using a plane mirror and one pin

The converging lens is placed on top of a plane mirror and the pin is fixed directly above the lens as shown in Fig. 1.32.

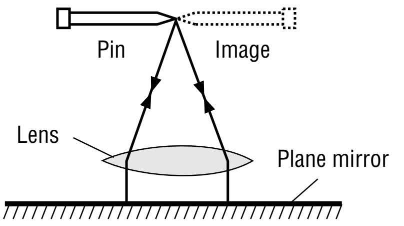


Fig. 1.32

The plane mirror must be level. It is desirable to know the approximate focal length of the lens before the apparatus is set up so that the pin can be fixed at about this height above the centre of the lens. The pin is moved until the position of no parallax between the pin and the image of the pin is found. The pin is now exactly at the focus of the lens so the distance of the pin from the lens is the focal length.

The light from the tip of the pin which is at the focus of the lens will be parallel to the principal axis of the lens when it passes through the lens. This light then strikes the plane mirror normally so it is reflected back along its own path and is parallel to the principal axis when it meets the lens a second time. Parallel light passing through a converging lens meets at the focus and forms the image. Both the image and the object are therefore at the focus of the lens.

APPENDIX

1. The Camera

A Simple Pin-Hole Camera

All that is required to make a simple pin-hole camera is the centre cardboard former from a kitchen roll, aluminium foil, greaseproof paper and a small elastic band. Cut across the centre of the cardboard roll but not right through; this provides a slot into which the greaseproof paper is inserted. Cover one end with aluminium foil and secure it with the elastic band. Pierce a small hole in the aluminium foil and view from the other end, Fig. 1.33.

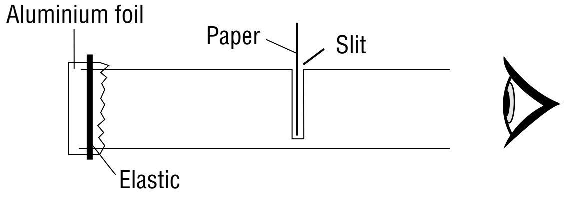


Fig. 1.33 *Simple pin-hole camera*

A Single Lens Camera

Using a camera with a fixed focal length lens show how the position of the lens is adjusted to give a sharp focus. Adjusting the aperture controls the amount of light entering the camera. The lens also has an f number. This number is equal to the focal length divided by the diameter of the aperture. Therefore a small f number indicates a wide aperture.

The shutter speed controls the time of exposure and therefore the total amount of light which reaches the film. The total exposure is therefore determined by the combination of shutter speed and f number selected. For pictures which contain moving objects the exposure time must be very small (the shutter speed must be high) to prevent multiple images or blurred images being formed on the film.

Depth of field is the distance in front of the camera within which there is sharp focus. Small lenses and small aperture openings give the greatest depth of field.

Photographic Film

Add a solution of hydrochloric acid to a solution of silver nitrate; a white precipitate is formed instantly. Leave this precipitate exposed to light and it turns black. This illustrates the principle of black and white film. Emphasise that the film inserted in the camera becomes the negatives which we get back after processing.

Television Camera

In the television camera the ‘film’ is a screen which is made up of very small, light-sensitive, electronic devices. When exposed to light they convert light energy to electrical energy. The camera scans each of these picture cells (pixels) in turn. The signal from each pixel is sent down the cable one after the other to give an electrical code which represents the picture on the screen. The screen is scanned 25 times a second and the coded signal is recorded or broadcast. (See also p. 36.)

2. The Eye

The two most important parts of the eye are the cornea at the front to focus the light and the retina at the back to record the image. The lens, by changing its shape, allows us to focus on objects at different distances. The iris, the coloured part of the eye, expands and contracts to regulate the size of the central opening, the pupil. The pupil determines the amount of light entering the eye. Covering the cornea is a thin delicate membrane called the conjunctiva. The conjunctiva is kept moist by a fluid produced in the tear gland under the eyelid. Between the lens and the cornea is a fluid called the aqueous humour while behind the lens is a gelatinous substance, the vitreous humour.

The retina is at the back of the eyeball and contains light-sensitive cells. There are two types of visual cells, rods and cones. Rods are highly sensitive and useful in dim light while the cones are less sensitive but provide high acuity and colour vision. The part of the retina responsible for

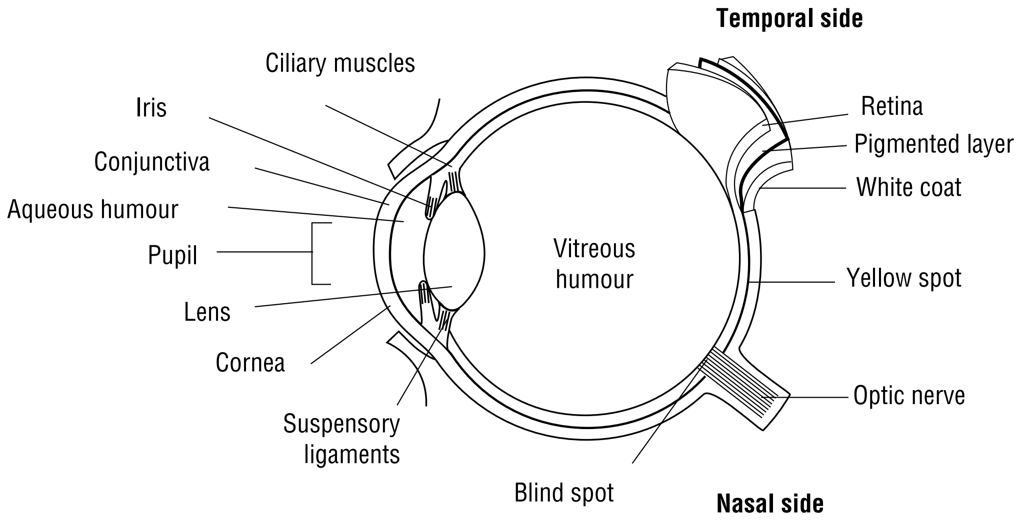


Fig. 1.34 *The eye viewed from above*

seeing things clearly is the yellow spot (macula lutea) right in the middle which is composed mainly of cone cells. Around this yellow spot the retina is composed mainly of rod cells which enable us to see in dim light only as they are desensitised in bright light. There are three different types of cone cell, each sensitive to one of the primary colours. The colour we see depends on how many of each type of cell are stimulated. A large optic nerve runs from the back of the eye to the brain. The point where the optic nerve is attached to the retina has no receptor cells, it cannot see things and so is called the blind spot.

The fluids inside the eye exert pressure which maintains the shape of the eye. The vitreous humour is nearly permanent while the aqueous humour is continually drained and replaced by new fluid. Failure of the drainage system leads to excess pressure in the eye. This condition is called glaucoma.

The cornea evolved from the skin of the head. It merges with the inner layer of the eyelid and so is continuous with the skin. The refractive index of the cornea is 1.376, so this is where the light is refracted most and made to converge on the retina. The lens has a refractive index of 1.42 while both the aqueous humour and the vitreous humour have a refractive index of 1.336. Refraction at these interfaces is very small but the lens does increase convergence slightly. When the eye is relaxed the image will be formed in the retina's photosensitive layer and so objects from infinity to about 6 m can be seen clearly. Below 6 m the lens increases its curvature which increases its converging power, i.e. it accommodates to give a clear image.

There is a limit to how close an object can be brought and still be accommodated by the lens. This limit is known as the least distance of distinct vision and is taken to be about 25 cm, but this increases with age. Age hardens the lens so it becomes less flexible and therefore less accommodating.

The eye, like the camera, forms an inverted image on the retina. The retina transmits the image to the brain where it is interpreted, the right way up. An interesting experiment was carried out using special spectacles which made everything appear upside down. Initially the images were upside down but gradually the brain made the necessary correction and everything appeared right way up again. The brain assumes that everything is lighted from above and adjusts pictures accordingly. If a picture of the moon is inverted the mountains appear as craters and the craters become mountains.

As we have two eyes we have binocular vision. Each eye sees a slightly different aspect of the object. The brain combines these two images to give a single three-dimensional view. Binocular vision enables distance to be judged more accurately.

Student Experiments

1. Examine the workings of the iris by closing the eyes for about ten seconds and then opening them in bright light.
2. Examine the blind spot by looking at a picture of two separate objects about 7.5 cm apart. Close the left eye and, while focusing on the left object with the right eye, adjust the distance of the picture from the eye. The right object should vanish when the image of it is falling on the blind spot.
3. Use a set of colour-blindness charts to check for colour blindness.

Malfunctions of the Eye

Short-sightedness is caused by the eyeball being too deep so that an object at infinity is focused in front of the retina in the vitreous humour, Fig. 1.35(a). Closer objects will focus normally, Fig. 1.35(b). A diverging lens will correct this problem, Fig. 1.35(c).

Long-sightedness is caused by the eyeball being too shallow or by the lens hardening, so that there is insufficient accommodation when an object is close to the eye. While distant objects are focused normally on the retina, Fig. 1.36(a), near objects are focused behind the retina, Fig. 1.36(b). The condition may be corrected by using a converging lens, Fig. 1.36(c).

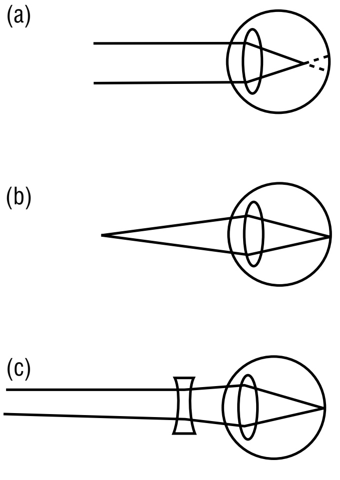


Fig. 1.35 *Short-sightedness*

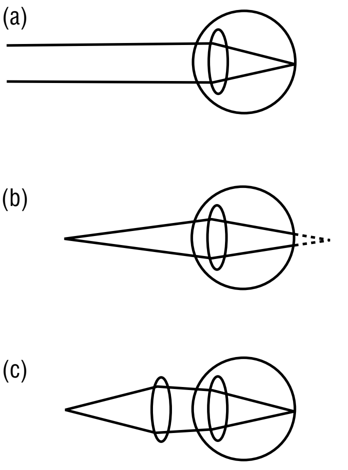


Fig. 1.36 *Long-sightedness*

Astigmatism occurs when the shape of the cornea is not spherical so that it has a different curvature in one direction than in another. This means it has a different focal length in each direction so there is no clear focus on the retina. An astigmatic lens with different focal lengths in different directions can be used to exactly balance the astigmatic cornea and achieve clear focus.

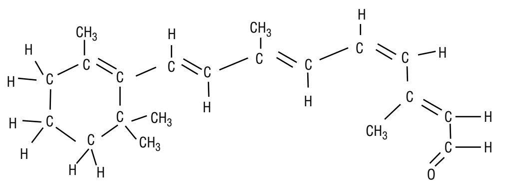
Chromatic aberration occurs in the eye, with blue light being focused to the front of the photo-sensitive layer while the red is focused at the back. If the photosensitive layer is too thin to accommodate both focal planes nothing can be done to correct the problem. If you view red and green stripes and focus clearly on the red, the green will be slightly out of focus due to chromatic aberration. The brain will interpret this as the green being at a different depth so the red stripes appear to stand out.

Cataracts occur when the lens becomes cloudy and stops letting light through. The lens can be removed and replaced with an artificial acrylic lens.

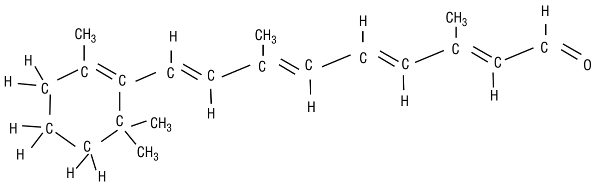
Detached retina occurs if a hole develops in the retina. The fluid from the eyeball gets through the hole and pushes the retina away from the underlying tissue. If untreated, partial blindness results. A laser shot at the edges of the hole can weld the retina to the tissues underneath and so prevent further damage.

How the Eye Detects Light

Light is electromagnetic radiation which can be detected by the retina of the eye. The visual cells in the retina contain rhodopsin, a deep red pigment which absorbs all wavelengths of light. The rhodopsin molecule consists of two parts, a protein molecule, opsin, and a simple organic molecule called retinal. The opsin molecule, like all protein molecules, is built up of many amino acids and contains several thousand atoms while retinal is simply $C_{20}H_{28}O$. In the dark the retinal is in the *cis* form,



but on absorption of light it is converted to the *trans* form,



The *trans* form splits from the protein and the pigment is bleached. In the dark the retinal spontaneously returns to the protein and reforms the rhodopsin. Retinal is found in many photopigments, in the rods and cones of the human eye, and in the visual cells of all animals.

3. The Telescope

An astronomical telescope consists of two converging lenses. To show the working of a telescope two lenses and a greaseproof screen may be used. Using lens 1 focus a distant object on the screen, Fig. 1.37. The image will be diminished, inverted and formed at the focus of lens 1 as the light from a distant object is nearly parallel. Lens 2 can now be used to magnify that image so the image from lens 1 must be at or inside the focus of lens 2.

The screen can be removed and this gives the arrangement of the lenses in a telescope. The final image will be enlarged, inverted and virtual. If the image from lens 1 falls on the focus of lens 2 the final image will be formed at infinity and so it can be viewed with a relaxed eye. This arrangement is called normal adjustment.

Lens 1 is called the objective lens and lens 2 is the eyepiece lens. In the telescope the object is very large, e.g. the moon, and is at a great distance. We have no desire to magnify the object; we want to bring the image closer and then magnify the diminished image. The light from a point on the object is effectively parallel, so the

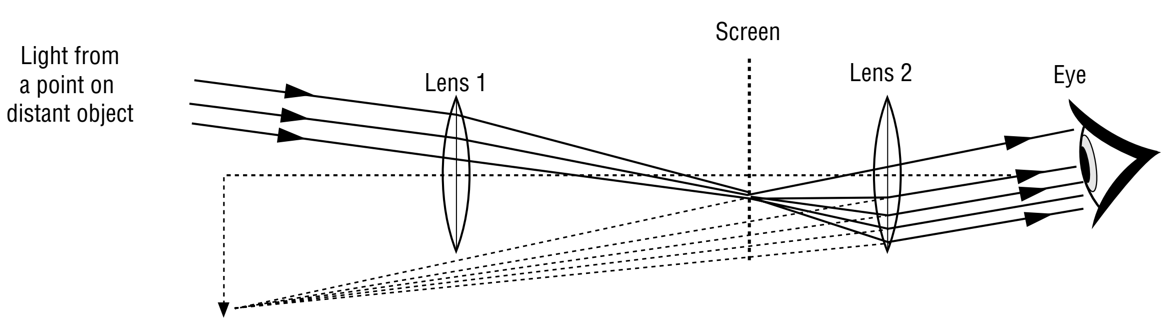


Fig. 1.37 *Astronomical telescope*

objective lens forms a real, diminished, inverted image at its focus. When the telescope is in normal adjustment this image falls exactly on the focus of the eyepiece lens, so the final image is produced at infinity. The eyepiece lens does the exact opposite of the objective lens but if the eyepiece lens has a shorter focal length it will be more effective at bending the light, so the angle subtended by the image will be greater than that of the object. The size of the image is determined by the angle subtended at the eye hence the size appears to increase.

The magnification is determined by the ratio of the two focal lengths. The shorter the focal length of the eyepiece lens and the longer the focal length of the objective lens the greater the magnification. There is a limit to how short the focal length of the eyepiece can be made and this limits the magnification of the telescope. The magnification is also limited by the diameter of the objective lens; the greater the magnification the greater the diameter required to illuminate the image. The diameter of a lens is limited by the difficulty of producing and maintaining a large lens of uniform refractive index (see also p. 38).

Note that chromatic aberration is apparent in the simple arrangement illustrated in Fig. 1.37.

CHAPTER 2

WAVE NATURE OF LIGHT

2.1 Introduction

This chapter deals with the principal wave properties of light, viz. diffraction, interference and polarisation. The use of the diffraction grating formula, $n\lambda = d \sin \theta$ is discussed and a number of worked examples are given.

Dispersion by a prism and a diffraction grating are discussed. This leads to a consideration of the three-colour theory and its application in television, etc.

The electromagnetic spectrum is considered in terms of the sources and applications of the radiations which constitute the spectrum. In the appendix aspects of electromagnetic radiation are dealt with in more detail.

Suggested Teaching Approach

The section of the syllabus covering waves must be studied before this section.

Order of topics

The spectrometer should be studied first as it will be used to show diffraction and interference.

The concept of interference should be stressed; constructive interference gives brightness, destructive interference gives darkness.

Young's slits experiment should be observed by everyone as this then gives an introduction to the use of the diffraction grating to find the wavelength of monochromatic light.

The extra number of slits in the grating makes the constructive interference fringe much brighter and sharper than in Young's slits due to destructive interference taking place over most of the region. The closeness of the slits in the diffraction grating gives a much greater separation of the constructive interference fringes and so the angles can be measured much more accurately.

Use white light with the diffraction grating to reinforce the fact that the position of the interference fringe depends on the wavelength.

Carry out the mandatory experiment to measure the wavelength of monochromatic light.

The other topics in this section can then be carried out in the order given in the syllabus.

2.2 Background

It has been known from antiquity that when sunlight is bent in a glass prism the colours of the rainbow are produced. It was first thought that these different colours were the result of light from different parts of the sun striking the prism at different angles. Newton studied this beam of colours by placing a barrier, with a small hole in it, in the beam. This enabled him to study each of the colours in turn using a second prism.

He found that the red light was bent least and the violet most. He called one of the colours indigo as it was a fashionable colour at that time. He stated two theorems.

1. Lights which differ in colour, differ also in degrees of refrangibility (refraction).
2. The light of the sun consists of rays differently refrangible.

He re-formed white light from this beam of separated colours by using a second prism and this convinced him that sunlight is a combination of colours. Newton was then able to explain colour and answer the question often argued, 'Is colour an integral part of the object or does it depend on the reflected light from the surface?' By examining different coloured objects in different coloured lights, he concluded that a red object reflects all colours but reflects red more strongly. In white light the object reflects the red most strongly and so appears red.

Newton held that light was made up of tiny particles travelling in straight lines like bullets (the corpuscular theory) but he was open to change as he was aware of work being carried out by others. An Italian Jesuit, Francesco Grimaldi (1618–1663) found that the shadows of small objects were smaller than they should have been if light travelled in straight lines and that the boundary of the shadows was fuzzy and coloured instead of being sharp.

Newton had discovered that a convex surface placed on a plane glass surface gave a series of concentric light and dark circles, now called Newton's rings, which cannot be satisfactorily explained in terms of the corpuscular theory.

The first systematic wave theory was put forward by a Dutchman, Christian Huygens of the Hague (1629–1695). He showed that a train of waves striking the edge of a barrier will be bent around the edge and will penetrate the region behind the barrier. He also explained Snell's law in terms of the velocity of the light. He did so by assuming that waves travel more slowly in a denser medium, e.g. glass, than in air.

Huygens's work was largely ignored until research carried out by the English physicist and physician, Thomas Young (1773–1829) revived interest in the wave theory. Young, a man of many talents, contributed to the fields of botany, philosophy, languages, and physiology and translated ancient Egyptian hieroglyphics during his career as a practising physician. He verified the wave theory of light by demonstrating interference effects and also showed that the different colours of light were due to different wavelengths.

There was not a mass conversion to the new wave theory, especially in Britain, where Newton's reputation ensured a strong following for the corpuscular theory. Work by the French physicist Augustin Fresnel (1788–1827) on diffraction, which produced results for the wavelength of light which agreed with Young's values strengthened the case for the wave theory.

The corpuscular theory assumed that light travelled faster in a more dense medium while the wave theory required that light travelled more slowly in the more dense medium. Foucault's measurement of the speed of light in water, which showed that light travelled more slowly in water than in air, sealed the case for the wave theory.

2.3 Do You Know

Why are soap bubbles coloured in white light?

The light waves reflected from the front and back surfaces of the bubble interfere, resulting in different colours being in phase at different angles.

What would a soap bubble look like in monochromatic sodium light?

A soap bubble would appear as yellow and black horizontal stripes due to constructive and destructive interference.

Why are compact discs coloured?

The disc acts as a diffraction grating. When the disc is illuminated with white light the reflected light from each 'groove' acts as a new source and, due to interference, coloured lines appear.

Why are Polaroid sunglasses of particular benefit to fishermen?

The reflected light from the surface of water is partially polarised horizontally, i.e. parallel to the surface, and the glare from this is blocked by the vertical polarisation of the glasses. The refracted light which passes into the water must be polarised vertically so when this light is reflected by a fish it can all pass through the glasses. The fish can be clearly seen but the glare is blocked.

How would a rainbow appear to passengers in an aeroplane when there is rain in the air above and below them?

It should appear as a complete circle as the rays reaching the passengers from all directions which are at an angle of 42° with the line from the sun to the aircraft will display the rainbow, Fig. 2.1.

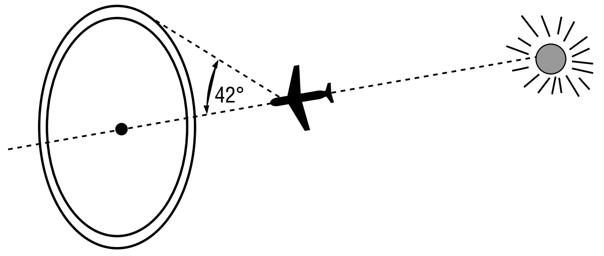


Fig. 2.1 *Rainbow seen from aircraft*

Is a rainbow polarised?

Yes. As the light is reflected within the rain drop it will undergo partial polarisation.

Why do materials purchased in a shop appear a different colour in daylight?

White light contains colours of all wavelengths from red to violet but they are not present in equal intensities. In sunlight the greatest intensities are in the green and blue portions of the spectrum while an incandescent bulb has greater intensity in the yellow to red portion of the spectrum. The colour observed is due to the light reflected from the material. This in turn depends on the light incident on the material. The difference in colour of the material is therefore due to the difference in intensities of the various colours in the artificial light compared with sunlight.

Why is the sky blue?

The gas molecules in the atmosphere scatter the light. The amount of scattering depends on the wavelength of the light, the shorter wavelengths being scattered preferentially.

What is scattering?

The oscillating magnetic and electric fields which constitute light moving across a molecule cause a sympathetic vibration of the same frequency in the molecule. In other words, the molecule absorbs light energy of that particular frequency. The oscillating molecule then re-emits the energy as radiation of the same frequency but in all directions. So the light which had been travelling in one direction has now been ‘scattered’ in all directions, including towards the earth. Since it is mainly the shorter wavelengths, i.e. the blue light, which are scattered the sky appears blue in all directions except when we look directly towards the sun. Scattering may also be caused by small particles, e.g. dust, in the atmosphere.

Why are sunsets orange?

At sunset the light rays from the sun are travelling through greater lengths of atmosphere so the amount of scattering of the blue light increases and the percentage of the reddish light reaching our eyes increases giving the setting sun an orange-red colour. Dust in the atmosphere also contributes to the scattering of the sun’s light. This accounts for the variety of shades of red in the evening sky. A particularly striking example of this occurred following the eruption of Mount Krakatoa in 1883. The dust from the eruption (some 20 km^3 of rock fragments were hurled up to 80 km into the atmosphere) resulted in spectacularly coloured sunsets around the world throughout the following year.

How does the eye see colour?

There are three different types of cone cells in the retina of the eye which allow colour vision. Each of these cells contains a different pigment labelled blue, green and red, as in the primary colours. Each cone absorbs over a range of wavelengths so there is considerable overlap and one wavelength of light may be absorbed by one, two or three different types of cone.

About 8% of the population are colour-blind, that is they are unable to distinguish red – green colours. Most colour-blind people have three different cone pigments but the ranges of wavelengths absorbed by the cones do not overlap sufficiently to enable them to respond to different hues. A minority of colour-blindness is due to the presence of only two types of pigment in the cone cells.

What causes after images and what colour are they?

If you stare at a bright colour for fifteen seconds and then look at a white wall you will see the after image in the complementary colour. The light coming from the colour tires that part of the retina of the eye sensitive to the colour so, on looking at the white wall, the part of the retina which is not tired dominates and the complementary colour is seen.

If black is a good absorber of heat why do Bedouins wear black robes in the hot desert?

The hotter black robe warms the air inside the robe by about six degrees Celsius more than a white robe would. This hot air rises and passes out through the fabric, with cooler air coming in at the bottom to take its place. This creates a continuous draught, due to convection, passing over the body. This cools the body, making a black robe more comfortable to wear than a white one.

Why is the filament of an electric light bulb coiled?

The filament is coiled to reduce heat loss, the higher the temperature the whiter the light. If the filament is uncoiled then it will only glow dimly.

Why is a firefly a more efficient light producer than a filament lamp?

The firefly produces 'cold' light so no energy is wasted in the production of heat, as occurs in the filament lamp. (This also explains why fluorescent lamps are more efficient than filament lamps.) The firefly is 100% efficient; one photon is emitted for each molecule oxidised.

Why is sunburn more likely to occur if you are up a high mountain or at the beach?

On a high mountain the path travelled by the ultraviolet radiation through the atmosphere is reduced, so the amount absorbed by the atmosphere is also reduced. This gives a greater concentration of ultraviolet radiation than at ground level and therefore more sunburn.

On the beach the ultraviolet radiation is reflected by the sand, so this again gives a greater concentration and more sunburn.

How do photosensitive sunglasses work?

The glass contains small crystals of a chemical, e.g. silver bromide. When the light falls on the glass the silver ions are converted to silver atoms which then darken the glass. In the dark the silver atoms will recombine and again form the colourless ions. The photosensitivity is thus due to chemical reactions occurring in the glass.

Glass lenses are used to focus light waves in the microscope. What is used to focus electron waves in the electron microscope?

As electrons are negatively charged the beam of electrons can be manipulated by electric and magnetic fields. Electric and magnetic fields are used to focus the electron waves in the electron microscope.

2.4 Conceptual Approach

Diffraction of Light

Diffraction causes the spreading out of waves which occurs after waves pass through a narrow opening in an obstacle or around the edge of an obstacle. It is an intrinsic property of waves.

Diffraction of light means that light can travel around corners but the angle through which it deviates from a straight line is very small except when the opening is very small, i.e. of the same order of magnitude as the wavelength of light (4×10^{-7} – 7×10^{-7} m). In most cases in everyday life we can say that light travels in straight lines.

Interference of Light

Two coherent sources of light can produce interference fringes. To achieve two coherent sources a single source illuminates two slits, now called Young's slits, Fig. 2.2. As a result of diffraction at the two slits a wave spreads out from each of them. Since these waves were originally part of a single wavefront the two waves spreading out from S_1 and S_2 are of the same frequency and in phase leaving the slits, i.e. S_1 and S_2 act as a pair of coherent sources.

When the waves from the two sources arrive at a point in phase, they will reinforce each other and brightness occurs (constructive interference). This will occur when the difference in the path lengths from each of the two sources to the point is zero or a whole number of wavelengths ($n\lambda$, where n is an integer and λ is the wavelength). When the waves from the two sources arrive at a point out

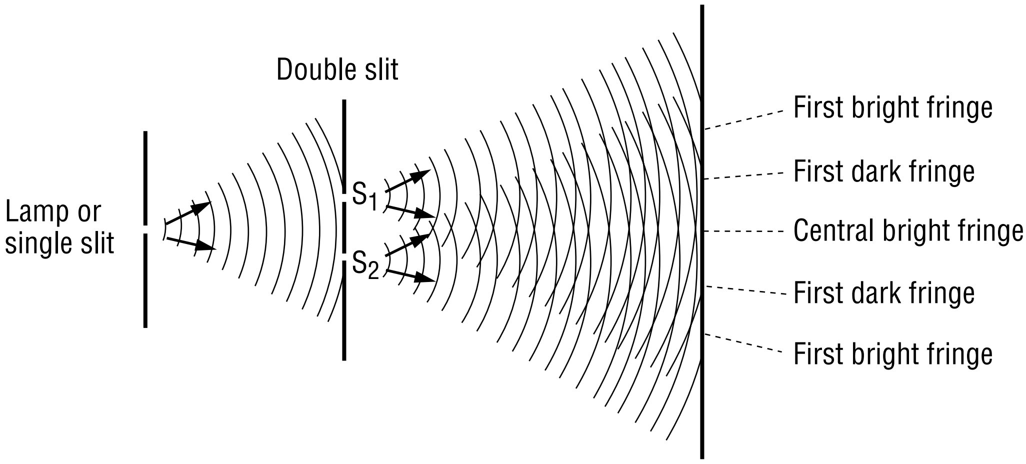


Fig. 2.2 *Young's slits*

of phase, a crest from one wave will coincide with a trough of the second wave; they cancel each other out and darkness results. This occurs when the difference in the path lengths from each of the two sources to the point is an odd number of half wavelengths $(2n - 1)\lambda/2$.

Interference in light was first demonstrated by Thomas Young in 1801. He allowed a narrow beam of light to fall on two parallel slits and observed a pattern of light and dark lines, or fringes. This showed that light travels by waves as had been suggested earlier by Christian Huygens (see p. 26).

The Diffraction Grating

The first grating was invented by Joseph von Fraunhofer and consisted of a series of parallel wires about 0.005 cm apart with the spaces between the wires acting as slits. School diffraction gratings are usually pieces of transparent plastic which have been cast or pressed from master plates, carefully engraved with parallel grooves. Although they are transparent the grooves, or lines, make them act as if they consisted of an array of opaque obstacles with transparent slits between them. Light waves passing through the lines on the grating are diffracted and then overlap, forming an interference pattern. The position of an interference maximum (constructive interference – brightness) is given by $n\lambda = d \sin \theta$, where n is an integer (called the order of the maximum), λ is the wavelength of the light, d is the distance between the lines on the grating, and θ is the angular displacement of the maximum from the normal to the grating. If, as is usual, d is very small (of the order of μm) it follows that $\sin \theta$ must have a large value for constructive interference to occur. This means that θ must be large, so the bright fringes will be well separated from each other. Because of the large number of lines on the grating there is destructive interference over most of the region and so the interference maxima are much sharper and brighter than in, for example, Young's slits. The angle θ between the bright fringes can be easily and accurately measured. Since d is known from the grating λ can then be calculated. Note that, for each value of n except 0, there are two bright fringes formed, one on either side of the normal, Fig. 2.3.

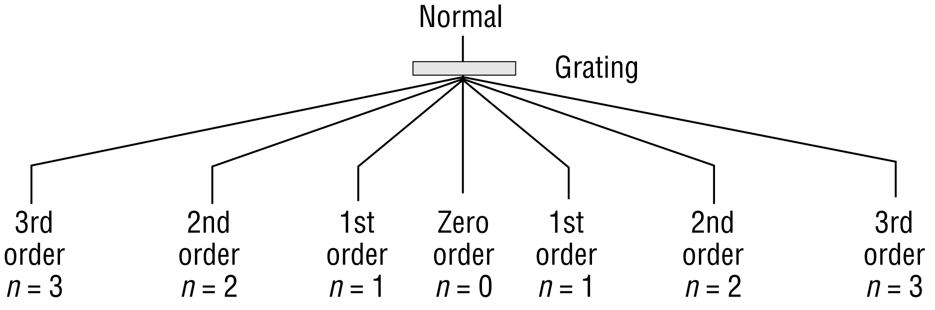


Fig. 2.3

When white light is used the direction for constructive interference is different for each wavelength, and so the interference maximum for each colour is formed at a different angle θ to satisfy the formula $n\lambda = d \sin \theta$. In other words the white light is dispersed; a spectrum is formed, Fig. 2.4. Spectra are formed for each value of n except 0. Since the zero order is formed on the normal all wavelengths are in phase and there is constructive interference for all wavelengths, so white light is formed. Mathematically, if n is zero θ is also zero for all values of λ. In each spectrum the violet is formed closest to the zero order as violet light has the shortest wavelength and so the smallest angle for constructive interference.

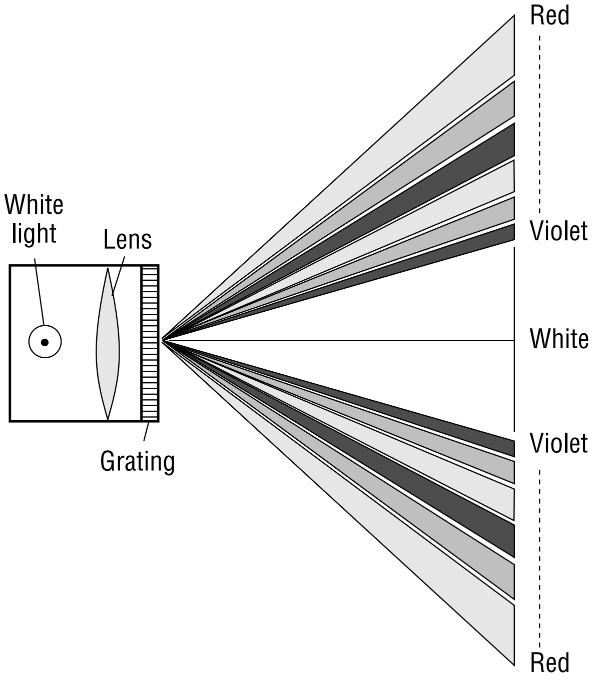


Fig. 2.4 *Dispersion using a diffraction grating*

Interference Colours

Very thin films of transparent colourless materials display bands of various colours in white light. These colours come from the interference of the light waves reflected from the front and back surfaces of the thin transparent film. The film thickness is of the order of the wavelength of the light. The incident light ray will be both reflected and refracted at the front surface, the refracted ray crosses the thin film and will be reflected and refracted at the back surface, Fig. 2.5. The reflected ray will again cross the thin film and be reflected and refracted at the front surface. The light initially reflected from the front of the film and the light finally refracted at the front surface will reach the eye.

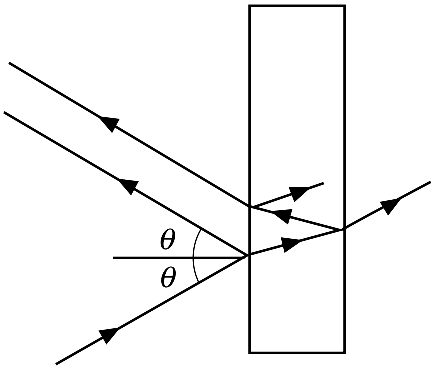


Fig. 2.5 *Interference in a thin film*

If they are in phase for a particular wavelength (colour) then a bright region in that colour occurs. If they are out of phase a dark region results or, if the thickness of the film is suitable, a bright region in another colour is produced.

In a vertical soap film or a bubble, the thickness of the film varies from very thin at the top to thicker at the bottom due to gravity. The top is so thin that destructive interference occurs and it appears dark. Lower down the colour will depend primarily on the wavelength at which the reflected light undergoes constructive interference, which in turn depends on the thickness of the film. At the bottom the fringes become progressively narrower and colours begin to overlap and fade.

Transverse Nature of Light Waves

Are light waves longitudinal or transverse? Transverse waves can be polarised but longitudinal waves cannot be polarised. If light waves can be polarised this would verify that light is a transverse wave.

Polarisation

Light is emitted when the outer electrons of atoms, which were excited to higher unstable energy levels, return to their original levels. Each atom acts as an independent light source so the total light given off consists of many independent waves whose planes of vibration are randomly oriented, Fig. 2.6. This light is called unpolarised light.

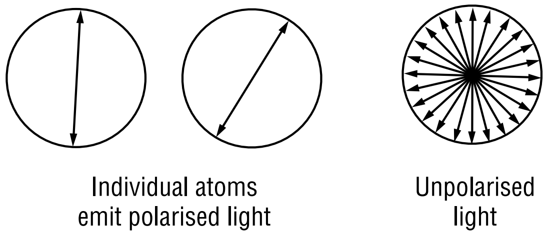


Fig. 2.6

Unpolarised light can be converted to polarised light by passing it through a polariser sheet. A polariser sheet can be made by embedding long chain molecules in a flexible plastic sheet. The sheet is then stretched and this aligns the molecules parallel to each other.

When light, which is a form of electromagnetic radiation, falls on the sheet, the long molecules, acting like aerials, absorb the electric field vectors which are parallel to them and only the perpendicular field vectors pass through. The unpolarised light is now polarised but, as 50% of the light is absorbed, the intensity is halved.

This polarised light can be tested with a second polariser sheet. If the plane of polarisation is the same in both sheets then light passes through but if the plane of polarisation of one sheet is perpendicular to the plane of polarisation of the second, no light passes, Fig. 2.7.

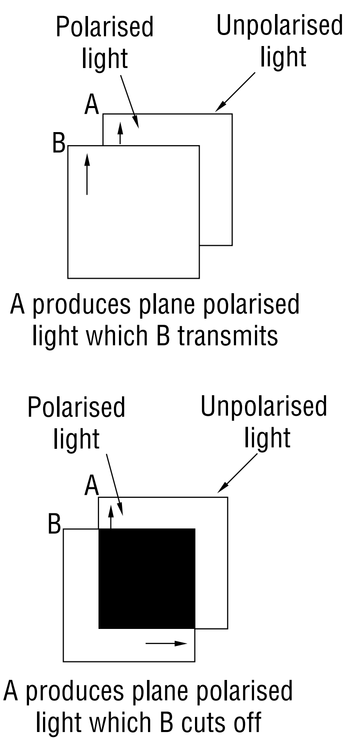


Fig. 2.7

Polarisation by Reflection

When unpolarised light is reflected by a surface it is found to be partially polarised. The electric field vectors which are parallel to the surface are reflected while those which are perpendicular are refracted into the medium or absorbed. (This can be compared to skimming stones off the surface of water. Only stones which hit the water with their surface parallel to the surface of the water are reflected and bounce.) In 1815 the Scottish physicist David Brewster (1781–1868) found that if the angle of incidence for glass is about 56° then the reflected ray is totally polarised and the reflected ray and the refracted ray are at right angles to each other, Fig. 2.8. The angle of incidence which gives rise to a totally polarised reflected ray is now known as Brewster's angle, B. From Fig. 2.8 it may be shown that $\tan B = n$, the refractive index of the glass. Thus, if $n = 1.5$ $B = \tan^{-1} (1.5) = 56.3°$.

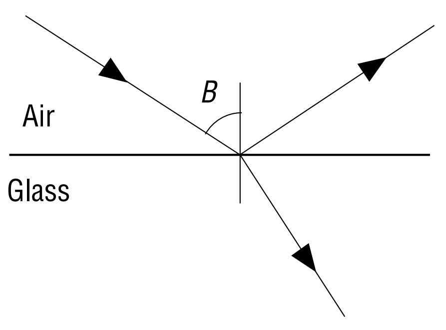


Fig. 2.8

2.5 Experimental Approach

Diffraction of Light

Facing a distant bright light hold up your left hand at arm's length with your first two fingers close together so that a narrow slit of light can pass through them. View this slit through the first two fingers of your right hand held close to the eye. Vertical bands of light much wider than the original slit will be seen. This shows that the light has been diffracted and that it is therefore a wave.

View a distant bright light bulb through a piece of cloth material, e.g. your shirt. The light bulb appears much larger due to diffraction occurring in all directions as the light passes through the spaces between the threads of the material.

Interference of Light

Cut two fine sharp slits about 0.5 mm apart in a piece of aluminium foil using a sharp scalpel. Hold the slits up close to the eye and view a bright illuminated white wall or screen, Fig. 2.9. A series of dark and bright fringes should be seen.

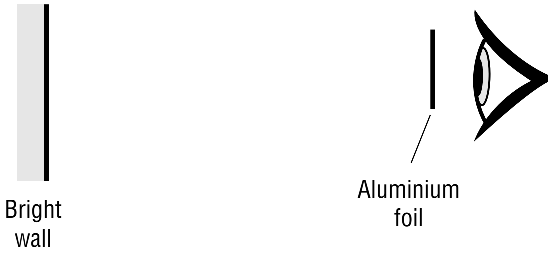


Fig. 2.9

A more accurate view of Young's fringes can be obtained using a spectrometer, a monochromatic (one wavelength only) light source such as a sodium lamp, and a Young's slits slide. Adjust the spectrometer for parallel light by focusing the telescope on a distant object. Illuminate the slit in the collimator with the sodium lamp. Place the slits slide in the holder on the table. Adjust the position and width of the collimator slit until the fringes are as sharp and clear as possible.

Instead of the sodium lamp and the spectrometer a laser may be used, with the fringes then being produced on a wall or screen. (See Chapter 3 for precautions to be observed when using a laser.) This method has the advantage that it may be done as a class demonstration.

If a white light source is used with the spectrometer instead of the sodium lamp coloured fringes should be observed. This is due to the different colours (wavelengths) being in phase at different angles (see p. 30).

Polarisation

Set up the apparatus as shown in Fig. 2.10. The lenses from Polaroid sunglasses may be used as polarising sheets.

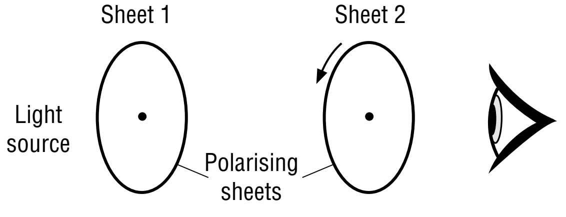


Fig. 2.10 *Demonstrating polarisation*

Rotate sheet 2 through 180° while observing the light source through both sheets. The intensity of the light should vary from a maximum (when the polarising planes of the sheets are parallel to each other) to zero (when the polarising planes are perpendicular to each other). This shows that light can be polarised so must therefore be a transverse wave.

This experiment can also be done very effectively as a demonstration on an overhead projector.

Polarisation by Reflection

Light, from a light bulb or other source, reflected by a polished bench surface will show polarisation, Fig. 2.11. Note: if a laser is used as a light source the reflected light must be directed onto a wall or screen. (See Chapter 3 for precautions to be observed when using a laser.)

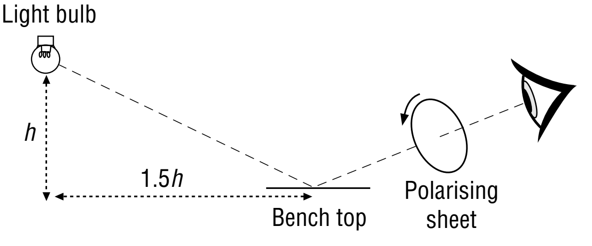


Fig. 2.11 *Light is polarised by reflection*

As the polarising sheet is rotated the intensity of the reflected light increases to a maximum and then reduces to a minimum. This shows that light is polarised by reflection.

Dispersion of Light

It has been known since antiquity that sunlight bent in a glass prism gives the colours of the rainbow. When light goes from air to glass it slows down and this causes the change in direction which we call refraction. The different wavelengths of light travel at different speeds in glass, the red light has the highest speed and the violet light has the lowest. It is this difference in speed which causes the separation of the colours, i.e. dispersion.

If a converging lens is held in the path of the dispersed light, Fig. 2.12, the colours recombine to give white light.

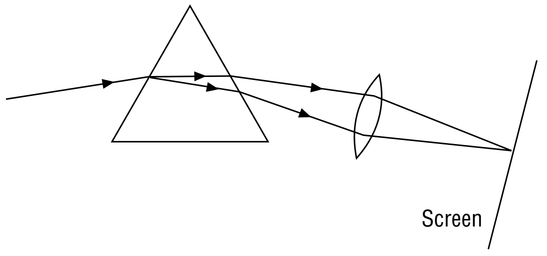


Fig. 2.12

A second prism, placed ‘upside-down’ compared to the first, will also recombine the dispersed light.

This dispersion of light occurs naturally in a rainbow. When white sunlight strikes a raindrop some of the light is refracted into the drop, this light is reflected at the back of the drop by total internal reflection and is again refracted as it emerges from the drop, Fig. 2.13.

To observe a rainbow we must be positioned between the sun and the falling rain, Fig. 2.14. In a rainbow the red light comes from drops that are observed at a larger angle with the horizontal than the drops giving the blue light. Every rainbow is unique to the observer. Another observer standing close by gets light from different drops.

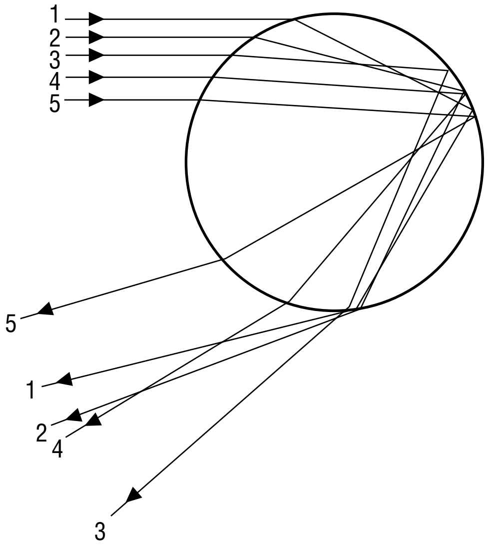


Fig. 2.13 *Refraction and reflection in a raindrop*

How We See a Rainbow

All the drops sending separated colours are angled at about 42° from a point directly opposite the sun, Fig. 2.14.

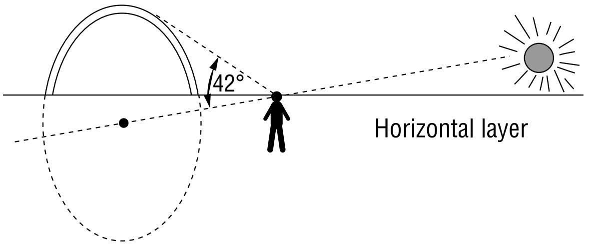


Fig. 2.14

If the sun is low in the sky, nearly at the horizon, a semicircular arc of colours can be seen.

The secondary bow which can sometimes be seen is due to the light rays being reflected twice in the raindrop before they are refracted out of the raindrop, Fig. 2.15.

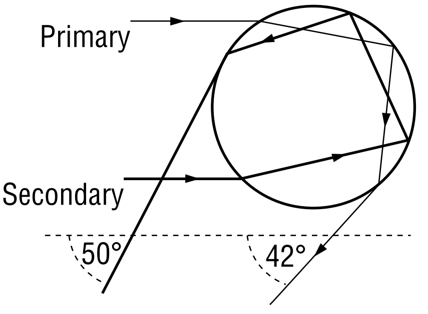


Fig. 2.15 *Formation of a secondary rainbow*

The loss in intensity on reflection accounts for the secondary bow being dimmer than the primary one.

A rainbow can be created using a garden hose with a fine spray on a sunny day. Remember to stand with your back to the sun !

Colour

White light can be created using just three colours, red, blue and green. These three colours are called the primary colours.

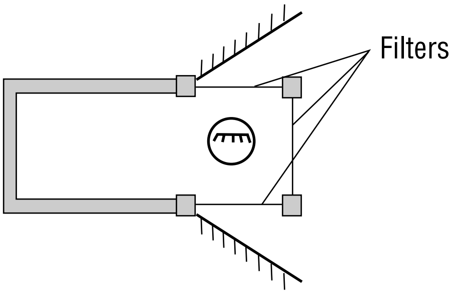


Fig. 2.16 *Colour mixing*

A lamp box with two mirrors on the sides can be used to demonstrate that the three primary colours give white light, Fig. 2.16.

Three filters, red, blue and green, are placed in a lampbox as shown in the diagram. Adjusting the positions of the wing mirrors reflects the side colours into the path of the front colour. Where two primary colours overlap a secondary colour is formed. This arrangement gives two of the three secondary colours. Further adjustment of the mirrors cause the three colours to overlap and white light should be formed. Changing the relative positions of the filters will give the third secondary colour.

This apparatus also demonstrates complementary colours, i.e. the formation of white light by the combination of a primary and a secondary colour. First mix two primary colours, by adjusting the appropriate mirror, to get a secondary colour, then add the third primary colour to the secondary to get white light. The colour triangle, Fig. 2.17, may be used as an aid to remembering the different colours.

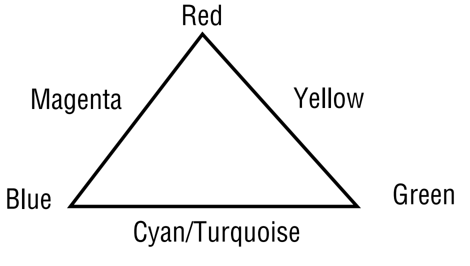


Fig. 2.17

Primary colours are at the apexes. Secondary colours are on the sides. Complementary colours are at an apex and the opposite side.

Jordon's colour-mixing apparatus can also be used to demonstrate primary, secondary and complementary colours.

Television and the films used in colour photography use the primary colours to give the whole range of colours which we see.

Paints and dyes work by reflecting light and this explains the different primary colours used in art. When white light is shone on a red surface, the dye absorbs nearly all the light except the red, which is reflected. If blue light is shone on this red surface it will appear very dark, nearly black, as the blue is absorbed and little light is reflected.

About 8% of the population are colour-blind. It is a sex-linked gene and only boys should be affected. Colour charts can be used to check for colour-blindness.

In conclusion, colour depends on four factors, viz. the incident light, the reflecting object, the eye and the brain, and so ultimately colour is a phenomenon of perception.

Electromagnetic Spectrum

The general properties of electromagnetic radiations may be listed as follows.

All are produced by accelerating electric charges.

All can travel through vacuum.

All travel at the same speed in vacuum.

All show the wave effects of interference and diffraction.

All are transverse waves and can be polarised.

Reflection at surfaces, refraction at boundaries, focusing by lenses and mirrors are shown by radio waves, microwaves, infrared radiation, light, and ultraviolet radiation, due to the fact that their wavelengths are long relative to the separation of atoms in solids.

A full description of the electromagnetic spectrum is given in the Appendix, p. 42.

If a mercury thermometer with a blackened bulb or a thermocouple thermometer is placed beyond the red end of a spectrum formed by a glass prism, Fig. 2.18, it shows an increase in temperature. This indicates the presence of infrared radiation.

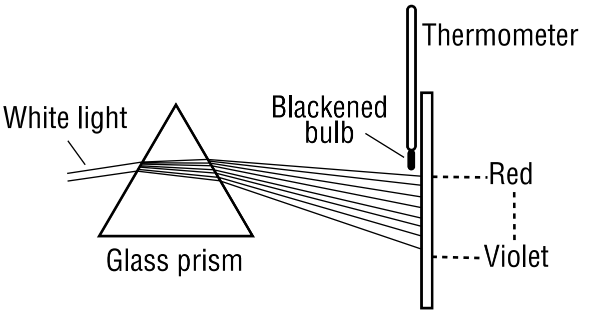


Fig. 2.18 *Detecting infrared radiation in a spectrum*

If the glass prism is replaced by a quartz prism and a white cloth, which has been washed in a modern detergent, is held just beyond the violet end of the spectrum, Fig. 2.19, it will glow due to fluorescence. This shows the presence of ultra-violet radiation.

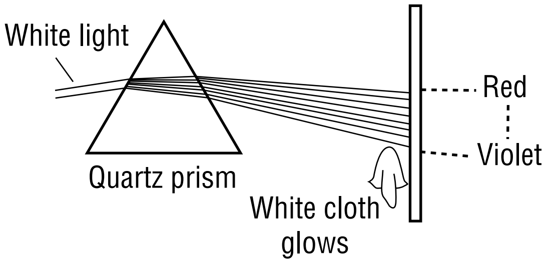


Fig. 2.19 *Detecting ultraviolet radiation in a spectrum*

The ultraviolet radiation (u.v.) is absorbed by the detergent and converted to light. Vaseline, starch and fluorescent paints will all glow in ultraviolet radiation.

Most of the electrical energy supplied to an incandescent light bulb is radiated as infrared radiation; only a small percentage of the electrical energy is converted to useful light. Fluorescent lamps are much more efficient since they do not produce as much infrared radiation.

2.6 Applications

Lenses

A thin film, used as an interference film, can reduce the amount of reflected light so increasing the transmitted light in the lenses of optical instruments. The lenses are covered with an antireflection coating by depositing a transparent layer, about one quarter wavelength thick, of calcium or magnesium fluoride on the surface of the lens. This layer reduces reflection by destructive interference, as explained on p. 11.

Windows

A thin film may increase reflection of certain wavelengths, hence reducing transmission, as in windows coated to reflect infrared while transmitting light. These windows help to keep the room warmer in winter and cooler in summer.

Butterflies

The iridescence seen in the top surface of the morpho butterfly wings is due to constructive interference of the reflected light by thin terraces of transparent cuticle-like material in the wings.

Television Aerials

As transmitted television signals are polarised, aerials must be set up parallel to the plane of polarisation of the signal. All aerials directed at a given transmitter should therefore have the same orientation, i.e. either vertical or horizontal.

Polaroid Sunglasses

Polaroid absorbs 50% of all unpolarised light so the intensity is halved. When light is reflected off a horizontal surface it is partially polarised horizontally (see p. 32). The glare of reflected light can be eliminated if the polarising plane of the sunglasses is perpendicular to the plane of the reflected light. To achieve optimum results the polarising plane in sunglasses should therefore be vertical.

Strains in Structures

Certain materials, such as perspex, have the property of rotating the plane of the polarised light passing through them when they are under stress. If a model of a structure is made in perspex and placed between crossed Polaroids (two pieces of Polaroid set with their polarising planes at right angles to each other) the transmitted light will indicate stress patterns. Engineers can find this a useful tool in the design of structures.

Optically Active Sugars

Optically active sugars rotate the plane of polarised light. This can be used to calculate the concentration of sugar solutions since the angle of rotation is proportional to the concentration. Other liquids, e.g. turpentine, also show optical activity.

Liquid Crystal Displays (LCDs)

These are used in watches, calculators and lap-top computers. Each segment of a display consists of a liquid crystal sandwiched between two polarising sheets and placed on a reflecting surface, Fig. 2.20. The polarising sheets are arranged so that their polarising planes are perpendicular to each other. The liquid crystal has properties intermediate between those of a solid crystal and of a liquid. The molecules of the crystal are rod-like. The molecules in the top layer of the crystal are arranged parallel to the polarising plane of the top polariser, while those in the bottom layer are parallel to the polarising plane of the bottom polariser. Thus, the plane of polarisation of light passing through the top polariser is gradually rotated through 90° so that it can pass through the bottom polariser.

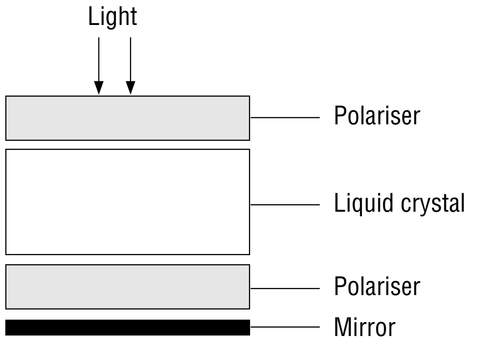


Fig. 2.20 *Liquid crystal display*

When no voltage is applied light travels through the first polariser, the liquid crystal, the second polariser, is reflected at the mirror and travels back out along its original path. When a voltage is applied between the top and bottom of the crystal the molecules line up parallel to the electric field and so no longer rotate the plane of the polarised light through 90°. The light cannot now pass through the second polariser since its plane of vibration is no longer parallel to the polarising plane of the polariser. So, while the electric field is applied to the crystal no light is reflected out of it and it appears dark.

Each digit in a display consists of seven segments. A particular digit is displayed by applying an electric field to the appropriate segments, e.g. a zero is displayed by applying a voltage to the six outer segments. LCDs use very little power as ambient light is used to read the display and the liquid crystal is a good insulator.

Gemstones

Gemstones have high refractive indices. Diamond has a refractive index about 1.5 times that of glass. This property is used by the diamond cutter, who cuts the diamond to ensure that it acts as a prism, splitting light into a spectrum. This spectrum then undergoes total internal reflection giving sparkle to the diamond, Fig. 2.21.

The largest diamond ever, was found in a mine in South Africa in 1905. It measured 11.5 cm by 6.5 cm and was called the Cullinan diamond. It was cut into nine major gems and ninety-six smaller ones which all form part of the British Crown Jewels. Real diamonds, in contrast to artificial ones, fluoresce in ultraviolet radiation.

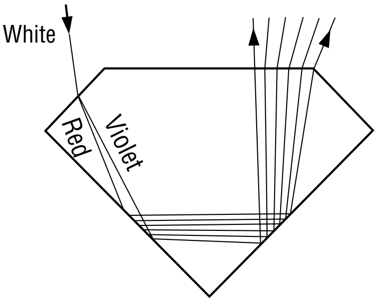


Fig. 2.21 *Reflection and refraction in a diamond*

Colour Photography

Colour film consists of three layers of emulsion on top of each other. Blue light affects the top layer, green light affects the middle layer and red the bottom layer. During development the top layer becomes yellow, the complementary colour of blue, the middle layer becomes magenta and the bottom layer cyan. The printing paper also has three layers, each sensitive to one colour, e.g. the yellow image gives blue, etc., to give a faithful reproduction of the original colour.

Television

In the television camera the film is a screen which is made up of a regular array of thousands of very small light-sensitive electronic devices; when exposed to light they convert light energy to electrical energy. The camera scans each of these picture cells, or pixels, in the order in which the eye scans the print in a book, line by line from left to right. The signal from each pixel is sent down the cable in turn to give an electrical code which represents the picture on the screen. The screen is scanned 25 times a second and the coded signal, a varying electric current, is used to modulate a carrier signal which is recorded or broadcast live.

This modulated signal is picked up by an aerial and fed into the cathode ray tube in the television set. An electron gun fires electrons at the screen ; the number of electrons striking a particular point on the screen is controlled by the original signal. Where the electrons strike the screen it fluoresces and the picture is formed from the series of bright and dark spots. The screen is scanned 25 times a second to give an exact replica of the original picture.

In the colour television camera three separate images are projected through red, blue and green filters onto three photosensitive layers, each of which gives a signal corresponding to the red, blue, or green component of the original picture. These separate signals are fed to three electron guns in the cathode ray tube of the colour television. The screen contains millions of phosphor dots which give red, blue or green light when struck by an electron. The dots are arranged in such a way that only the 'red' dots are struck by electrons from the 'red' electron gun, and similarly with the other two guns. The intensities of the electron beams are determined by the signals received from the camera. (The intensities may also be varied by the controls on the television.) The colours on the screen are therefore an exact replica of the original picture.

Ultraviolet Radiation

Ultraviolet radiation causes tanning and sunburn of the skin, see Appendix, p. 44.

A sun lamp is a convenient source of ultraviolet radiation. It is made entirely from quartz glass so that the ultraviolet radiation can escape (ordinary glass is opaque to most ultraviolet radiation). The lamp can be converted to a source of visible light only, by surrounding it with an envelope made of ordinary glass.

The lamp can also be converted to a source of 'black light' by surrounding it with an envelope of dyed quartz. Only the ultraviolet radiation will escape. This can be used to produce theatrical effects, as only fluorescent materials will be visible; a white sheet washed in modern detergent could give the ghost.

Ultraviolet radiation can also be used to test eggs; fresh eggs have a reddish colour while stale ones are blue/violet.

A further application of ultraviolet radiation is to produce 'invisible' signatures/security marks on, for example, bank deposit books and banknotes. The signatures are produced in an ink which does not reflect light but which fluoresces under ultraviolet radiation.

The Ozone Layer

The ozone layer lies in the region of the atmosphere known as the stratosphere. It is very beneficial as it absorbs harmful ultraviolet radiation. The ozone layer can be up to 50 km thick and it only allows radiation of wavelength greater than 290 nm to reach the earth's surface.

Ozone is also produced by the burning of petroleum in internal combustion engines. Under the extreme conditions in the cylinder the nitrogen is oxidised to nitrogen monoxide.

$$N_2 + O_2 \longrightarrow 2NO$$

The nitrogen monoxide is expelled through the exhaust to the atmosphere where it is immediately oxidised in the air to nitrogen dioxide.

$$2NO + O_2 \longrightarrow 2NO_2$$

Sunlight splits up the nitrogen dioxide.

$$NO_2 + \text{light} \longrightarrow NO + O$$

The free oxygen atom can react with another oxygen molecule to form ozone.

$$O + O_2 \longrightarrow O_3$$

Ozone is a toxic gas with a chlorine-like odour which causes severe respiratory problems. As it is highly reactive it is both a component of, and a contributor to, the formation of smog.

Chlorofluorocarbons, or CFCs, are inert non-toxic gases which have been used as cooling fluids in refrigerators and air conditioning systems as well as propellants in aerosol spray cans. When released to the atmosphere CFCs move slowly upwards on air currents into the stratosphere. Here they react with the ozone in a free radical chain reaction, so one CFC molecule is capable of destroying 100 000 ozone molecules. A 1% decrease in the ozone gives a 2% increase in the ultraviolet radiation reaching earth. This can cause a 4% to 6% increase in some types of skin cancer.

The stratosphere over the latitude band that includes Dublin, Moscow and Anchorage had about 8% less ozone in January 1986 than it had in January 1969 and this has been attributed to the effect of CFCs. Even with an immediate ban on CFCs it will take centuries for the atmosphere to recover from the effects of them. Ironically, the ozone layer was not in the early atmosphere and the intense ultraviolet radiation may have been instrumental in catalysing the early chemical reactions which led to the formation of primitive life.

Infrared Radiation

As near infrared will affect special photographic plates and will pass through glass it can be used to take photographs. It penetrates mist and haze much better than visible light and so is of use in meteorology.

Infrared searchlights – a white searchlight covered with dyed plastic sheet to eliminate the visible light – and telescopes were developed during the 1939–1945 war. The telescope contained photocells which converted the infrared to electrical currents which in turn were converted to light as in a television. Today, infrared binoculars (‘night sights’) are used in military and security applications.

Infrared radiation can be used to speed up the healing process in injured muscles. The infrared will penetrate to the cells where the energy stimulates the chemical reactions of the cells, accelerating their natural healing powers.

Greenhouse Effect

Solar radiation reaching earth has its maximum intensity in the visible region (wavelength 400 to 700 nm) of the electromagnetic spectrum. This radiation passes through the atmosphere and is absorbed by the earth’s surface. It is then re-emitted as infrared radiation with wavelengths some 20 times greater than the incident radiation. This emitted radiation has frequencies which correspond with the natural frequencies of vibration of the carbon dioxide, methane, water and ozone molecules. These molecules absorb the radiation and re-emit it in all directions, including back to the earth’s surface.

The net effect is that the atmosphere effectively traps some of the sun’s energy and prevents the earth from cooling rapidly at night. This effect results in an average increase of over 30 °C in the earth’s temperature; without it the average temperature would be about a chilly –10 °C.

This is the same principle as the greenhouse. Visible radiation enters through the glass, is absorbed and emitted as long wavelength infrared which cannot pass out through the glass. The greenhouse is a radiation trap.

THE IRISH TIMES – Saturday, June 25, 1994

WEATHER EYE

Seeing through the fog

SEA fog is a common phenomenon at this time of year. It occurs when the pressure pattern is such as to bring warm moist air from low latitudes up northwards over cold seas that have not yet had time to absorb the summer heat. As the moving air is cooled by contact with the cold waters underneath, its moisture condenses into tiny water droplets — and the result is fog.

It is difficult to forecast fog at sea, not least because weather observations over the ocean are few and far between. It helps, however, if the forecaster knows where fog exists already; not only is it a very obvious sign that conditions are suitable for its formation, but it also highlights adjacent danger areas — places currently clear but into which fog may shortly be advected by the wind.

Satellite pictures are useful for this purpose; two types are used in combination — those taken with instruments sensitive to radiation in the visible part of the electromagnetic spectrum, and those that react only to the infrared.

The visible picture records more or less what might be seen by an ordinary black and white camera. The clouds are brilliant white, because they reflect back to the satellite a large proportion of the sunlight. Fog shows up in the same way, and for this reason is often difficult to distinguish from the cloud. Land and sea, on the other hand, being less reflective, appear darker.

In the case of infrared images, the sensor on the satellite reacts to temperature rather than to visible light; objects at low temperatures appear white and those at high temperatures, black. During the day in summertime, for instance, the land is usually much hotter than the ocean, while on a winter’s night it may be much colder, and there is therefore often a stark contrast between land and sea on infrared images. Moreover, since high clouds are very much colder than low clouds, they appear as a much brighter shade of white, allowing the forecaster to distinguish between the two.

But let us return to fog. On a visible picture, fog stands out clearly in contrast to any land or sea in the vicinity, but may be indistinguishable in appearance from cirrus 20,000 feet above the ground — a feature of no immediate interest to the forecaster.

On an infrared image, however, fog can hardly be seen at all, because there is little or no contrast in temperature between it and the land or sea a few feet underneath. By comparing the two pictures, the forecaster can easily distinguish fog at sea for what it is.

Brendan McWilliams

Gases that can absorb the infrared radiation, preventing its escape from the earth's surface, are known as greenhouse gases. The excessive trapping of infrared by increased concentrations of certain atmospheric gases is known as the greenhouse effect.

The extreme example of this effect is on the planet Venus. The atmosphere on Venus is very dense, with a high concentration of carbon dioxide. The resulting greenhouse effect has produced an average temperature on the surface of Venus of about 500 °C.

The carbon dioxide content of the Earth's atmosphere is increasing due to the burning of fossil fuels. Methane (natural gas) is also produced in the anaerobic treatment of sewage and rubbish where the bacteria break down the sewage, producing methane. Peat bogs and rice paddies also generate methane and our cattle give off large quantities of the gas (a cow fed on grass produces 350 litres per day) which, as it is difficult to trap, adds daily to the greenhouse effect.

Scientists fear that the increase in greenhouse gas concentration in the atmosphere could lead to climatic changes, with greater melting of the ice caps and resulting flooding in low-lying countries.

2.7 Worked Examples

1. Monochromatic light of wavelength 6.0×10^{-7} m was passed through a diffraction grating having 400 lines per mm. Find the angles of deviation of the first, second and third order maxima.

   Formula $\quad n\lambda = d \sin\theta$

   d = the distance between slits

   $$= \frac{1}{400} \text{ mm}$$
   $$= \frac{1}{400} \times 10^{-3} \text{ m}$$
   $$= 2.5 \times 10^{-6} \text{ m}$$

   λ = wavelength of light = 6.0×10^{-7} m

   First order maximum, $n = 1$

   $$\sin\theta = \frac{n\lambda}{d}$$
   $$= \frac{(1 \times 6.0 \times 10^{-7})}{(2.5 \times 10^{-6})}$$
   $$= 0.24$$
   $$\theta = 14°.$$

   Second order maximum, $n = 2$

   $$\sin\theta = \frac{n\lambda}{d}$$
   $$= \frac{(2 \times 6.0 \times 10^{-7})}{(2.5 \times 10^{-6})}$$
   $$= 0.48$$
   $$\theta = 29°.$$

   Third order maximum, $n = 3$

   $$\sin\theta = \frac{n\lambda}{d}$$
   $$= \frac{(3 \times 6.0 \times 10^{-7})}{(2.5 \times 10^{-6})}$$
   $$= 0.72$$
   $$\theta = 46°.$$

2. What is the wavelength of light which is deviated in the first order through an angle of 20° by a diffraction grating having 6000 lines per cm? What is the angle of the second order deviation?

   Formula $\quad n\lambda = d \sin\theta$

   d = distance between slits

   $$d = \frac{1}{6000} \text{ cm}$$
   $$= \frac{1}{6000} \times 10^{-2} \text{ m}$$
   $$= 1.67 \times 10^{-6} \text{ m}$$

   First order angle = 20°, $n = 1$

   $$\sin\theta = \sin 20°$$
   $$= 0.3420$$

$$\lambda = \frac{d \sin\theta}{1}$$

$$= 1.67 \times 10^{-6} \times 0.3420$$

$$= 5.71 \times 10^{-7} \text{ m}$$

Wavelength = 5.7×10^{-7} m.

Second order angle

$$\sin \theta = \frac{n\lambda}{d}$$

$$\sin \theta = \frac{(2 \times 5.71 \times 10^{-7})}{(1.67 \times 10^{-6})}$$

$$= 0.684$$

$$\theta = 43°.$$

3. What is the longest wavelength that can be observed to the fourth order maximum through a diffraction grating of 5000 lines per cm?

Formula $n\lambda = d \sin \theta$

d = distance between slits

$$d = \frac{1}{5000} \text{ cm}$$

$$= \frac{1}{5000} \times 10^{-2} \text{ m}$$

$$= 2 \times 10^{-6} \text{ m}$$

Maximum value of θ is 90°

Maximum value of sin θ is 1

Fourth order maximum, n = 4

$$\text{maximum } \lambda = \frac{d \sin \theta_{(\text{max})}}{n}$$

$$= \frac{(2 \times 10^{-6} \times 1)}{4}$$

$$= 5 \times 10^{-7} \text{ m}$$

Maximum wavelength which can be observed

$$= 5 \times 10^{-7} \text{ m.}$$

4. The atomic hydrogen spectrum is a line spectrum. One line is of wavelength 4.10×10^{-7} m and another is 6.56×10^{-7} m. Calculate the angular separation, in degrees, of the second order maximum of each of these lines when the light is passed through a diffraction grating having 4000 lines per cm.

Formula $n\lambda = d \sin \theta$

d = separation of the slits

$$d = \frac{1}{4000} \text{ cm}$$

$$= \frac{1}{4000} \times 10^{-2} \text{ m}$$

$$= 2.5 \times 10^{-6} \text{ m}$$

Second order maximum, n = 2

Wavelength, $\lambda = 4.1 \times 10^{-7}$ m

$$\sin \theta = \frac{n\lambda}{d}$$

$$= \frac{(2 \times 4.1 \times 10^{-7})}{(2.5 \times 10^{-6})}$$

$$= 0.328$$

$$\theta = 19.15°$$

Wavelength, $\lambda = 6.56 \times 10^{-7}$ m

$$\sin \theta = \frac{n\lambda}{d}$$

$$= \frac{(2 \times 6.56 \text{ x } 10^{-7})}{(2.5 \times 10^{-6})}$$

$$= 0.5248$$

$$\theta = 31.65°$$

Angular separation

$$= 31.65° – 19.15°$$

$$= 12.5°.$$

5. A beam of monochromatic light of wavelength 6.0×10^{-7} m falls normally on a diffraction grating having 5000 lines per cm. A screen is placed 0.50 m from the grating. Find the distance between the two second order maxima on the screen.

This problem has two parts. First find the angular separation of the two maxima then use the mathematics of the triangle to find the linear separation of the maxima.

Formula $n\lambda = d \sin \theta$

d = separation of the slits

$$d = \frac{1}{5000} \text{ cm}$$

$$= \frac{1}{5000} \times 10^{-2} \text{ m}$$

$$= 2 \times 10^{-6} \text{ m}$$

Second order maximum, $n = 2$

$$\sin \theta = \frac{n\lambda}{d}$$

$$= \frac{(2 \times 6 \times 10^{-7})}{(2 \times 10^{-6})}$$

$$= 0.6$$

$$\theta = 36.9°$$

The maxima are formed on the screen at either side of the central zero order maximum as shown in Fig. 2.22.

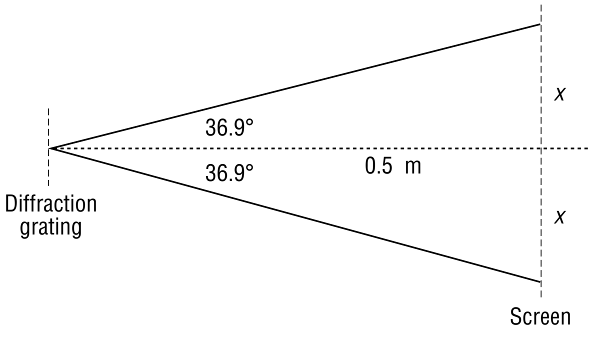


Fig. 2.22

$$\text{Tan } 36.9° = \frac{x}{0.5}$$

$$x = 0.5 \times \tan 36.9°$$

$$x = (0.5 \times 0.75)$$

$$x = 0.375 \text{ m.}$$

The separation of the second order maxima is

$(2 \times 0.375) = 0.75$ m.

2.8 Student Experiment

Measurement of the Wavelength of Monochromatic Light

Detailed instructions for carrying out the adjustments to the spectrometer are given in Chapter 3.

Start with the telescope in the 'in line' position. Then slowly move the telescope to the left, counting the number of bright images. Focus the telescope on the last image which can be seen clearly and record the order of the image and record the angular position of the telescope accurately using the vernier scale. Move the telescope slowly to the right, recording the order and position of each image as far as the last distinct image on the right.

Suppose θ_1 and θ_2 are the angular positions of the two first order images. The angle between these two positions is twice the actual angle between the zero order and first order images.

$$\theta = \frac{(\theta_1 - \theta_2)}{2}$$

The wavelength of the light can be calculated using the formula

$$n\lambda = d \sin \theta.$$

$n = 1$ as this is the first order image, d is the grating constant.

The calculation is repeated for the other order images and an average value is found for the wavelength of the light.

APPENDIX

1. Electromagnetic Radiation

Light travels as a wave through vacuum or through a transparent medium. If light is a wave which can travel through a vacuum what oscillates? The electric and magnetic fields oscillate.

It had been known from the work of Michael Faraday (1791–1867) and others that an oscillating electric charge has oscillating electric and magnetic fields associated with it. In the middle of the nineteenth century the Scottish physicist James Clerk Maxwell (1831–1879) showed that an accelerating charge radiates energy. This energy will spread out from the source (the oscillating charge) through the creation of oscillating electric and magnetic fields, i.e. electromagnetic fields. An electric charge or a small magnet in the path of an electromagnetic wave will experience a force and will be forced to oscillate.

The electromagnetic wave travels at a unique speed in vacuum. The frequency of the wave is characteristic of the source of the wave. The wavelength and velocity of the wave depend on the medium through which it is travelling. The three properties, frequency, wavelength and velocity are related by the equation:

$$\text{velocity} = \text{frequency} \times \text{wavelength}.$$

Electromagnetic waves are produced in a range of frequencies which constitutes the electromagnetic spectrum. The spectrum is divided into different regions on the basis of the source and/or the frequency of the radiation. In order of increasing frequency the regions of the electromagnetic spectrum are: radio waves; microwaves; infrared; visible; ultraviolet; X-rays; gamma rays.

Radio Waves

These waves were discovered by the German physicist Heinrich Hertz (1857–1894).

A radio transmitter causes electrons to oscillate in the broadcasting antenna, i.e. an alternating current flows in the antenna. The basic electromagnetic wave is of fixed frequency and is called the carrier wave. It is modified (modulated) by the varying electric currents fed in by the studio microphones (see module on Waves, p. 2). These electromagnetic waves (radio waves) spread out from the transmitting antenna and may, depending on their frequency, be reflected around the earth by the ionosphere.

When the wave reaches a receiving antenna, which must be parallel to the transmitting antenna (see p. 35), the oscillating electric and magnetic fields cause minute oscillating currents in the antenna. As electromagnetic waves from all transmitting stations will reach the antenna and cause their own currents, a tuning circuit is necessary to select out the desired frequency. When the natural frequency of the tuning circuit is adjusted to that of the chosen carrier wave, resonance occurs and the desired signal is selected. The receiver separates the carrier wave from the audio frequencies which are then amplified and fed to the loudspeakers.

Microwaves

Microwaves are produced in a magnetron. The magnetron was the basis of the World War 2 radar system. It consists of a central cathode, an anode with eight circular chambers, an output loop of wire from one of the chambers and a large permanent magnet which produces a magnetic field at right angles to the plane containing the anode and cathode. The cathode is heated and electrons are emitted by thermionic emission. The electrons are accelerated towards the anode but the permanent magnet and the varying magnetic fields induced in the chambers of the anode cause the electrons to travel in four rapidly rotating shafts. The electrons are deposited on the anode causing pulses of current which induce magnetic fields within the chambers, which in turn induce electric fields. These electric fields are picked up by the loop which transmits them to the antenna which emits the microwave.

How do Microwave Ovens Cook?

Water, sugar and fat molecules all contain polar O—H bonds. The hydrogen atom has a slight positive charge while the oxygen has a slight negative charge. When microwaves strike the O—H group in the food the oscillating electric field causes the polar O—H group to line up with the field. In so doing bonds are broken, the energy being supplied by the microwave. New bonds will be formed immediately and the energy released will increase the temperature of the food, cooking the food.

Infrared Radiation

This radiation is emitted from all objects and absorbed by all objects. Hot bodies are better emitters than cool ones and cool bodies are better absorbers than hot ones. A perfectly black body is the best emitter and the best absorber of infrared radiation at a given temperature. The closest approximation to a black body is obtained by punching a small hole in a closed metal container. When the container is heated, radiation is emitted from the hole and the nature of this radiation is found to be independent of the material of the container; it depends only on temperature.

Visible Radiation (Light)

How is visible electromagnetic radiation produced? Light is produced by the outer electrons of the atoms in a material gaining energy and being promoted to higher energy levels where they are unstable. In returning to the ground state they must emit the excess energy, that is the exact energy difference, *E*, between the two energy levels. The energy is emitted in the form of electromagnetic radiation according to the equation

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

The emitted energy will therefore have a definite frequency and therefore a definite colour. If all frequencies are produced simultaneously, white light results (if the intensities of the various wavelengths are in the correct proportions).

Very hot objects (white hot) emit light. Examples include the sun and the hot filament of an ordinary light bulb. This is called incandescence.

Some living organisms such as bacteria, insects (e.g. firefly), fish and squid are bioluminescent. A chemical reaction involving oxidation occurs in their bodies and the product molecules are in an excited state and so are unstable. These molecules return to the ground state and in so doing emit the excess energy in the form of light. Animals use luminescence as a mating signal or as bait to attract the next meal.

Gas discharge tubes, e.g. household fluorescent light tubes, also emit white light but there are two steps in its production. Fluorescent tubes contain mercury vapour at a reduced pressure. An electric discharge through the vapour excites the outer electrons of the mercury atoms, these are then unstable and return to the ground state with the emission of ultraviolet radiation and a little light. The ultraviolet radiation is trapped, as glass is opaque to it. The second stage involves the absorption of this ultraviolet by the fluorescent material which coats the inside of the tube. The outer electrons in the atoms of the material are promoted to higher energy levels. Here they are unstable and return to the ground state with the emission of electromagnetic radiation in the visible region. As there is no high temperature involved a fluorescent lamp is much more efficient than a filament bulb.

Ultraviolet Radiation

Any object that is hot enough will emit ultraviolet radiation. The sun is a good emitter of ultraviolet. It is also emitted by the filament of an electric light bulb but it does not escape as glass is opaque to ultraviolet. Discharge tubes made of quartz and containing mercury, hydrogen or helium are also good sources of ultraviolet, as are arc and spark discharges.

Ultraviolet radiation results when electrons in atoms are promoted to much higher energy levels and then return to the ground state, emitting large amounts of energy, according to $E = hf = hc/\lambda$. Since *E* is large the radiation emitted has very high frequency and short wavelength.

Ultraviolet radiation is readily absorbed by many substances, causing them to fluoresce. Diamonds, paraffin oil, dayglow paints and some washing powders, fluoresce in ultraviolet radiation. In ultraviolet radiation natural teeth are bright white due to fluorescence while some artificial ones are dark.

As ultraviolet is absorbed by proteins and nucleic acids it will kill bacteria, and so is used to sterilise objects. Ultraviolet radiation can cause mutations in cells by damaging the DNA in the cell nuclei, resulting in skin cancers. Fair skin is particularly susceptible to damage by ultraviolet radiation so sun bathing and the use of sun beds must be very carefully regulated. Ultraviolet radiation tans light-coloured skin by oxidising a colourless pigment, it activates tyrosinase (an enzyme), which oxidises tyrosine (an aromatic amino acid which is found in proteins), and leads to the formation of brown or black pigments, melanins. The melanin protects the nuclei of skin cells by forming a layer over the cells which filters out the ultraviolet.

Sun blocks are creams which will absorb all the ultraviolet. Thus they provide no tanning but do protect sensitive skin. Sun filter creams will selectively absorb the shorter wavelengths which cause sunburning and allow the longer wavelengths through to tan the skin and promote the formation of vitamin D in the skin.

The ozone layer absorbs the very short wave-lengths of ultraviolet which cause serious damage to the skin resulting in skin cancers.

X-rays

X-rays are produced when electrons, moving very fast, strike a solid metal target, Fig. 2.23. An X-ray tube is an evacuated glass tube containing a filament, a cathode and a metal anode. Electrons are emitted from the hot filament and are accelerated onto the cooled target in the anode. The target is made of a metal with a high melting point, usually tungsten, and is set in a copper anode. X-rays are emitted from the area struck by the beam of high speed electrons and emerge from the tube through a window.

The spectrum of radiation emitted is a broad continuous spectrum containing distinct peaks of definite wavelength which are characteristic of the target material. The continuous spectrum is produced when the fast-moving electrons are slowed down and brought to rest in the target. An accelerated charge radiates energy as electromagnetic radiation, the wavelength of the

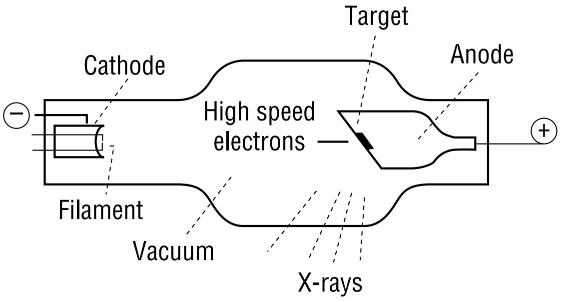


Fig. 2.23 *X-ray tube*

radiation depending on the magnitude of the acceleration. Since the acceleration of an electron depends on how closely it approaches a target nucleus a range of accelerations, and therefore a range of wavelengths, results. This radiation is known as bremsstrahlung, meaning 'braking radiation' in German.

The characteristic X-rays are produced by the high speed electrons knocking out electrons from the lowest energy levels in the target atoms. Other electrons from higher energy levels fall to these vacated levels and in so doing emit radiation of characteristic wavelengths.

It was through the study of the characteristic X-ray spectrum of elements that Moseley concluded that there was a fundamental quantity in atoms, more fundamental than atomic mass, namely the charge on the nucleus. This was called the atomic number and from this we have our present periodic table of the elements.

The medical use of X-rays depends on their considerable penetrating ability. They are readily absorbed by bone which contains the element calcium, while they pass through organic material which contains carbon, hydrogen and oxygen.

X-rays are also used to check for metal fatigue in the aircraft industry.

Gamma rays

These rays are emitted by the nuclei of atoms. A stable nucleus may be raised to an excited energy level in a number of ways, e.g. by the absorption of radiation, by capturing a neutron or by emitting an alpha or a beta particle. The unstable nucleus will return to the stable state by the emission of radiation of very short wavelength, gamma rays. Gamma rays are only slightly absorbed by matter and so are very penetrating. Short exposure to intense radiation will kill living cells – radiation burns result which can cause death. A low level exposure over an extended period may result in leukaemia, a disease in which there is an excess of white corpuscles in the blood.

Gamma radiation will sterilize chemicals, instruments, etc. It can also be used to detect leaks in underground pipes by adding a gamma-emitter to the liquid in the pipe and following its progress above ground with a suitable detector.

CHAPTER 3

EXPERIMENTAL WORK IN OPTICS

3.1 The Spectrometer

The spectrometer is an instrument for measuring angular deviations of light, usually through a prism, double slit or diffraction grating. It is used with a diffraction grating to measure the wavelength of light and with a prism or a diffraction grating to study spectra. It consists essentially of four main parts: the collimator, the telescope, the table, and the scale.

The Collimator

The collimator serves to provide a parallel beam of light from the source. It consists of a tube with an achromatic converging lens at one end and an adjustable slit at the other. The width of the slit may be varied by moving one jaw relative to the other by means of a small thumbscrew. When the slit is illuminated it acts as a source of light. The distance between the slit and the lens can be varied by a focusing screw until the slit is at the focus of the lens. This gives a parallel beam of light emerging from the collimator.

The collimator is mounted on a fixed pillar attached to the base of the instrument. The lens has a focal length of approximately 180 mm and a clear aperture of 25 to 32 mm. The adjustable slit is 6 to 7 mm long.

The Telescope

The telescope, which can be rotated, has a vernier scale fitted where it adjoins the table, enabling their relative movements to be measured to 0.1° in the case of the standard spectrometer or less in the case of other spectrometers. The function of the telescope is to receive the parallel light from the collimator and bring it to a focus at the cross-wires. The optic axes of the collimator and telescope intersect on the common axis of rotation of the table and telescope.

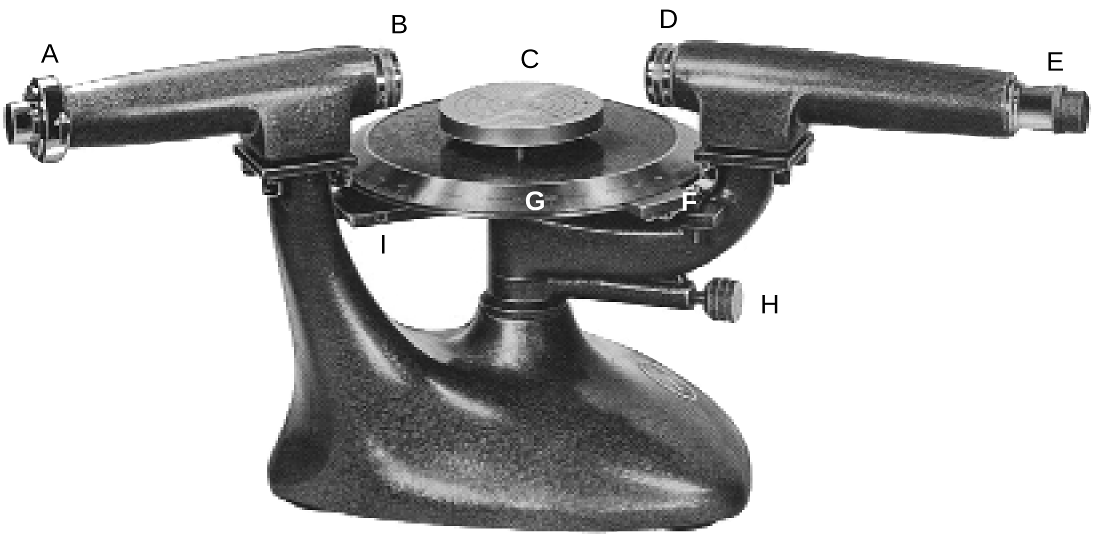
The position of both the table and telescope can be locked and then adjusted, generally over a limited range, by means of fine adjustment screws. The telescope is mounted on a moveable pillar and fitted with an approximately 178 mm focus, approximately 25 mm clear aperture, achromatic objective lens. Also fitted is a x8 to x15 Ramsden's eyepiece with cross-wires. The telescope has rack and pinion focusing by means of a screw on its side.

The Table

The table, which can be rotated, forms a suitable mount for a prism, double slit or diffraction grating. The circular edge of the table has a scale graduated in degrees, half-degrees or thirds of a degree depending on the type of spectrometer. The height of the table may be adjusted and there are also three levelling screws. The table, marked with lines to assist placing the prism with respect to the levelling screws, has screwed holes for the attachment of interchangeable clamping units for prism and diffraction grating. Large-diameter locking screws are fitted for moveable control and fine adjustment is provided for the telescope and table.

Accessories

The spectrometer is usually supplied with standard accessories: 1 prism clamp, 1 dense flint glass prism, 1 diffraction grating holder, 1 watchmaker's eyepiece, 1 tommy bar for adjusting the optical axis, and 1 plywood case.



A	Adjustable Slit	D	Focusing Screw on Telescope with Lens	G	Main Scale
B	Focusing Screw on Collimator with Lens	E	Eyepiece	H	Telescope Locking Screw
C	Table	F	Spring Loaded Vernier	I	Table Locking Screw

Fig. 3.1 *Standard spectrometer with decimal vernier*

The Standard Spectrometer

The standard spectrometer, Fig. 3.1, is capable of reading to 6 minutes of arc. The scale, usually around 170 to 180 mm diameter, is independently rotatable with a locking screw. There is a spring-loaded single decimal vernier scale attached to the telescope mount. The main scale, Fig. 3.2, is marked in 1° divisions but it is often only numbered at 10° intervals. The vernier is marked at intervals of 0.1° (6 minutes of arc), but usually only numbered at 0.5° intervals.

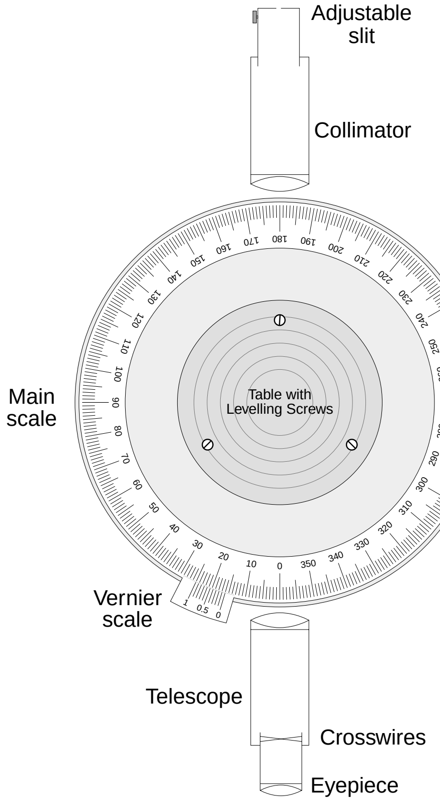


Fig. 3.2 *Standard spectrometer with decimal vernier*

When reading the angle corresponding to a particular setting of the telescope, first read the main scale, to the nearest 1° division before the position corresponding to the zero on the vernier scale. Then follow the vernier scale until one of its divisions coincides exactly with a main scale division. This is the correct vernier reading in the decimal part of a degree. The angle is thus the main scale reading plus the vernier reading in the decimal part of a degree. In Fig. 3.3 reading 1,

$$\begin{aligned} \text{main scale reading} &= 130° \\ \text{+ vernier reading} &= 0.1° \\ \hline \text{overall reading} &= 130.1° \\ &= 130° \, 6' \end{aligned}$$

A spectrometer with a 60 minute vernier is capable of reading to 1' of arc. The main scale is marked in 1° divisions, but is often only numbered at 10° intervals. The double ended verniers are marked at intervals of 1' but are usually only numbered at 30' intervals. When reading the angle corresponding to a particular setting of the telescope, first read the main scale to the nearest 1° division before the position corresponding to the zero on the vernier scale. Then follow the vernier scale from the zero mark until one of its divisions coincides exactly with a main scale division. This is the correct vernier reading in minutes. The angle is thus the main scale reading plus the vernier reading in minutes. In Fig. 3.4, reading 1A

main scale reading = 78°
+ vernier reading = 40'
overall reading = 78° 40'

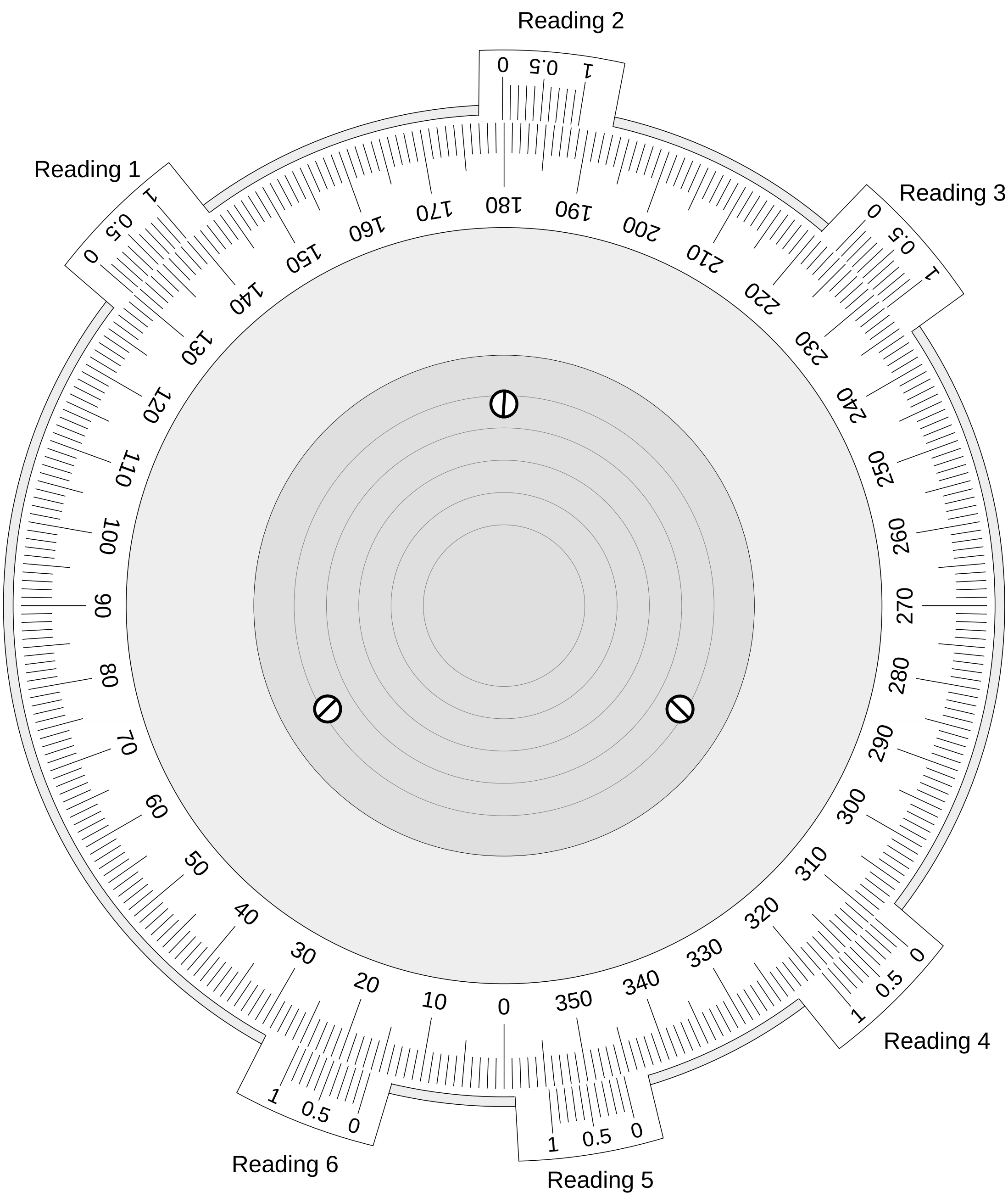


Fig. 3.3 *Examples of various standard spectrometer readings with decimal vernier*

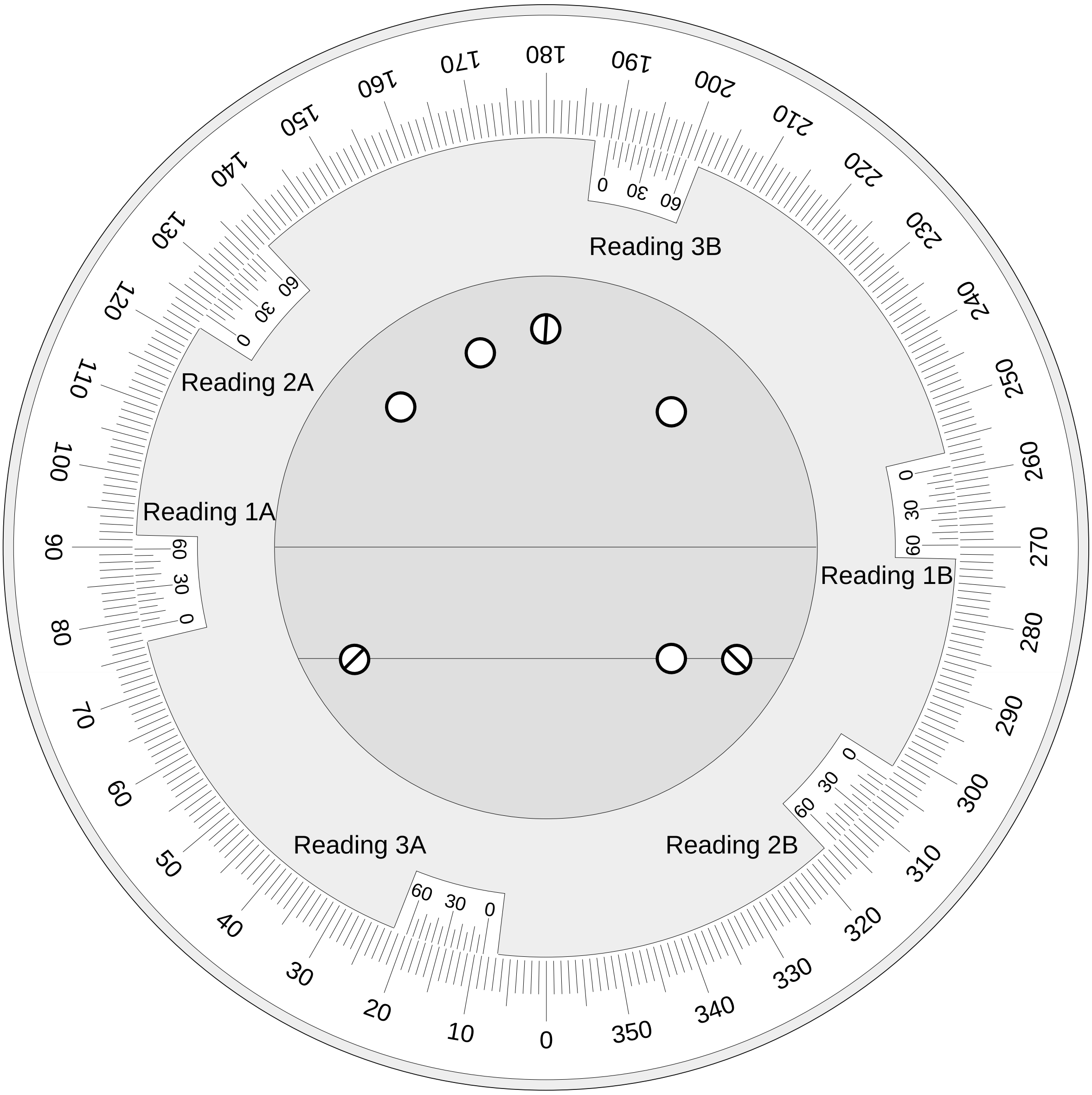


Fig. 3.4 Examples of various spectrometer readings with 60' vernier scale

Adjustments to the Spectrometer

In brief, the essential adjustments to the spectrometer are as follows. Firstly, the eyepiece should be adjusted to give a good sharp image of the cross-wires. It should not be adjusted subsequently. Secondly, the telescope should be adjusted to give a good sharp image of a distant object. Again, this adjustment should not be changed subsequently. Finally, the collimator should be adjusted so that a good sharp image of the slit is obtained when viewed through the telescope. (Note that, in some very basic instruments, the collimator cannot be adjusted.) For more accurate work the spectrometer is adjusted as described in the following paragraphs.

1. Cross-wires

The eyepiece is first adjusted so that the cross-wires can be seen distinctly. The eyepiece is adjusted by sliding it in and out of its tube by hand until the cross-wires are clearly seen with one eye, whilst a distant object is viewed with the other eye. (This may be done more easily with the cross-wires in the $\times$ rather then the + position.) The image of the cross-wires is then effectively at infinity. This adjustment depends to some extent on the observer and it may vary slightly from observer to observer. If the observer wears spectacles it is best to keep them on during this adjustment. Care must be taken to ensure that this adjustment is not changed subsequently.

2. Telescope

Take the spectrometer outside or point the telescope out an open window. Unlock the telescope and point it towards a distant object at least 50 to 100 m away. The focusing screw is adjusted by hand until the distant object is in focus, i.e. there is no parallax between the image of the distant object and the cross-wires. The telescope is now focused to receive parallel light. Do not make any subsequent adjustments to the telescope.

Remember to unlock the locking screws before attempting to move either part, and never use force for any of the adjustments of a spectrometer. The vernier scale should not be marked or defaced in any way as this would make it difficult to read.

3. Collimator

The spectrometer is now returned to the laboratory bench. The source of light to be viewed is placed close to the adjustable slit of the collimator. The telescope is rotated until it is in line with the collimator so as to view the light from the slit. Make sure at this point that the slit is open. The position of the slit is adjusted relative to the lens using the focusing screw on the side of the collimator until the image of the slit is in sharp focus when viewed through the telescope, i.e. there is no parallax between the image of the slit and the cross-wires. Since the telescope had previously been adjusted for parallel light it follows that the collimator is now giving parallel light.

Schuster’s method of focusing

When the above method of focusing is impracticable the spectrometer can be adjusted by Schuster’s method. The telescope is set to receive light from a prism placed on the prism table, set at an angle greater than minimum deviation, and focused to give a sharp image of the slit. Since the prism is not set at the angle of minimum deviation, there will be another position of the prism which will give an image in the telescope. The prism is rotated to this position, and now the collimator is focused to produce a sharp image. The above steps are continued until the image of the slit is equally sharp in both positions.

4. Levelling the table

Place the prism (60°) on the table so that one face AC is perpendicular to the line XY joining two of the three levelling screws XYZ, Fig. 3.5. The table is marked with lines to assist in placing the prism with respect to the table levelling screws.

Adjust the height of the table so that it is at the correct height. Rotate the table so that the refracting angle faces the collimator and light from the illuminated slit falls on two of the faces AB and AC. If the prism has a ground face, the refracting angle will usually be opposite to this face. It should be possible to locate the two reflected images of the slit from faces AB and AC using the telescope. If either of these images is not visible it is likely that the table is at the incorrect height, so

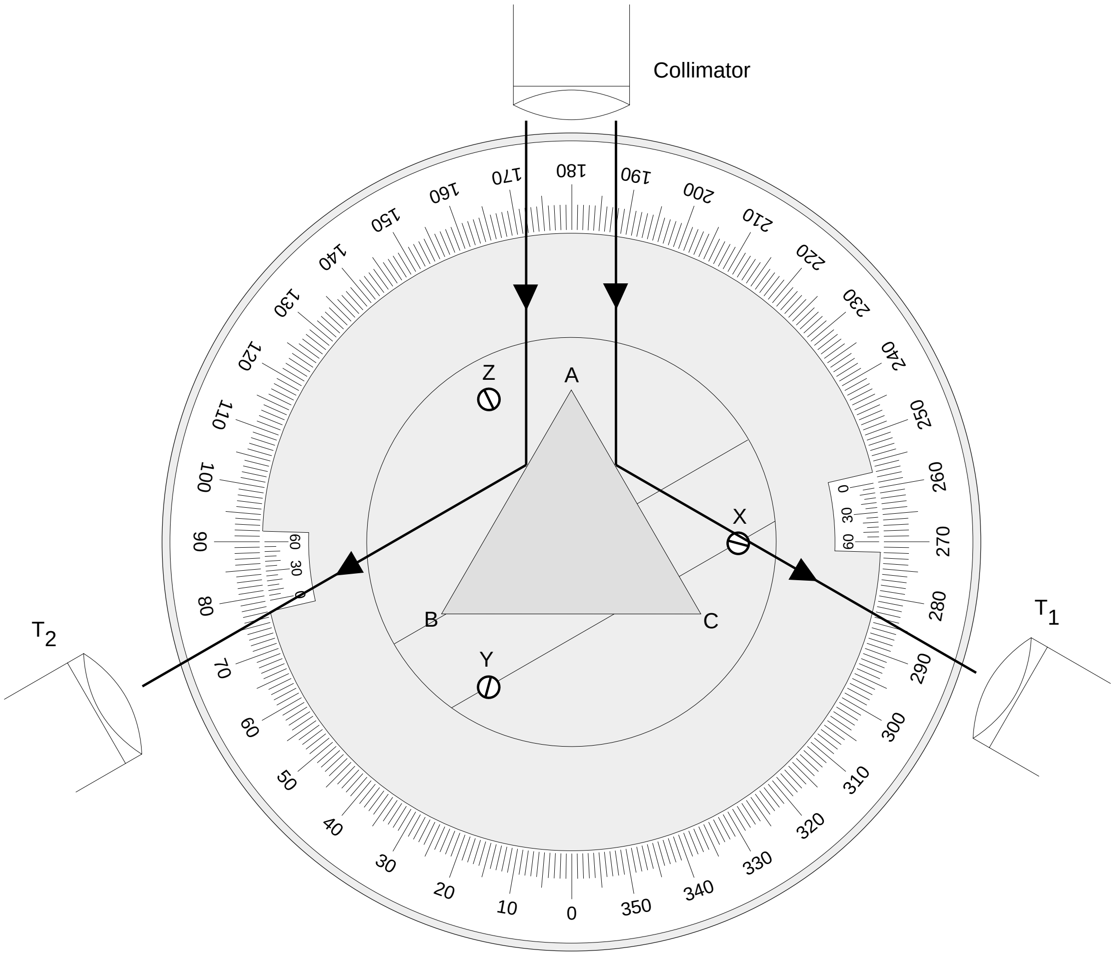


Fig. 3.5 *Levelling the spectrometer table*

adjust the height of the table so that at least one of the images can be seen. If only one of the images can be seen the table is not level. Rotate the telescope to position T_1 to receive light reflected from the face AC and adjust either of the screws X or Y to tilt the face AC until the centre of the image is at the centre of the cross-wires. Turn the telescope to position T_2 to receive light reflected from the face AB and adjust the screw Z until the centre of the slit is at the centre of the cross-wires. Return the telescope to position T_1 and repeat until the image is central whichever face is viewed. The prism and table are now level.

It is a useful precaution to locate the images first by eye, without using the telescope. It should then be possible to adjust the height of the table correctly. If the images are not visible, the refracting angle of the prism is too near the collimator, so move it farther away so that the images can be located.

5. Adjustment of the plane of the diffraction grating

Rotate the telescope to the straight-through position and centre the image of the slit on the cross-wires. Lock the telescope and table and centre the image of the slit exactly on the cross-wires using the fine adjustment screw. Note the scale reading. Unlock the telescope and rotate it until the scale reading has changed by 90°. Lock the telescope again and adjust the fine adjustment screw to do this exactly. The telescope is now at right angles to the collimator. Screw on the diffraction grating holder and place the diffraction grating G on to the table of the spectrometer, making sure that its plane is perpendicular to one side, e.g. XY, of the triangle formed by the three levelling screws XYZ, Fig. 3.6. Unlock the table and rotate it until an image of the slit can be seen by reflection from one face of the grating. Adjust one or both of the levelling

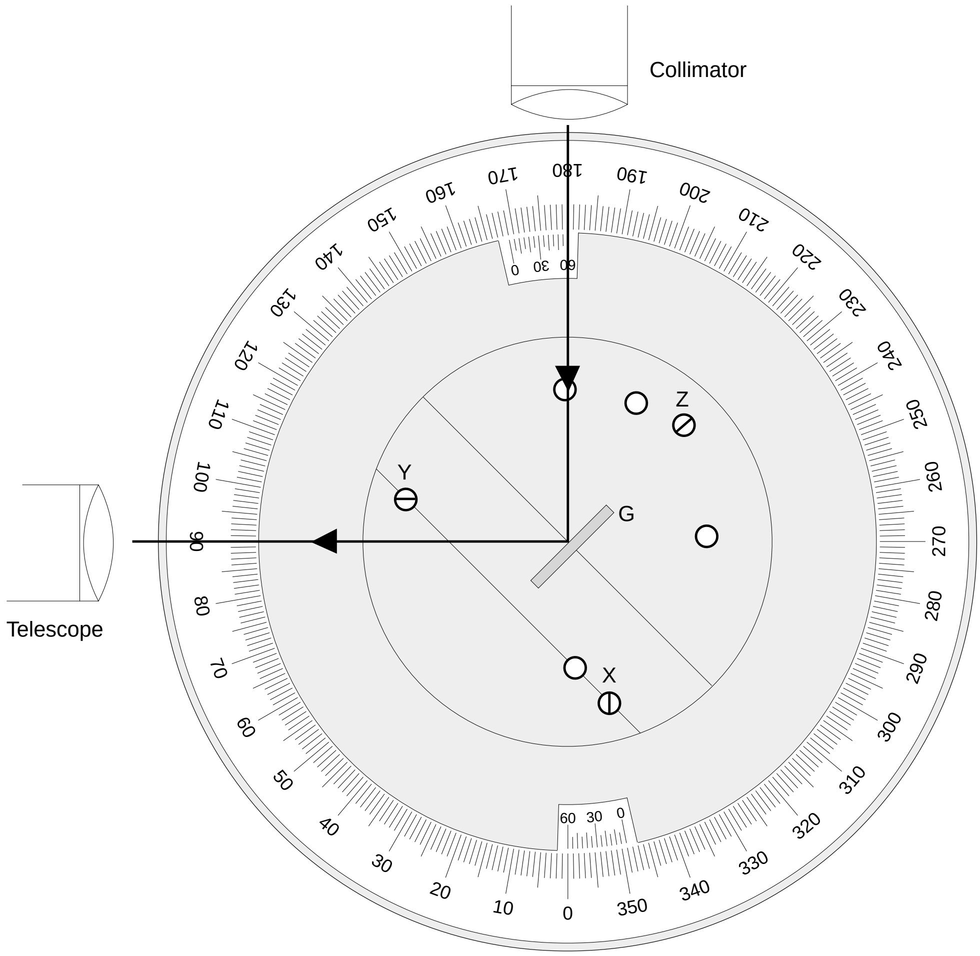


Fig. 3.6 *Telescope at 90° to collimator and diffraction grating (G) at 45° to collimator*

screws X and Y until the image of the slit is in the centre of the field of view. Lock the table in this position and adjust the fine adjustment screw until the image is centred exactly on the cross wires. Note the scale reading. The table and the grating are now at 45° to the collimator.

Unlock the table and rotate it so that the scale reading changes by exactly 45°, locking the table to do this exactly with the fine adjustment screw. Light from the collimator now strikes the grating normally.

6. Adjustment of the lines of the diffraction grating

Unlock the telescope and rotate it so as to receive light from one of the first order maxima Fig. 3.7. If the grating lines are not perpendicular to the plane of rotation of the telescope the image may not be at the centre of the field of view.

Adjust screw Z until the image is at the centre of the field of view. Rotate the telescope to receive light from the other first order maximum and check that the image is still in the centre of the field of view. If it is not, repeat adjustments 5 and 6.

Production of a Pure Spectrum Using a Spectrometer

Place the prism on the table so that one face is roughly parallel with the collimator and rotate the telescope to pick up the spectrum, Fig. 3.8.

A diverging beam of white light emerging from the slit of the collimator is made parallel by the lens of the collimator and falls on the prism. Refraction through the prism splits up the light into separate parallel beams of different colour, each travelling in a slightly different direction and each of which is brought to its own focus in the focal plane of the

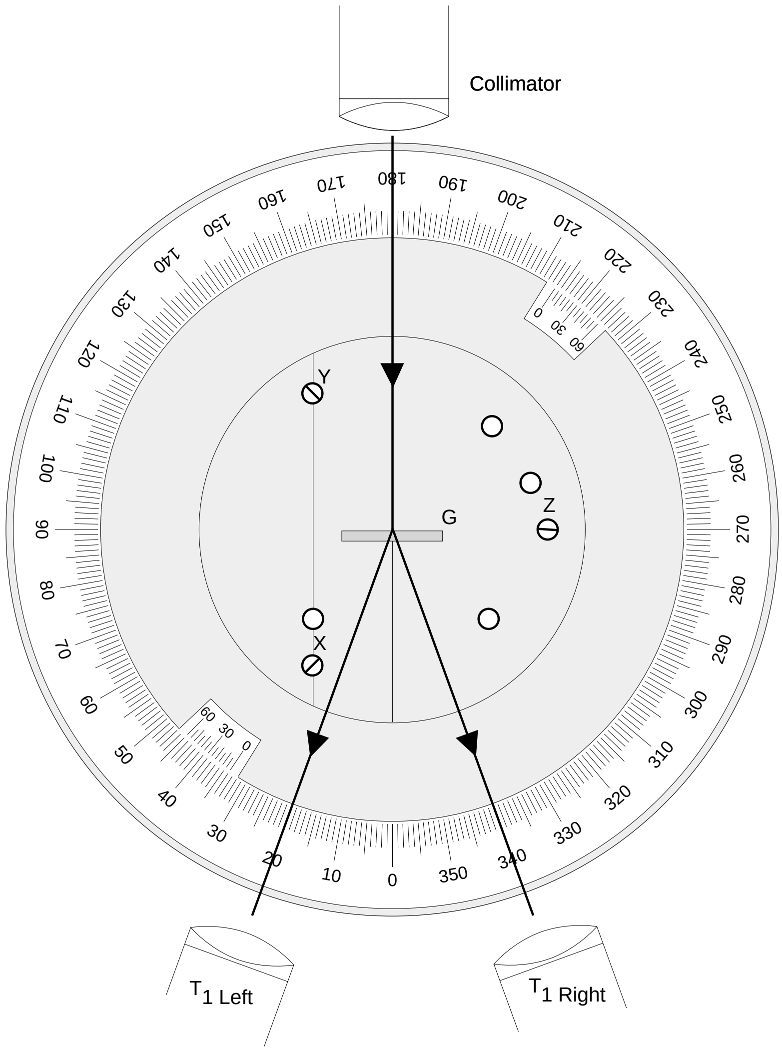


Fig. 3.7 *Diffraction grating at 90° to collimator and lines of grating parallel to slit*

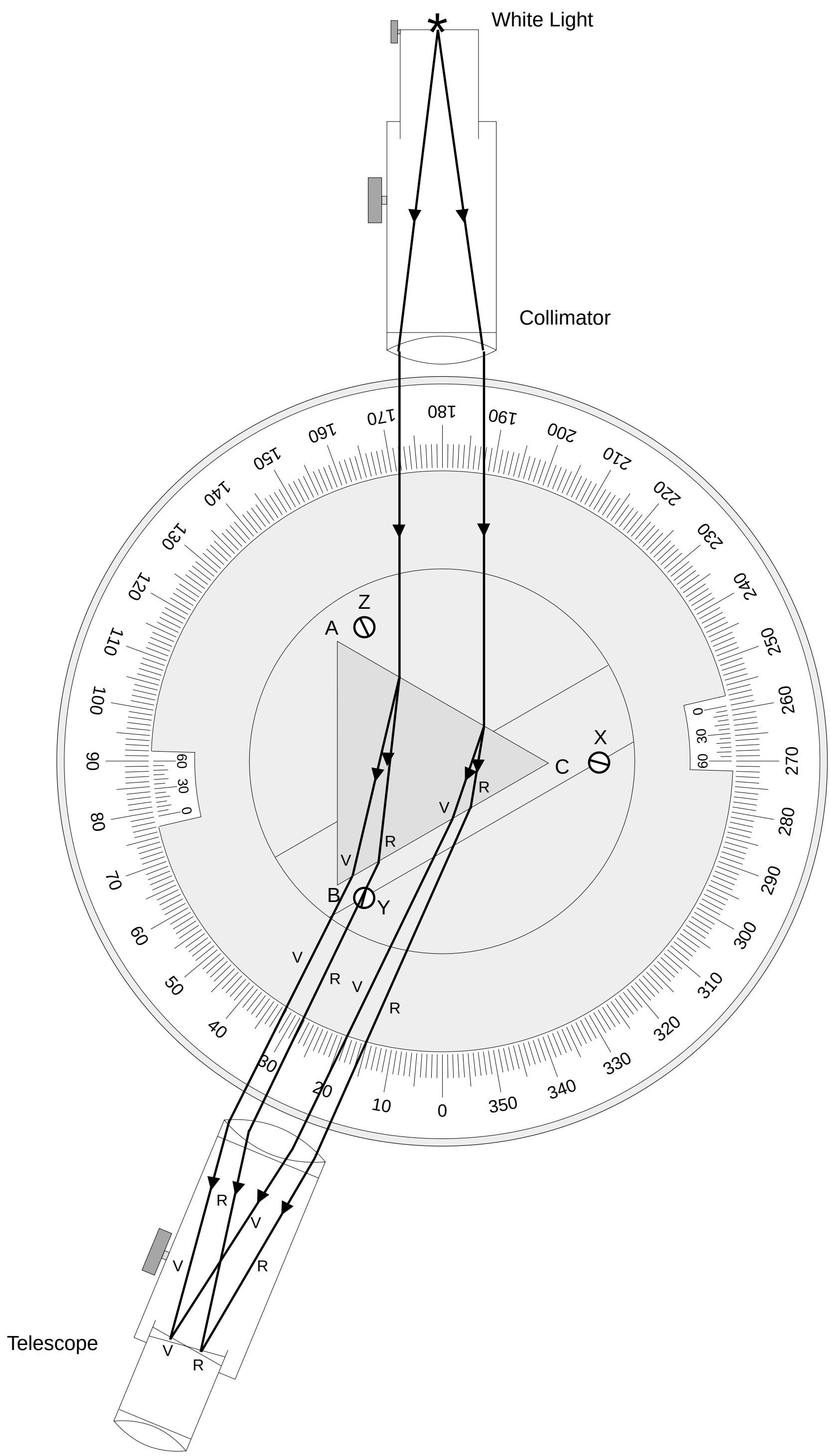


Fig. 3.8 *Formation of a pure spectrum using a spectrometer*

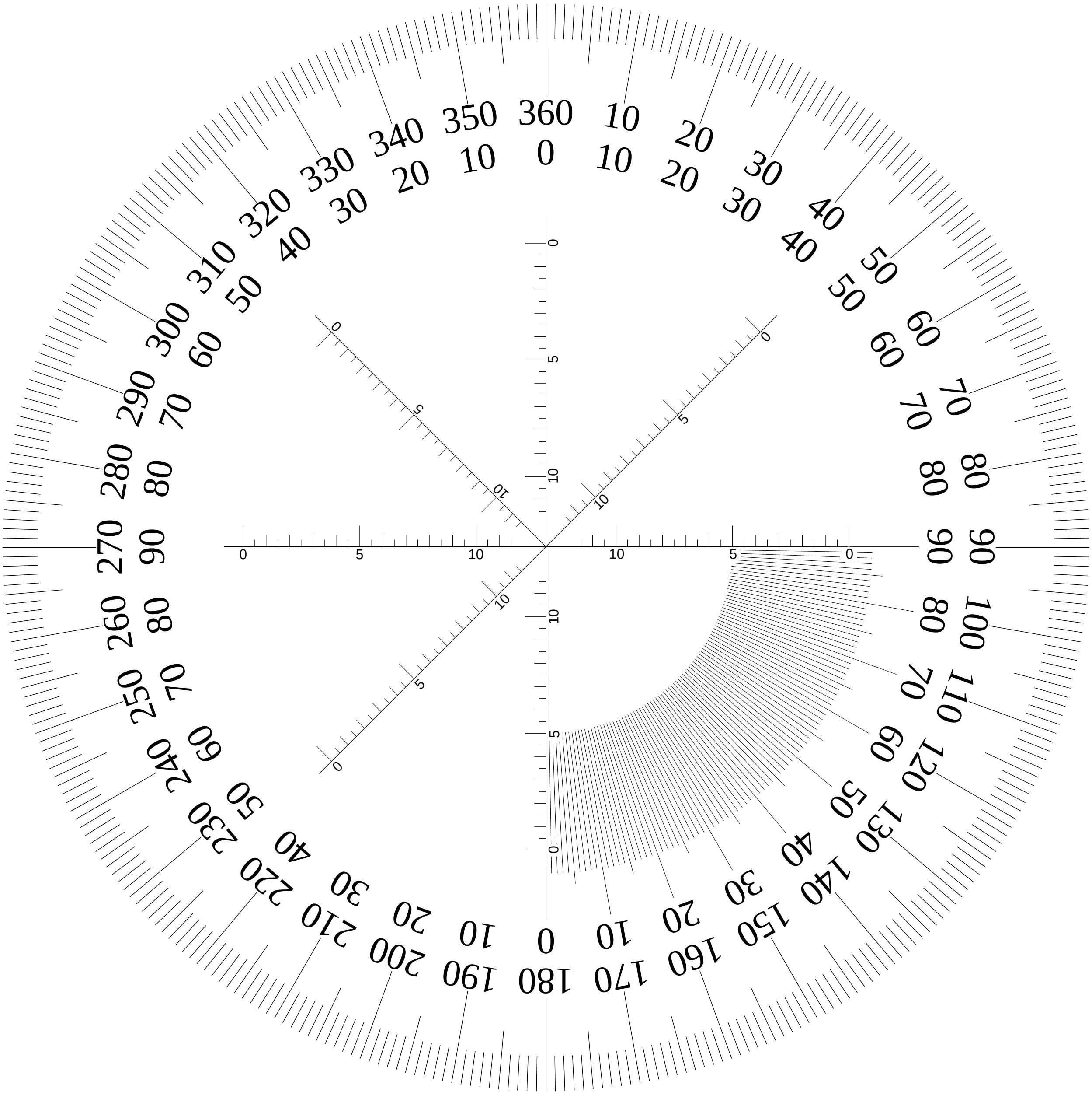


Fig. 3.9 *0–360° protractor with quadrant markings*

objective of the telescope. The spectrum, viewed through the eyepiece of the telescope, is a series of monochromatic images of the slit; the narrower the slit the purer the spectrum.

Light Sources

1. Glass rod held over a lighted bunsen burner.
2. Sodium flame pencils held over a bunsen flame.
3. Cotton wool soaked in salt solution held in a tongs over a lighted bunsen flame.
4. Spectral lamp in spectral lamphouse with transformer. After switching off leave untouched until it gets cold.
5. Spectrum tubes and power supply.
6. Four or five ultra bright LEDs connected in parallel. Red, green and yellow, and now blue, colours are available, giving lights of different wavelengths.
7. Outdoor sodium lamps with choke.

For experiments using a spectrometer see pp. 32 and 41.

3.2 Ray Box

A ray box is an instrument which can produce single or multiple rays of light, the bottom part of which is visible on paper. The bulb and lens can be moved relative to each other so that the emergent rays may be made convergent, divergent, or parallel if the bulb is at the focus of the lens. Some types of ray box have hinged mirrors for colour mixing (see p. 33).

Experiments with a ray box are best carried out in a darkened room or in a darkened corner so that the light rays, reflected and refracted, show up well on white paper. Experiments are easier to arrange if cylindrical mirrors and lenses are used instead of spherical ones. Since the rays are confined to one plane this does not invalidate the experiment.

Recording Ray Paths

To record ray paths, mark the position of the mirror (plane or spherical), lens, or prism by outlining its position with a sharp pencil. Then mark the position of the centre of the incident ray being observed with two pencil marks, one near

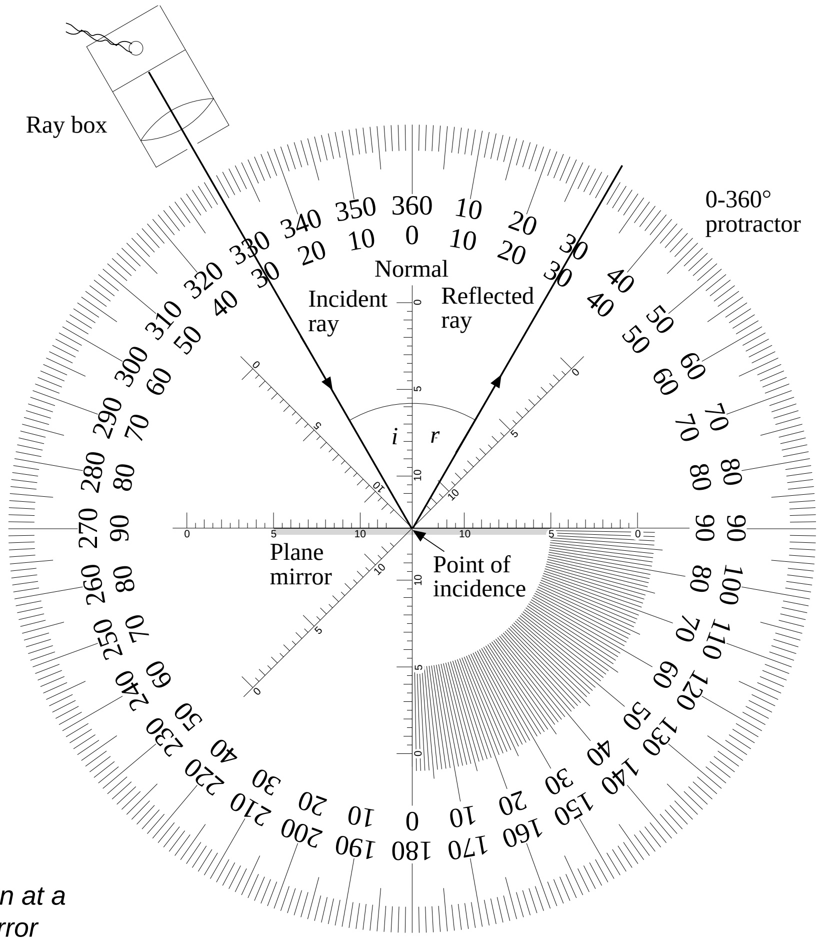


Fig. 3.10 *Reflection at a plane mirror*

the ray box and the other near the optical device. Mark the ray reflected or refracted by the device with two other pencil marks. Number the marks if the ray pattern is complicated. Remove the mirror, lens or prism and join up the same-numbered points to show the ray paths to and from the device. Mark arrow heads on the lines to indicate their direction through the device.

Light Travels in Straight Lines

Aim a single ray of light along a sheet of white paper. The edges of the ray are well defined so it may therefore be concluded that light travels in straight lines.

1. Plane Mirror

In a darkened room place a photocopy of the 0–360° protractor with degree divisions, Fig. 3.9, on a bench and aim a single ray of light from the ray box along the 360–180° line. Place a plane mirror vertically and centrally on the 90–270° line so that the back of the mirror, i.e. the reflecting surface of the mirror, is on the line and the reflecting side is facing 360°, Fig. 3.10. The incident ray now strikes the mirror normally, at the intersection of the 180–360° line and the 90–270° line, and is reflected back along its own path.

Rotate the ray box, keeping the point of incidence the same, until the incident ray is along the 10° line. Record the angle of incidence *i* and the angle of reflection *r*. Repeat for other angles of incidence and record the angle of incidence and corresponding angle of reflection in each case.

Observations

1. The angle of incidence is equal to the angle of reflection.
2. The incident ray, the reflected ray and the normal all lie in the same plane.

Optical lever

In a darkened room place a photocopy of the 0–360° protractor on the bench and place a plane mirror vertically on it so that the back of the mirror, i.e. the reflecting surface of the mirror, is central on the 90–270° line, Fig. 3.11. Markings on this

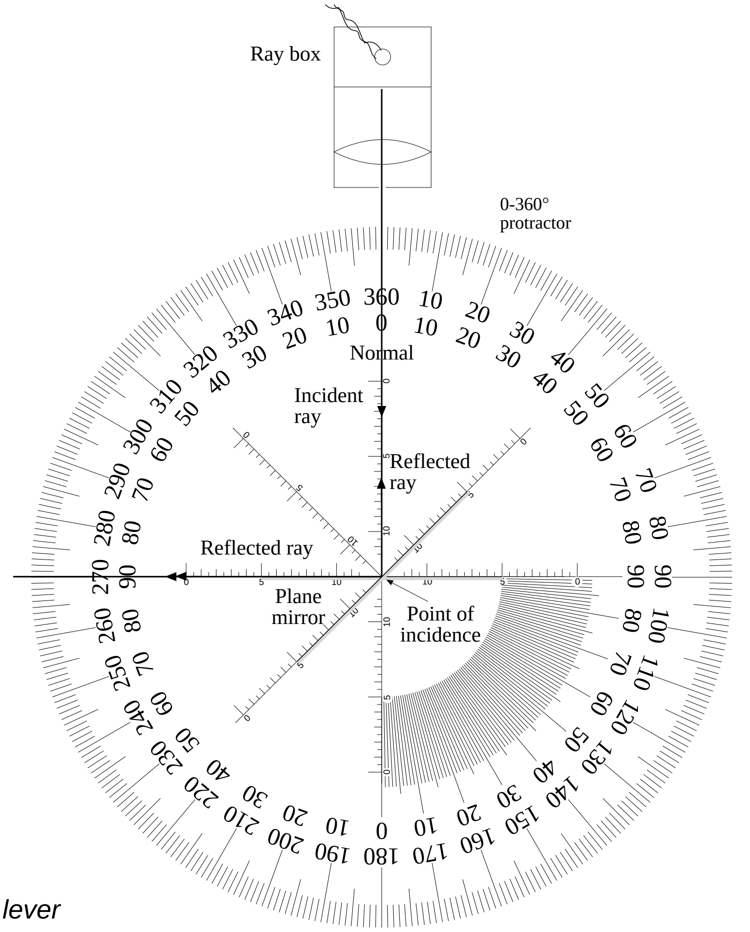


Fig. 3.11 *The optical lever*

line assist in the placing of the mirror.

Aim a single ray of light from the ray box along the 360–180° line so that it strikes the mirror normally, at the intersection of the 180–360° line and the 90–270° line, and is reflected back along its own path. The mirror is now at right angles to the ray box. Rotate the mirror through 45° about this point of reflection into a new position, so that the back of the mirror is central along the 45–225° line. Note that the reflected ray is now along the 90–270° line at right angles to the ray box, i.e. the reflected ray is rotated through 90°.

Repeat for other rotations of the mirror to verify that if a mirror is rotated through an angle θ the reflected light is rotated through an angle 2θ.

Image in a plane mirror

Draw a straight line in the middle of a white sheet of paper. Place the plane mirror vertically on this line so that the back of the mirror is along the line, Fig. 3.12. Place a dot on the paper in front of the mirror and aim a single ray of light from the ray box through this dot.

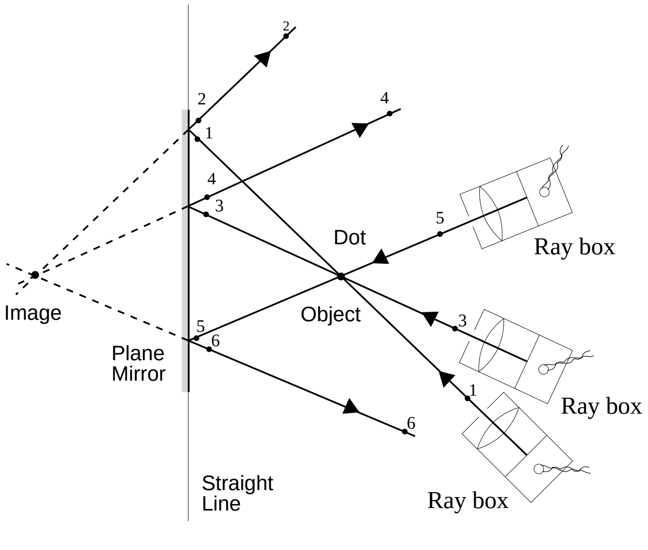


Fig. 3.12 *Locating the image of an object in a plane mirror*

Mark the position of the incident and reflected rays with two numbered marks, each of the same number. Move the ray box to a new position and again aim the ray through the dot to the mirror. Mark the incident and reflected rays, again with numbered marks. Repeat this procedure once more. Remove the mirror and join up the points which are numbered the same. Continue the reflected rays backwards to meet at one point. This point is the image of the original dot. Measure the distance between this point, the image, and the mirror, and between the mirror and the original dot, the object. They will be found to be equal, i.e. the image is as far behind the mirror as the object is in front of the mirror.

2. Spherical Mirrors: Radius of Curvature and Focal Length

To find the centre of curvature of a concave mirror place the mirror on a sheet of graph paper and mark the reflecting surface (front or back depending on whether the mirror is polished metal or silvered glass, respectively) with a sharp pencil. Aim a single ray of light from the ray box at the mirror and adjust its position until it is reflected back along its own path, Fig. 3.13.

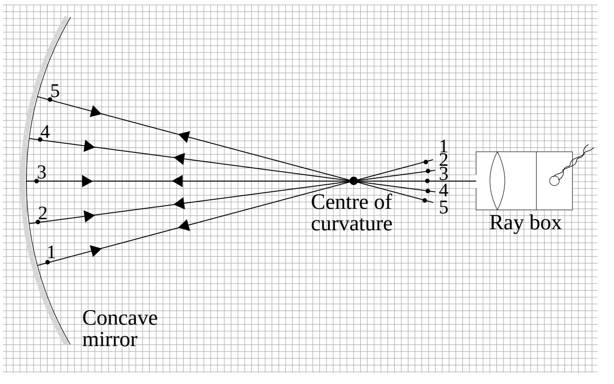


Fig. 3.13 *Finding the centre of curvature of a concave mirror*

Mark the position of this ray with two pencil marks, both numbered the same, one close to the mirror and the other close to the ray box. Move the ray box to a new position so that the ray of light from the ray box is again reflected back along its own path. Mark the position of this ray with two pencil marks, numbered differently from the previous two marks. Repeat this procedure three more times. Try to have at least two rays on either side of the principal axis.

Remove the mirror and join up the points which are numbered the same. The point where all the rays meet is called the centre of curvature. Measure the distance between this point and the mirror. This is the radius of curvature of the mirror.

The focal length is equal to half the radius of curvature and so may be easily found once the radius of curvature is known. It may also be found as follows.

Place a sheet of graph paper on the bench and place the ray box on it. Aim a set of parallel rays, three or four, along the lines of the graph paper and adjust the position of the bulb relative to the lens of the ray box until the rays are parallel to the lines on the graph paper. The rays of light from the ray box are now parallel to each other.

Place the mirror on the graph paper so that the rays hit the inside curve of the mirror near the centre parallel to the principal axis of the mirror, Fig. 3.14.

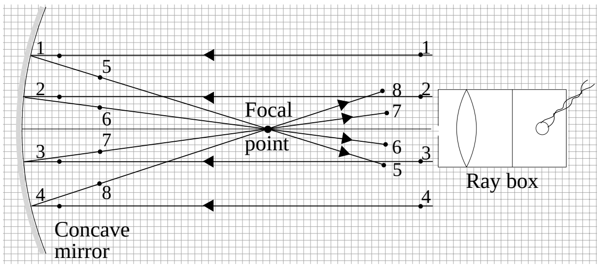


Fig. 3.14 *Finding the focal length of a concave mirror*

Mark the reflecting surface of the mirror, the incident rays and the reflected rays (with numbered points), and the point where the reflected rays meet. This point is called the focal point. Measure the distance between this point and the pole of the mirror. This distance is equal to the focal length of the mirror.

To find the focal length of a convex mirror arrange the ray box as for the concave mirror. Place the convex mirror on the graph paper so that the rays hit the mirror, near the centre, parallel to the principal axis of the mirror, Fig. 3.15.

Mark the reflecting surface of the mirror, and the incident rays and the reflected rays with numbered points. Remove the mirror and join up the same-numbered points. Continue the reflected rays backwards to meet at one point. This point is the focal point. Measure the distance between this point and the pole of the mirror. This distance is equal to the focal length of the mirror.

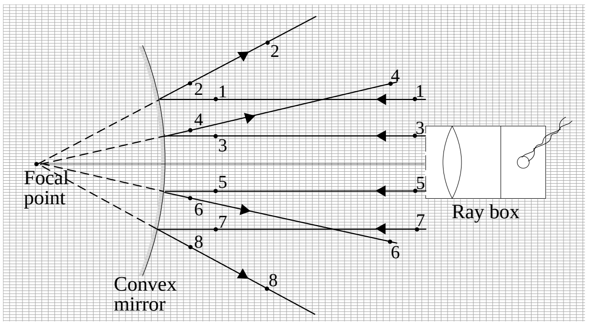


Fig. 3.15 *Finding the focal length of a convex mirror*

Ray diagrams

Place a concave mirror on a sheet of graph paper and mark the reflecting surface with a sharp pencil. Find the focal point F and the centre of curvature C and mark these points on the graph paper. Place a dot on the graph paper in front of the mirror. Aim a single ray of light from the ray box at the centre of the mirror and adjust its position until it is reflected back along its own path, ray A, Fig. 3.16.

Mark the position of this ray with two pencil marks 1,1, one close to the mirror and the other close to the ray box. Move the ray box so that ray B through the dot is parallel to the principal axis. Mark the incident and reflected rays with numbered points 2,2 and 3,3 respectively. Rotate the ray box so that ray C passes through the dot and the focal point F. Mark incident and reflected rays with numbered points 4,4 and 5,5. Now rotate the ray box so that ray D through the dot hits the mirror at an angle θ on one side of the principal axis. Again mark incident and reflected rays with numbered points 6,6 and 7,7, respectively.

Observations

Ray A. Incident rays of light which strike a spherical mirror normal to the surface (ones passing through the centre of curvature) are reflected back on themselves and pass through the centre of curvature.

Ray B. Incident rays of light travelling parallel to the principal axis are reflected by the mirror through the focal point.

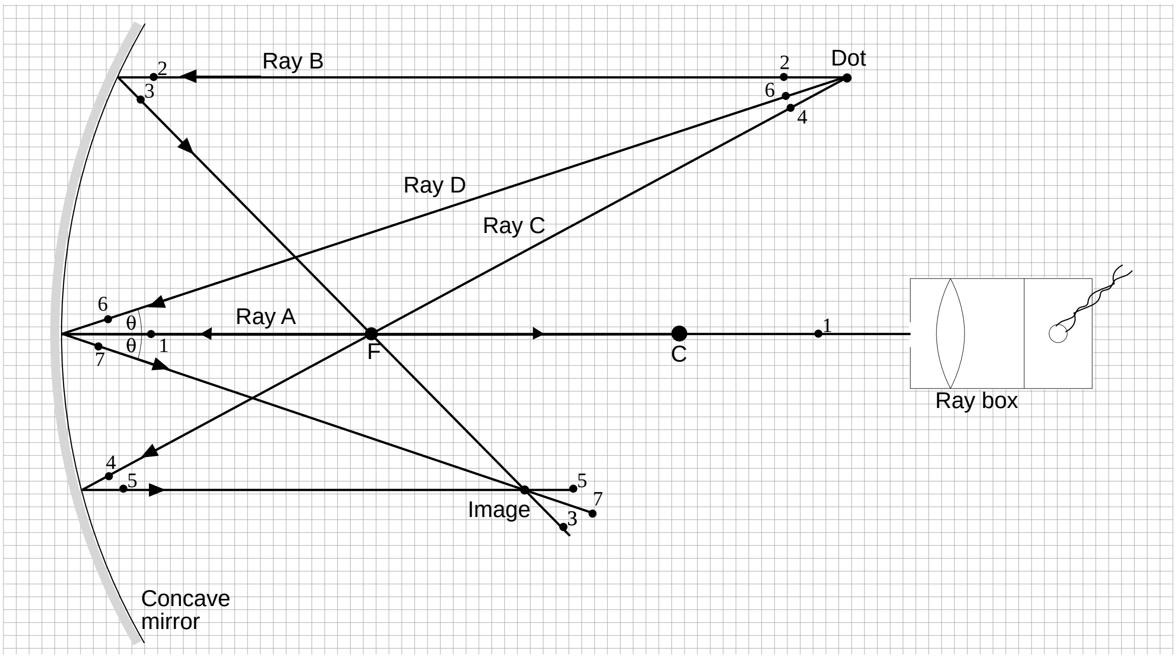


Fig. 3.16 *Ray diagram for a concave mirror*

Ray C. Incident rays of light passing through the focal point are reflected from the mirror parallel to the principal axis.

Ray D. A ray of light incident at the pole of a mirror at an angle θ is reflected by the mirror on the other side of the principal axis at the same angle θ.

Similar diagrams may be drawn for a convex mirror Fig. 3.17.

3. Refraction in a Rectangular Glass Block

Place a photocopy of the 0–360° (Fig. 3.9, p. 56). protractor on the bench and aim a single ray of light from the ray box along the 360–180° line. Place the rectangular glass block on the protractor so that one of the long sides is centrally along the 90–270° line, Fig. 3.18, perpendicular to the incident ray.

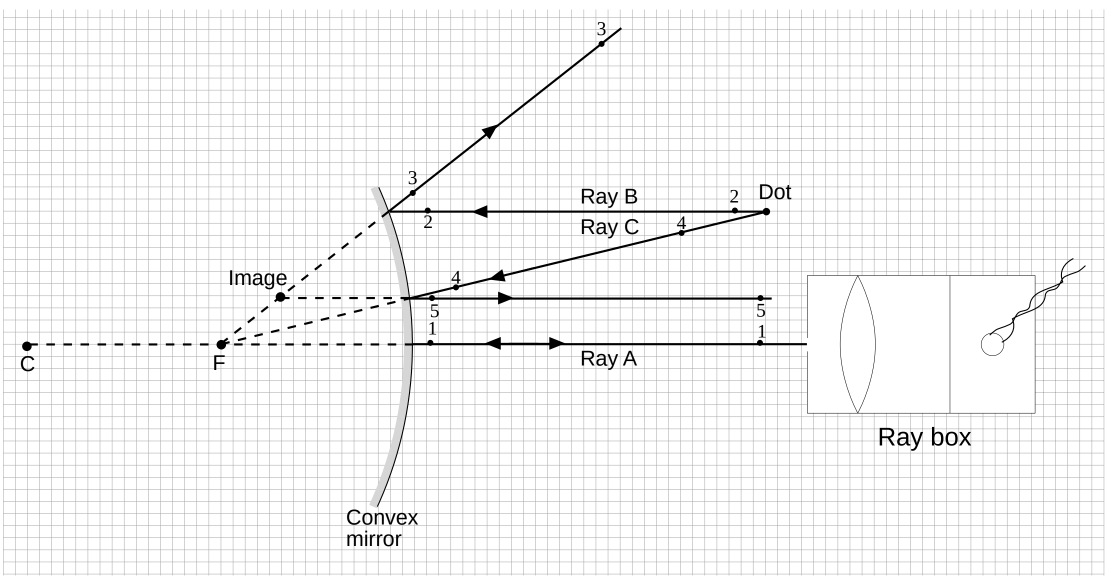


Fig. 3.17 *Ray diagram for a convex mirror*

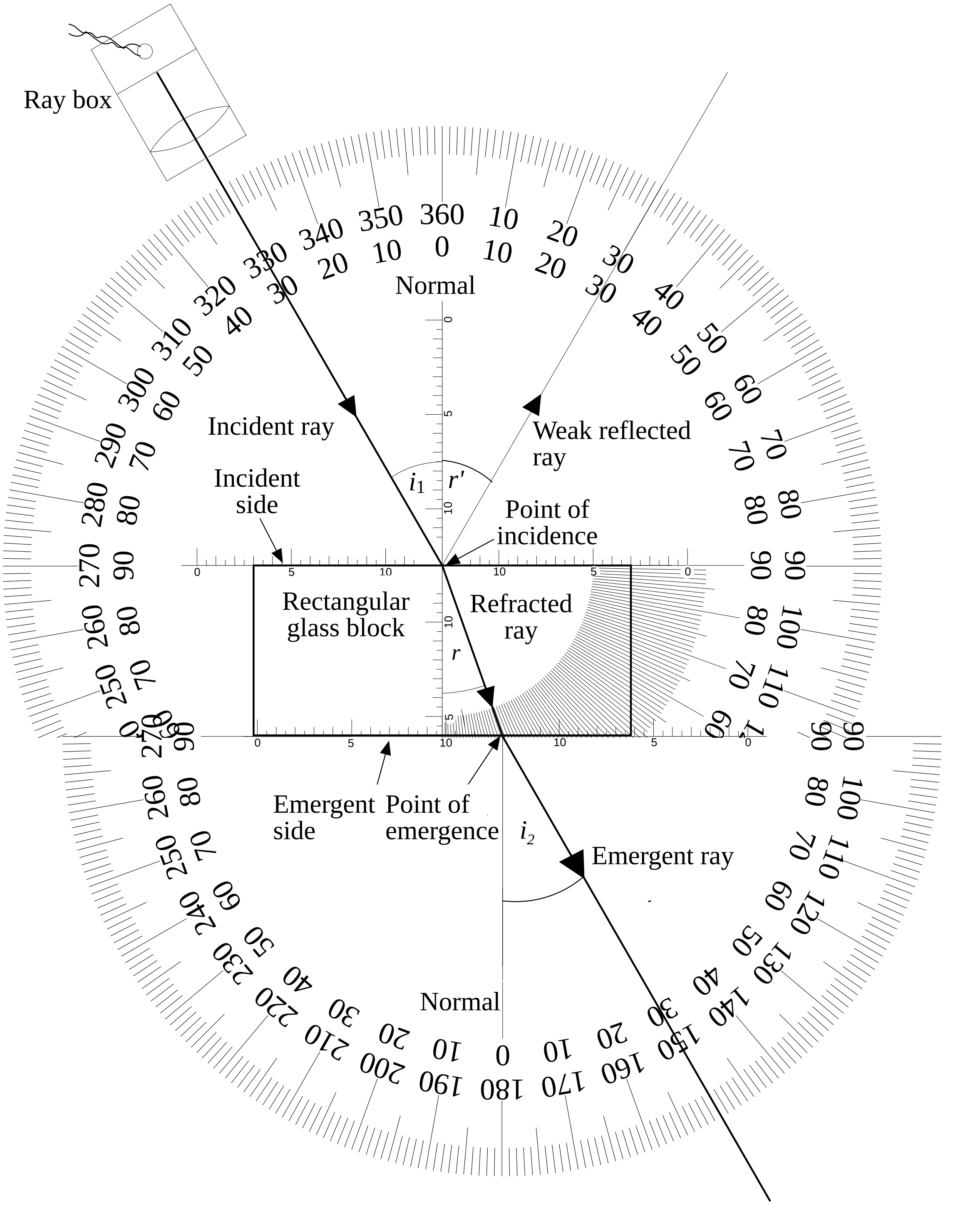


Fig. 3.18 Refraction in a rectangular glass block

Rotate the ray box, keeping the point of incidence the same at the intersection of the 180–360° and 90–270° line, until the incident ray is along the 10° line. Record the angle of incidence i_1 and the angle of refraction r. Calculate the ratio $\sin i_1/\sin r$ which is the refractive index of air to glass. Now place a 0–180° protractor (cut a 0–360° protractor along the 90–270° line) so that the 180° line is at the point of emergence of the second face. Measure the angle of emergence i_2 and calculate the ratio $\sin r/\sin i_2$ which is the refractive index of glass to air. This should be the reciprocal of that for air to glass.

Repeat the above procedure for different angles of incidence and calculate the two refractive indices in each case.

4. Refraction in a Semicircular Glass Block

In a darkened room place a photocopy of the 0–360° (Fig. 3.9, p. 56) protractor on the bench and place the semicircular glass or perspex block on it so that its flat side is centrally on the 90–270° line and facing 360°, Fig. 3.19. Markings on this line assist in the placement of the block.

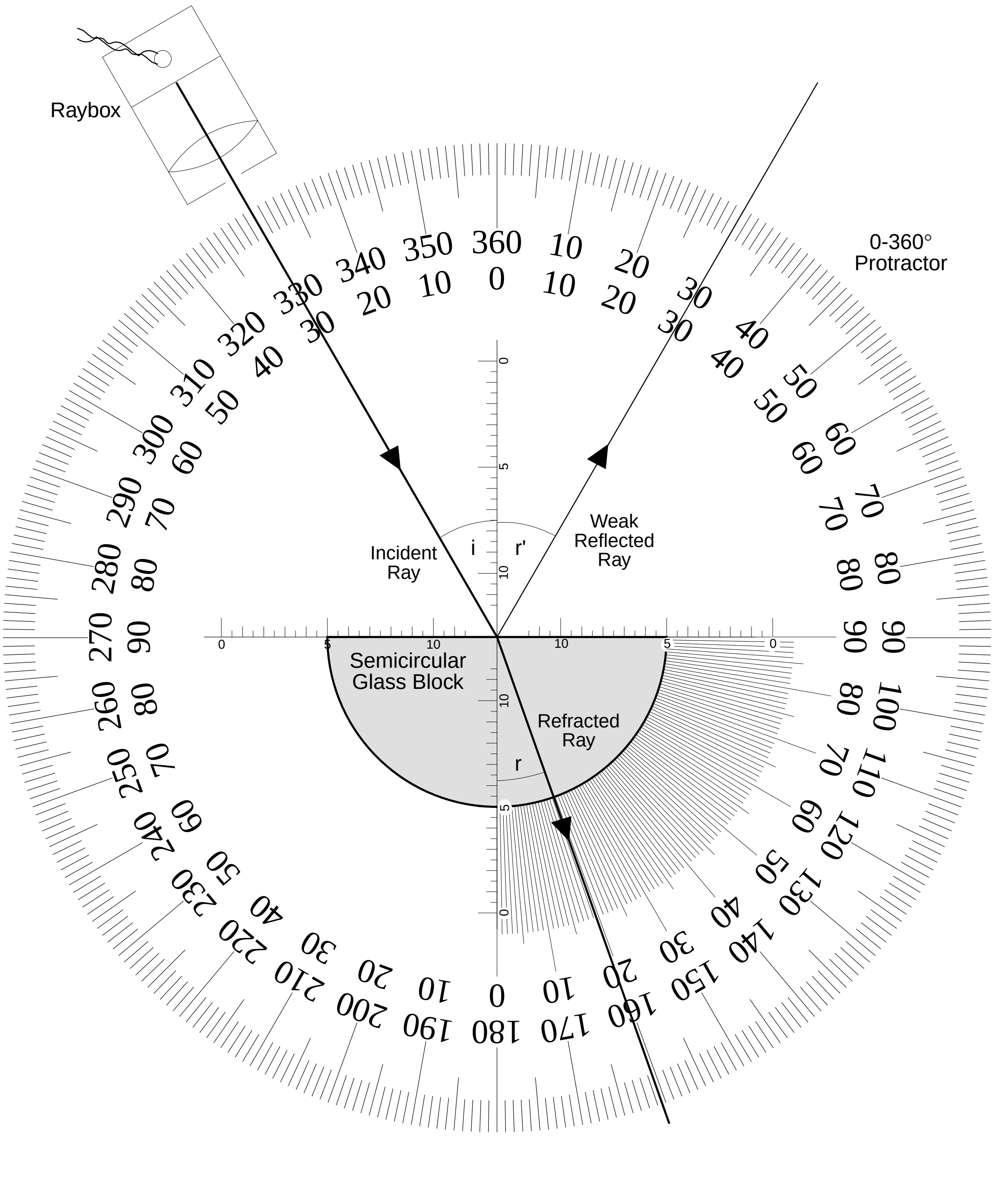


Fig. 3.19 *Refraction in a semicircular glass block*

Aim a single ray of light from the ray box along the 360–180° line so that it strikes the flat side at its centre and passes through the centre of the curved side. Notice that there is no refraction but a very weak reflected ray, reflected back along the incident ray.

Rotate the ray box, keeping the point of incidence the same at the intersection of the 90–270° and the 180–360° lines, to the 10°, 20°, 30°, etc., up as far as the 90° line. Notice the bright refracted ray and the weak reflected ray, and note that, as the angle of incidence increases, the brightness of the reflected ray also increases. Measure the angles of incidence *i*, angles of refraction *r* and angles of reflection *r*'.

Calculate the ratio sin *i*/sin *r* for the various angles of incidence and refraction. This ratio is the refractive index. Note: at small angles of incidence, unless the room is darkened, the reflected ray may not be visible.

Observations

1. As the angle of incidence increases, the angle of refraction also increases, and the ratio sin *i*/sin *r* is a constant which gives the refractive index.
2. At small angles of incidence, the reflected ray is very weak.
3. The angle of incidence is equal to the angle of reflection.

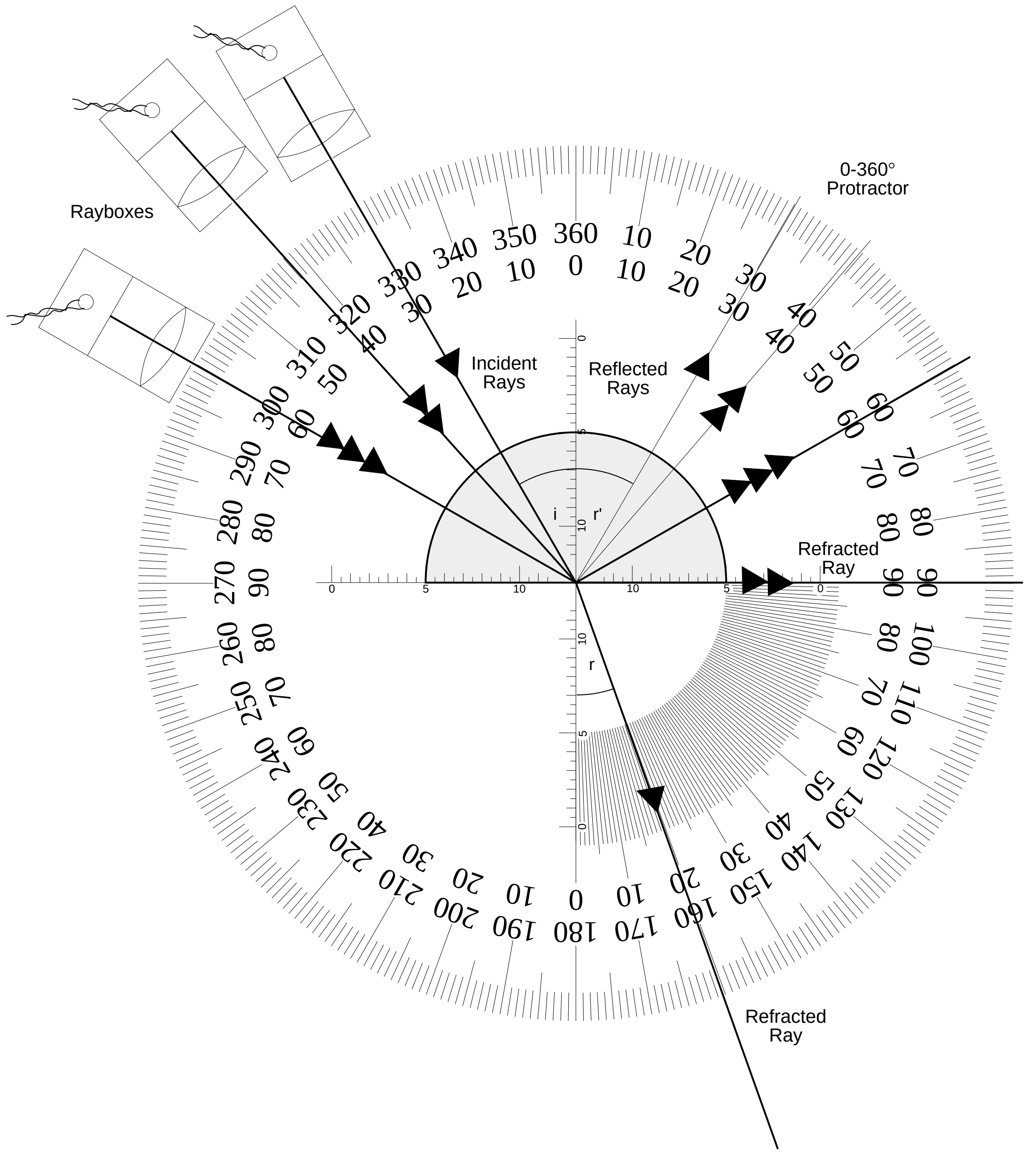


Fig. 3.20 *Total internal reflection in a semicircular glass block*

5. Total Internal Reflection in a Semicircular Glass Block

In a darkened room place a photocopy of the 0–360° (Fig. 3.9, p. 56) protractor on the bench and place the semicircular glass or perspex block on it so that its flat side is centrally on the 90–270° line and facing 180°, Fig. 3.20. Markings on this line assist in the placement of the block.

Aim a single ray of light from the ray box along the 360–180° line so that it strikes the curved surface at its centre and passes through the centre of the flat side. Notice that the ray of light passes straight through the block.

Rotate the ray box, continuing to aim the ray at the centre point, towards the 10°, 20°, 30°, and 40° lines. Notice the bright refracted ray and the weak reflected ray and as the angle of incidence increases, the brightness of the reflected ray also increases. Measure the angles of incidence *i*, angles of refraction *r* and angles of reflection *r'*.

Continue to rotate the ray box increasing the angle of incidence further until the refracted ray is along the flat side and there is a bright reflected ray. Measure the angle of incidence *i,* which is the critical angle, and the angle of reflection *r'*.

Continue to rotate the ray box beyond the critical angle and notice that there is no longer a refracted ray but a bright reflected ray. Measure the angle of incidence *i* and the angle of reflection *r'*.

Observations

1. As the angle of incidence increases the angle of refraction also increases, up to the critical angle of incidence.
2. At a certain angle of incidence, called the critical angle, the refracted ray disappears and a bright reflected ray is obtained at an angle of reflection equal to the angle of incidence. Note that the sine of the critical angle is the reciprocal of the refractive index of the material in the block.
3. The angle of incidence at all times is equal to the angle of reflection whether refraction or total internal reflection is taking place.
4. At small angles of incidence, the reflected ray is weak but gets much brighter as the critical angle is approached.

6. Converging Lens

Aim a single ray of light along the middle line of a sheet of graph paper. Mark both ends of this ray with a pencil. Place the converging lens centrally in the path of the ray and adjust its position so that the light goes through the lens undeviated, Fig. 3.21. This line now represents the principal axis of the lens. Mark the outline of the lens with a pencil and mark the rays on either side of the lens with marks of the same number.

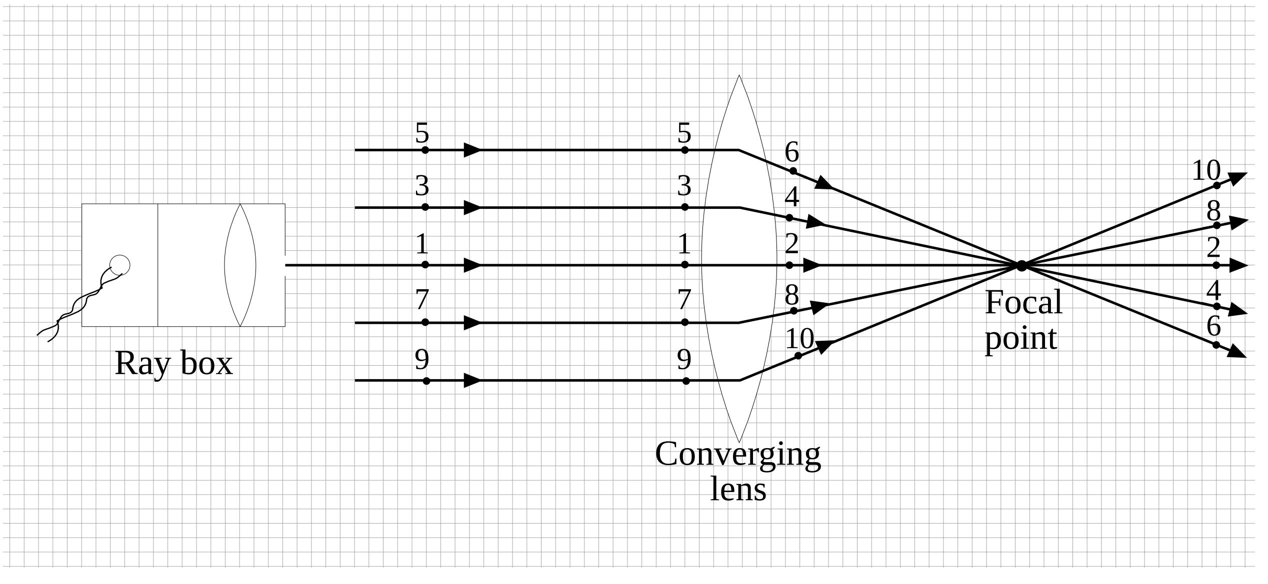


Fig. 3.21 *Refraction through a converging lens*

Move the ray box to a new position, not too far away from the principal axis, so that the ray is parallel to the principal axis. Mark the rays on either side of the lens with two pencil marks of the same number. Repeat above for one more ray. Now move the ray box to the other side of the principal axis so that the ray is still parallel to the principal axis. Mark the two rays as before and repeat for one more ray. Remove the lens and join up the same-numbered points on both sides of the lens. The point where the rays meet after passing through the lens is the focal point and the distance between this point and the axis of symmetry of the lens is the focal length.

Place the lens on a sheet of new graph paper and place the lens centrally on one of the lines of the graph paper near the middle. Aim three parallel rays onto the lens so that the middle ray passes undeviated through the optic centre. Mark the point on the principal axis where the three rays meet. This point is the focal point.

Now rotate the ray box, at a small angle to the principal axis, so that the middle ray still passes through the optic centre undeviated, Fig. 3.22. Mark the point where the rays cross after passing through the lens. Repeat this on the other side of the lens.

Aim the rays at larger angles from the axis. For each position, mark the point where the rays cross after passing through the lens. It is possible to draw a straight line through these points. The vertical plane through the focal point is the focal plane.

7. Diverging Lens

Place a sheet of graph paper on the bench and aim four or five parallel rays from the ray box along it. Adjust the position of the lens of the ray box relative to the bulb until the rays are parallel to the lines on the graph paper. Place the diverging lens centrally on the graph paper and adjust its position until the middle ray passes through undeviated, Fig. 3.23. Mark each of the rays on either side of the lens with two numbered pencil marks.

Remove the lens and join up the same-numbered pencil marks on either side of the lens. Continue the emergent rays back until they meet at one point. This point is called the focal point of the lens. It is a virtual focus since rays of light do not actually pass through it.

Ray diagrams for spherical lenses

Place a converging lens on a sheet of graph paper and mark its outline with a pencil. Place a dot on the graph paper, Fig. 3.24, and aim a single ray of light, ray B, through the dot so that it passes through the optic centre of the lens undeviated.

Mark the position of this ray with two numbered pencil marks, 1, 1, one on either side of the lens. Rotate the ray box so that ray A through the dot is parallel to the principal axis of the lens. Mark the incident and refracted rays with numbered points 2, 2 and 3, 3, respectively. Now rotate the ray box so that ray C through the dot passes through the focal point in front of the lens. Mark the incident and refracted rays with numbered points 4, 4 and 5, 5, respectively. Join up the same-numbered points in front of and behind the lens.

Observations

Ray A. An incident ray of light parallel to the principal axis is refracted so as to pass through the focal point.

Ray B. An incident ray of light passing through the optic centre of a lens is refracted along a line which is parallel to the incident ray and only slightly displaced from it (the thinner the lens the smaller the displacement).

Ray C. An incident ray of light passing through the focal point of a converging lens emerges on the other side of the lens parallel to the principal axis.

Repeat the above procedure using a diverging lens, Fig. 3.25.

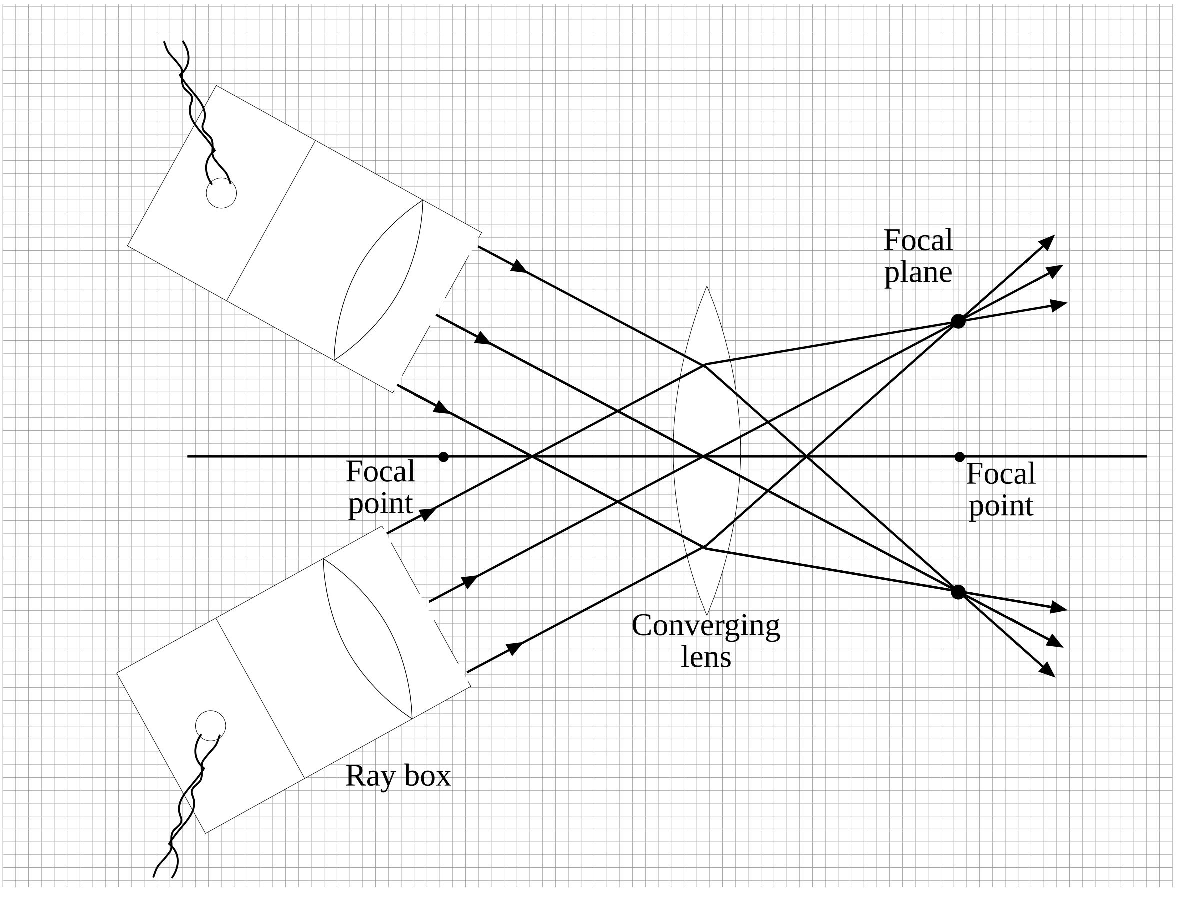


Fig. 3.22 *The focal plane of a converging lens*

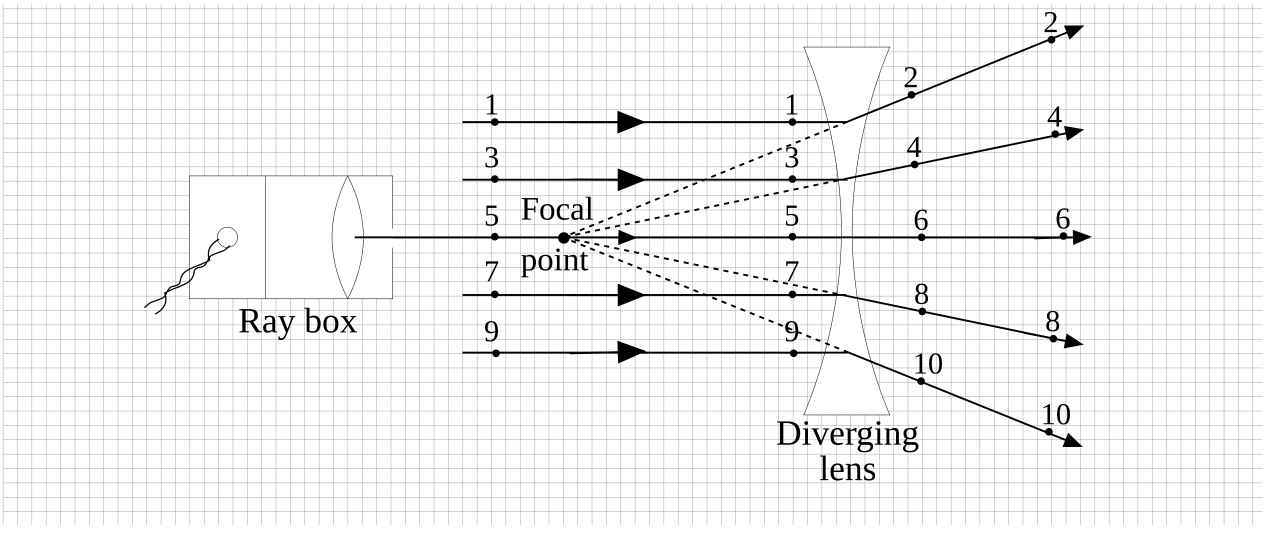


Fig. 3.23 *Refraction through a diverging lens*

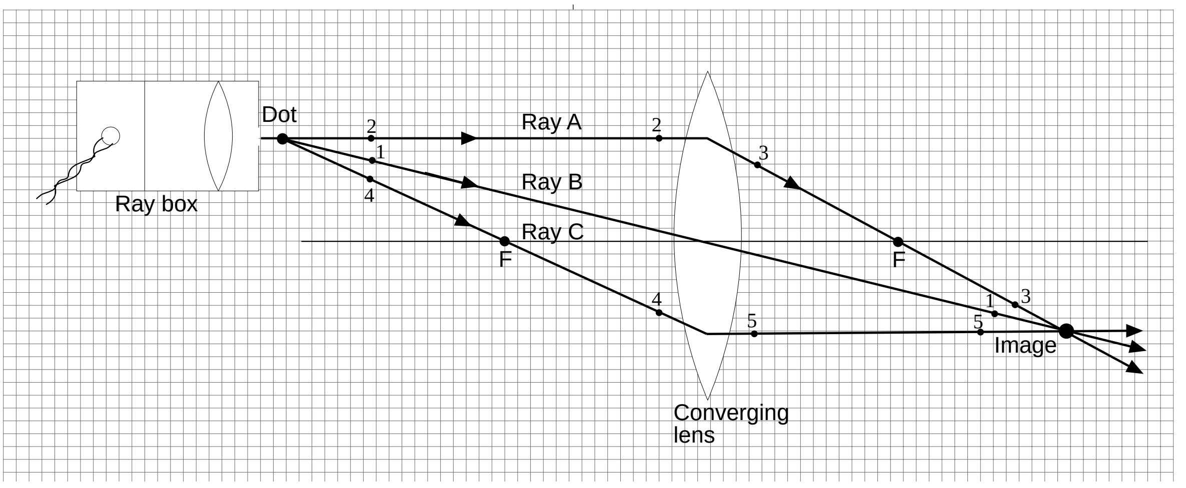


Fig. 3.24 *Ray diagram for a converging lens*

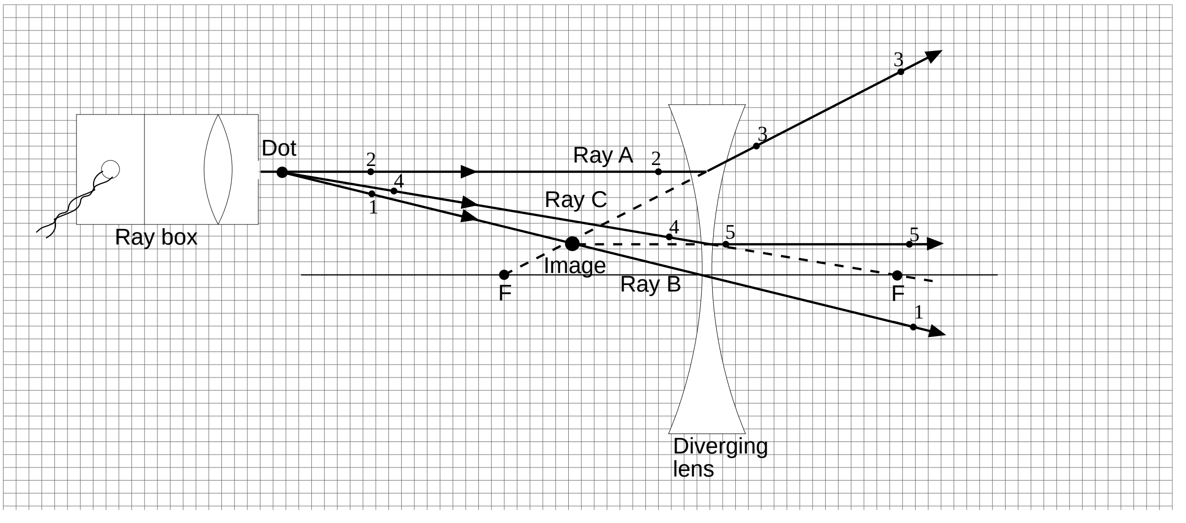


Fig. 3.25 *Ray diagram for a diverging lens*

8. Re-combining the Colours of the Spectrum Using Two Prisms

Aim a single ray of light from the ray box at one side of the equilateral triangular prism in a position of minimum deviation to produce a clear spectrum on a screen placed vertically to intercept the light. Place a second similar prism, the opposite way round to the first, Fig. 3.26, in the path of the dispersed beam, making sure that the sides of the prisms are close and parallel to each other.

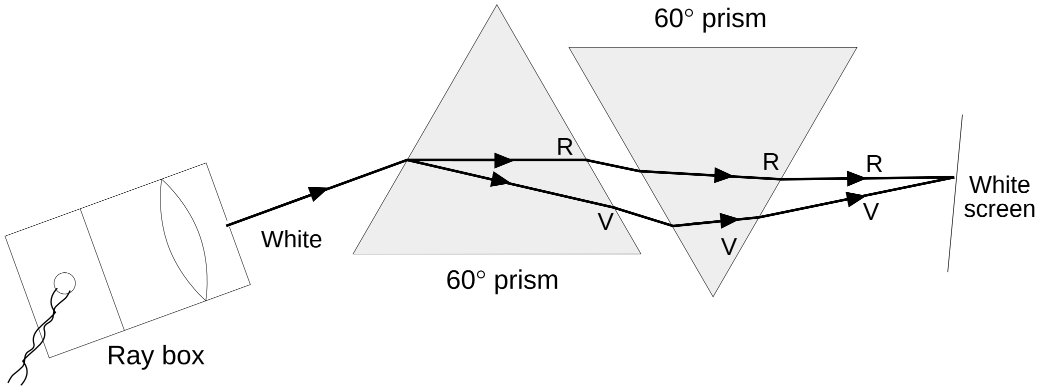


Fig. 3.26 *Recombining of colours*

The second prism re-combines the colours of the spectrum to form white light on the screen.

To produce a purer spectrum

Aim a narrow beam of light from the ray box along a sheet of white paper and place a white screen vertically in the path of the beam. Place a converging lens so as to form an image on the screen. The prism is now placed in the path of the light from the lens. Move the screen (or the lens) until the colours do not overlap and the spectrum is almost pure.

To produce a pure spectrum

Adjust the ray box to give a divergent beam on a sheet of white paper. Place a converging lens between the ray box and the prism so that the beam of light incident on the prism is parallel. Rotate the prism to find the approximate position of minimum deviation. Place the second converging lens between the prism and a white screen (white card) and adjust its position to focus the various colours of the spectrum on to the screen.

This also illustrates the principle of the spectrometer.

Observations

In all experiments it is essential that the optical elements, objects, images, mirrors, lenses, screens, etc., are all aligned on the same optical axis, which is parallel to the scale along the optical bench if one is used. To achieve this bring all optical elements in their holders together and adjust their vertical heights until they are all approximately the same. If all the optical elements have the same vertical height it should only be necessary to tilt a mirror or lens slightly in its holder so that they all have a common optical axis. The simple theory of mirrors and lenses is valid for rays close to the optical axis only.

3.3 Optical Illusions

1. Real Image Using Concave Mirror

Forming a real image by using a concave mirror is a means of producing an optical illusion. Place a concealed illuminated object, such as a lighted bulb, Fig. 3.27, at the centre of curvature of a large-aperture (about 20 cm) concave mirror.

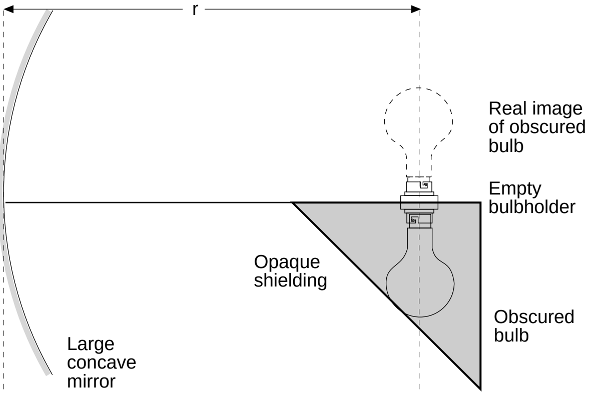


Fig. 3.27 *Concave mirror producing a real image*

Adjust the position of the illuminated object until the object and image coincide by the method of no parallax. Make sure that the shielding is painted matt black to reduce reflections to a minimum.

2. Pepper's Ghost

This is a method by which ghosts may be 'created' on the stage of a theatre. A large sheet of very clean polished glass is set at an angle of 45° to the front of the stage and acts as a mirror but at the same time allows objects on stage to be seen through it. With the footlights out the glass is invisible to the audience. An actor, dressed as a ghost, is hidden in the wings at the front of the stage and is illuminated strongly. Everything else around the actor is painted matt black so that only an image of the actor in the glass is visible to the audience.

3. Candle 'Burning' in Water

This is a laboratory illustration of Pepper's ghost. Fix an unlit candle to the bottom of an empty beaker by melting some candle wax and conceal a lighted candle behind a shield. Both candles should be the same length. Place a large sheet of very clean polished glass vertically at right angles to the line joining the candles and half way between them, Fig. 3.28.

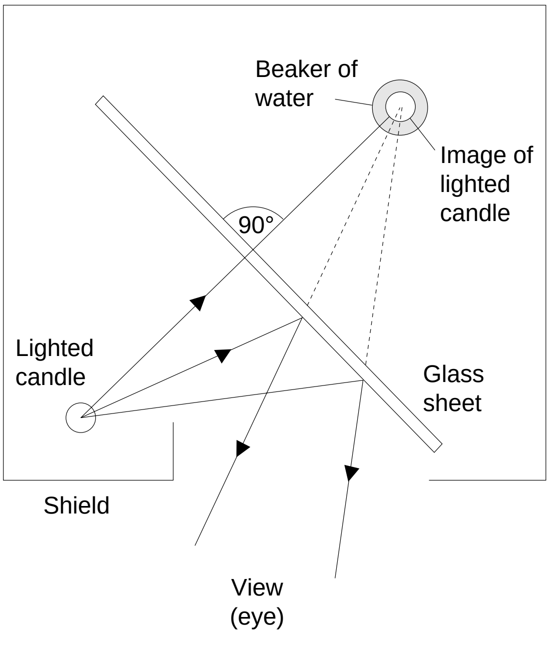


Fig. 3.28 *Pepper's ghost*

Adjust the position of the lighted candle until the unlit candle in the beaker appears to be alight. Now pour water into the beaker until the candle and the 'flame' are submerged and the candle appears to be burning in the water.

3.4 Lasers

The emergence of the laser as a scientific tool has greatly changed the practice of science. Lasers are used in geo-astronomy (lunar ranging) biology (optical trapping of bacteria, fluorescent tagging), electrical engineering (optoelectronics, optical computing), physics (laser-induced fusion, atomic spectroscopy), meteorology (air velocity measurements, wind shear detection), and more. Laser applications include bar code reader, laser beam printing, counter measurement, size measurement, laser surgery, surveying (light source for laser theodolite), optical card memory, cutting and welding, optical disc memory, information storage and retrieval from CD-ROMS, medical equipment (laser acupuncture, analysis) and fibre optic communication. The laser provides an opportunity to demonstrate a wide range of optical principles and phenomena in the study of light. Many of these principles and phenomena are difficult or impossible to demonstrate to a group of students by other means.

Safe Use of Lasers

Although laser light consists of the same electromagnetic radiation as ordinary light it does possess several distinguishing properties, e.g. extreme brightness, low divergence, and substantial phase coherence.

The following rules should be observed when using lasers.

1. Never look directly into the laser beam or into any direct reflections of the beam or aim it into the eyes of anyone else. Everyone in the room should be made aware of the dangers and should be told to close their eyes and look away should they be dazzled. The presence of the laser beam should always be detected by a white non-reflective card and never with the naked eye.

2. Watch out for reflections of the beam off shiny objects if the laser is being moved around. Cover or remove any bright reflective objects that might reflect the laser beam. Since the path of the direct beam is so well-defined eliminate any stray reflections or block them off if possible with non-reflective objects such as dark-coloured paper.

3. Be careful when placing optical devices , e.g. lenses, mirrors, etc., in the path of the beam. Move slowly, and try to predict the direction of the beam after it ‘interacts’ with the device.
4. When the beam is used in a horizontal plane, and must travel long distances, keep it away from eye level. Use as high a level of background lighting as is consistent with the experiment being performed. This reduces the size of the eye’s pupil to a minimum and provides some protection. Where work in darkness or minimum illumination is inevitable then easily accessible light switches should be provided. Unfortunately, laser goggles which block the 632.8 nm radiation of the helium neon laser are not practical since many of the demonstrations cannot be seen if laser goggles are worn.
5. Beware of the high voltages used in He–Ne lasers, typically 2000 V. It is important never to expose the laser tube when the laser is switched on. Drinks should not be allowed in the laboratory, and great care should be taken when performing experiments involving liquids.

Safe On/off Operation of a Laser

1. Before plugging in the laser check that the on/off switch (usually a key operated switch) is in the off position, locate the exit port from which the beam will emerge, and make sure that the shutter is closed (some lasers do not have this feature).
2. Set the laser on a firm surface, and place a ‘beam block’ about 10 cm away from the exit port and in line with the laser axis. The block should be made of wood or metal and painted matt black.
3. Switch on the laser and open the shutter to check that the laser is operating.
4. With the laser operating, the beam block can be moved back from the laser, always blocking the beam, to allow room for the experiment to be performed.
5. When work with the laser is complete, the shutter should be closed, the laser switched off and the key removed for safe keeping.

Precautions in respect of the use of lasers should be displayed prominently in the laboratory.

Laser Diodes

Laser diodes emit red light that is clearly visible to the human eye. (They produce visible light in the wavelength range 660–670 nm.) There is a substantial reduction in size, mass and power consumption, when compared to He–Ne lasers. They also have a high reliability.

Laser Diffraction Kits

Laser accessories and diffraction kits usually contain some or all of the following items.

Six slides of 1 to 6 slits, diffraction gratings, single slit – tapered, double slit – tapered, metal gauze, circular apertures, hologram, polaroids, adjustable single slit, slide holders, lens holders, Young’s slits, set of lenses, slide holder, air cell, microscope slides, laser goggles, letter slides, beam splitter, glass scale and holder, educational diffraction grating.

Screen

The screen can be a large projector screen, self-supporting like the one used for overhead projectors or slide projectors, or a white card held in the hand. In all cases the screen should be matt white and not shiny or glossy.

Making the Beam Visible

A white card held perpendicularly in the beam will show the point where the beam strikes the card as a red spot. If the card is held almost parallel to the beam at glancing incidence the beam will show up as a red line. Smoke or chalk dust scattered in the path of the light will show up the beam in air. In water, a few drops of milk, or Dettol or fluorescein can be added to the water to scatter the beam and show up its position. It is best to use distilled or filtered water.

Specular (Regular) and Diffuse Reflection

In a darkened room, aim the laser beam at objects with a rough appearance. Observe the spot of reflected light that appears on a screen.

This is irregular or diffuse reflection. These objects both reflect and scatter the light. The light is reflected from the object in all directions. This is also illustrated by the fact that the spot where the laser beam strikes the object can be seen from every point in the room.

Now aim the laser beam at objects that have a smooth surface and observe the spot on a screen. This is regular or specular reflection and does not scatter or diffuse the light.

Substitute a plane mirror for the screen and observe any reflections on a wall or ceiling. Make sure the reflected light is directed away from students. This is another example of regular reflection.

Reflection and Absorption

Aim the laser beam at white, and various coloured, objects. Observe the intensity of the reflected light from these illuminated objects. Try to deduce which surfaces are the best reflectors and which surfaces are the best absorbers.

Verification of the Law of Reflection

Place a photocopy of the 0–360° protractor on a bench and place a plane mirror, front surfaced if possible, vertically and centrally on the 90–270° line so that the reflecting surface faces 360°.

Place the laser along the 180–360° line. Aim the laser at the point of intersection of the 90–270° line and the 180–360° line, so that the reflected light comes back along its own path to the laser aperture. The position of the laser should be adjusted so that a red line shows up on the paper. This is the normal to the mirror surface. We will use this point of incidence for other incident rays. Aim the laser beam at this point for various angles of incidence. Record the angle of incidence i and the corresponding angle of reflection r in each case for the laser beam. These angles will be found to be equal.

If it is difficult to get the red lines to show up grazing along the paper, a pin can be placed vertically to intercept the laser beam and a hole marked on the paper at that point directly under the pin. Two holes, 1 and 2 can mark the incident ray and two more holes, 3 and 4, the reflected ray. Points 1 and 2 are joined to form the incident ray and 3 and 4 to form the reflected ray.

Multiple Reflections

A plane mirror is usually made of a layer of silver or an amalgam on the back of a plane sheet of glass – the glass protects the silver from scratches and tarnishing. This give rise to multiple reflections. Not only is the light reflected from the front and rear surfaces of the glass but also within the glass giving other reflected rays.

Plane Mirror

Aim the laser beam at the mirror, Fig. 3.29, at an angle of incidence around 45° and pick up the reflected light on a white card acting as a screen. Notice that the second image, the reflection from the back of the mirror, is the brightest.

Glass Sheet

Aim the laser beam at the sheet of glass in the same way as the plane mirror. Observe the reflected and refracted images on white cards, Fig. 3.30. A glass sheet also gives rise to multiple reflections in much the same way as a back-silvered mirror. Since only about 10% of the incident beam is reflected, the first two reflections have about the same intensity and interference takes place in the region where they overlap on the screen. The first refracted image is extremely bright and the rest are extremely weak.

Absorption

Shine the laser beam through various colour filters, coloured plastic, and various liquids, and observe the intensity of the emergent laser beam. Vary the thickness of the filter and observe what happens.

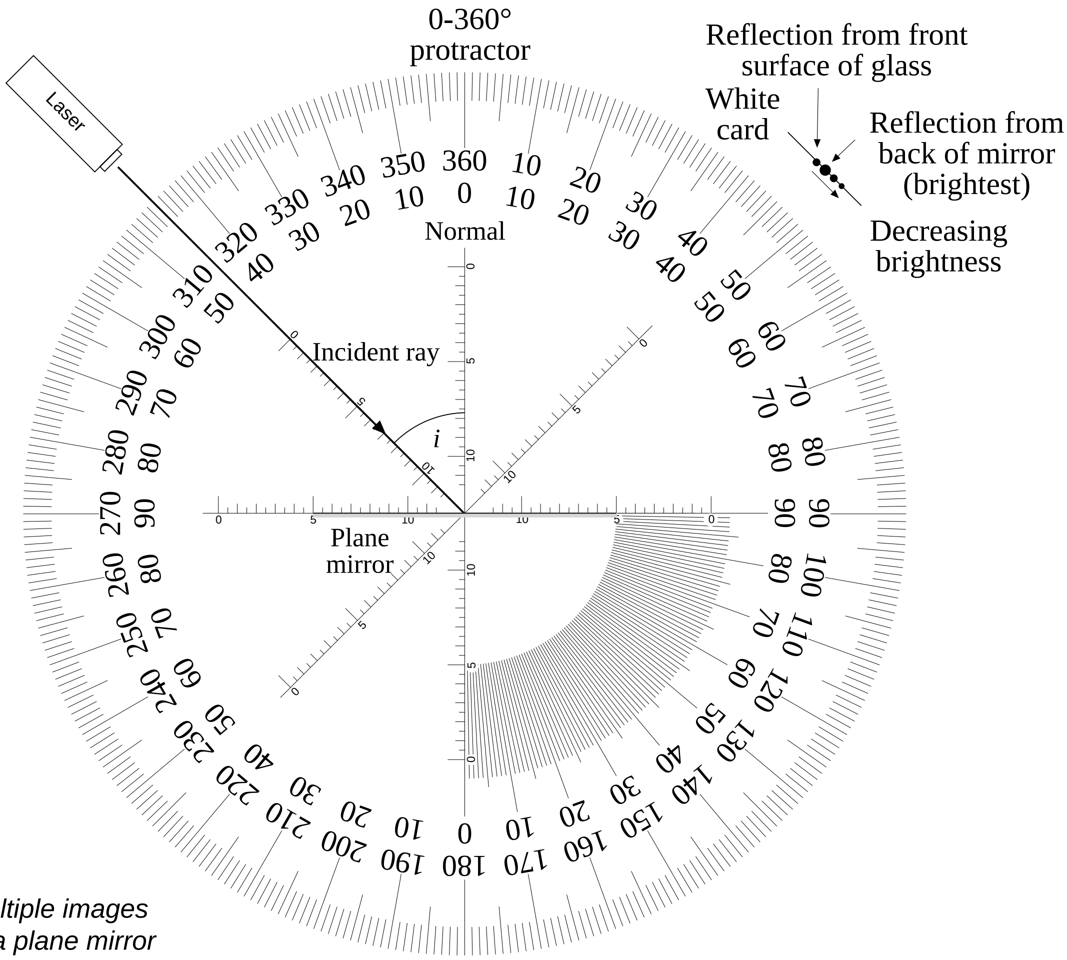


Fig. 3.29 *Multiple images in a plane mirror*

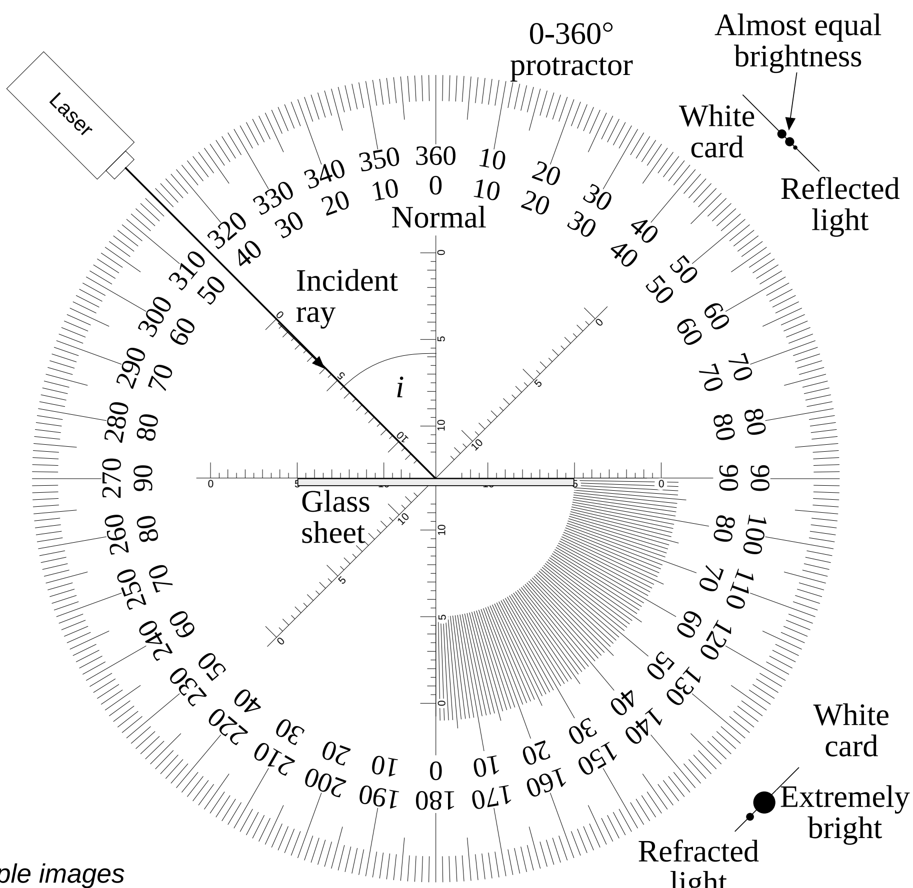


Fig. 3.30 *Multiple images in a glass sheet*

The special green filters used in laser goggles attenuate the laser beam more than one thousand times. Because of this attenuation it is not possible to see the beam under most conditions.

Ice Sheet

Aim the laser beam at a sheet of ice at an angle of incidence of about 45° in a darkened room. The reflected and refracted laser light will give various patterns depending on the thickness and crystal structure of the ice.

Slide the edge of the sheet of ice part way into the path of the laser beam and observe the pattern on a screen. It should be possible to obtain a streak of light on the screen by adjusting the ice.

Refractive Index of Glass or Perspex

Place a rectangular glass or perspex block in the centre of a white sheet of paper and carefully outline its position with a sharp pencil. Aim the laser beam so that it enters the glass at an angle as shown in the diagram, Fig. 3.31.

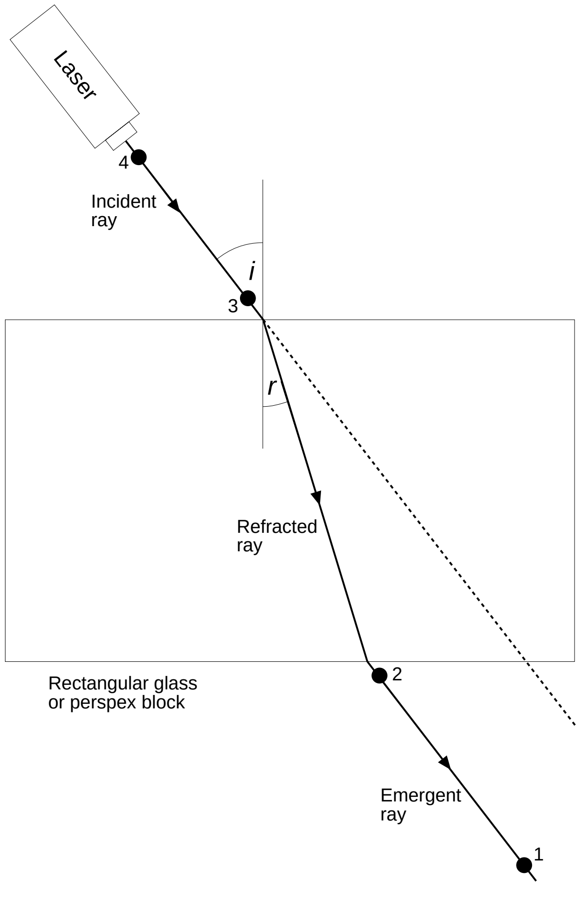


Fig. 3.31 *Verification of the law of refraction*

Place four pins vertically into the paper to intercept the laser beam in the order 1, 2, 3, and 4. Order is important since if pin 4 were inserted first it would block the laser light and it would not be possible to insert the other pins. Make sure the pins are inserted vertically and that the laser beam hits the pins near the paper.

Remove the glass block and draw lines joining pins 1 and 2 to indicate the emergent ray and 3 and 4 to indicate the incident ray. Draw a line also on the paper to show the path of the laser beam when it was inside the glass block, i.e. a line joining the point of incidence to the point of emergence.

With the aid of a protractor (a transparency of the 0–360° would do), construct the normal at the point of incidence. Measure the angle of incidence i and the angle of refraction r. Calculate the refractive index from the ratio $\sin i/\sin r$. Vary the angle of incidence and measure the angle of incidence and refraction for a range of angles. Calculate the average refractive index.

Plot a graph of $\sin i$ (on the y-axis) against $\sin r$ (on the x-axis). The slope of this graph is the refractive index.

The experiment may be repeated using a semi-circular glass block.

Refractive Index of a Liquid

Place a few drops of Dettol into a tank of distilled or filtered water. Stick a 0–360° plastic protractor to one side of the tank so that the middle line of the protractor is along and parallel to the level of the water. Aim the laser beam into the water just inside the tank at an angle of incidence *i,* Fig. 3.32.

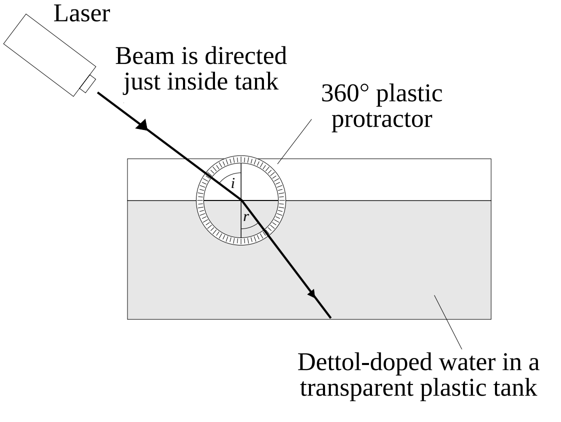


Fig. 3.32 *Refraction in a rectangular tank of Dettol-doped water*

Note the angle of incidence and angle of refraction from the protractor.

Calculate the refractive index from the ratio sin i/sin r. Repeat for different angles of incidence and calculate the average refractive index.

Plot a graph of sin i (on the y-axis) against sin r (on the x-axis). The slope of this graph is the refractive index.

Continuous Refraction

If a tank of liquid varies in density, and therefore in refractive index, with depth then a ray of light passing through it will follow a curved path. When the refractive index increases with depth, the light is bent downwards and when the refractive index decreases with depth, the light is bent upwards towards the surface. This is what happens when a mirage is seen.

Place a layer of ordinary sugar about 1 cm thick (a row of sugar cubes is ideal for this) at the bottom of a transparent rectangular tank. Place about 15–20 cm of Dettol-doped water on top of the sugar layer without disturbing the sugar, Fig. 3.33.

Allow the sugar to dissolve undisturbed for a few days. The rate of diffusion of sugar into the water is slow.

Aim the laser beam upwards into the side of the tank. The laser beam will bend into a very pronounced curve as it encounters changes in the refractive index of the solution. It may be necessary to leave the tank for a week, or even more, to get the desired effect.

This phenomenon may also be demonstrated using a hotplate. Aim a laser beam horizontally at a wall or screen, Fig. 3.34, and mark its position with a pencil. Place a hotplate directly and closely underneath the laser beam. Mark the new position of the laser beam on the screen or wall. The beam bends upwards due to the different refractive indices of the air above the hotplate.

Total Internal Reflection in a Round-bottomed Flask

Half fill a 500 ml round-bottomed flask with Dettol-doped water and fill the remainder with smoke. Cover the top with a piece of clear plastic or glass, Fig. 3.35. Aim the laser beam normally through the bottom of the flask.

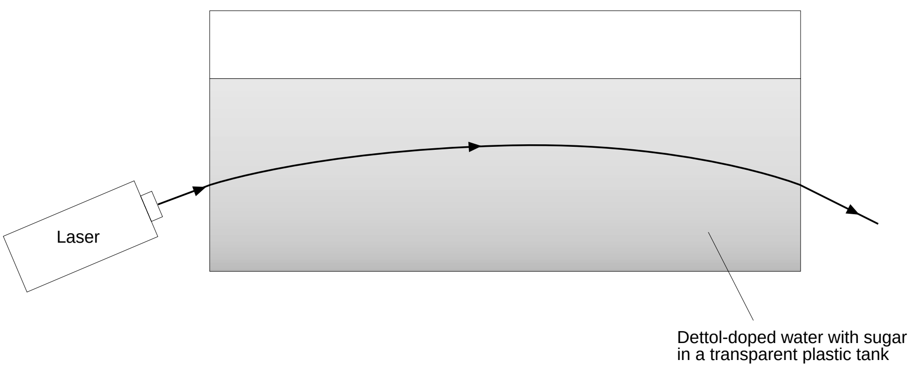


Fig. 3.33 *Downward curve*

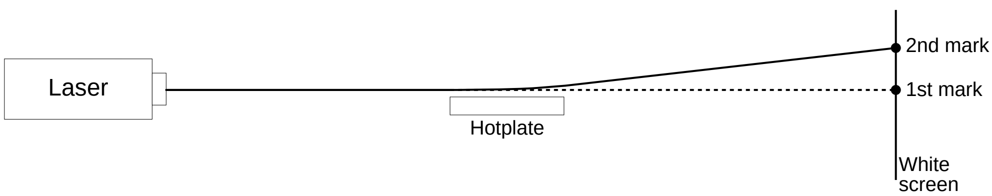


Fig. 3.34 *Upward curve*

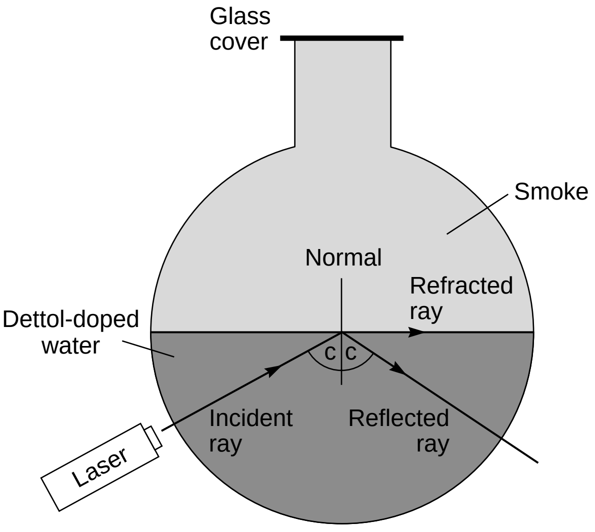


Fig. 3.35 *Total internal reflection in a round-bottomed flask half-filled with Dettol-doped water*

Vary the angle of incidence, keeping the point of incidence the same and observe the weak reflected ray and the bright refracted ray. Notice that at the critical angle the refracted ray is along the boundary between the water and the smoke. Measure this angle. It should be around 49°. Increase the angle of incidence further and notice that the refracted ray disappears and a bright reflected ray is obtained. Use a protractor to measure the angle of incidence and angle of reflection. They should be the equal.

Total Internal Reflection in a Test-Tube

Fill a test-tube with Dettol-doped water. Place the test-tube in a retort stand so that it is at an angle of about 45°. In a darkened room aim the laser beam so that it enters the bottom of the test-tube and emerges at the top. Now slowly vary the angle at which the laser beam hits the test-tube, Fig. 3.36.

It should be possible to see the beam bouncing along the tube. Notice how, in general, the path of the beam rotates or corkscrews along the tube axis. By careful adjustment of the laser beam this corkscrewing can be eliminated.

Total Internal Reflection in a Water Jet

Place a glass tube about 0.5 cm internal diameter in the bottom side socket of an aspirator bottle of about 5 litre capacity, Fig. 3.37. Fill the bottle with

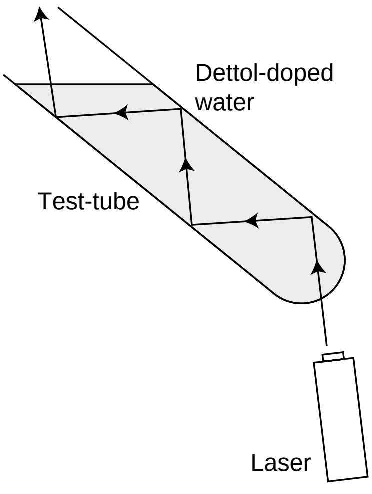


Fig. 3.36 *Dettol-filled test-tube illustrating total internal reflection*

Dettol-doped water and place a stopper in the neck.

In a darkened room aim the laser beam so that it enters the side of the bottle and emerges through the glass tube. Remove the stopper and adjust the laser beam so that it is totally internally reflected each time it tries to leave the edges of the water jet and will change its arc as the water pressure decreases in the bottle. By using different size glass tubes and adjusting the stopper in the neck of the bottle, various sizes and arcs of water jets can be investigated. If it is not possible to obtain an aspirator bottle then a one and a half litre soft drinks bottle with a small round hole punched in its side will do.

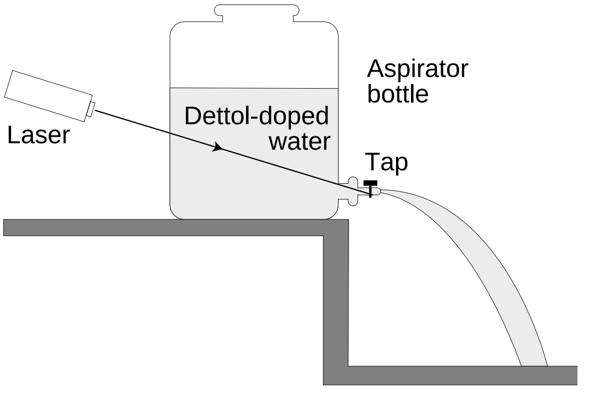


Fig. 3.37 *Total internal reflection in a Dettol-doped water jet*

Deviation and Dispersion

Aim a laser beam at a white screen and mark its position with a pencil. Place a prism in the path of the beam so that the base of the prism is roughly parallel with the laser and mark the new position of the spot on the screen. It will be observed that the light is deviated but is not dispersed (see p. 32).

Polarisation

Aim the laser beam at a white screen and note the brightness of the spot. Place a polaroid filter in the path of the laser beam and again note the brightness of the spot. Rotate the filter through 360°. The polaroid filter absorbs at least half the intensity, but when rotated does not prevent light reaching the screen. Hold a second polaroid by hand, Fig. 3.38(a), and rotate it in the path of the beam. Observe the variations in intensity of the spot as the second polaroid is rotated. When no light reaches the screen the polaroids are at right angles, Fig. 3.38(b).

Diffraction at a Single Slit

Aim the laser beam at a white screen about 4 m away. Place a single slit vertically, centrally and at right angles between the laser beam and the screen. Several maxima and minima will be visible on the screen. Observe the effects on the diffraction pattern as the width of the slit is changed using either a single tapered slit or a slit of variable width. A very narrow slit gives a wide smeared out pattern becoming broader as the slit is made narrower. The diffraction pattern will consist of a wide central bright fringe with a series of less intense narrower fringes on either side. A wider slit gives a very compact pattern.

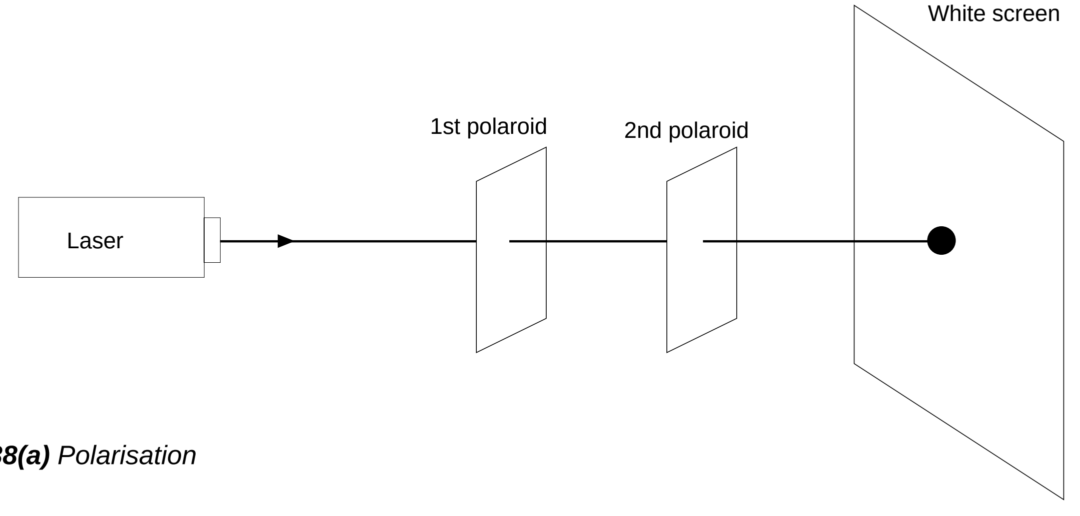


Fig. 3.38(a) *Polarisation*

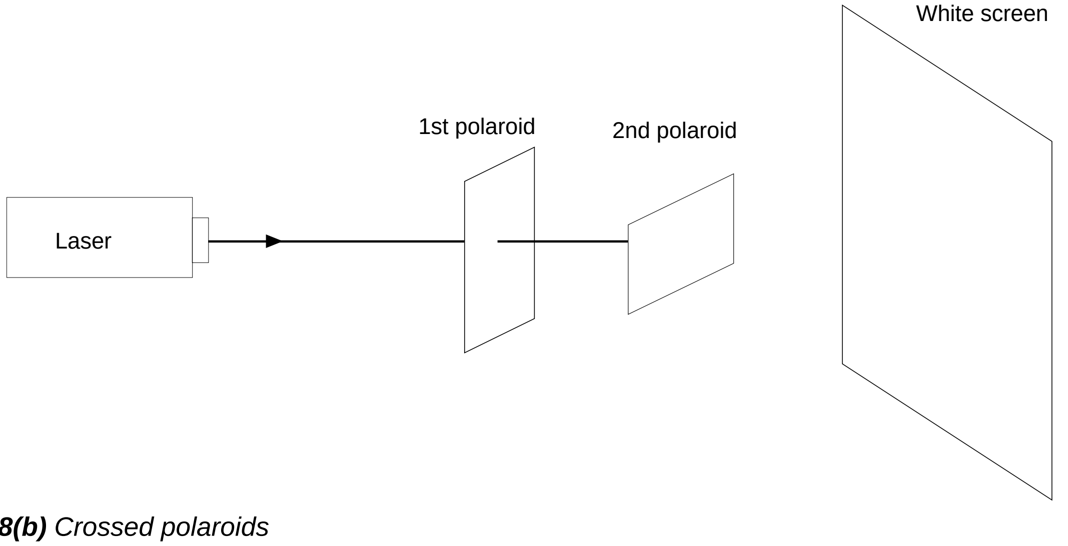


Fig. 3.38(b) *Crossed polaroids*

Diffraction Using a Shiny Steel Ruler

Aim the laser beam at the bottom of a white screen a few metres away. Mark the position of the spot on the screen. Move the ruler into position so as to intercept the laser beam at a grazing angle, Fig. 3.39, so that the laser light is diffracted by the mm divisions of the ruler.

Diffraction at a Double Slit

Aim a laser beam at the centre of a white screen 5 m away with the double slit placed vertically in the beam. A series of red dots is produced on either side of the centre line. Measure the average distance, x, between successive bright dots, the distance D between the double slit and the screen, and the distance d between the slits using a travelling microscope. Calculate the wavelength, λ, using the formula $\lambda = dx/D$.

The distance between the slits, d, can also be found by projecting an image of the slits onto a screen using a single converging lens. Measure the distance between the images of the slits, s, the object distance, u, and the image distance, v. Since $s/d = v/u$, the value of d may be calculated.

Wavelength of Laser Light Using a Diffraction Grating

Aim a laser beam at the centre of a white screen 5 m away with the diffraction grating placed vertically in the beam half way in between, Fig. 3.40. A series of red dots is produced on each side of the centre line.

Measure the average distance x between successive bright dots for the first order diffracted pattern, the distance D between the double slit and the screen, and record the number of lines per metre, N, on the grating. The slit spacing is the reciprocal of this number, i.e. $d = 1/N$.

Calculate the wavelength, λ using the formula $n\lambda = d \sin \theta$. The angle θ can be calculated from $\tan \theta = x/D$.

Measure the distance x for the second, third, and fourth orders (the fourth order may not exist if d is less than about 2.6 μm). Calculate an average value for the wavelength of the laser light.

Diffraction by Single Objects

A. Ball bearing

Aim a laser beam at the centre of a white screen a few metres away. Place a small ball bearing in the path of the beam. Unless the object is small the laser beam will have to be expanded using a diverging lens or a reversed telescope or a medium power microscope objective. This gives rise to a circular shadow surrounded by bright and dark rings. Note that there is a bright spot in the centre of the shadow.

B. Razor blade

Aim the laser beam at a white screen a few metres away. Slide the edge of a new razor blade part way into the path of the laser beam and observe the pattern on the screen. A series of bright and dark fringes parallel to the edge of the razor blade will be seen on the screen.

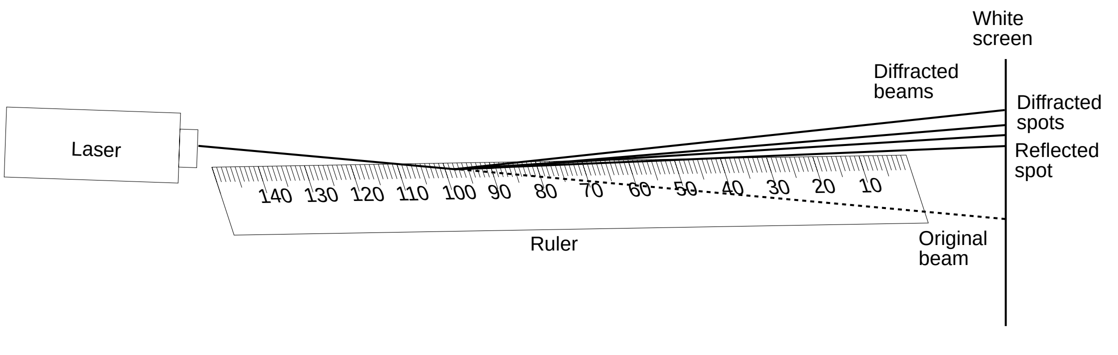


Fig. 3.39 *Ruler as diffraction grating*

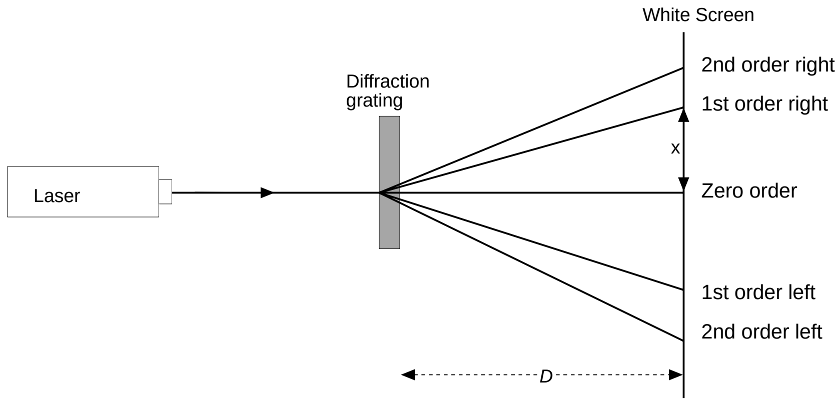


Fig. 3.40 *Measuring the wavelength of laser light using a diffraction grating*

C. Circular aperture

Make a very small pin hole in a fresh piece of aluminium foil. Aim the laser beam at the hole and observe the diffraction pattern on a screen a few metres away. If the hole is small enough a bull's-eye pattern will be formed on the screen. Experiment with holes of different size. It will be found that the smaller the pin hole the larger the bull's-eye on the screen. Experiment with many holes pierced in the foil.

The converse of this experiment is to sprinkle some lycopodium powder on a microscope slide and aim the laser beam through it.

D. Mesh grating

Allow the laser beam to pass through a wire mesh and observe the pattern on the screen. The pattern will have a chequered appearance. Replace the wire mesh with netted curtain material, nylon coffee strainer, etc. Observe the pattern on the screen for each.

3.5 Interference Patterns

Photocopy Fig. 3.41 and Fig. 3.42 onto acetate sheets. Make two copies of each. If the two copies of either diagram are placed one on top of the other an interference-type pattern will be obtained similar to that in Fig. 3.43.

The effect of changing the distance between the centres (sources of waves) and of using widely spaced or closely spaced circles (different wavelengths) can be investigated.

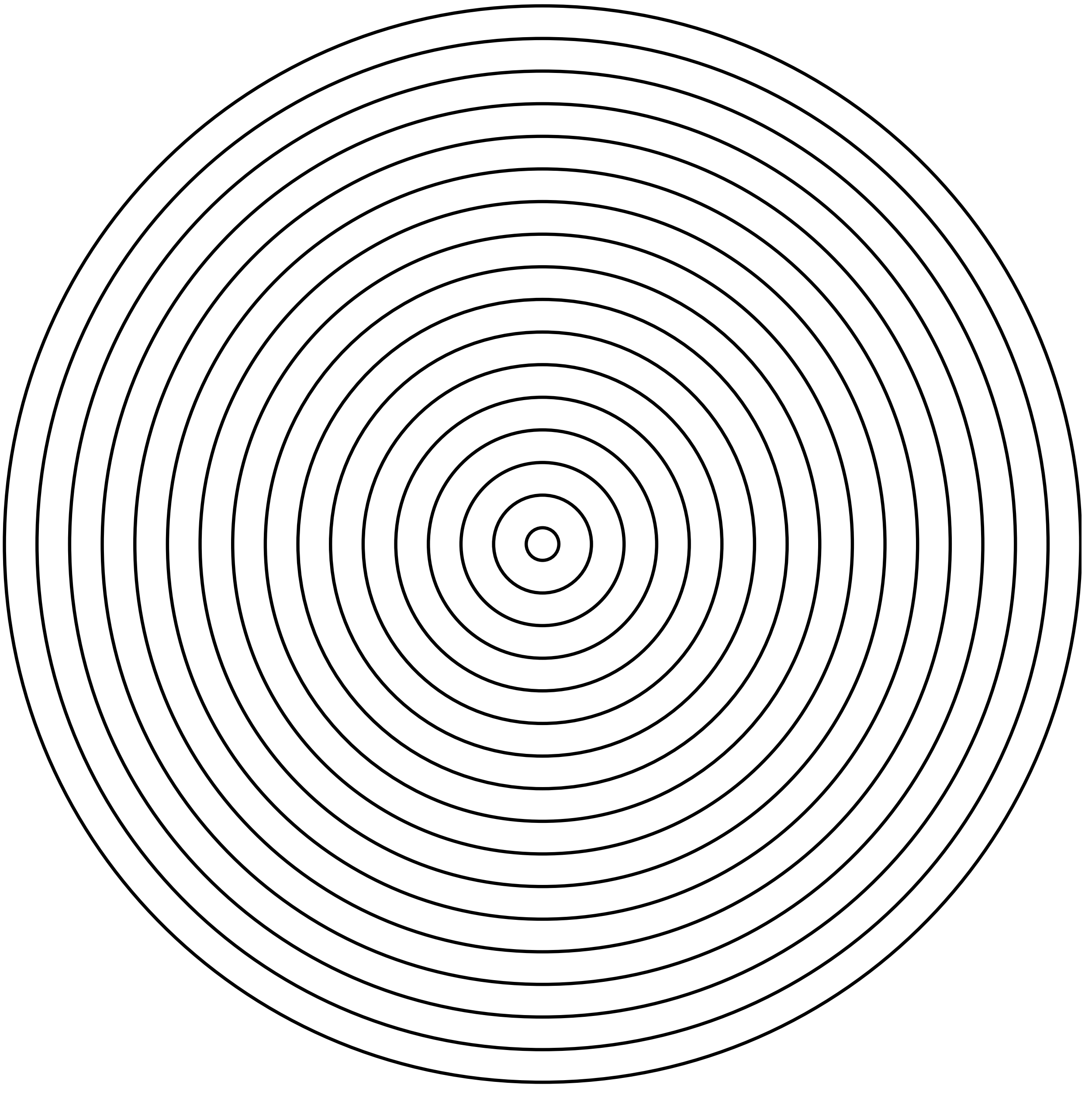


Fig. 3.41

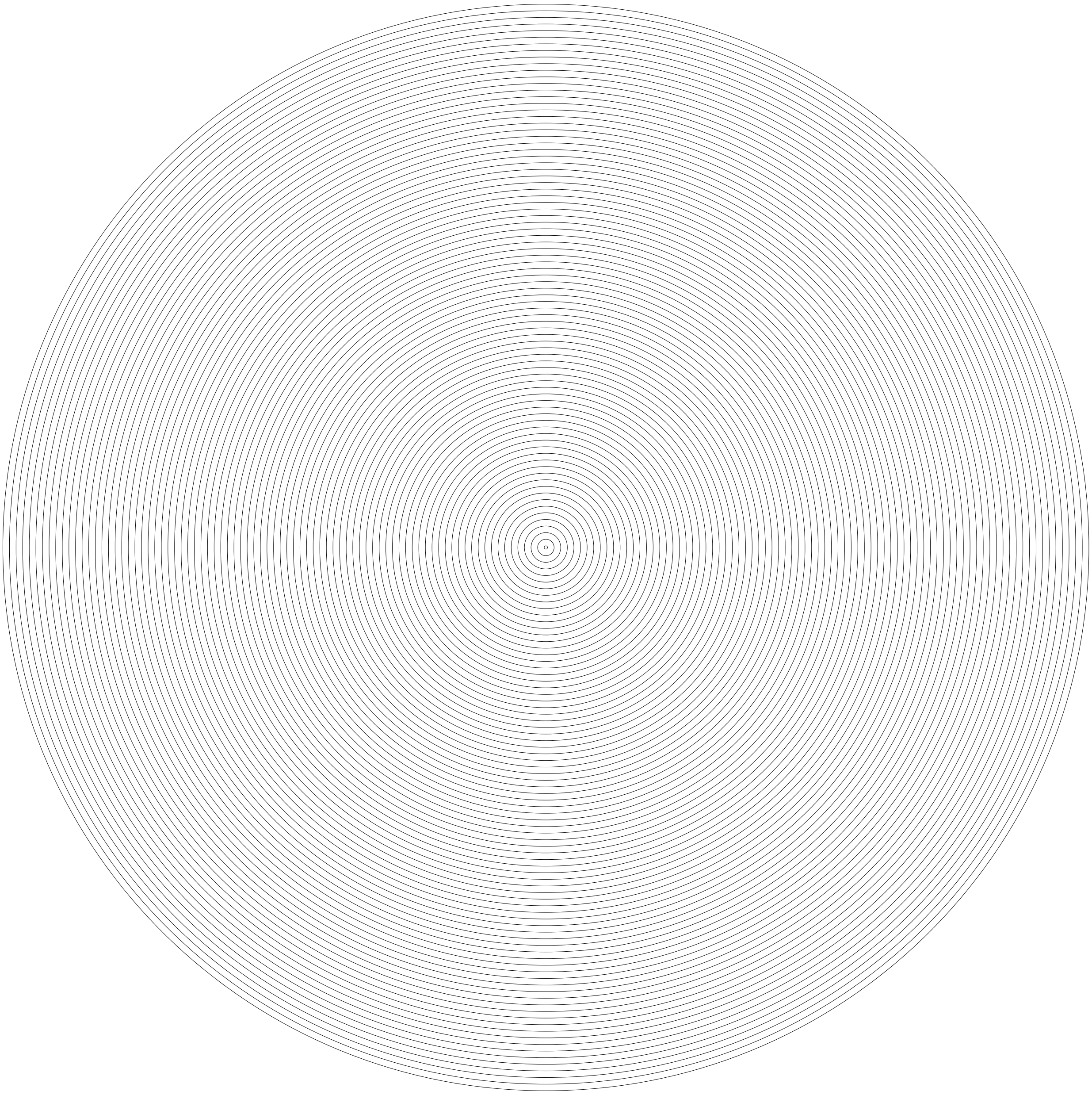


Fig. 3.42

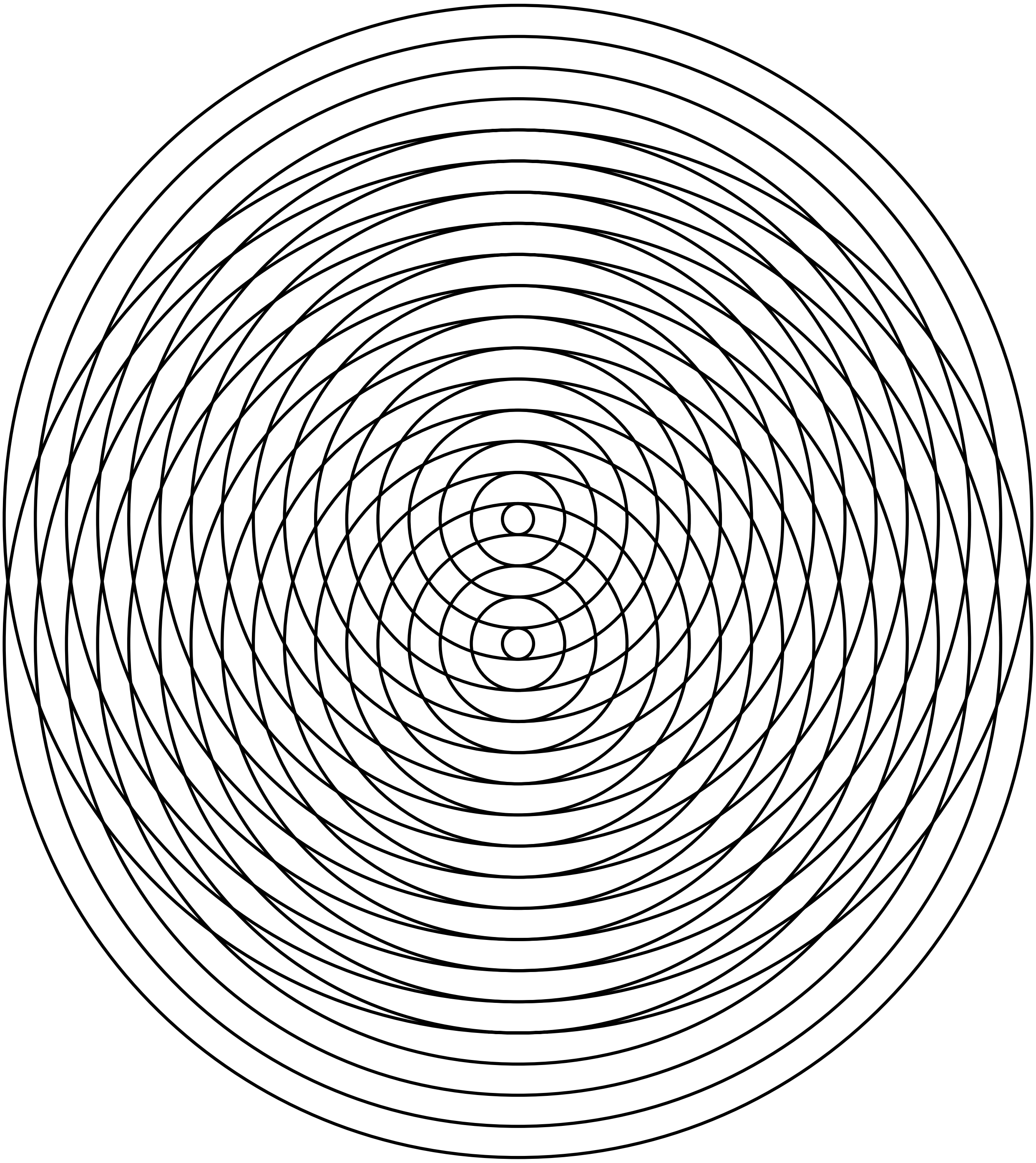


Fig. 3.43

MODULE 2

Waves

CONTENTS

CHAPTER 1 WAVES

1.1 Background

Waves may be divided into two major groups.

(1) Mechanical Waves
These include water waves, waves on a stretched string, sound waves, seismic waves (i.e. shock waves produced either artificially from explosions used by geologists in explorations or naturally occurring due to earthquakes).

(2) Electromagnetic Waves
These include radio waves, microwaves, infrared waves, light, ultraviolet rays, X-rays, and gamma rays which are due to oscillating electric charges.

While ocean waves were always known to man, as was visible radiation, many parts of the electromagnetic spectrum were not discovered until the latter years of the 19th century and the beginning of the 20th century. Radio waves were first produced and detected by Hertz in 1887, infrared rays were discovered by Herschel in 1800, ultraviolet rays were discovered in 1801 by Ritter, X-rays were discovered by Röntgen in 1895 and gamma rays were discovered and named by Rutherford in 1900.

When we think of waves, the power and might of ocean waves often come to mind and the concept of a wave as a means of transporting energy is thus part of our experience. These waves, which are due to a combination of gravitational and wind effects, are a vast reservoir of natural energy. In Edwardian times it was even thought that it might be possible to propel ships by using wave energy and patents on trying to harness this energy date back for more than a century. Since then, and particularly following the big oil price increases in 1973, ocean waves have been identified as a large unexploited source of renewable energy. However, it requires considerable ingenuity to convert the heaving motion of waves efficiently into the smooth rotation of a generator, not to mention the cost of reliably engineering large structures for survival in the worst ocean weather. As an example of the difficulties involved a small prototype plant operating off the coast of Norway was destroyed in a storm. Queen’s University, Belfast has developed an air turbine with two sets of blades which can rotate a shaft continuously in the same direction whichever way the air is blowing. This is used in oscillating-column wave power devices in which the up and down wave motion forces air in and out of a large steel or concrete chamber.

In the Republic of Ireland there are a number of wave-powered navigational buoys in use around the coast but these are only of low power. The technical difficulties and prohibitive cost have inhibited the harnessing of wave power to date but it remains a challenge for the future, particularly for island communities. Ireland is richly endowed in having some of the highest coastal wave power levels in the world. In the early eighties a survey concluded that the available power is 25 000 MW, of which 10 000 MW could be recovered. This is four times the current ESB peak carrying capacity of about 2 500 MW.

The officially highest recorded wave at sea was calculated to be 34 m. It was measured by Lt. Margraff of the US Navy from the USS *Ramapo*, while travelling from Manila in the Philippines to San Diego, California on the night of 6–7 February 1933 during a 126 km/h hurricane.

Tsunamis, (which is the correct term for what is generally called a 'tidal wave') can travel at up to 790 km/h. A tsunami 85 m high off Ishigaki Island, Ryukyo Chain on April 24 1771 threw a 750 tonne block of coral a distance of 2.5 km!

Telecommunication, meaning the transfer of information over large distances, is possible only if energy, which carries the information we want to send, can travel from one place to another. Sound signals, light signals, electrical signals, and radio signals – forms of energy used in telecommunications, all travel in waves, so an understanding of waves is important for telecommunications. Radio waves were first predicted by Maxwell in 1864 and in 1887 the German scientist Heinrich Hertz demonstrated radio waves but it was not until about 1894 that Marconi (whose mother was Irish – one of the Jameson family from Wexford) used these waves to send simple messages over short distances. He called the equipment a wireless telegraph. On 12 December 1901 he made the first transatlantic transmission of a wireless signal (the letter S in Morse code) from Poldhu in Cornwall to St John's, Newfoundland, Canada. His wireless station in Clifden was the second largest in the world in its time and its staff were the first to meet Alcock and Browne after their transatlantic flight in June 1919.

Radio waves are electromagnetic waves with high frequencies, from about 10 kHz to about 1000 GHz, and are split into different **wavebands** for different users. (See the tuning guide in the RTE Guide.) **LF** stands for low frequency with channels from about 30 to 300 kHz (also called the long waveband). Transmissions using LF can travel up to 1000 km by day and further at night, depending on the power of the transmitter. A frequency of 254 kHz in the LF band is assigned to Ireland which, at a power of 500 kW, can reach all of Wales, the western half of England and southern Scotland.

Many radio stations use **MF** (medium frequency or medium wave-band from about 300 kHz to about 3 MHz), which is the longest-established of the broadcasting bands. The range of most MF transmitters does not usually extend beyond their country of origin during the day but at night, if they are sufficiently powerful, they may be heard many thousands of kilometres away. Stations using this frequency band may thus suffer from interference after darkness from another powerful station on the same frequency as this band is quite overloaded in Europe. **Short Waves** or **HF** (from 3 MHz to 30 MHz) are used for distant communication, being able to travel further than long, medium or VHF transmissions. They may be used for international broadcasting but in Ireland they are used mainly by hobbyists and can be transmitted many thousands of kilometres even at low power, depending on the time of day, the season, the frequency used and atmospheric conditions. In the **VHF** (very high frequency – from 30 MHz – 300 MHz) wave-band 87.5 MHz to 108 MHz is used for sound broadcasting. For instance, when we hear of 98 FM that station is broadcasting on 98 MHz. VHF is not as good at penetrating hilly terrain, i.e. bending around hills, etc., (diffraction) as MF. **ILS** (Instrument Landing Systems) used by aircraft for landing at an airport occupy the frequency band just above 108 MHz. If there is a broadcasting service near an airport in the 106 MHz to 108 MHz band it may have to be limited in power to prevent interference with the ILS. Passengers on airplanes are asked not to use laptop computers whose frequencies could also interfere with the ILS.

AM and **FM** refer to amplitude modulation and frequency modulation respectively. In modulation some characteristic of a high frequency sine wave, which is called a carrier, is varied in accordance with the instantaneous value of a signal. The sine wave may be represented by the equation $E = E_0 \sin(\omega t + \phi)$ where E is the instantaneous value of the carrier, E_0 is its maximum amplitude, ω is its angular velocity, and ϕ is its phase relationship with respect to some reference. Amplitude, frequency, or phase modulation refer to variation of E_0, ω, or ϕ, respectively, of the carrier wave by the modulating signal.

Why is there a need for a modulated carrier to transmit signals over long distances in the radio channel? Firstly, for transmitting signals in the audio range (20 Hz to 20 kHz approx.) the transmitting and receiving antennas would have to have heights comparable to one quarter of a wavelength of the wave used, which at 10 kHz would be 7500 m! It would not be practical to have a vertical antenna this size. Secondly, since all audible sounds are within the 20 to 20 000 Hz band, if these frequencies were transmitted directly instead of using a carrier all the sound from different sources would be totally mixed up. Hence different sources are allocated different portions of the radio waves section of the electromagnetic spectrum. These waves act as carrier waves for the various signals to be transmitted and either the frequency, amplitude or phase of the carrier is varied in proportion to the instantaneous value of the modulating signal. The rate of the variation is the frequency of the modulating signal. As well as allowing the separation of different transmissions modulation also helps to overcome the problem of poor radiation at low frequencies.

A receiver uses a tuned circuit to ensure that only the desired part of the spectrum is admitted and, since the tuning can be varied, different transmissions within a certain range can be selected. The process of extracting the signal from the carrier is known as de-modulation.

LF and MF are the two common amplitude modulated wavebands with the VHF broadcast signals usually using frequency modulation (e.g. 98 FM) although other VHF signals may use AM or some other form of modulation.

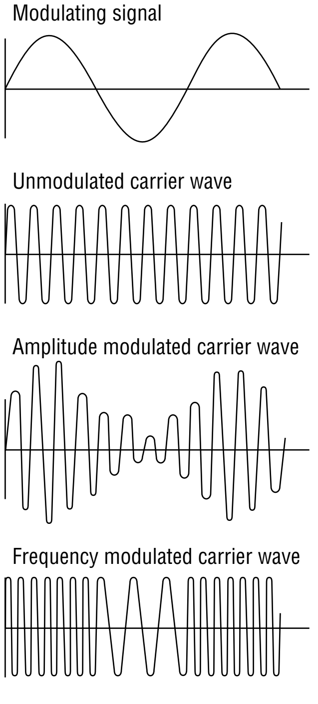


Fig. 1.1 *Modulation*

UHF (ultra high frequency, from 300 MHz to 3000 MHz) is used by television channels and is even less easily diffracted than VHF. **Microwaves** (frequencies between 3 GHz and about 30 GHz) are used mainly for satellite transmissions, telecommunications and in radioastronomy. They are dependent on almost direct line of sight path between transmitter and receiver. Microwaves of very short wavelength (a few mm) are used for radar and for cooking.

The first trials using 'wireless waves' to detect moving objects took place in the 1930s in the UK, although as far back as 1904 **Hullsmeyer** in Germany developed a primitive form of radar (RAdio Detection And Ranging) but did not develop any application.

In 1916 **Marconi** and **Franklin** picked up reflections of radio waves of wavelength 2 m from objects in the path of the waves.

In 1935 **Robert Watson-Watts's** research team from the Radio Research Laboratory at Slough succeeded in detecting aircraft, using a BBC transmitter and observing the beats produced by the interference of direct and reflected radio waves. The reflected waves are slightly shifted in frequency (Doppler shifted) due to reflection from a moving source. Radar can also be used to study bird navigation and for monitoring locust swarms.

Waves as energy carriers and their application, whether it be transporting information in the case of radio waves, killing germs for example in the case of ultraviolet radiation, helping to locate broken bones by being able to penetrate the human body in the case of X-rays, or killing cancer cells in the case of gamma rays, are an essential part of modern living.

A wave may be described as a means of transferring energy through a medium or through space in which there is a large-scale coordinated disturbance of the particles in a material medium (mechanical waves) or a disturbance in the form of varying magnetic and electric fields in which case a medium is not necessary (electromagnetic waves).

In progressive (travelling) waves energy is transferred from the source to its surroundings as a result of a disturbance moving from the source. There is no net transfer of matter. The energy may be transferred by means of a longitudinal wave motion or a transverse wave motion, as well as other wave motions which are beyond the scope of this book.

1.2 Do You Know

How does a scorpion detect the exact location of a beetle moving in the sand at night?

The moving beetle disturbs the sand and both longitudinal and transverse waves are sent out along the surface of the sand. The longitudinal wave travels much faster than the transverse wave so the scorpion, with its eight legs out-stretched, receives the longitudinal wave first and this gives the direction of the beetle. The time lapse between the longitudinal and the transverse pulse gives the distance to the beetle. Armed with this information the scorpion can precisely locate its prey.

How are radio waves transmitted around the curvature of the earth?

When radio waves are transmitted by a transmitting aerial one of the ways the radiation can reach the receiving aerial is by means of sky waves which may leave a transmitting aerial at many different angles. This is the main means of propagation for amateurs in the bands from 3 MHz to 30 MHz. (Ground waves follow the curvature of the earth and progress along its surface and space waves travel in the troposphere, the portion of the atmosphere closest to the ground, travelling in straight lines and striking the ground between the aerial and the horizon.) The sky waves appear to be reflected by one of the layers of the ionosphere (layers of ionised gas, see below) and

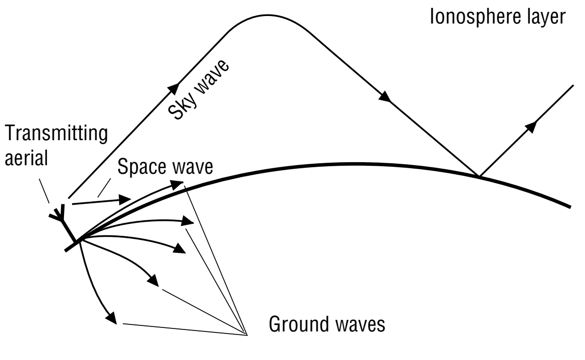


Fig. 1.2 *Transmission of radio waves*

so are returned to earth well beyond the horizon. To reach receivers on the opposite side of the earth they must be reflected by the ground and the ionosphere many times. The process repeats until the wave is completely attenuated. In fact the mechanism involved is *refraction* with the wave being bent gradually further and further away from the normal as the refractive index of each layer of the ionosphere reduces and eventually it is bent downward towards the earth.

In 1925 Sir Edward Appleton's experimental work showed that the energy from the sun is sufficient to ionise the molecules of the atmosphere into positive and negative ions. He showed that there were several layers of ionisation at different heights which reflected high frequency waves (below a certain critical frequency) back to earth which might otherwise have escaped into space. His work in fact verified what had previously been suspected. These layers extend from about 70 km above the earth's surface to about 400 km and vary with the time of day, the seasons and the eleven year sun-spot cycle. Some of the layers disappear at night due to the absence of radiation from the sun while the topmost layer, which is the most highly ionised, persists at night. Hence the ionosphere is somewhat 'unreliable' for commun-ication purposes, especially in the short wave-band, as the strength of the signal at the receiver point may vary enormously due to the afore-mentioned variations in conditions. Artificial satellites are not only more dependable but can also boost the signals they are receiving and reflecting.

Why do modern telecommunication systems use dish aerials to send messages over very long distances?

Dish aerials are parabolic in shape and use **reflection** of waves.

Ordinary aerials send signals in all directions so the signal gets weaker the further it travels

A dish aerial does not send energy in all directions. The signal starts out in the opposite direction to its final direction, and is *reflected* off the dish into a parallel beam.

Likewise incoming signals which may be weak having travelled long distances – too weak for

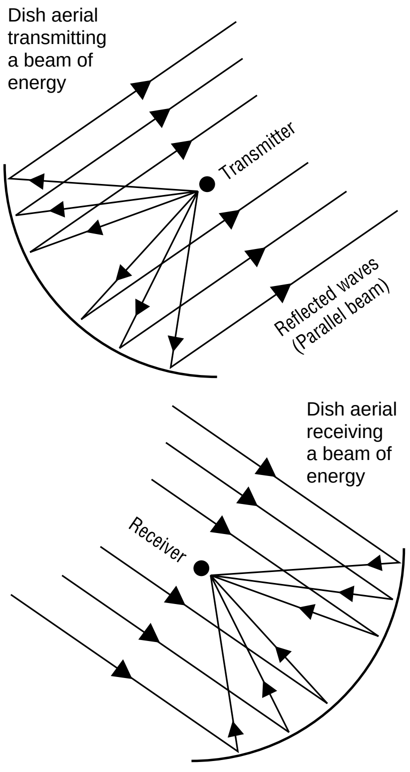


Fig. 1.3 *Dish aerials*

ordinary aerials to detect – are *reflected* off the dish and all the energy is focused to one point. The receiver must face the dish and the dish must point towards the satellite or other source.

How is it possible to transmit radio and television signals over mountains and into valleys?

This is possible only because waves can bend around corners, i.e. **diffraction**, Fig. 1.4.

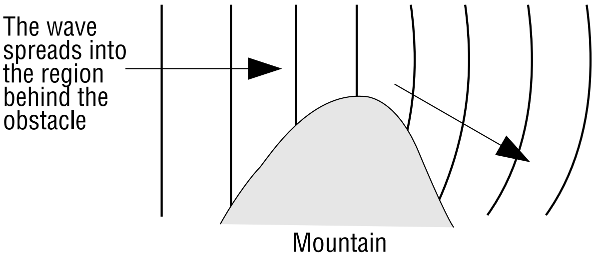


Fig. 1.4 *Diffraction: bending signals over mountains and into valleys*

Why is a throbbing sound sometimes heard when aircraft are flying overhead?

This is due to the fact that the frequencies of the engines may be close but not exactly the same, hence **beats** are produced.

Why is it that when walking across the floor with a full cup of tea, it invariably 'sloshes' over the sides despite the fact that you are watching it carefully and actively trying to stop it spilling over?

The problem partly lies in watching it, as you unwittingly make corrections which add to the amplitude of the wave. As the tea moves to one side you raise that side thus adding energy at just the right time, when it is about to move to the other side. When it moves to the other side and rises there you automatically raise that side too at just the right time to add energy in phase with the wave motion again.

Even though the person's movements are very small the wave generated becomes relatively large in amplitude so that the tea spills.

This is because when a crest is being reflected at one end, that side is raised up so that the crest is reflected as a crest and by constructive interference the two crests add to give a larger crest. Likewise at the same time a trough exists at the opposite side and this is reflected as a trough when the opposite side is tipped up – the two troughs combining to give a larger trough. In the centre the reflected crests and troughs meet, cancelling each other by destructive interference.

How can one make stationary waves in the bath?

Moving oneself in the bath so as to make the water higher at one end and lower at the other end causes the water to rush from the higher to the lower end. Moving in the opposite direction *at just the right time* to make the new pile of water higher, and therefore the new 'low' lower, causes the water to rush faster from the higher to the lower end than in the first instance. Energy is added to the wave at just 'the right time'. A stationary wave is formed due to the superposition of incident and reflected waves with an antinode at each end and a node in the centre.

The wave will increase in amplitude until the bath overflows or until the energy lost due to friction exceeds the energy input.

How do bats find their way around?

Bats use sonar (SOund Navigation And Ranging) instead of radar (RAdio Detection And Ranging). They send out high frequency sound waves which echo off their prey, e.g. a moth.

How does the bat know if the moth is moving towards or away from him?

If the moth is moving towards the bat or the bat moving towards the moth the pitch of the reflected sound wave is higher than if the moth were moving away.

How does the bat distinguish between a moth and a small inedible object fluttering in the breeze?

Since the moth's wings flutter, the sound reflected off the wings is heard by the bat as if the moth were moving slightly towards the bat, and then slightly away from him/her.

This causes *minute variations in pitch*, which a bat's ear has evolved to be able to detect and which the bat "knows" is characteristic of a moth and would not occur with an inedible object.

How does the bat determine the direction and distance away of its prey?

It determines direction by the time difference between the echoes received by *each ear* and the distance is judged by the time it takes the last sound emitted to return.

How may a foetal heart be monitored?

It is possible to hear the foetal heart after about twelve weeks and foetal heart monitoring is the most common use of Doppler-shift ultrasound. The transmitter and receiver signals are electronically mixed and the output is filtered leaving only the Doppler-shift frequency to be amplified. This frequency is in the audio range and to the alarm of expectant mothers and fathers can sound like galloping horses!

1.3 Experimental Approach

Waves, their properties and wave phenomena may be demonstrated using the following.

A. The Slinky.
B. The wave machine.
C. The ripple tank.
D. The microwave apparatus.

Waveforms may be displayed using **the cathode ray oscilloscope**.

Transverse Waves and Longitudinal Waves on a Slinky

The Slinky is stretched along the bench or on the floor.

1. A **transverse wave pulse** is sent along the spring by a quick flick of the wrist to the right or left at right angles to the spring and then back to the original position. The pulse travels along the spring and is reflected at the fixed end.

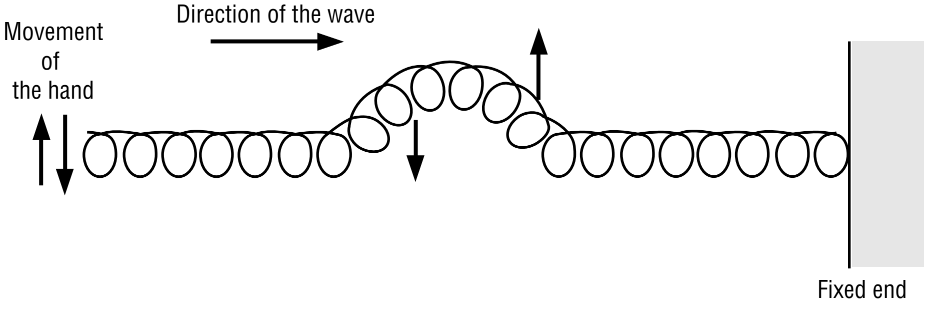


Fig. 1.5 *Transverse pulse on the Slinky*

2. A **longitudinal pulse** is sent along the spring by a quick push forward followed by a pull back of the hand in line with the Slinky. Again, the pulse is seen to be reflected at the fixed end.

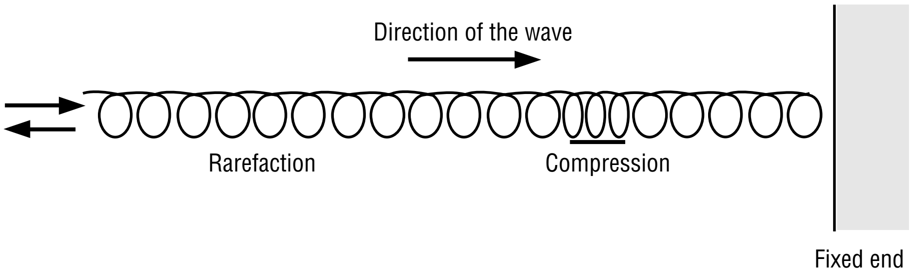


Fig. 1.6 *Longitudinal pulse on the Slinky*

3. To produce a transverse wavetrain the hand is moved at a constant rate, left then right, at right angles to the Slinky. The longer the Slinky the better as the crests and troughs can be seen travelling continuously down the spring. When the wave is reflected at the fixed end the reflected wave interferes with the incident wave, making the progress of the wave difficult to discern.

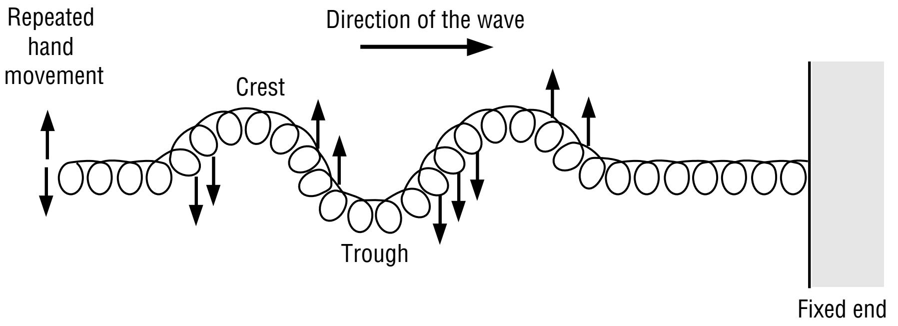


Fig. 1. 7 *Continuous transverse waves on the Slinky*

4. To produce a longitudinal wavetrain the hand is pushed back and forth at a constant rate in line with the Slinky. **Compressions** (regions of high coil density) followed by **rarefactions** (regions of low coil density) can be seen continuously travelling along the spring until the wave motion reaches the fixed end.

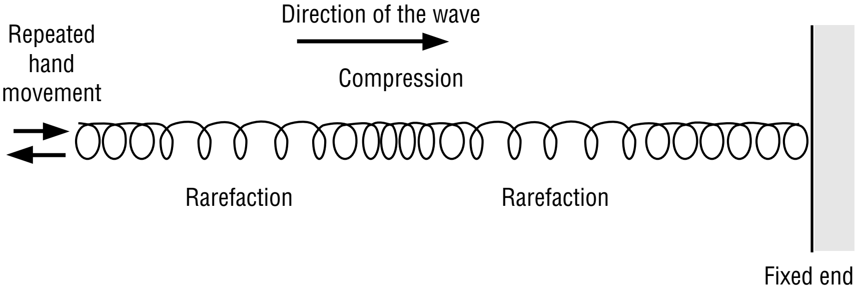


Fig. 1.8 *Continuous longitudinal waves on the Slinky*

Observations

1. The amplitude of the wave decreases slightly as it travels along the Slinky – energy is lost due to friction. The amplitude may be varied by varying the energy supplied. The amplitude of the reflected pulse is also less than that of the transmitted pulse – some energy is absorbed at the fixed end on reflection.
2. Tie a piece of coloured thread (thin, to avoid straining the coils) to a coil and note the direction of movement of a single coil of the Slinky.
3. By varying the amount by which the Slinky is stretched note the effect on the speed of the waves. Varying the amplitude or shape of the wave does not alter its speed.
4. When the transverse wavetrain is reflected at the fixed end a stationary wave may be set up (see page 13). Therefore the Slinky is limited in demonstrating progressive longitudinal and transverse wavetrains and the wave machine is more useful.

Transverse and Longitudinal Wave Motions on a Wave Machine

By turning the handle the arrangement of the rods results in a transverse progressive wave motion at one side and a longitudinal progressive wave motion at the other side.

Frequency, Amplitude and Wavelength using the Slinky and the CRO

Moving the end of the Slinky faster gives more waves per second, i.e. higher frequency and shorter wavelength, whilst the opposite is true when the hand is moved more slowly.

The cathode ray oscilloscope connected to the signal generator (loudspeaker on) may be used to observe variations in amplitude, frequency, and wavelength of waveforms and to look at different waveforms, e.g. sine, square, and triangular, while listening to the changes in loudness and pitch and quality of the sound produced by the loudspeaker. (See Multipurpose Apparatus – The Cathode Ray Oscilloscope.)

Speed

One complete vibration of a source generates one wave and the disturbance spreads out a distance of one wavelength from the source. If f vibrations per second are produced by a continuously vibrating source, then f waves will be produced per second and the disturbance will spread out *a* distance f wavelengths $(f\lambda)$ in one second. Since distance travelled per unit time equals speed,

$$\text{wave speed} = f\lambda.$$

The speed of the wave depends on the properties of the medium through which it travels. (See *Speed of Mechanical Waves* in the Reference Section.)

Reflection Using the Ripple Tank

1. Reflect a circular wave pulse off the long straight barrier (a straight metal strip). The reflected pulse appears to originate from a point behind the barrier.

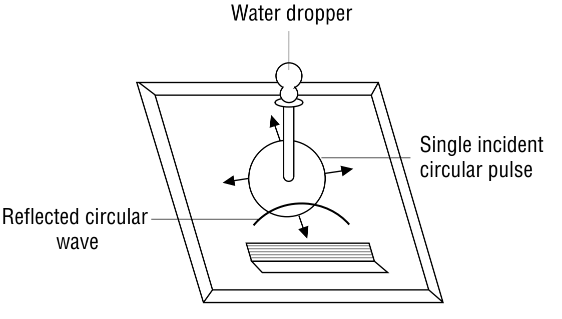


Fig. 1.9 *Reflection of a circular pulse at a long straight barrier*

2. Reflect a straight wave pulse off the long barrier first normally and then at an angle.
3. Reflect continuous circular waves off the straight barrier. A *standing wave pattern* is set up between the reflecting barrier and the source and the waves appear stationary due to interference.
4. Reflect continuous straight waves off the straight barrier.
5. The curved barrier may be used as a concave or convex reflector, reflecting straight waves or circular waves.

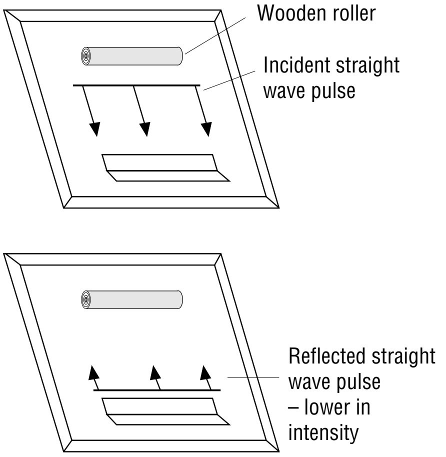


Fig. 1.10 Reflection of a straight wave incident normally on the barrier

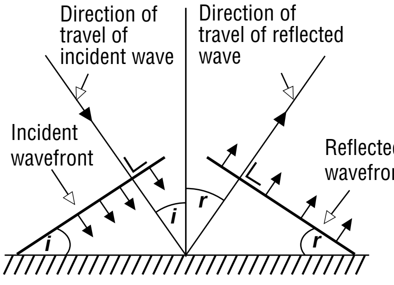


Fig. 1.11 *Reflection of a straight wave pulse incident at an angle other than 90°*

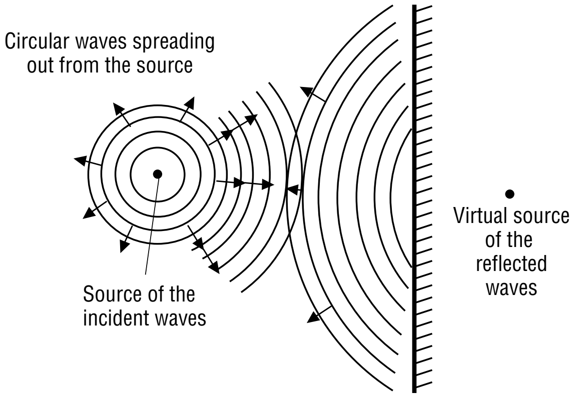


Fig. 1.12 *Reflection of continuous circular waves at a long straight barrier*

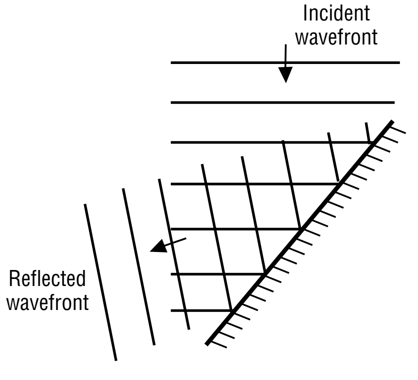


Fig. 1.13 *Reflection of continuous straight waves at a straight barrier*

In all cases the incident and reflected waves both travel at the same speed, have the same wavelength, and obey the laws of reflection.

Refraction Using the Ripple Tank

It is important to check first that the tank is level.

Refraction is demonstrated by placing a clear plastic rectangular plate in the water to create a shallow region. A thin metal washer placed under each corner of the plate facilitates its removal.

The wave velocity and the wavelength are decreased in the shallow region.

A straight wave pulse may be used first and then the pattern displayed using continuous waves.

To be convincing the water above the block must be very shallow, only about 1–2 mm. Such a thin film of water may break due to surface tension effects which may be avoided by adding a very small quantity of detergent (too much will cause bubbles, destroying further observation).

The film of water over the block must not be too thin either, or reflection from the leading edge of the block may be more obvious than wave movement across it.

When the wavefronts are parallel to the boundary there is a reduction in wavelength only, Fig. 1.14.

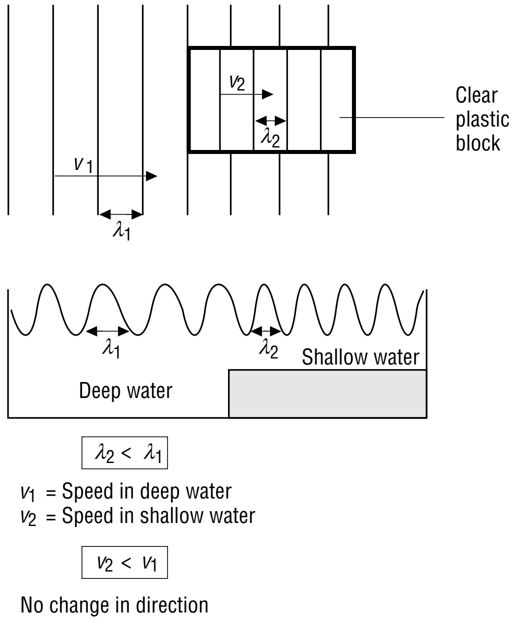


Fig. 1.14 *Refraction (ripple tank): wave fronts parallel to the boundary*

When the wavefronts are at an angle to the boundary vary the angle of incidence of the waves and see the change in the angle of refraction.

Note. The angles need to be fairly large for a convincing demonstration.

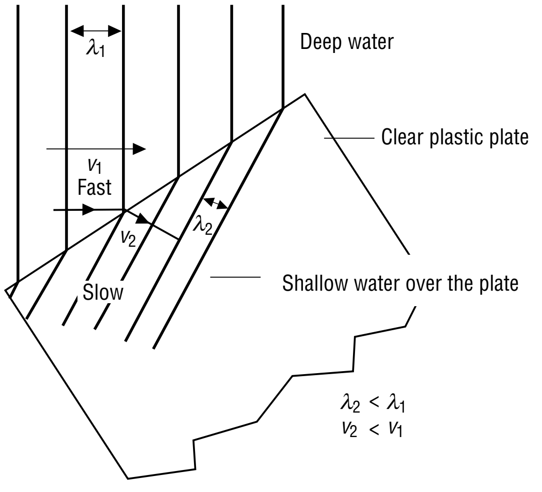


Fig. 1.15 *Refraction: wavefronts at an angle to the boundary*

Both the waves in the deep water (longer λ) and in the shallow water (shorter λ) are 'frozen' by the same stroboscope frequency at the same time showing that frequency does not change on refraction. Since $v = f\lambda$, and wavelength is reduced in the shallow water waves travel more slowly in shallow than in deep water.

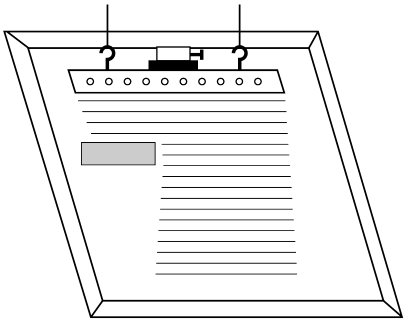
(Contrast with Doppler effect where there is a change in both frequency and wavelength while the speed of the wave is unchanged.)

Diffraction Using the Ripple Tank

The ripple tank is a very convincing and easy way of displaying the spreading out of waves on passing through openings or around obstacles, i.e. diffraction.

Use straight pulse waves first with the following steps 1 and 2 and then set the ripple beam to produce continuous straight waves.

1. Place a long barrier parallel to the ripple beam with its vertical edge in the middle of the tank. Notice the diffraction or bending of the waves at different wave frequencies (varying motor speed). The shorter the wavelength the less the amount of bending, Fig. 1.16.



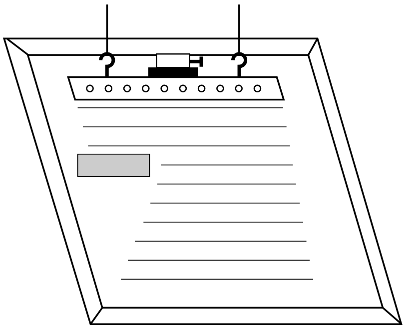


Fig 1.16 *Diffraction at the edge of an obstacle – the lower the motor speed the longer the wavelength and the more diffraction*

2. Position the second barrier about 3 cm from the first one to form a slit. Investigate the effect on the resulting diffraction pattern of varying slit width for a fixed wavelength and of varying wavelength for a fixed slit width.

   The greatest spreading into the area of geometrical shadow behind the slit occurs when the slit width is comparable to, or smaller than, the wavelength of the waves, *Fig. 1.18.* A small amount of spreading occurs into the region behind the barriers.

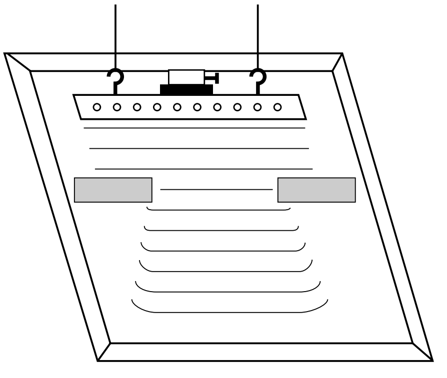


Fig. 1.17 *Diffraction through an opening which is wide compared to the wavelength*

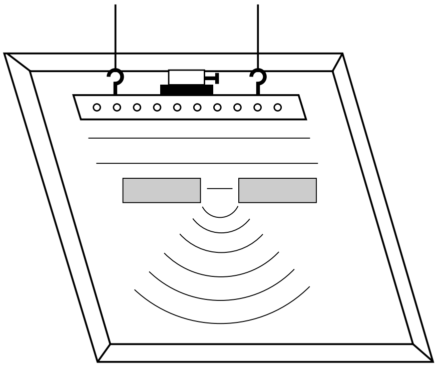


Fig. 1.18 Diffraction through an opening which is comparable with the wavelength

Interference Using the Ripple Tank

How waves combine is described by the principle of superposition which may be stated as follows.

The net displacement at a given place and time caused by a number of waves traversing the same space is the algebraic sum of the individual displacements at that point.

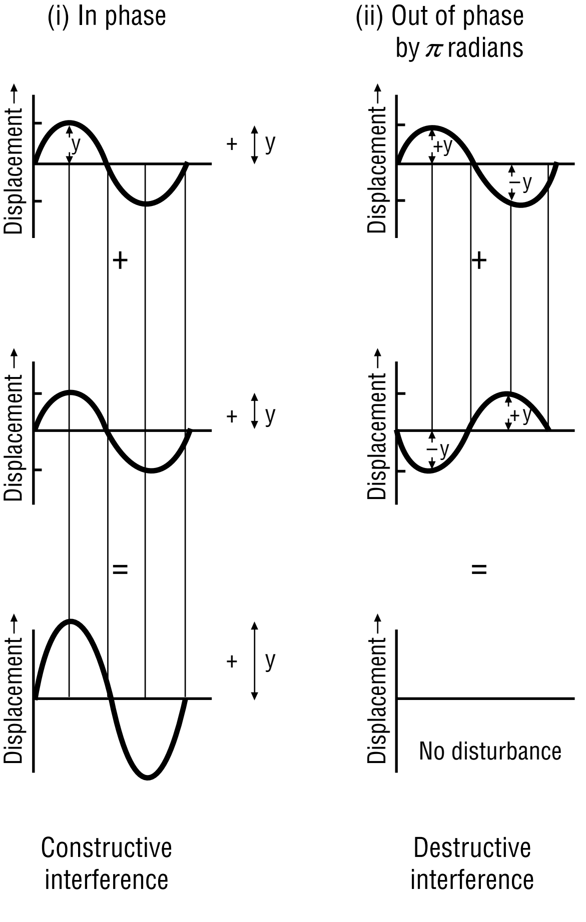


Fig. 1.19 *Superposition of two waves*

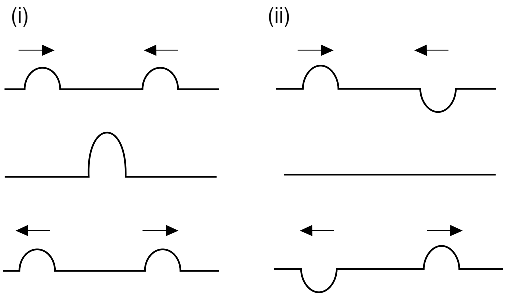


Fig. 1.20 *On the Slinky – two wave pulses meeting, travelling in opposite directions, superposing, and then continuing on*

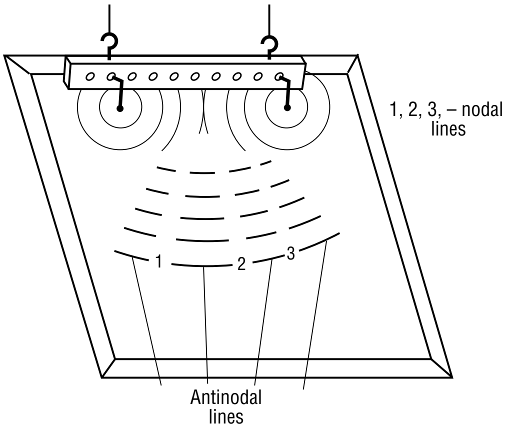
1. Raise the ripple bar and use two spherical dippers in adjacent holes on the same ripple bar to provide two coherent sources, i.e. two sources with the same frequency and amplitude and in phase or with a constant phase difference between them, Fig. 1.21.
2. Run the motor slowly and observe the effects when the two wavetrains overlap, i.e. interference.

   Nodal lines (positions of destructive interference) can be seen radiating outwards from between the spheres with regions of constructive interference (antinodal lines) in between.
3. Speed up the motor and use the stroboscope to 'freeze' the wave pattern at high speeds and then allow it to run forward at various speeds. The static position of the interference fringes within the moving pattern becomes more apparent.
4. Vary the distance between the spheres – as it decreases the separation of the nodal lines increases.

Note 1. A stationary wave pattern is set up on a line joining the two spheres.

Note 2. A similar interference pattern is obtained by passing either straight or circular waves (using just one dipper) through a barrier with two small apertures (using two long barriers either side of a short barrier stood upright on its shorter section). The waves are diffracted on passing through the two slits which act as coherent point sources, Fig. 1.22.

The first method of demonstrating interference is more successful than the latter using two slits, but the latter is useful for comparison with Young's slits experiment with light waves.



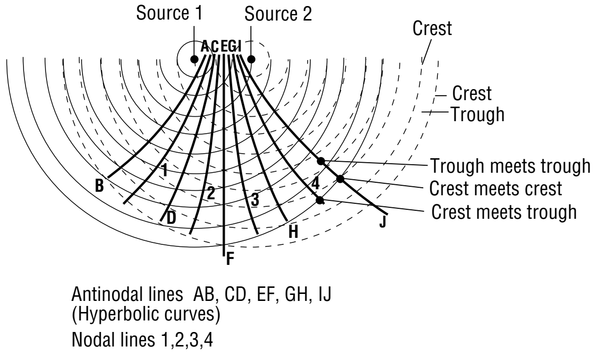


Fig. 1.21 *Interference using two spherical dippers in the ripple tank*

The interference pattern obtained is due to waves being in phase in some places (along the antinodal lines) and out of phase in others.

Illustrating Interference

Cut two sets of sine waveforms of the same amplitude and frequency out of acetate or card. Place the two waveforms on the overhead

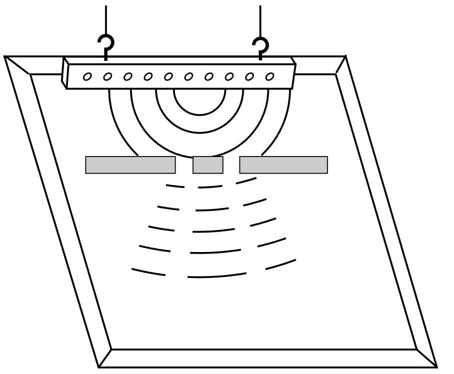


Fig. 1.22 *Interference between two sets of diffracted waves obtained from circular waves from a single source passing through two small apertures*

projector or on the bench, one above the other, between two parallel lines so that they are in phase. Anchor them with the first two fingers of the left hand at A and B so that they are free to rotate about those points, Fig. 1.23. Move the waveforms so that they meet at a point O which is equidistant from A and B. They meet in phase, i.e. crest opposite crest or trough opposite trough.

Keeping the distance AB constant rotate the two waveforms so that they meet at point W. They now meet out of phase, i.e. crest opposite trough. Rotating the two waveforms again and keeping distance AB constant, when they meet at X they are now in phase again.

The variation in phase between two waves is due to the different path lengths travelled by the different waves to a particular point. If the difference in path length is an odd number of half wavelengths then the waves meet out of phase (destructive interference) and if the difference in path length is a whole number of wavelengths then the waves are in phase (constructive interference).

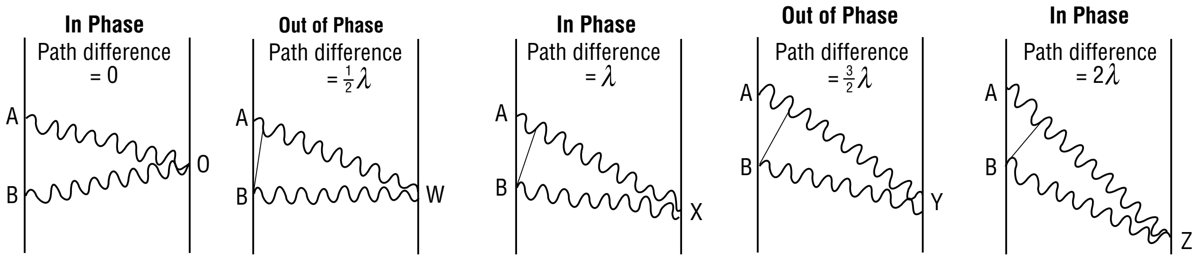


Fig. 1.23

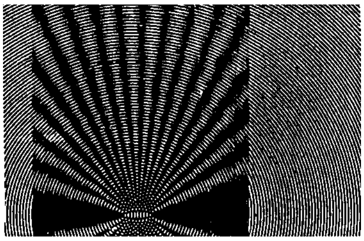


Fig. 1.24 *Interference using ring plates*

Ring plates, Fig. 1.24, may be used to illustrate interference between circular waves. Both plates have a series of concentric circular black grooves of increasing radius. The plates may be completely or partly overlapped showing a continuous interference pattern between the two centres.

Interference Between Non-Coherent Sources – Beats

Set two tuning forks of slightly different frequencies vibrating at the same time with similar amplitudes and press the stems against the bench top if they are not already mounted on sounding boxes.

A regular rise and fall in loudness is heard, called beats, Fig. 1.25.

The waves add up according to the principle of superposition. Both waves arriving at the ear are alternately in and out of phase with each other, producing alternate constructive and destructive interference.

The beat frequency equals the difference in frequency of the two forks.

Two tuning forks of equal frequency may be used with the frequency of one of them slightly reduced

(i) by loading its prongs with tiny pieces of wax or plasticine, or

(ii) by wrapping an elastic band **tightly** around the prong or,

(iii) by attaching a small metal collar (called a slider) to a prong which can be moved up or down the prong thus varying the natural frequency of the fork.

Loading the tuning fork decreases its frequency.

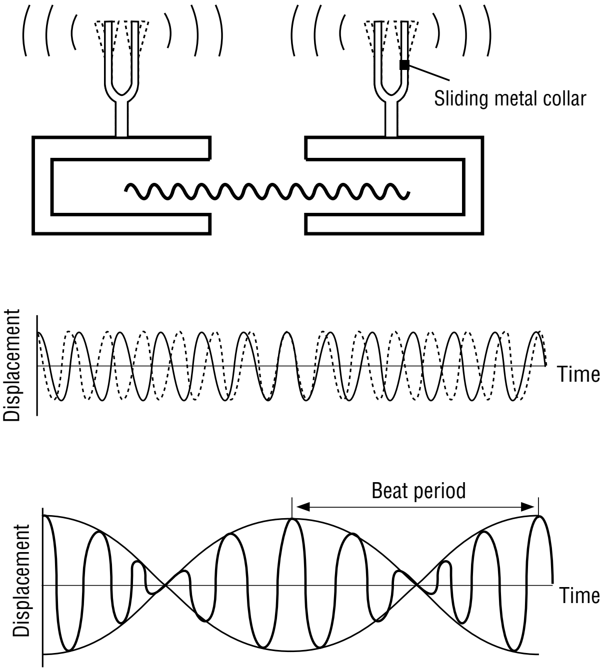


Fig. 1.25 *Two tuning forks of slightly different frequencies mounted on sounding boxes, set vibrating to demonstrate beats*

Beats may also be produced using two signal generators, each connected to a loudspeaker, to produce two notes of almost the same frequency. If two generators are not available, use the signal generator plus loudspeaker to produce a note of a particular frequency and simultaneously sing a note of a slightly different frequency. Listen for the beats produced. Playing the lowest C and the C# beside it on the piano simultaneously will also result in beats.

Stationary Waves

Stationary waves result from the superposition of two wavetrains of equal amplitude and frequency travelling at the same speed but in opposite directions. This is a special case of interference.

Stationary waves were referred to earlier:

1. on the Slinky;
2. on a line joining the two sources in the interference demonstration in the ripple tank.

Three possible demonstrations of stationary waves are described here. (See also stationary waves in a stretched wire using the sonometer – Sound section.)

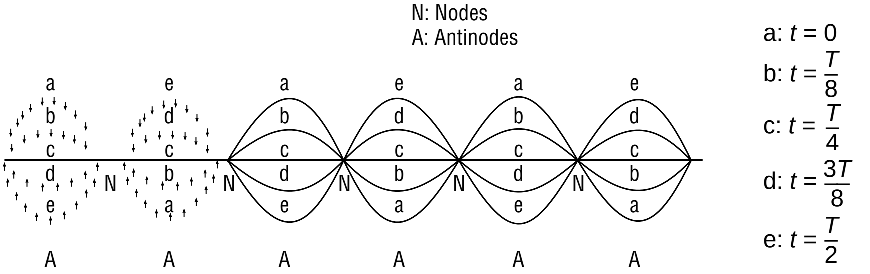


Fig. 1.26 *Stationary waves*

Stationary Waves on the Slinky

Stretch out the Slinky on the floor with one end fixed.

Move the free end continuously from side to side at right angles to the spring to generate a progressive transverse wave.

When it is reflected at the fixed end as a progressive wave the incident and reflected progressive waves meet, their displacements adding according to the principle of superposition, and one or more loops of large amplitude can be seen.

Since continuous reflection occurs a transverse stationary wave is set up which does not travel along the string.

Observations

1. There are points called nodes at which there is always totally destructive superposition – displacement always zero.
2. Amplitude of displacement varies according to position, being zero at the nodes and a maximum at the antinodes.
3. All particles between two adjacent nodes oscillate in phase, i.e. they all have their maximum displacement at the same time.
4. Particles in adjacent loops have a phase difference of π radians.
5. The frequency of vibration of the standing wave is the same as that of the two progressive waves which form it.
6. Wavelength equals twice the distance between two adjacent nodes or antinodes.
7. The waveform does not advance (no net transfer of energy).
8. The Slinky becomes straight twice within each cycle.

Energy is redistributed in standing waves. Energy at the minima goes to the maxima. Stationary waves only exist in systems with boundaries.

Transverse stationary waves are set up in, for example, stretched strings at resonance (see p. 48).

Longitudinal stationary waves are set up in, for example, closed and open pipes at resonance (see Sound section).

Melde's Experiment

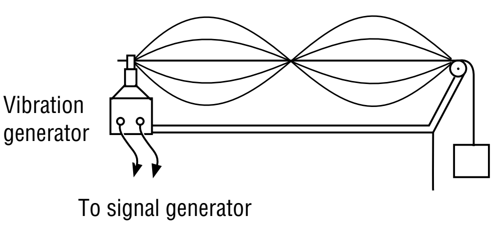


Fig. 1.27

Apparatus

Rubber cord about 0.5 m stretched to 1 m (or cotton).

Vibration generator.

Signal generator.

The string is set vibrating by a vibration generator attached to an audio-frequency signal generator.

Starting at low frequencies (c. 10 Hz), as the frequency of the signal generator is slowly increased stationary waves are set up in the string with one or two or more loops.

Since the string is fixed at both ends the ends are displacement nodes.

Resonance occurs only for waves whose wavelength match or 'fit' into the length of the string. In this case the incident waves and the waves reflected at the pulley or clamp combine to form the stationary wave.

By reducing the ambient light and illuminating the vibrating string with a flashing lamp stroboscope at a frequency close to that of the vibrator the string can be seen moving up and down slowly, troughs changing to crests and back to troughs again. Adjacent loops can also be seen to vibrate out of phase with each other.

Stationary Waves in Air

Apparatus

Signal generator + amplifier, using low-impedance output.

Sensitive microphone (connected to Y-input of CRO).

Loudspeaker (8/16 Ω).

Cathode ray oscilloscope (50 mV/div, 0.1 ms/div, Auto).

Reflecting surface, e.g. hard even wall or sheet of metal or wood on a stand.

Metre stick.

Screened lead about 1m long.

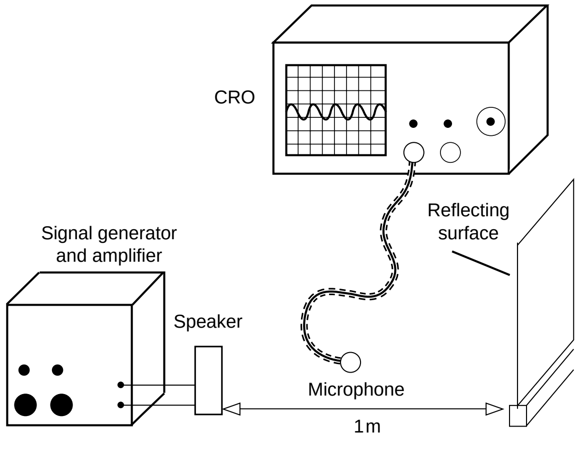


Fig. 1.28 *Stationary waves in air*

Set the signal generator to 1000 Hz and the loudspeaker and reflecting surface 1 m apart. (Frequencies in the range 1 kHz to 3 kHz may be used.)

As the microphone is moved along a line parallel to the ruler and about 1 cm from it, so that the sound travels across rather than into the microphone, the trace on the screen is seen to increase and decrease in amplitude.

Standing waves have been set up between the speaker and the reflecting barrier, Fig. 1.29. The wave from the speaker is interfering with the reflected wave.

Fig. 1.29 assumes that the amplitude of the reflected pulse is equal to that of the incident pulse. In practice, the minima in intensity will not be zero as the reflected wave has a smaller amplitude than the incident wave.

The trace may not be exactly sinusoidal as the microphone will detect stray signals across the laboratory.

As the frequency is increased more minima will be detected. At 1000 Hz the minima should be about 15 cm apart and four or five should be detected.

If the frequency is too low fewer minima are detected and the one nearest the loudspeaker is not clearly defined. Use a frequency which is high enough to give several minima well away from the loudspeaker, especially if using this experiment to measure the speed of sound.

The reflecting barrier and loudspeaker may be placed on two different benches 1 m apart to cut down on unwanted reflections from the bench or else a strip of felt may be placed on the bench between them.

If n minima are detected the average distance between two of them is $d = (n-1)\lambda/2$, $\lambda = 2d/(n-1)$.

Since $v/f = \lambda$, plotting a graph of λ (y-axis) against $1/f$ (x-axis) gives a straight line through the origin of slope v = speed of sound.

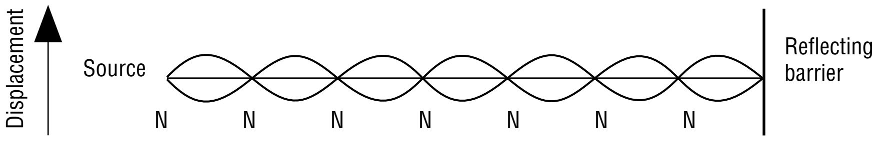


Fig. 1.29 *Formation of stationary waves between the source and the reflector*

Comparing Progressive and Stationary Waves

	Progressive Wave	Stationary Wave
Waveform	Moves forward with the wave velocity.	Does not advance. Vibrating object passes through rest position twice within each cycle.
Energy	Energy is transferred in the direction of travel of the wave.	No net transfer of energy but energy is associated with the wave.
Amplitude	The displacement of all particles varies in amplitude from zero to a maximum displacement.	Varies from zero at the nodes which are permanently at rest, to a maximum (twice the amplitude of each progressive wave) at the antinodes.
Wavelength	Distance between two adjacent particles in phase.	Twice distance between two adjacent nodes or antinodes.
Phase	Within one wavelength all particles have different phases.	Between two adjacent nodes all particles oscillate in phase but with different amplitudes, i.e. they all have their maximum displacement at the same time. Particles in adjacent loops have a phase difference of 180 degrees.

Doppler Effect

Whenever there is relative movement between the source of a wave and an observer the frequency of the wave measured by the observer differs from that emitted by the source. (This effect was first explained by the Austrian physicist Christian Johann Doppler (1803–1853) in 1842.)

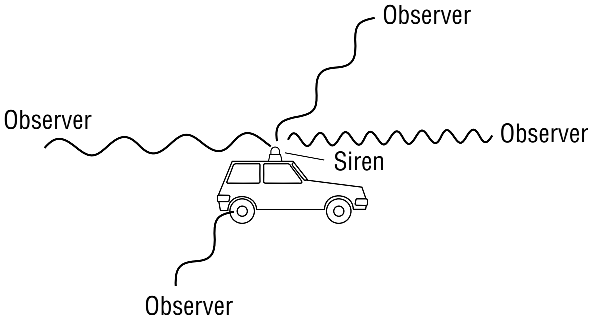


Fig. 1.30 *Doppler effect*

The pitch of a siren on a fast-moving car sounds higher to a stationary observer in front of it and lower to a stationary observer behind it.

To demonstrate the Doppler effect with sound, place a whistle in a long piece of rubber tubing. Whirl the tube in a horizontal circle above the head while blowing the whistle through the tube.

The open end of the tube acts as a moving sound source and an observer hears the pitch rising and falling as the open end revolves.

Note. If a source of sound is approaching an observer directly with constant velocity the apparent pitch does not increase as the source gets nearer – it is constant but higher than the pitch emitted by the source. The apparent *change in pitch* heard by the observer occurs suddenly as the source passes.

Doppler shift also occurs with electromagnetic radiation like light but we don't notice it in everyday life because the speed of even the fastest racing car is negligible compared with the speed of light ($c + v_S \approx c \approx c - v_S$). The effect is applied, using microwaves, in speed checks, etc.

1.4 Applications

Speed Traps

Microwaves encountering a moving object are reflected from it, and the frequency of the reflected signal is changed (Doppler shifted) relative to the emitted signal as there is relative motion between the source (reflecting object) and the observer (i.e. the receiver). The reflected wave is compared with a sample of the original wave and beats occur due to interference of the slightly different frequencies.

The beat frequency gives the difference in frequency of the two waves which depends on the relative target velocity which can thus be calculated.

Gardaí and race-track officials at Le Mans both find this figure to be of interest.

Instrument Landing Systems

The instrument landing system of an aircraft can land it without the pilot being able to see through the windscreen. Aircraft instruments calculate the speed of the plane by transmitting a radar beam at a certain frequency and because the plane is moving the wave reflected from the ground is Doppler shifted in frequency. From the amount of shift, the instruments calculate the speed of the plane relative to the ground. Since the instrument landing system can also calculate how high the plane is, from the time it takes the transmitted wave to return, and in which direction it is going, it can land the aircraft without the pilot being able to see through the windscreen.

Doppler Shift in Medicine

When ultrasonic waves are reflected from a moving part, e.g. blood, the reflection is Doppler shifted and computers use this to build up images of unseen moving objects, e.g. the beating of an unborn baby's heart. The speed at which the heart muscle is contracting may be calculated from the change in frequency of the reflected wave and this can be used to gauge its state of health.

A blockage due to a clot (i.e. thrombosis) or a narrowing of the blood vessel walls due to plaque is immediately apparent as a sudden change in the Doppler-shift frequency.

Velocity of Stars

If a light source, e.g. a star, is retreating from the earth then the observed frequency is less than the emitted frequency when at rest. Similarly, if a star is approaching the earth the observed frequency is greater than the emitted frequency.

The positions of certain wavelengths or spectral lines due to an identifiable element in the star's spectrum are compared with their positions in a laboratory-produced specimen of the element.

If the spectral lines on the star's spectrum are displaced towards the red end of the spectrum, i.e. frequency reduced, then this *redshift* indicates that the star is moving away from the earth. Similarly, a blueshift indicates that the star is approaching the earth. The amount of the shift in either case allows the speed of the star to be calculated.

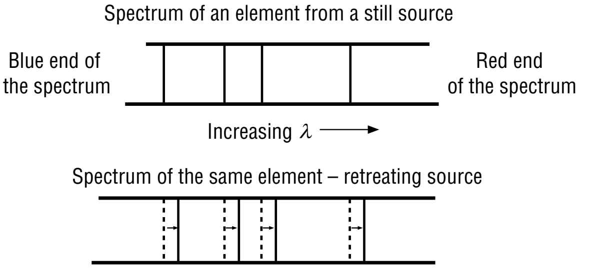


Fig. 1.31 *Redshift*

Speed of Rotation of the Sun

The sun shows equal but opposite frequency shifts for Fraunhofer lines received from opposite ends of an equatorial diameter as one edge of the sun is moving towards the earth and the other edge is receding due to rotation. This frequency difference is used to calculate the speed of rotation of the sun.

Are Saturn's Rings Rigid Discs?

Doppler effect observations on the sun's light reflected from Saturn's ring show that they are not rigid discs since different parts of the rings have different angular velocities.

Temperature Measurement

Doppler effect is one cause of the broadening of spectral lines, i.e. a small spread of wavelengths about the mean position of each line. The thermal motions of atoms in a source which is emitting light are random in direction and widely distributed in magnitude, i.e. the atoms are moving towards and away from the observer with many different speeds and hence the frequency of a particular spectral line is shifted both up and down by varying amounts.

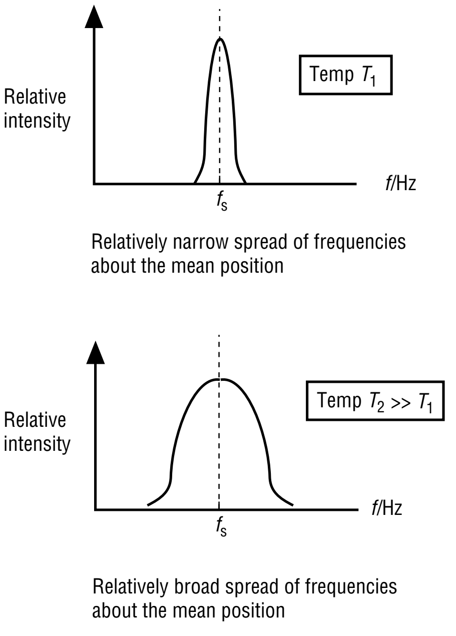


Fig. 1.32

The magnitude of the frequency shift, Δf, is proportional to the square root of the absolute temperature of the gas since the average kinetic energy of the molecules is proportional to the absolute temperature of the gas. Therefore broadening of spectral lines increases with increasing temperature, Fig. 1.32. This can be used to measure the temperature of very hot gases or plasma used in thermonuclear fusion experiments, where there is present a gas which is not totally ionised.

Δf is the thermometric property. Atoms of the glowing gas are moving towards and away from the observer at very high speeds and hence due to the Doppler effect the frequency of a particular spectral line is apparently shifted up and down. Atoms responsible for one edge of the line are moving towards the observer producing an apparently increased frequency whereas atoms responsible for the other edge are moving away from the observer producing an apparently decreased frequency. Hence the line is broadened.

The width of the line can be measured using a diffraction grating and the velocity of the atoms can then be calculated. Since this is approximately the root-mean-square velocity, which is proportional to the square root of the absolute temperature then the temperature of the gas can be calculated.

Beats

Beats can be used to tune a stretched string to the frequency of a tuning fork by sounding the two of them together and listening for beats. As the frequencies of both come closer together the number of beats per second decreases until the combined sounds have a constant amplitude when their frequencies are equal – no beats.

1.5 Worked Example

Calculate the beat frequency heard by a stationary observer when a source of sound of frequency 150 Hz moves directly away from her with a speed of 10 m s^{-1} towards a vertical reflecting surface. (Speed of sound in air = 340 m s^{-1}.)

Beats are heard because of interference between the sound reaching the observer directly from the source and the sound reflected from the vertical surface.

The apparent frequency of the sound heard directly is less than 150 Hz since the sound source is moving away from her, while the apparent frequency of the reflected sound is greater than 150 Hz. This is because the source is moving towards the vertical surface and the frequency of the waves incident on it is greater than 150 Hz and therefore to the stationary observer the reflected waves appear to be coming from a source moving towards her.

The apparent frequency of sound reaching the observer directly is

$$f_O = \frac{f_s v}{(v + u)}$$

$$= \frac{150 \times 340}{340 + 10}$$

$$= 145.7 \text{ Hz}$$

The apparent frequency of the reflected sound is

$$f_O = \frac{f_s v}{(v - u)}$$

$$= \frac{150 \times 340}{340 - 10}$$

$$= 154.5 \text{ Hz}$$

Beat Frequency $= 154.5 - 145.7$

$= 8.8$ Hz.

Appendix I Reference Material

Make a Transverse Wave Model

1. 39 strips, 2.5 mm wide and 40 mm high are drawn on thin card and every second one is coloured in. Fold down the top and bottom to make a guide.

2. A wave curve 2.5 mm wide and 220 mm long is then cut out of card using a sharp knife.

3. The waveform is inserted into and drawn along the guide and as the wave moves to the right the vibration of the particles (coloured strips) perpendicular to the wave direction can be seen.

Make a Longitudinal Wave Model (Crova's Disc)

1. Draw the innermost circle with a radius of 7 mm on a sheet of card.

2. Mark off 8 equidistant points on the circumference, labelling them consecutively 1 to 8.

3. Using each of these points in turn as centres draw a series of circles (at least 11) with progressively increasing radii so that successive circles are about 1 mm apart at their closest positions.

4. Using the original centre draw a circle enclosing all the circles already drawn and cut out the resulting disc.

5. Pin a strip of card with a slit in it to the centre of the disc. (The length of the slit should be approximately equal to the radius of the disc and the width of the slit should be 5–7 mm.)

6. Rotate the disc and through the slit view compressions followed by rarefactions moving across the slit.

Each circle representing a layer of air does not however move across the slit but oscillates about its mean position.

Refraction – A Possible Analogy

Refraction could be explained in terms of a line of children, holding a long bar, representing a wavefront, and running in a field containing both long and short grass.

(a) *Wavefront moving parallel to the boundary between 2 media*

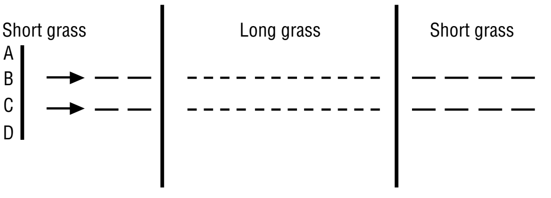


Fig. 1.33

The children, represented by the letters A, B, C, D, are running in step in the direction shown by the arrow.

All the children will reach the long grass at the same time and so they are slowed down together and will continue in a straight line with the bar ('wavefront') still parallel to the boundary between the two types of grasses.

(b) *Wavefront at an angle to the boundary between two media*

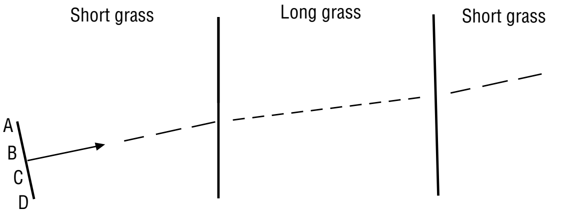


Fig. 1.34

In this case child 'D' will reach the long grass first and so will be slowed down before the other three children. Child 'C' will next reach the long grass, then 'B' and finally 'A'. The net effect is that the bar they are holding (representing a wavefront) changes direction at the boundary. When they are all in the long grass they will all continue at the same speed in this new direction as shown in the diagram.

On reaching the short grass child 'D' will reach it first and speed up first, then 'C', then 'B' then finally 'A'. The net effect is that the bar (representing a wavefront) changes direction again at the boundary but in the opposite direction to the change at the first boundary.

Speed of Mechanical Waves

Mechanical waves, i.e. waves requiring a medium for their propagation, are transmitted by vibrations of particles of the medium. The medium must have mass and elasticity in order to be able to vibrate.

Density and elasticity determine the speed of a wave.

For transverse waves on a taut spring

$$v = \sqrt{T/\mu}$$

where T = tension,

and μ = mass per unit length.

The numerator is the 'elasticity term' ($T = kx$, where x is the extension of the spring and k is the elastic constant). The denominator is the density term.

For longitudinal waves in a bar of material $v = \sqrt{E/\rho}$, where E is Young's modulus (the elasticity term) and ρ is the density. (See also p. 64.)

CHAPTER 2 SOUND

2.1 Background

The source of all sounds consists of vibrations of an object sending out waves through the medium in which it is placed.

In a BBC Television Series on Science Topics the film opens with an interview with a dancer who is deaf.

This ballerina is able to dance to music because she feels the vibrations of sound through her body (BBC Television Series Science Topics 1984).

At a disco one can easily feel the vibrations of the sound. We are continually immersed in sound stimuli which affect our activities.

Humans produce sound vocally, i.e. in the respiratory system. Frogs also produce sounds by moving air past vocal cords between the mouth and lungs and the sounds are specific to a particular species. The vocal sounds of birds are produced by the 'syrinx' which is a specialized region at the lower end of the trachea (in contrast to the upper end in mammals).

Many animals produce sounds by friction between body parts or with some item in the environment. Crickets produce their familiar sounds by rubbing together rasp-like structures on their legs. Beavers slap the water surface with their tails sending sound signals through the underwater passageways leading to their lodges. The tail tip of the rattlesnake is specially modified to produce sound.

Sound communication between animals can supersede other forms especially in darkness, across large distances, in dense forest or underwater.

From the beginning man must have known that striking a solid object in air would produce sound. Experimenting with various materials and techniques to produce pleasant or harmonious sounds led to the development of music. **Pythagoras**, the 6th century BC Greek philosopher is credited with the idea that the pitch of a sound in some way depended on the frequency of vibration of the vibrating object. **Aristotle** believed that motion of the air had something to do with the transmission of sound although he may have thought of it as a streaming of the air towards the ear rather than a wave motion. **Marcus Vitruvius Pollio**, a Roman architect, understood sound to be a wave motion, drawing comparisons between sound waves and water waves. His ideas were largely forgotten during the Middle Ages when a type of corpuscular theory for the transmission of sound prevailed, i.e. that sound consisted of a stream of small invisible particles which somehow affected the ear. The corpuscular theory was revived in the 17th century by the French mathematician and philosopher **Pierre Gassendi**. It was also in this century that attention was paid to calculating the speed of propagation of sound and in 1635 Gassendi made one of the first recorded measurements of the speed of sound in air at 478 $m\,s^{-1}$. He timed the delay between seeing the powder flash from a gun and hearing the noise. Assuming that he saw the flash instantaneously, and knowing time and distance from the source, he was able to calculate speed. Later on in the 17th and 18th centuries many other measurements were carried out and this value was shown to be too high. About 1750 the Academy of Sciences in Paris directed measurements in windless, open air using a cannon as the source of sound about 30 km from the observer. The result was 332 $m\,s^{-1}$ calibrated for 0 °C. Subsequent measurements differ from this by less than 1 percent. In 1808 the speed of sound in an iron pipe was established to be about 5 000 m s^{-1}.

In 1826 the speed of sound in water was found to be 1 435 m s^{-1} at 8 °C, more than four times the speed in air. The experiment was carried out in Switzerland by **Daniel Colladon**, a Swiss physicist, and a French mathematician, **Charles-François Sturm**. A bell immersed in Lake Geneva, was struck and at the same instant gunpowder was fired. Miles away the time between seeing the flash and the later arrival of the sound through the water was recorded and from time and distance speed was thus calculated.

Sound as a Wave Motion

Newton (1642–1727) attempted to apply the wave theory, which had been around for some centuries, to calculate the speed of sound theoretically but the formula he used gave too low a value. **Laplace** made corrections to it based on the assumption that the compressions and rarefactions do not take place at constant temperature and the corrections gave values which agreed with experimental results.

In 1747 **d'Alembert** proposed a mathematical expression for wave motion which led many scientists to come up with theories about mechanical waves in continuous media. Much of the mathematical work carried out by **Leonhard Euler**, a Swiss mathematician, and **Joseph-Louis Lagrange** of France made it possible to understand the vibrations of strings, rods, organ pipes and musical instruments in general, employing calculus methods.

In 1830 **Felix Savart** established frequency limits of audibility which are similar to values established with modern sophisticated equipment. In 1870 **August Toepler** and **Ludwig Boltzmann**, Viennese physicists, measured the threshold of audibility to be 10^{-7} W m^{-2}. The accepted value now is taken to be 10^{-12} W m^{-2}.

Ohm, a German physicist, better known for Ohm's Law in electricity, suggested in 1843 that the ear can analyze a complex sound wave into its component pure notes just as, mathematically, a complex waveform can be analyzed into its component parts using Fourier analysis. Another German physicist, **Helmholtz,** continued Ohm's work, and wrote a book called 'Sensations of Tone'.

The current theories of sound can be said to date from a book which was published in 1877 called 'The Theory of Sound', by **Lord Rayleigh** (John William Strutt), an English physicist.

This century has seen huge developments in sound recording and reproduction and the exploitation of frequencies outside the audio range for medical as well as industrial purposes.

2.2 Do You Know

Sound of certain wavelengths shows potential in the non-chemical control of insects.

Adult Indian-meal moths exposed to these wavelengths during the egg-laying stage had their reproduction rate reduced by 75%.

Rumbling of thunder is due to reflection of sound from various obstacles.

Thunder is caused by lightning, which occurs when charges move from cloud to cloud or from a cloud to the ground. These moving charges heat the air to such very high temperatures that the air molecules glow, giving off light. The air is heated so rapidly that it expands violently outwards all along the lightning bolt, compressing the air around it. Hence a sound wave is generated as sound is a compressional wave.

Near the lightning one loud sound is heard due to sound travelling directly to the listener. **Rumbling** is heard if you are further away due to sound from different parts of the lightning bolt arriving at different times as a result of reflections from the ground, hills, or clouds thus prolonging the sound, Fig. 2.1.

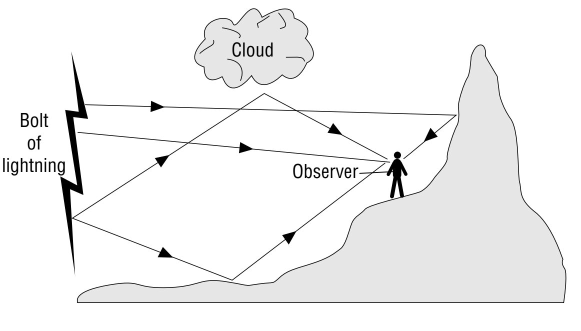


Fig. 2.1 *Thunder reaching the observer by different paths*

Why do we sometimes see lightning but not hear thunder?

The lightning still causes thunder but it doesn't reach you even though the lightning is visible because the sound is refracted upwards and is not audible on the ground. Under the right conditions lightning can be seen up to 150 km away but thunder can only be heard up to about 15 km away. The air is cooler higher up in the atmosphere and sound travels more slowly in cooler air. The section of the wavefront which is lower down travels faster than that higher up and so the wave is refracted upwards, Fig. 2.2.

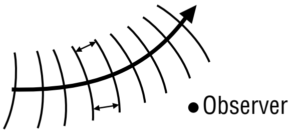


Fig. 2.2 *Sound of thunder refracted upwards*

How can relatively low sounds (e.g. normal conversation) travel long distances over water?

On a calm evening the water temperature is lower than the air temperature higher up so it cools the air next to it. Since cool air is denser than warm air it stays near the water surface provided there is no wind. This is the opposite situation to the thunderstorm. Sound is refracted downwards because it travels faster higher up in the warmer air, and when it hits the water surface it is reflected upwards only to be refracted downwards again. It continues thus, travelling across the water and does not lose its energy as quickly as normal as it is not spreading upwards but only in a flattened circle close to the water and so it can travel great distances. Conversations can be heard up to 1.5 km across a lake under such conditions. According to the Guinness Book of Records there is a recorded case of a human voice being detectable at 16.9 km across still water at night.

Why does sound travel further on cool nights than on warm days?

At night the air near the ground is cooler than the air higher up. Sound travels faster in the warmer air. The section of the sound wave higher up travels faster than the section near the ground so the wavefront is bent downwards, Fig. 2.3.

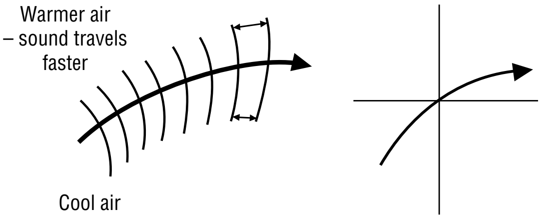


Fig. 2.3 *Sound refracted towards the ground at night*

During the day the reverse is true. The air near the ground is warmer than the air higher up so the wavefront is bent upwards over the head of the listener, Fig. 2.4.

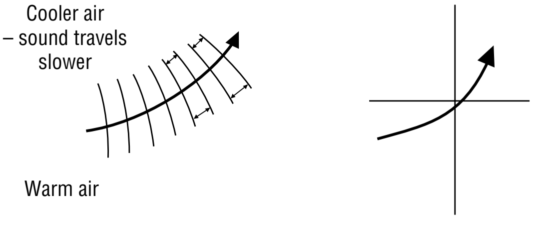


Fig. 2.4 *Sound refracted upwards on a warm day*

High frequency sounds do not spread out as much as low frequency sounds.

High pitch sounds are heard best in front of a loudspeaker rather than at the side or behind it.

Bats use high frequency, short wavelength waves which will be reflected off their prey rather than diffracted around it.

The range of diffracted beams is less than the range of narrow beams. This is why ships' fathometers use high frequency waves which can be confined to narrow beams which will travel greater distances as they are diffracted less.

Why does a stereo system not produce an interference pattern?

A domestic stereo system does not usually show such regular changes in sound level since it operates over a wide range of frequencies and therefore you wouldn't get an interference pattern as for a single frequency. Reflections from walls, ceilings and various objects in the path of the waves also tend to even out the sound.

What happens when an aircraft crosses the sound barrier?

The sound barrier represents a sharp rise in aerodynamic drag that occurs as an aircraft approaches the speed of sound. It used to be an obstacle to supersonic flight.

An aircraft such as Concorde is capable of flying faster than the speed of sound. As it approaches the speed of sound it begins to catch up on the sound waves travelling in front of it and they are pushed closer and closer together, Fig. 2.5. A barrier of compressed air is thus formed in front of the aircraft. As the aircraft reaches the speed of sound it overtakes this high pressure air which spreads out behind it in a powerful shock wave. In any time interval the source moves further than the wavefront and a cone (Mach cone) encloses all the waves emitted by the source (similar to the conical envelope of the bow wave pattern of a fast-moving boat.) On the ground the shock wave is heard like a clap of thunder called a sonic boom which may have enough energy to shatter windows.

v_s/v = Mach number, where

v_s = speed of the source, and

v = speed of sound.

Mach number > 1 is supersonic.

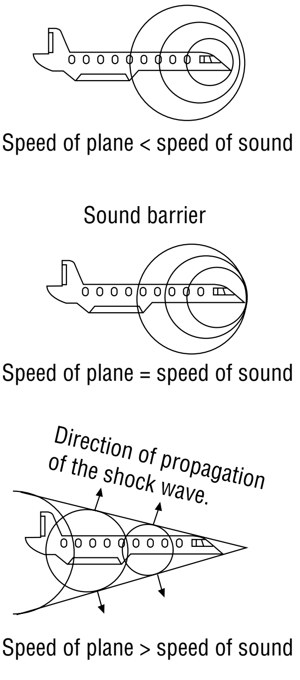


Fig. 2.5 *Shock waves and the sonic boom*

Astronauts cannot speak to each other on the moon without the aid of radio waves because of the lack of air.

In one film, when the microphones broke down, the two astronauts put their helmets together, thus transmitting the sound vibrations through the helmets. Sometimes, in popular science fiction films, we see and hear an explosion in space which is incorrect as the sound waves, unlike light, cannot travel through empty space.

The fact that the speed of sound varies with temperature is an important factor in playing wind instruments.

If the temperature of the air column changes then the frequency of the note and hence its pitch will change since

$$\text{speed of a wave} = \text{frequency} \times \text{wavelength}$$

and the wavelength is determined by the length of the vibrating air column.

Therefore it is necessary for the player to warm up the instrument before playing it so that its temperature remains constant.

When an instrument is being played outside in cold weather, breaks between playing cause problems due to the instrument cooling down very quickly.

The speed of sound may be used to estimate distance from a thunderstorm.

Seeing lightning before hearing the thunder is due to the relatively low speed of sound compared to that of light. Since sound travels approximately 0.3 km in 1 second, by measuring the number of seconds between the flash of lightning being seen and the thunder being heard, the distance away of the storm may be estimated and by repeating the procedure it can be determined whether or not the storm is approaching or retreating.

A baby's cry contains frequencies corresponding to the most sensitive part of the human ear's frequency range.

Loudness depends on intensity but it doesn't follow that it is proportional to intensity as the ear varies in its sensitivity to sounds of different frequency, being most sensitive to frequencies between 2 and 4 kHz, Fig. 2.6. In general the greater the intensity the greater the loudness.

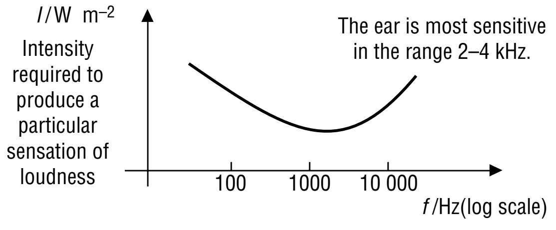


Fig. 2.6 *Frequency response of the human ear*

Driving a car with an open window may generate an infrasonic resonance.

Infrasonic waves are generated whenever a large high-speed flow of air occurs into an enclosed box or room. The effects on motorists driving with open windows is currently of interest.

Elephants respond to very low frequencies; other animals are sensitive to very high frequencies.

Elephants use very low frequency sounds produced in their nasal passages which can be heard over several kilometres by other elephants and are used to coordinate herd movements.

Dogs can hear up to 35 kHz, some bats up to 80 kHz, and whales up to 200 kHz.

The human voice can produce a wide range of frequencies.

The average fundamental frequency of the female speaking voice is about 230 Hz, with a singing range of about six notes up on the scale and six down. However Marita Gunther, trained by Alfred Wolfsohn, has produced all of $7^1/_4$ octaves above the lowest note on the piano, an 'A', although only six of these octaves were considered to have musical value. The average fundamental frequency of the male speaking voice is about 150 Hz with a singing range also of about six notes up on the scale and six down. Again extremes have been reached with Roy Hart, also trained by Wolfsohn, reaching notes below the lowest note on the piano.

Laboratory tuning forks cannot be used to tune musical instruments.

Most tuning forks found in laboratories are in scientific pitch which is not the same as the pitch of corresponding notes on an equally tempered scale and are thus unsuitable for tuning musical instruments.

Scientific scale

Note	C	D	E	F	G	A	B	C
Frequency/Hz	256	288	320	340	384	427	480	512

Equally tempered scale

Note	C	D	E	F	G	A	B	C
Frequency/Hz	262	294	330	349	392	440	494	523

Standard concert pitch is based on a frequency of 440 Hz for A. (See Reference Section – Different musical scales and standard concert pitch.)

Voice prints can be used to help catch criminals.

Police in Los Angeles are building up files of voice prints. For example a suspected criminal is asked to repeat some phrase recorded previously, perhaps from an anonymous threatening telephone call. The recorded voices are converted to digital pulses by the computer and voice prints are produced which are then compared to see if a suspect's voice print matches the original. The method of course is based on the uniqueness of each person's voice print.

2.3 Experimental Approach

To Demonstrate that Sound-Producing Objects are Vibrating

1. Touch the prongs of a vibrating tuning fork off the surface of water in a container and it will cause splashing.
2. Suspend a small pith ball or piece of polystyrene from a string. It will be kicked aside by a vibrating tuning fork.

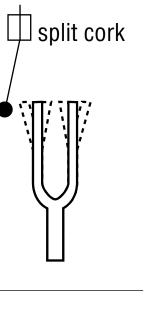


Fig. 2.7

3. Hold a metre stick or ruler over the edge of the bench and twang the free end so as to make it vibrate. This produces a sound which may be varied in frequency by varying the length of the free end.

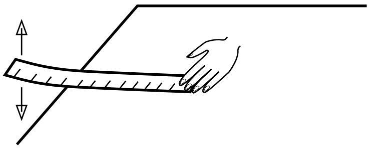


Fig. 2.8

4. Pluck the string of a sonometer to make it vibrate. Vary the frequency of the sound produced by varying length (distance between the bridges) or tension.

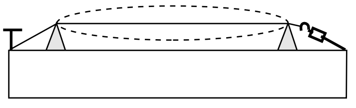


Fig. 2.9

5. Feel your throat when speaking. The vibrating vocal cords produce sound.

6. Connect a signal generator to a loudspeaker. At low frequencies the vibrations of the cone can be seen, felt with the fingertips, and heard.

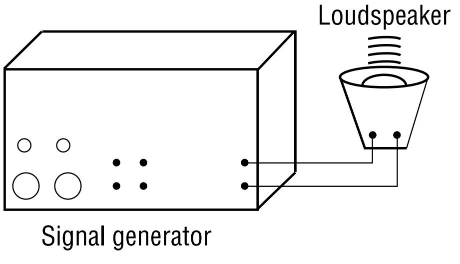


Fig. 2.10

7. Look at a vibrating tuning fork in the light from a stroboscope and the prongs can clearly be seen to move back and forth.

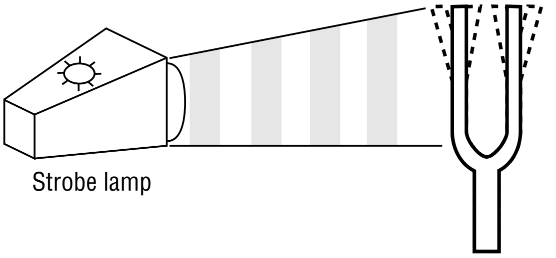


Fig. 2.11

Note. When the prongs of the tuning fork move outwards the air molecules outside each prong are squeezed together forming a compression. When the prongs move inwards the air molecules outside them spread out or are thinned out to form a rarefaction. (See compressions and rarefactions formed on the Slinky, Fig. 1.8, p. 7 and Crova's disc, p. 20)

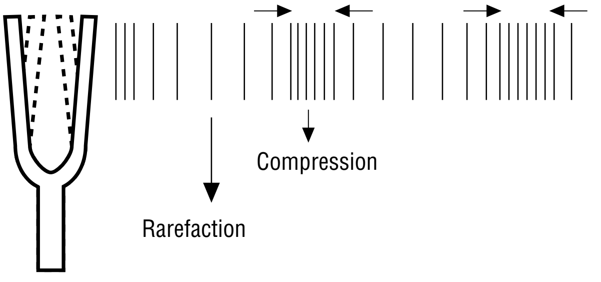


Fig. 2.12

Sound is a longitudinal pressure wave propagated by vibrations of particles of the medium through which it is transmitted. The wave travels through a medium when a section of the medium is compressed and the compressed section expands into the adjacent section which compresses the next section and so on.

Sound waves may be progressive, e.g. sound waves from the vibrating tuning fork, or stationary, e.g. sound waves in an organ pipe.

Reflection of Sound

Apparatus

Signal generator, small loudspeaker, 2 cardboard tubes, microphone, cathode ray oscilloscope, reflecting surface, e.g. smooth wall or drawing board, barrier, sheet of white paper.

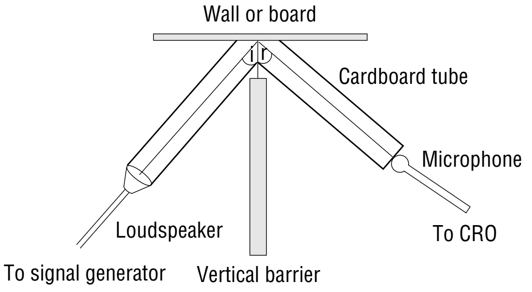


Fig. 2.13

Sound from the loudspeaker is reflected from the wall. The position of the tube with the microphone is varied until the maximum trace is seen on the oscilloscope. Alternatively, if using a drawing board as a reflector, its position may be varied instead of the tube until the trace is a maximum. The directions of the incident and reflected waves are marked on the sheet of paper, a normal drawn and the angles of incidence and reflection measured. Sound waves obey the laws of reflection.

Make a Tube Telephone

Apparatus

A few metres of rubber tubing, e.g. garden hose pipe, 2 funnels.

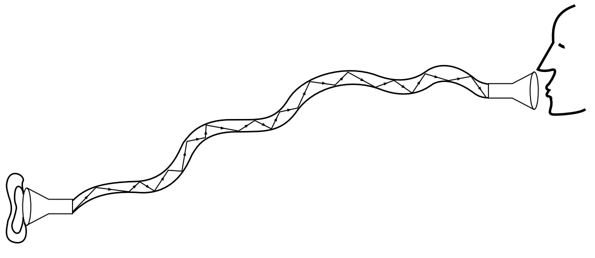


Fig. 2.14 *Tube telephone*

Two people can whisper messages to each other from one room to another but it is important not to shout as it could damage the receiver's hearing.

Diffraction of Sound

Sound bends around corners and when it passes through an open door it doesn't just remain in a narrow beam but fans out in all directions. The wavelength of sound waves varies from about 20 cm to 10 m. The wavelengths which are similar in size to, or larger than, the aperture they are passing through are diffracted most. A sound of frequency 300 Hz has a wavelength around 1 m which is about the width of most doorways so it would be diffracted a lot.

High pitched sounds are more directional or do not spread out as much as low frequency sounds.

Interference of Sound

Since sound is a wave motion sound waves combine rather than collide when they meet, according to the principle of superposition. They pass through each other (like ghosts!?) and the effects are additive.

(1) Using a Tuning Fork

Slowly rotate a vibrating tuning fork near the ear. When the prongs move inwards a compression is produced between them, Fig. 2.15. At the same time a rarefaction is produced outside them.

The compressions and rarefactions travel outwards as approximately spherical wavefronts and cancellation (destructive interference) occurs in, for example, the directions CC' and DD' where compressions and rarefactions meet.

Thus destructive interference occurs four times in each rotation. Constructive interference also occurs, e.g. in the directions AA' and BB'.

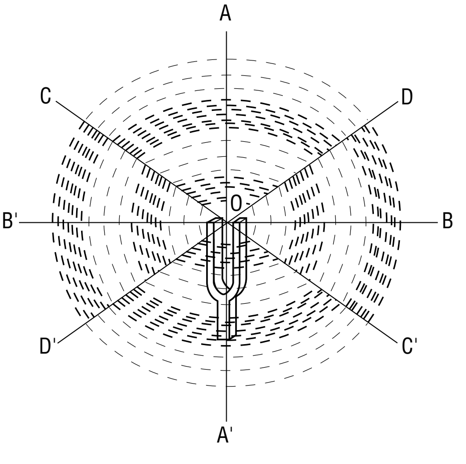


Fig. 2.15

(2) Using Two Loudspeakers

Apparatus

Signal generator + amplifier, 2 loudspeakers (e.g. 64 Ω, 0.3 W, diameter 6 cm).

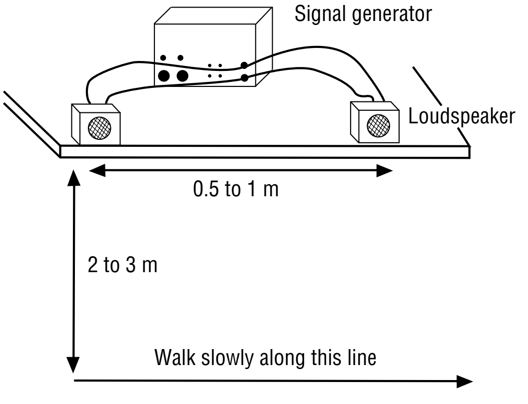


Fig. 2.16

Method

1. Place the two speakers facing the same way, about 0.5 m to 1 m apart and use frequencies of around 3000 Hz.
2. Walk slowly parallel to the speakers and a couple of metres from them, with no intervening objects to cause unwanted reflections, and listen for the variation in loudness due to alternate constructive and destructive interference.

The distance between consecutive positions of constructive interference increases if

(a) the distance between the speakers is decreased;

(b) the wavelength is increased;

(c) the distance between the listener and the speakers is increased.

In both of the above demonstrations the two sources of sound, i.e. the two prongs of the tuning fork and the two loudspeakers are coherent sources. Waves from coherent sources have the same frequency and leave the sources with a fixed phase relationship .

(3) Using Non-Coherent Sources

Apparatus

2 signal generators (not necessarily the same type), 1 loudspeaker (e.g. 64 Ω, 0.3 W, diameter 6 cm).

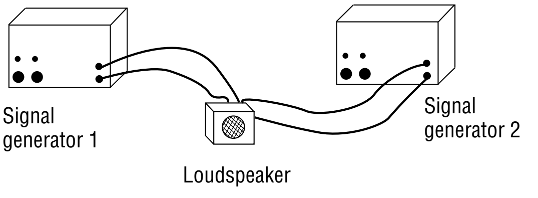


Fig. 2.17

Method

1. Connect the signal generators in parallel to the loudspeaker.
2. Set the frequencies from each generator to slightly different values, e.g. 300 Hz and 308 Hz, and the amplitudes equal.
3. Listen to the combination of the sound waves from the speaker.

At the given frequencies 8 beats per second will be heard. Reducing the difference between the two frequencies reduces the number of beats per second until the sound has a constant amplitude when the two frequencies are equal.

If the two signal generators are connected in parallel to the Channel 1 input of the CRO and the time base frequency set to around half the beat frequency the combination of the two frequencies may be viewed showing the alternate constructive and destructive interference, Fig. 2.18.

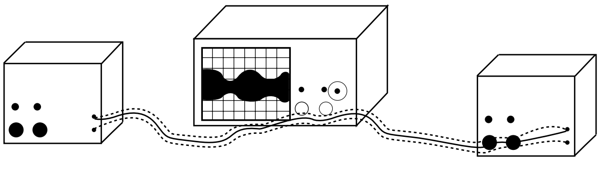


Fig. 2.18

From the diagram the beat occurs every 4 divisions and if the oscilloscope time base setting is 50 ms/div then the beat period is $4 \times 50 = 200$ ms. Therefore the beat frequency is 5 Hz.

Transmission of Sound

Sound being a mechanical wave needs a medium, either solid, liquid, or gas for its transmission.

(See speed of a mechanical wave – Reference Section on Waves.)

There is a resistance to the passage of sound through a medium which is called acoustic impedance. (Electrical resistance resists the passage of current.) Acoustic impedance matching is important to ensure that sound is transmitted from one medium to another and not reflected at the interface between the two media. To ensure that there is no air between the skin and the transducer when having an ultrasound scan a gel or oil is first rubbed onto the skin which improves impedance matching.

Apparatus

Bell in an evacuated glass jar.

As the air is removed the sound gets fainter and fainter until it can no longer be heard even though the hammer can still be seen striking the gong. Some vibrations may travel out through the rubber bands by which the bell is supported giving a faint burring sound. Coiled flexible wires are used to connect to the power supply which will absorb rather than transmit the sound to the outside.

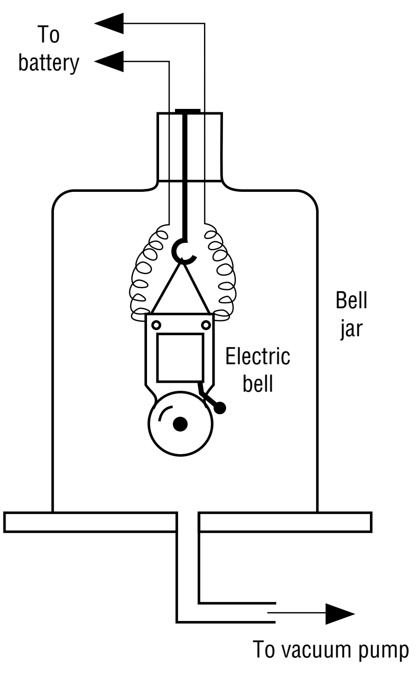


Fig. 2.19

Speed of Sound

By the time we hear the kick on a football at a match the ball is already in the air indicating that the transmission of sound is not instantaneous.

The nature of the transmitting medium affects the speed with which vibrations are passed from one particle to the next and hence the speed of sound.

Any medium which has particles that can vibrate can transmit sound. In general, the speed of transmission depends on the density of the medium and on the elasticity of the medium.

Transmission of sound through different media

1. Listen to a ticking clock through a bar of wood, a bar of metal, and the same distance from the clock through air. Sound is concentrated in the solid rods and is louder than if it were only transmitted through the air where it spreads out in all directions. It travels fastest through the metal.

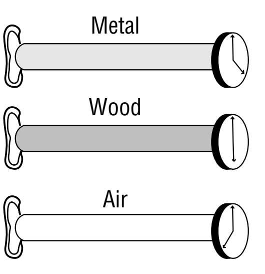


Fig. 2.20

2. Make a string telephone with two yoghurt pots and some string. The string must be kept taut. The vibrations of one's voice cause the pot to vibrate which causes the string to vibrate and it passes the vibrations on to the other pot where they are heard as sound.

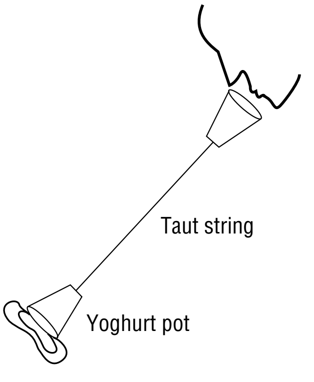


Fig. 2.21 *String telephone*

3. Using a long iron fence, one person places their ear to the fence while another taps the fence with a stone as far from the listener as possible. The sound is heard through the fence first, followed by another sound heard through the air. Sound travels fifteen times faster through iron than through air.

   The central heating pipes could also be used instead of an iron fence.

4. Listen through the base of a glass pressed against a wall to sounds being made in an adjacent room. The sounds cause not only the air to vibrate but the walls and doors also. The vibrating wall causes the air in the glass to vibrate passing the sound to your ear.

The stronger the bonding between the molecules the greater the speed of sound in the medium, the bonds acting like springs between the molecules.

Hence the speed of sound in solids > speed of sound in liquids > speed of sound in gases.

Speed of sound in iron = 5000 m s^{-1}.

Speed of sound in water at 25 °C = 1498 m s^{-1}.

Speed of sound in air at 0 °C = 331.3 m s^{-1}.

Factors Affecting the Speed of Sound in Gases

(a) The speed of sound in gases is independent of pitch and loudness

When listening to an orchestra all sounds, regardless of pitch or loudness, which are produced simultaneously, arrive at the listener simultaneously. However, shock waves, for example from a violent explosion, travel faster near the source.

(b) The speed of sound is independent of pressure changes at constant temperature
(See Reference material.)

(c) The effects of temperature

The speed of sound in air is 331 m s^{-1} at 0 °C and increases by 0.6 m s^{-1} for every 1 degree Celsius rise in temperature. At room temperature its speed is about 340 m s^{-1}.

(See Reference material.)

(d) The effect of the type of gas

The lower the gas density the greater the speed of sound.

Connect a tin whistle to a gas tap (natural gas) with several metres of rubber tubing. On turning on the gas the whistle emits a note whose pitch (and hence its frequency) rises as air is blown out of the whistle and is replaced by the less dense gas from the tap. Now take in a deep breath of air, disconnect the tubing from the tap, and blow **down** the tube (**don't inhale**). The pitch of the note now falls as the gas in the whistle is replaced by denser air.

The speed of sound is given by $v = f\lambda$. Wavelength depends only on the length of the tin whistle which does not change. So, if speed increases, then frequency increases.

Therefore, since sound travels faster in the less dense gas the pitch of the note is higher in the less dense gas.

(e) The effect of humidity

The speed of sound increases with humidity.

(See Reference material.)

(f) The effect of wind

The velocities of the wind and of the sound must be added vectorially as the wind is moving the air through which the sound is travelling. If they are both travelling in the same direction the effect is an increase in the velocity of sound and if they are travelling in opposite directions the effect is a decrease in the velocity of sound.

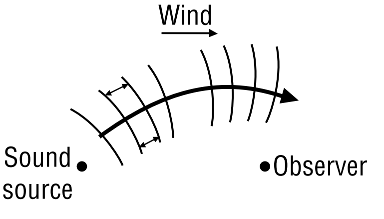


Fig. 2.22 *Wind and sound travelling in the same direction*

In the first case, as the velocity of the wind increases with height above the ground the upper part of the sound wavefront is moving faster than the lower part and the wavefront is distorted so that the sound is easily heard downwind, Fig. 2.22.

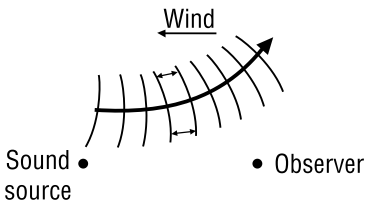


Fig. 2.23 *Wind and sound travelling in opposite directions*

When the wind and sound are travelling in opposite directions the upper part of the wavefront is moving more slowly than the lower part and the higher up the greater the decrease in velocity. Hence the wavefront is bent away from the ground, Fig. 2.23.

Experiment 2.1

To Measure the Speed of Sound Using Echoes

(Students may do this experiment outside of school time. Results will be only approximate due to inaccuracies in timing.)

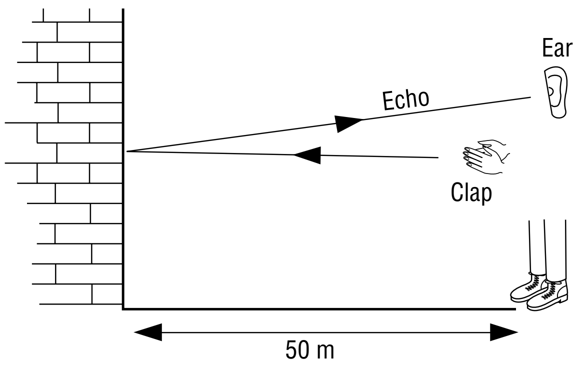


Fig. 2.24 *Measuring speed of sound using echoes*

1. Find a building with a large flat wall in a quiet place.
2. Measure the distance walked in 10 paces and use this to measure distance from the building keeping the paces even in size.
3. Move about 50 m or further from the building, as long as there are no intervening obstacles and an echo can be clearly heard at the particular distance.
4. Clap hands to make a sound or bang two boards together.
5. As soon as the echo is heard repeat the sound and continue repeating the sound so as to overlap the echo. When the correct rate of clapping has been achieved after some practice a second observer times about 30 of them.

The first clap counted when the stopclock is started is counted as zero.

A large number of claps must be counted to minimise the timing error as the time taken for the sound to travel to the wall and back is less than 1 s if the journey is less than about 330 m.

Calculations

Distance to the wall = d

Time for 30 claps = t

Time taken by the sound to travel to the wall and back = $t/30$

Speed of sound = distance/time = $2d/(t/30)$

The experiment may be repeated several times and a mean value calculated.

Experiment 2.2

To Measure the Speed of Sound in Air Using a Resonance Tube

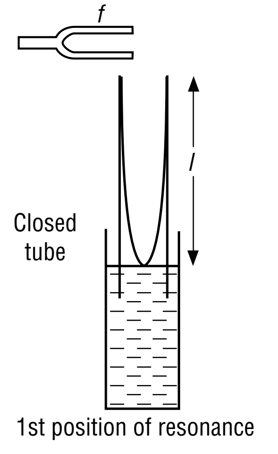


Fig. 2.25

Apparatus

250 ml graduated cylinder, Wavin pipe of about 2.5 cm diameter or a combustion tube, selection of tuning forks, vernier calipers, metre stick.

● The tuning fork must not be struck too hard otherwise it will not vibrate at a single frequency and resonance may be obtained for one of its harmonics. Resonance positions must be for the same note as the tuning fork and not for one of its harmonics.

● A short length of air outside the open end of the tube is also set vibrating and so the length of the vibrating air column is $l + e$, where l is the length of the air column and e is the end correction. The end correction is approximately equal to 0.3 times the diameter of the tube.

● Using an approximate value for the speed of sound of 300 m s^{-1} calculate an approximate value for the first position of resonance for the 512 Hz tuning fork. Speed = $f\lambda$ and $l = \lambda/4$.

● Keeping the movable tube as near vertical as possible set the length of the air column about 4 or 5 cm shorter than the calculated value and holding the vibrating tuning fork above it slowly raise the tube out of the water until resonance is obtained. Clamp the tube in that position and measure the length of the air column with a metre rule. Repeat the procedure with the same tuning fork and measure l again. Take the average of the measurements for l.

$$\lambda = 4(l + e).$$

● Repeat the above procedure for tuning forks of decreasing frequency – hence the resonance lengths will be increasing and calculate the average value of the speed of sound from:

$$v = 4f(l + e).$$

Calculating the speed of sound from a graph

● Speed of sound $v = 4f(l + e)$. Manipulating the equation gives $l = (v/4f) - e$.

A graph plotted of l on the y-axis against $1/f$ on the x-axis yields a straight line of slope $c/4$ the intercept on the y-axis being $-e$, Fig. 2.26. Hence the speed of sound is equal to four times the slope. The graph will show up any doubtful points which can then be rechecked.

Sample results

f/Hz	l/m
512	0.156
480	0.17
426	0.19
384	0.216
341	0.238
320	0.26
288	0.29
256	0.319

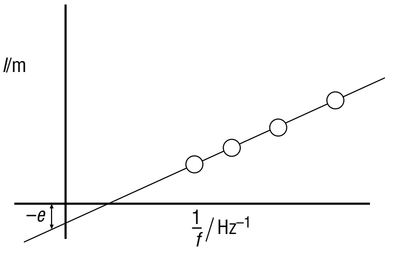


Fig. 2.26

The Characteristics of a Musical Note

Loudness is a subjective measurement of sound power which is determined by:

(i) the sensitivity of the listener's ear;

(ii) frequency (see graph of frequency response of the human ear, Fig. 2.6, p. 27);

(iii) the intensity of the sound.

Very low and very high frequencies need greater amplification to produce the same sensation of loudness as other frequencies of equal intensity.

The situation is analogous to the frequency response of amplifiers, which should amplify all frequencies equally but whose amplification falls off at the higher frequencies (above 10 kHz) and below about 100 Hz, Fig. 2.27. The treble control can be used to vary the intensity of the upper frequencies according to personal taste and likewise the bass control to vary the intensity of the lower frequencies.

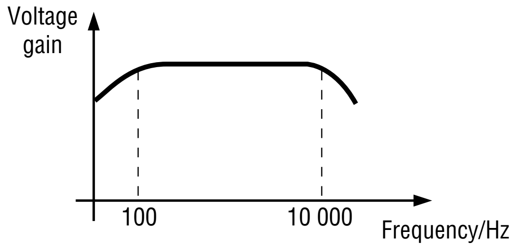


Fig. 2.27 *Frequency response of amplifiers*

Pitch is a sensation experienced by a listener which depends on frequency.

It determines the position of a particular note in a musical scale and is analogous to colour in the visible spectrum.

To Demonstrate Loudness and Pitch

Apparatus

Cathode ray oscilloscope, signal generator + amplifier, loudspeaker.

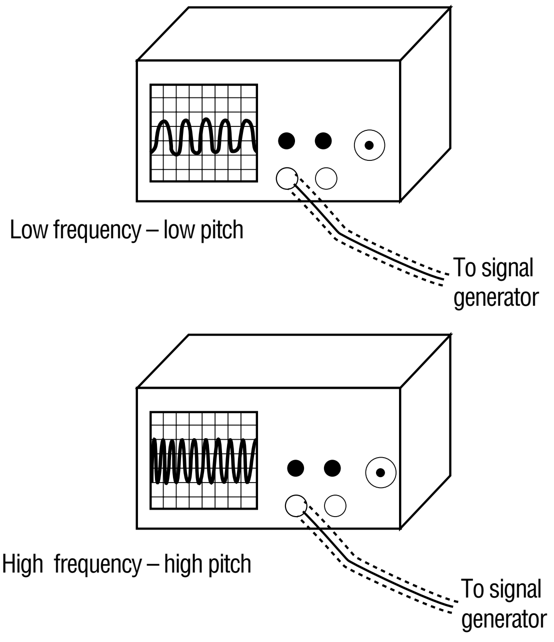


Fig. 2.28

Method

Attach the signal generator to the oscilloscope and the loudspeaker. Vary the frequency setting of the generator and adjust the timebase setting of the oscilloscope (see p. 67) to give a small number of waves on the screen, Fig. 2.28. Note that as frequency is increased as seen on the oscilloscope pitch also increases and that as the amplitude increases the loudness increases.

To Demonstrate the Frequency Limits of Audibility

Apparatus

Audio signal generator + amplifier connected to a loudspeaker and to Channel 1 input of the CRO.

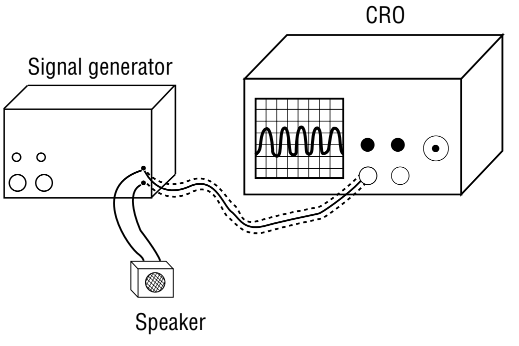


Fig. 2.29

Set up the CRO as explained on p. 67.

Start with the lowest frequency on the signal generator applied to the loudspeaker and the CRO. Note the frequency at which the signal first becomes audible. This is the lower limit of audibility. Continue increasing the frequency, observing the waveform on the screen and listening to the sound. Note that at some frequencies the sound appears louder than at others due to variation in the ear's response to different frequencies.

At a certain high frequency, the sound disappears while the trace is still visible on the screen. This is the upper limit of audibility.

The range of audible frequencies decreases with age and/or as a result of exposure to very loud noises and varies among individuals. The maximum range for humans is about 20 Hz to 20 kHz. Only small children have this range; most adults cannot hear above 16 kHz. Frequencies above 20 kHz are called ultrasonics.

Infrasound, i.e. frequencies below the audible frequency range are produced for example in earthquakes. These frequencies can be felt rather than heard.

Some geological vibrations complete a cycle in 100 seconds or more, i.e. have a frequency of 0.01 Hz or less!

Quality

Quality distinguishes two notes of the same pitch played on different instruments. One of the factors which determines the quality is the overtones present along with the fundamental and the relative amplitude of each overtone to the fundamental.

The fundamental component of a note has the lowest pitch and usually has the greatest amplitude. It is the other frequencies present, i.e. the overtones, which contribute to the quality of the sound.

Combining the fundamental with a higher frequency to build a complex waveform

Add, using the principle of superposition, the fundamental frequency and the third harmonic of less than half the amplitude, Fig. 2.30, (or other combinations).

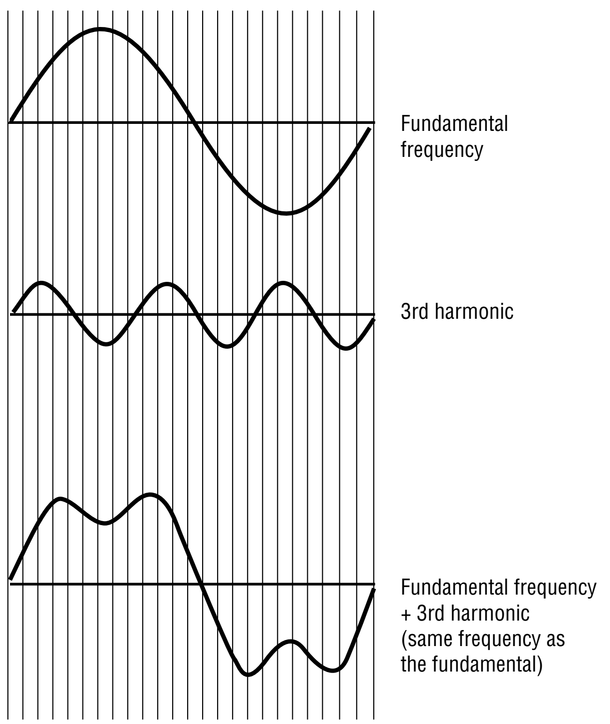


Fig. 2.30 *Combining frequencies*

The sounds made by a full orchestra at any particular instant add up to a very complex waveform yet the paper cone of a loudspeaker is able to reproduce it exactly.

The opposite exercise, i.e. reducing a complex waveform to its component sine waves of related frequency is known as Fourier analysis.

To Demonstrate the Waveforms of Different Notes

Apparatus

CRO, microphone.

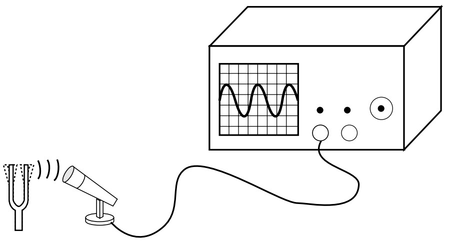


Fig. 2.31

● Look at the waveform from a vibrating tuning fork.

● Different students sing the same note into a microphone connected to the CRO.

● Play the same note from different musical instruments (or alternatively make recordings of the same note played on different musical instruments) into a microphone connected to the CRO and observe the waveforms.

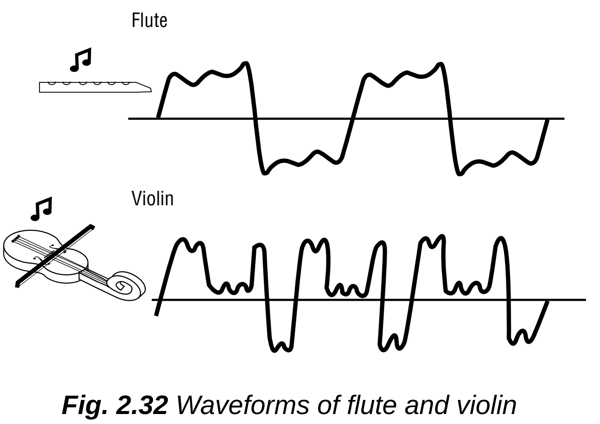


Fig. 2.32 *Waveforms of flute and violin*

● Change the shape of the lips or tongue while singing a particular note into the microphone.

● Compare the above waveforms with that produced by the tuning fork which provides a pure note consisting of one frequency and a smooth regular curve.

● Play a tape recording of perhaps an orchestral piece of music and display it on the oscilloscope by picking it up on a microphone connected to the CRO.

● Connect an electronic organ to an oscilloscope and again observe the waveforms for different musical instruments.

2.4 Applications

Speaking Tubes

Speaking tubes on ships are made of metal with a funnel at each end. Sound travels along the tube, even if it's bent, by total internal reflection, as long as the angle of incidence exceeds the critical angle. The sound travels more slowly in the air inside the tube than in the metal walls of the pipe. The situation is analogous to light travelling along optical fibres.

Echoes

Echoes are reflected sounds. They must be heard at least 0.1 s after the original sound to be distinguishable from it. A surface will reflect sound regularly if it is flat to within a fraction of the wavelength of the sound. As speech sound waves may have wavelengths several metres long even rough surfaces such as cliffs can give regular reflection thus giving echoes.

In open spaces where sound spreads out in all directions echoes are fainter than in enclosed spaces such as a tunnel where the reflected sound cannot escape.

Echoes of ultrasonic waves are used by ships to:

(i) calculate ocean depth;

(ii) detect shoals of fish;

(iii) detect submarines or shipwrecks.

This is called SONAR, i.e. Sound Navigation And Ranging. The time between sending out a pulse of ultrasound and receiving the echo can be used to calculate the position and shape of objects under the ship. It can show the difference between a large fish and a shoal of small fish. Sonar can also be used to draw up a map of the seabed which is very important for the laying

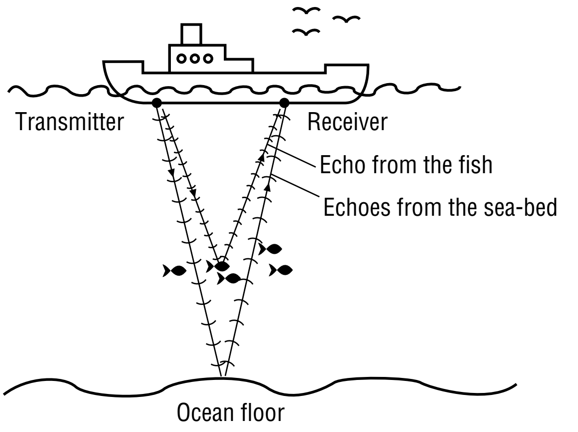


Fig. 2.33 *Ship using echo sounding*

down of optic fibre cables for the purpose of transatlantic communication.

Echoes of ultrasonic waves are also used by:

dolphins to find objects underwater;

bats to locate their prey;

visually impaired people fitted with special equipment to locate obstacles.

Vibrato or Tremolo Effect in the Organ

Large organs may have many thousands of pipes, varying in length from a few centimetres to over nine metres. Some groups of pipes may have two pipes for each note, with one slightly detuned, hence producing beats when sounded.

Active Noise Cancellation (ANC)

Noise in cars from the engine and road surfaces which interferes with music from the car radio can be cancelled out by destructive interference with its own mirror image using ANC.

The noise is sampled by a microphone attached to a microcomputer and the original noise and its mirror image are transmitted through two speakers placed behind the driver's head and so the noise is cancelled out, Fig. 2.34.

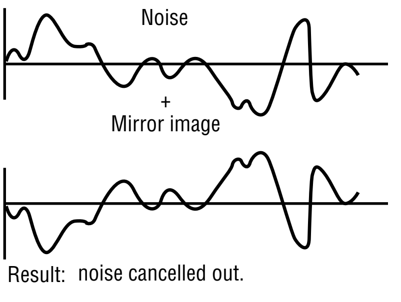


Fig. 2.34 *Active noise cancellation*

Stethoscopes

The stethoscope directs sound into the doctor's ear. Sound normally spreads out in all directions and this is how quiet sounds are soon lost. When the faint sounds of lungs and heart are closely confined to the path down the tube this enables the doctor to hear them. (Experiment: make one using a short piece of rubber tubing and a funnel in one end, Fig. 2.35.)

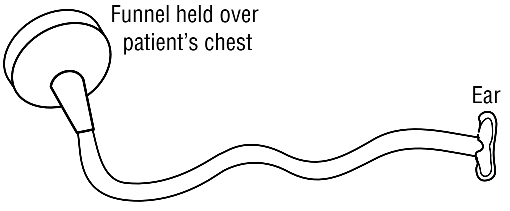


Fig. 2.35 *A simple stethoscope*

Doctors used to use thin rods with broadened ends as stethoscopes and, like modern types, one end was placed against the patient's chest and sounds therein were transmitted through the wood to the doctor's ear.

Motor mechanics may use wooden rods to trace sounds of knocking in engines. The sounds are concentrated in the rod as explained earlier and other noises are blocked out.

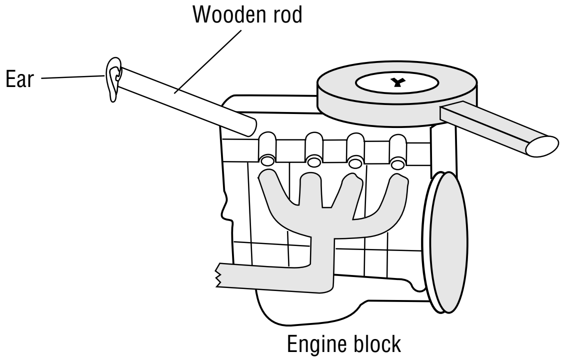


Fig. 2.36 *Motor mechanics may listen for knocking sounds in engines using wooden rods*

Native American scout listening through the ground for oncoming horse's hooves

The hooves strike the ground directly so sound enters the ground where it is produced. The vibrations that occur cause air molecules to vibrate also. Hence the sound can reach the listener's ears through the ground or through the air. The sound travelling through the air may meet obstacles and even though sound can bend around obstacles, the energy becomes more spread out and it would not be as loud as if heard directly. Also, if the air near the ground is warmer than that higher up (usually what occurs in daytime) the waves bend upwards over the listener's head and the sound is not heard.

The sound that travels through the ground goes directly to the scout's ear, rather than having to go around obstacles, compressing a section of the rock which expands and compresses the next section and so on.

This works particularly well in the American West with its thin rocky soil, with lots of rock outcrops in contact with the bedrock. Softer soils would absorb a lot of the sound's energy.

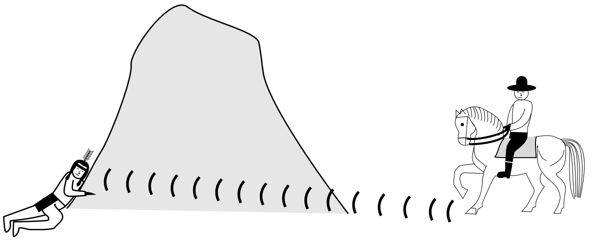


Fig. 2.37 *Native American scout listening for oncoming horses hooves*

Attracting Fish

Fishermen use transistor beepers in the water which are claimed to attract fish over a radius of 1.5 km. The fact that fish were attracted by the sound of a bell underwater was discovered over 300 years ago by Otto von Guericke who also discovered that sound could not travel through a vacuum.

Synthesisers and Electronic Organs

Varying electric currents rather than mechanical vibrations produce the sounds. Several harmonics are added to the fundamental and their amplitudes are adjusted until the waveform produced electronically matches the natural waveform of a particular instrument. These electronic codes are then stored in memory and when a particular instrument is selected and notes played on the keyboard the correct type of waveform for that instrument is reproduced from memory. In synthesisers, as opposed to electronic organs, the shape of the waveforms may be changed, thus synthesising new sounds. An amplifier and loudspeaker convert these electronic signals into sound waves. Backing tracks or complete tunes may be stored in memory and played at the press of a button.

The waveforms shown in Fig. 2.32, p. 36, only represent a very small interval of time, of the order of hundredths of a second and so the waveforms appear regular. However if looked at over longer periods of time they would show irregularities. These irregularities or 'wobbles' enable us to distinguish from the real instrument electronic sounds which attempt to emulate the sounds of 'real' musical instruments. In the early days of electronic organs the distinction was very obvious but now even the irregularities, and variations in the irregularities, can be synthesised, bringing about vast improvements in the electronic production of sound.

The Vocal Cords

Vibrations of the vocal cords, which are controlled by muscular tension, give rise to the human voice and these vibrations determine the fundamental pitch of the voice. The vocal cords are bands of connective tissue in the larynx. Several sets of muscles can pull these cords with various degrees of tension into the path of air coming from the lungs. They then vibrate as violin strings vibrate. Pitch is varied by varying tension in the vocal cords while volume is varied by varying the force of the air current. Vocal cords are shorter in women than in men.

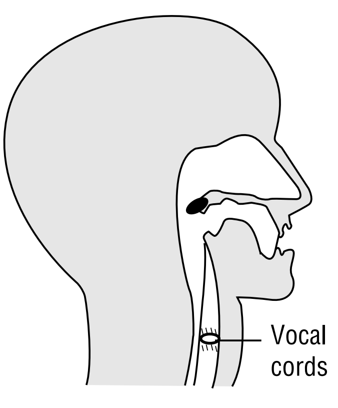


Fig. 2.38

The Quality of the Human Voice

The size and shape of the throat, mouth and nasal cavities determine the frequencies of the overtones produced along with the fundamental and hence these factors contribute to the quality of the voice. Sound echoes back and forth in the throat and mouth cavities. At each end some of the wave is reflected and waves of some frequencies superpose to give a large-amplitude standing wave while waves of other frequencies cancel out by destructive interference and are thus not heard. Just as longer, larger tubes give lower notes large people tend to have lower voices as they have bigger and longer voice boxes. Good quality singing voices place emphasis on two sets of resonant harmonics, a pharynx natural frequency of around 500 Hz and resonant frequencies of the throat, mouth and nasal cavities of between 2400 and 3200 Hz.

Change in the Pitch of the Voice on Inhaling Helium

NB. It is very dangerous to inhale gases other than air but it is safe to inhale helium for just one breath provided air is breathed in immediately afterwards. Gases other than helium could cause explosions or death.

Exhaling completely and then filling the lungs with helium gas for just one breath makes a person's voice sound strange and high pitched. Helium molecules have much less mass than air molecules and therefore at a given temperature, and therefore fixed kinetic energy, they will move faster than air molecules. Since the speed of sound is related to the velocity of the gas molecules transmitting it the sound travels faster in helium than in air. Since the speed is higher while the wavelength remains the same the frequency, and therefore the pitch, are higher. The speed of sound in helium is 981 m s^{-1} at 0 °C and in air is 331 m s^{-1} which is almost three times less, thus making the frequency of the voice almost three times higher in helium and somewhat less if air is mixed with it. The harmonics from the resonating mouth, throat and nasal cavities are raised about one and a half octaves and even the fundamental frequency due to the vibrating vocal cords is higher than normal. They move faster since they have a smaller mass of gas to move and being unused to this a person does not correct for it by decreasing tension, and thus pitch is higher.

Musical Sounds vs Noise

Sound of a regular frequency or consisting of related frequencies such as fundamental and harmonics, with a periodic waveform is a musical sound.

Sound with a constantly varying pitch and with no pattern or repetition in the waveform and consisting of a random mixture of many unrelated frequencies is noise.

'Hiss' (high pitch noise) produced in the electronics of hi-fi systems, and 'rumble', i.e. sound produced by the bearings in a turntable motor which is picked up by the stylus (lower frequency noise) can be a problem in older hi-fi systems, but have been eliminated by CD and DAT (Digital Audio Tape) systems.

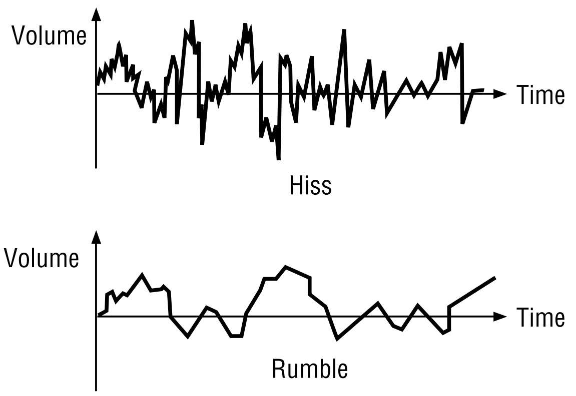


Fig. 2.39 *Hiss and rumble*

Musical scales are based on the relative frequencies of different sound waves. Scales are chosen to produce the greatest amount of harmony or pleasing combinations. Notes which have a simple ratio between their frequencies (e.g. 1:2 in the case of an octave) sound pleasant when played together and they are said to show concord. When the ratio is not a simple one the notes show discord when sounded together and produce a harsh, disagreeable sound.

The relationship between notes in a scale depends on the ratio of the frequencies rather than their differences. For example a ratio of 1 : 2 is an octave regardless of whether it is 256 : 512 or 512 : 1024. (This is similar to the ear's response to the ratio of two intensities rather than to their differences.)

The ratio of the frequencies of two notes is called a musical interval. A frequency ratio of 9 : 8 is a major tone, 10:9 is a minor tone, and 16 : 15 is a semitone.

Speech Therapy for Deaf People

Deaf people have to try and copy sounds by just looking at the shapes people make with their lips and mouth. Since the fundamental frequency is produced by the vibrations of the vocal cords, which they cannot see in order to imitate the sound, it is thus very difficult for them to reproduce the correct sound. To help them the waveforms of vowels and consonants are reproduced electronically and displayed on a screen. The person then continually alters his/her voice which is also displayed on the screen, until it has the desired form.

The Quality of Different Musical Instruments

A lot of skill is required in making musical instruments in order to suppress undesirable harmonics and enhance desirable ones. When a note is played, other higher frequencies are produced along with the fundamental frequency. The presence and relative intensities of these higher frequencies contribute to the quality of the sound and distinguish the instruments from each other. Using the mathematical technique of Fourier analysis as mentioned above the frequencies and amplitudes of the harmonics present in a note can be calculated and a sound spectrum obtained.

Brass instruments allow all the harmonics to be heard, especially the higher frequency ones which have relatively large amplitudes. This is because of their hard smooth interiors which do not absorb the harmonics.

The trumpet's quality depends on the presence of high frequency harmonics of large amplitude. When a mute is placed over a trumpet it absorbs the higher frequency harmonics so that the notes played sound more mellow.

In the piano most of the strings are struck by the hammers at approximately one seventh of their length from one end. This allows up to six harmonics to be produced in any note and suppresses the seventh which does not harmonise with the fundamental. The strings for the top $1^1/_2$ octaves or so are very short and in order to obtain the best tone from them they are struck as near to one end as possible.

The quality of the violin sound is due in part to the presence of a wide range of high frequency harmonics of high intensity which give it a piercing tone.

CHAPTER 3 RESONANCE, INTENSITY LEVEL AND ULTRASONICS

3.1 Background

Resonance occurs when an object is forced to vibrate at its own natural frequency and results in large amplitude oscillations and maximum transfer of energy between the forcing agent and the driven oscillator.

Natural frequency is the frequency of a freely vibrating body.

Resonance can occur in mechanical systems, in sound and in electrical circuits.

Engineers must be aware of possible sources of mechanical resonance in their design of bridges, buildings, aircraft and other mechanical systems.

Vibrations in Pipes

Longitudinal standing waves are produced in pipes when they sing out in resonance with a forcing frequency. Resonance occurs for those waves whose wavelengths are correctly matched to the length of the pipe.

A displacement node and a pressure antinode occur at the closed end and a displacement antinode and a pressure node at the open end.

Vibrations in air columns are set up by a disturbance at one end, the other end being closed or open. The progressive sound waves travel to the end of the pipe where they are reflected. The incident and reflected waves interfere and if constructive interference occurs, as when a compression arrives back at the source as it is producing another compression, a longitudinal stationary wave is set up.

The diagrams in Fig. 3.1 are drawn as if the waves were transverse stationary waves since it is not possible to draw a picture of a longitudinal standing wave. The possible modes of vibration are drawn separately but the actual mode of vibration is the result of the superposition of these various modes. The displacement antinode occurs a distance e, called the end correction, beyond the open end ($e \approx 0.3$ diameter).

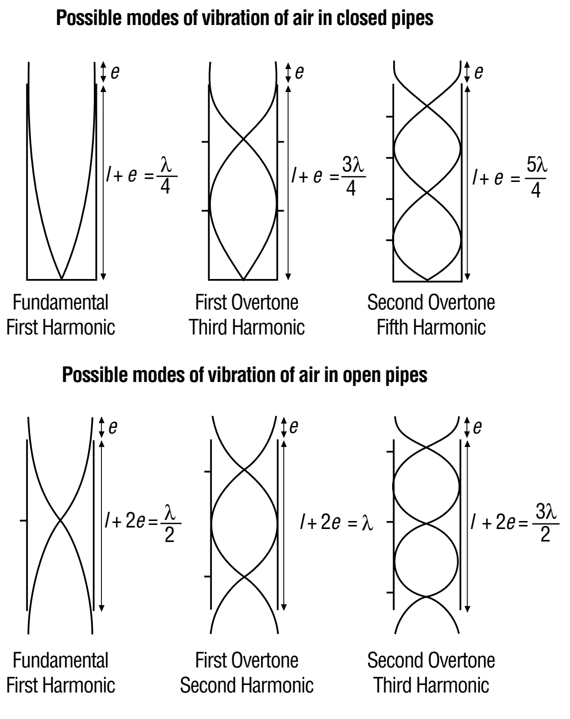


Fig 3.1 *Modes of vibration in closed and open pipes*

The trombone is a long tube which can be lengthened or shortened to change the pitch of the note produced. When the player blows into the mouthpiece causing 'white noise', i.e. a sound composed of many frequencies, the trombone selects a particular frequency with which it has resonance and a stationary wave of that particular frequency is set up in the tube with antinodes at each end.

The length of the vibrating air column in a trumpet, cornet, tuba, etc., is varied by means of valves. A valve is a piston which, when pressed down, adds an extra length of tube. The valves direct the air around longer or shorter pipes.

In the flute, recorder and tin whistle different holes are either covered or uncovered, thus changing the effective length of the pipe. The pitch can also be changed by 'overblowing', i.e. blowing harder. This encourages the higher harmonics and the fundamental frequency does not exist with this kind of blowing.

Intensity and Intensity Levels

Intensity is an objective measurement of sound power (see standard definition).

Intensity is inversely proportional to the square of the distance from a point source and is proportional to the amplitude of vibration squared. It increases as the mass of air set into vibration increases.

The bel is used to measure relative increases in intensity and sound intensity levels. The ear responds to ratios of intensities rather than their differences. Increasing the intensity of a quiet sound by a small amount causes a much bigger increase in loudness than increasing the intensity of a loud sound by the same amount. Hence a whisper can appear very loud in a quiet study room but is not noticeable at a disco.

The scale of intensity level mirrors the response of the human ear in that it is based on the ratios of sound intensities rather than their differences.

The energies involved in common sources of sound are very small. The power of a full orchestra is about 70 W. However the ear is sensitive to a very wide range of intensities, from 10^{-12} W m^{-2} to 1 W m^{-2}.

If the intensity of a source changes from I_1 to I_2 then the relative increase in intensity in bels = $\log_{10}(I_2/I_1)$.

The bel is quite a large unit and it is more common to use the decibel.

The intensity level of a source of sound is the log of the ratio of its intensity to some agreed 'zero' level of intensity.

The agreed 'zero' is the lowest audible sound, at a frequency of 1 kHz. This is taken to have an intensity of 10^{-12} W m^{-2}, which is called the threshold of audibility.

A sound with an intensity level of 80 dB is 8 B higher than the threshold of audibility and therefore its intensity must be 10^8 times greater. Therefore its intensity is $10^8 \times 10^{-12}$ W m^{-2} i.e. 10^{-4} W m^{-2}.

The minimum change in sound intensity which the human ear can detect is a change of approximately 25%. It corresponds to a relative increase or decrease in intensity of approximately 1 dB.

Sound Level Meters

Circuits are built into these meters so that they respond more to frequencies in the 2 to 4 kHz range and less to frequencies at the higher and lower extremes and in that way they respond similarly to the human ear. This frequency-weighted scale is called the A-scale and the units are called dB(A).

Everyday sound intensity levels in decibels

0	Threshold of hearing
10	Almost silence
20	Quiet room, e.g. library, or blood pulsing
30	Whisper
40	Quiet street
50	Quiet conversation
60	Normal conversation
70	Loud conversation
80	Door slamming or alarm clock at 1m
90	Everyone talking in a classroom
95	Radio playing very loud (1m away)
110	Pneumatic drill
120	Threshold of pain
140	Aircraft at 25 m
160	Rifle close to the ear (rupture of the eardrum)

Suppose the intensity level is 80 dB at a distance of 10 m from a point source. If the distance is doubled, i.e. increased to 20 m, the intensity will be 4 times less, and the relative decrease in intensity in decibels is 10 log 4, i.e. approximately 6 dB. Therefore, the intensity level at 20 m is (80 – 6) = 74 dB.

A drop of 6 dB means that the sound energy is approximately quartered.

A drop of 3 dB would mean that the sound energy is approximately halved (10 log 2).

Noise

Noise is a form of pollution in that it can spoil our enjoyment of pleasant sounds or disturb our tranquility or indeed it can permanently damage our hearing. We think of it as unwanted sound. The danger to our hearing increases as the intensity level and the length of exposure time increase.

Regular exposure to noises greater than 90 dB may be experienced by people, e.g. in a factory, driving a large vehicle, attending a rock concert, listening to a personal stereo, or operating a pneumatic drill. The intense noise due to overflying aircraft is also a source of concern, especially to those living near airports.

The effect of such exposure may be only temporary damage to hearing but the possibility of permanent serious damage exists, especially with continuous exposure.

One of the warning signs is ringing in the ears after the experience and very loud noise can even cause temporary blindness and nausea.

Persistent noise, even though it may not be excessively loud, can reduce concentration, and cause irritability, short temper, and tiredness.

Exposure to 110 dB for more than 2 minutes can result in permanent hearing loss.

Noise Reduction

- *Wearing ear muffs.*
- *Moving away from the source of sound.*

For a point source intensity is inversely proportional to distance squared. An intensity level of 100 dB at 10 cm is reduced to 88 dB at 40 cm. (Distance is increased by a factor of 4, hence intensity is reduced by a factor of 16. Relative decrease in intensity in decibels is 10 log 16 which is approximately 12 dB.)

- *The formation of stationary waves to suppress high levels of noise from a generator*.

The generator noise is detected and is produced as a sound wave from a loudspeaker. A stationary wave is formed due to interference between this wave and the wave reflected from a barrier. The loudspeaker is moved so that an interference minimum exists at the generator where personnel would be working perhaps. Elsewhere of course there would be interference maxima.

- *Double glazing and heavy curtains reduce noise from the outside.*
- *Trees or an earth bank or a wall between a house and busy traffic absorb noise.*
- *Carpets and soft wall coverings reduce echoes by absorbing rather than reflecting sound.*
- *Noisy machines may be enclosed in airtight containers sitting on, e.g. a rubber mat, and kept away from contact with the walls of the container.*
- *Improving the design of machines by preventing unnecessary vibrations which might lead to resonance.*

Ultrasonics

Ultrasonics are frequencies greater than 20 kHz. Up to about 1910 they were little more than a curiosity. They were used in the early forms of sonar to detect submerged submarines and are used in measuring the depth of the sea in ship's fathometers (see Fig. 2.33 and Echoes).

3.2 Do You Know

In an earthquake why may all buildings of a certain size be damaged while taller and smaller buildings remain undamaged?

A natural frequency of these buildings is the same as a major frequency of the tremor. Resonance occurs and the buildings vibrate with a large amplitude and are destroyed.

An opera singer can shatter glass by singing at its natural frequency.

'Pinging' a glass with one's finger causes it to vibrate at its own natural frequency. A singer singing at this frequency can cause the glass to vibrate so much in resonance with the forcing frequency that it cracks.

It is possible to get a response from a piano by singing to it.

With the damper (right) pedal down, so that all the strings are free to vibrate, singing into the piano, with the lid lifted, causes the string with the same natural frequency as the forcing frequency of the voice to vibrate in resonance with the voice and will go on vibrating after the singer stops singing.

Resonance caused problems in the early production days of motor cars.

'Drumming' sounds were heard at certain speeds due to resonance between the engine vibrations and the body panelling. A plastic coating on the inside of the panels reduced the natural frequency of vibration of the panels and also absorbed the vibrations.

Tides of up to 15 m occur in, e.g. the Bay of Fundy (Nova Scotia).

These very high tides are due to resonance. The natural frequency of this body of water (and some others) is very close to the frequency of the tides. Hence every time the tide comes in it is in step with the last tidal flow. Constructive interference occurs increasing the amplitude of the wave. It doesn't increase indefinitely due to the limiting effect of friction but is still spectacular.

Resonance can enhance the voices of bathroom singers.

If the natural frequency of the air in an enclosure approximates to the frequency of a sound being produced in the enclosure then resonance will occur and the frequency being produced will be reinforced through the response of the air within the enclosure.

Most rooms where music is performed are large enough so that their natural frequencies of vibration are lower than the lowest musical pitches being produced but in small bathrooms the natural frequencies of the room may be within the pitches of the male singing voice thus supporting it and making it appear richer.

Why do personal stereos not need a very large power output?

The ear is extremely sensitive and would be damaged by high intensity of sound. Also, nearly all the sound goes directly into the ears.

How sensitive is the human ear?

Standing 1 m away from a 40 W light bulb (low power) the total energy intensity would be approximately 3 W m^{-2}. The equivalent amount of sound intensity would be extremely painful and damaging to the human ear as it is so sensitive. 1 W m^{-2} is the threshold of pain for the human ear.

Blue whales emit the loudest sounds of any living source.

The loudest sounds emitted by any living source are the low frequency pulses made by blue whales communicating with each other. These pulses have been measured up to intensity levels of 188 dB. They have been detected up to 850 km away.

3.3 Experimental Approach

Barton's Pendulums

The driving pendulum is set into oscillation. It sends small amounts of energy through the supporting thread and this causes the other pendulums to oscillate. The pendulum c, Fig. 3.2, whose length is the same as that of the driver pendulum, oscillates with the largest amplitude, i.e. it shows resonance, because its natural frequency of vibration is the same as the forcing frequency.

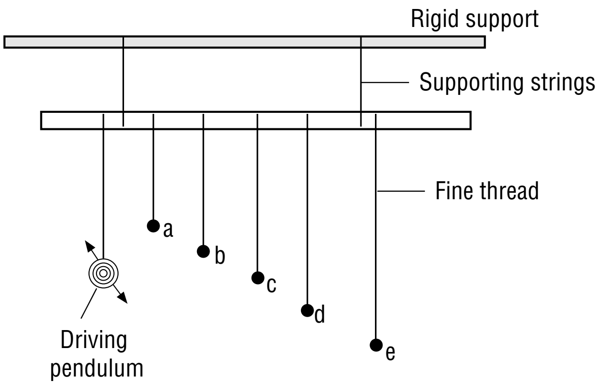


Fig 3.2 *Barton's pendulums*

Note: the pendulums do not all vibrate in step, i.e. they have different phases.

Making a Loudspeaker Cone Resonate

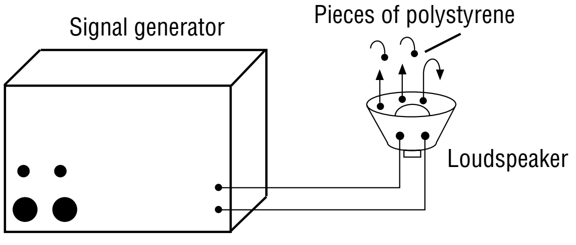


Fig. 3.3

Connect a loudspeaker to a signal generator (*low resistance output*).

Put small pieces of polystyrene into the cone.

Switch on the signal generator at low frequency. The cone vibrates.

Gradually increase the frequency.

At a certain frequency the amplitude of vibration of the cone increases greatly and the pieces of polystyrene are thrown into the air.

The signal generator has forced the cone to vibrate at its natural frequency, resulting in large amplitude vibrations, i.e. resonance occurs.

Minimal resonance however would be high on any hi-fi speaker designer's list of priorities as a good quality speaker must add as little as possible to a signal.

Tuning a Sonometer String with a Tuning Fork

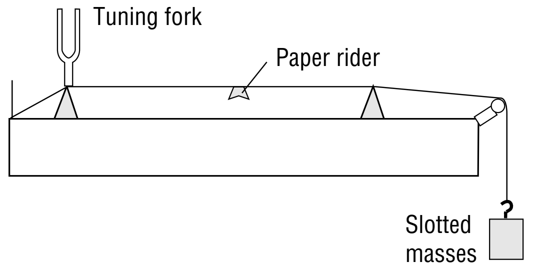


Fig. 3.4

Hold a vibrating tuning fork on one of the bridges. It forces the string to vibrate. If the natural frequency of vibration of the string is made equal to that of the fork, by changing its length or tension, then the string will vibrate with a large amplitude, shown by a paper rider placed at the centre of the wire being thrown off.

Resonance has occurred and the frequency of the string then equals that of the tuning fork.

Caution. Ensure that the slotted masses and holder are attached securely to the steel wire. It could cause serious injury if released suddenly. Wear goggles and do not peer too closely at the wire.

Resonating Tuning Forks

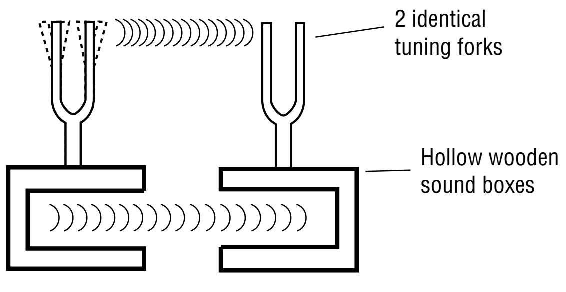


Fig. 3.5

If two identical tuning forks are mounted on two boxes and one of them is set vibrating then, as the two tuning forks have the same natural frequency of vibration, the second fork responds to the forcing frequency of the first fork with large amplitude vibrations and can be heard if the first one is stopped, i.e. resonance occurs.

If the experiment is repeated but this time changing the natural frequency of the second fork by wrapping elastic bands tightly around it or loading it with plasticine then resonance no longer occurs.

Stationary Waves in a Stretched String (1)

Apparatus

Sonometer, strong U-shaped magnet or 4 Magnadur magnets on a yoke, signal generator with an amplifier.

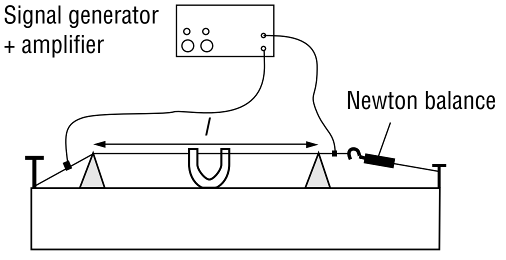


Fig. 3.6

Note. Connect to the low resistance terminals of the signal generator and turn down amplitude control if the wire is getting warm.

A current-carrying wire experiences a force F, when placed at right angles to a magnetic field of flux density B, the direction of the force being perpendicular to the field and the current I.

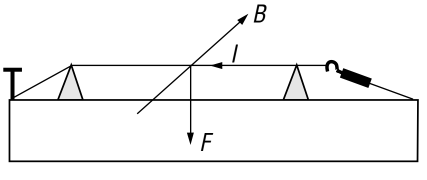


Fig. 3.7

If an a.c. current of frequency f is used then the force F will alternate with the same frequency.

The wire thus vibrates transversely with frequency f. If the frequency f is equal to the natural frequency of vibration of the wire, i.e. its frequency of vibration when plucked, then resonance occurs and can be heard and seen. Vary the applied frequency or else the tension in the wire to achieve this.

A loud sound and large amplitude vibrations result provided the tension is not too great. Stationary waves are set up as a result of the superposition of incident waves and waves reflected at the fixed ends.

If the magnetic field is applied at the centre of the wire an antinode exists at the centre and a node at each end, i.e. the fundamental, Fig. 3.8(a).

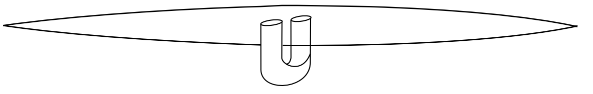


Fig. 3.8(a) *Fundamental*

For example, for a steel wire, 24 s.w.g., 0.56 mm diameter, tension 3 N, distance between the bridges = 50 cm, resonance occurs at 40 Hz approximately.

If the magnet is at a node of a particular harmonic, vibrations will not occur and it must be moved to an antinode.

To see two loops, double the frequency (second harmonic) for the same length and tension, and place the magnet one quarter way from the end of the wire. Adjust the frequency slightly until the wire vibrates.

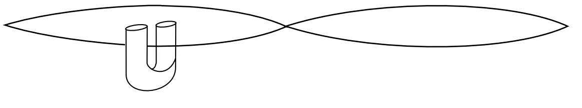


Fig. 3.8(b) *Second harmonic*

The third harmonic may also be demonstrated by increasing the generator frequency to three times the fundamental frequency and placing the magnet back at the centre near an expected antinode.

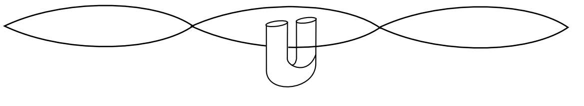


Fig. 3.8(c) *Third harmonic*

Note. The amplitude of vibration is greater at the lower resonant frequencies.

Stationary Waves in a Stretched String (2)

Apparatus

3 m of insulated tinned copper wire (28 s.w.g. 0.376 mm diameter), 2 V a.c. power supply, 12 Ω rheostat (max. current 1.5 – 2 A), strong U-shaped magnet.

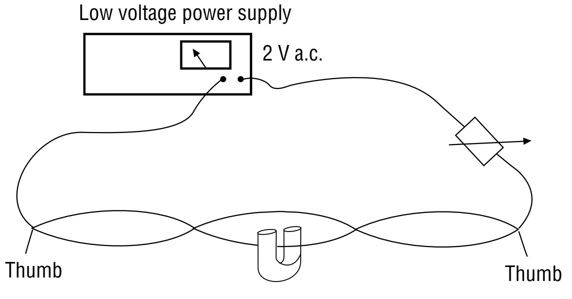


Fig. 3.9

The centre of the wire is placed between the poles of the magnet so that its magnetic field is perpendicular to the wire.

About 1 m of the wire is held stretched between the thumbnails.

At resonance, transverse stationary waves are set up due to the interference between the incident waves and the waves reflected at the fixed ends and the travelling wave emitted by the string is perceived as sound.

Varying the tension in the wire allows stationary waves with one or more loops to be seen. There are always nodes at the fixed ends and an antinode at the centre.

The fundamental is the lowest frequency of vibration of the string with a node at each end and an antinode in the centre, Fig. 3.8(a). It may be part of a more complex mode of vibration and will have the greatest amplitude as well as the lowest frequency.

Harmonics are whole number multiples of the fundamental frequency.

An overtone is a note which has a higher frequency and usually has a lower amplitude than the fundamental. Overtones are the frequencies above the fundamental which are actually obtainable from a particular vibrating system. If a system only vibrates with odd harmonics, for example, then the first overtone is the third harmonic, the second overtone is the fifth harmonic, and so on. Note that some systems, e.g. two-dimensional membranes like drumheads, produce overtones which are not harmonics.

A stretched string produces both odd and even harmonics and can vibrate in several modes simultaneously, which add up to a complicated waveform. Notes from musical instruments are not pure and the component frequencies and their relative intensities depend on the disturbance producing them and the shape of the instrument and its resonant cavities and this determines the quality of the note.

Strings may be plucked, struck or bowed (guitar, piano, violin). Vibrations are transferred to the air inside the body of the instrument or to the sounding board in the case of the piano, which sets a larger mass of air into vibration, otherwise the sound would be very weak.

Resonance occurs when an exact number of half wavelengths fits into the length of the string.

Speed of sound in a stretched string = $\sqrt{\dfrac{T}{\mu}}$

where T is tension and μ is mass per unit length.

Since speed = $f\lambda$,

the fundamental frequency = $\dfrac{1}{2l}\sqrt{\dfrac{T}{\mu}}$.

Waves in Pipes

1. Using increasing levels of water in 8 test tubes and blowing across the top try and make a musical scale by adjusting the water levels (use a pipette).

2. The effect of the length of the pipe on the pitch of the note.

Cover all the holes of a tin whistle. Place the open end into a container of water and blow into it steadily while gradually pulling it out of the water and note the falling pitch as the vibrating air column increases in length. Repeat, but this time push it further into the water as you blow into it and note the rising pitch.

3. Vibrations in a pipe closed at one end.

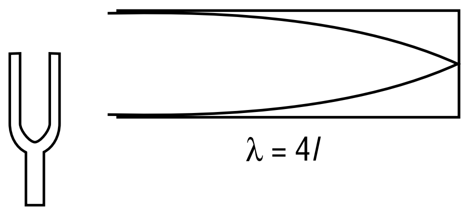


Fig. 3.10

Using a resonance tube, e.g. a combustion tube in a tall graduated cylinder of water forming a pipe closed at the water end, gradually raise the tube out of the water while holding a vibrating tuning fork above the air column. Find the first position of resonance corresponding to the fundamental frequency (note the length) and, raising the tube higher, find the second position of resonance (3 times the first length).

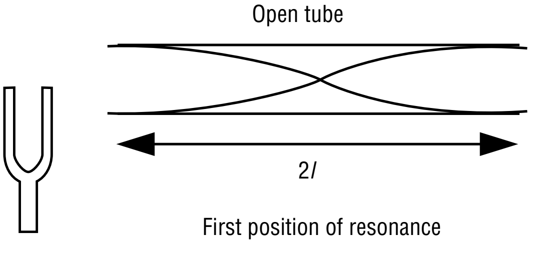


Fig. 3.11

4. Vibrations in a pipe open at both ends.

Use two tubes, one capable of sliding inside the other, and using the same tuning fork show that the length of the tube must be twice as long as for the closed pipe in order to get the first position of resonance. The second position of resonance corresponds to the second harmonic (2 times the first length). This harmonic is not produced by the closed pipe which only produces odd harmonics.

Organ pipes may be either closed (stopped) or open. Open pipes produce a more brilliant tone as they are capable of producing more harmonics than closed pipes.

To Show that Loudness Depends on Amplitude

Apparatus

CRO, signal generator, loudspeaker.

Connect the output of the signal generator to both the CRO and the loudspeaker.

Note the increase in loudness of a particular frequency as the amplitude of the wave is increased. High frequencies can be unpleasant so pick one which is reasonably ear-friendly.

Experiment 3.1

To investigate the relationship

(i) between frequency and tension (length and mass per unit length constant)

(ii) between frequency and length, (tension and mass per unit length constant).

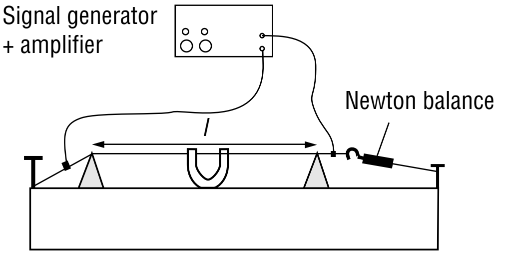


Fig. 3.12

The sonometer is fitted with steel wire which should be attached securely at one end.

Method 1

Apparatus

Signal generator, strong U-shaped magnet, sonometer with a newton balance and tension key or with a hanger + weights.

(i) Frequency and Tension

The experiment may be carried out using a signal generator with an amplifier, and strong U-shaped magnet as described on *page 48.* This method works very well and is visually and aurally convincing. Use the output from the amplifier if there is also an unamplified output.

The tension should be kept low so that stationary waves of large amplitude are set up. Always place the magnet at the centre of the wire. Set the frequency to a low value and adjust the tension slowly until resonance occurs. You can hear an increase in loudness approaching resonance and then a very slight adjustment of tension if using a tension key will achieve the maximum response from the wire. If using weights over a pulley it might be more convenient to vary frequency provided it can be read to the nearest Hz.

Note. Use a multimeter to measure frequency as the signal generator will not give an accurate reading of frequency.

If the wire is getting warm turn down the amplitude.

Wear goggles in case the wire snaps, and don't lean over it.

Example

Distance between the bridges – 0.5 m, steel wire 24 s.w.g., 0.56 mm diameter.

f/Hz	*T*/N
40	3
80	12
etc.	

When tension is increased by a factor of 4 (from 3 N to 12 N) frequency increases by a factor of $\sqrt{4}$, i.e. from 40 Hz to 80 Hz.

Plot a graph of frequency versus square root of tension and draw the best straight line through the points. It should be a straight line through the origin, showing that frequency is proportional to the square root of the tension.

(ii) Frequency and Length

The same method may be applied to investigate the relationship between frequency and length. Keeping tension and mass/unit length constant, start with the longest possible length between the bridges to minimise percentage errors in length measurements. The magnet should always be at the centre of the wire, i.e. at the antinode of the fundamental mode of vibration. Vary the frequency until resonance occurs.

Repeat the procedure for different lengths of wire and plot a graph of frequency versus inverse of length which should be a straight line through the origin. Hence frequency is inversely proportional to length.

Example

Steel wire, 24 s.w.g., tension = 4 N.

f/Hz	*l*/m
40	0.6
60	0.4
etc.	

The product of frequency and length is constant.

Method 2

Apparatus

Sonometer, a selection of tuning forks of varying frequency, small piece of paper, block of wood.

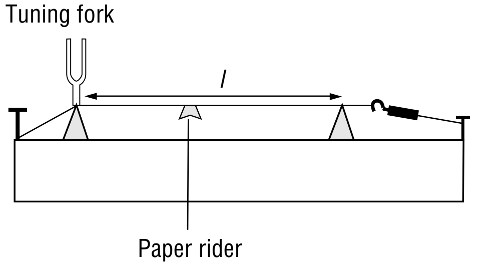


Fig. 3.13

In this method the wire is forced to vibrate by placing the base of a vibrating tuning fork on the wire on one of the bridges and adjusting either tension (length and mass/unit length constant) or length (tension and mass per unit length constant) until resonance occurs. Detection of resonance is aided by placing a small piece of paper on the centre of the wire. The paper is seen to vibrate at resonance. It is a more tedious method than the first one but the equipment may be more readily available. Small blocks of wood on which to strike the tuning forks should be provided and they should not be struck with excessive force.

Example (i)

Steel wire 26 s.w.g., length = 0.17 m.

f/Hz	T/N
256	10
512	40
etc.	

Example (ii)

Steel wire 26 s.w.g., tension = 35 N.

f/Hz	l/m
256	0.32
512	0.16
etc.	

3.4 Applications

A swing pushed at its natural frequency

A child on a swing left to swing freely will always make the same number of oscillations in a given time.

The system is vibrating at its natural frequency which is the frequency at which it will vibrate freely after being given a single push.

Due to friction the amplitude of the vibrations decreases unless energy is given to the system by pushing.

Small pushes given at exactly the right time, i.e. at the swing's natural frequency, will cause large amplitude vibrations, i.e. resonance occurs.

Rattling Windows

Loose window panes vibrate at their own natural frequency.

Passing vehicles producing vibrations containing this frequency cause the windows to resonate.

Passing vehicles can cause walls, floors and whole buildings to vibrate if they produce the natural frequency of vibration of these objects.

The windows and doors of a bus may also rattle as its engine speeds up to certain frequencies corresponding to their natural frequency of vibration.

Tacoma Narrows Bridge Collapse – Washington 1940

The wind produced an oscillating resultant force in resonance with a frequency equal to a natural frequency of the bridge, causing it to vibrate with such large amplitude oscillations that it completely collapsed. *(Amateur video of incident available.)*

Cooling towers and factory chimneys can also be set oscillating by the wind at their natural frequencies.

Bouncing a Car to Move It

If people lift a car at the natural frequency of oscillation of the springs a large amplitude bounce can be built up as a result of resonance. When bouncing, it can be moved to the side more easily due to reduced friction between the road and the wheels.

Vibrations in Washing Machines

Washing machines may vibrate violently at particular speeds because resonance occurs when the frequency of the rotating drum equals a natural frequency of the body of the machine. There are usually several natural frequencies at which resonance can occur. Some machines have a block of concrete attached to the outer drum and the consequent increase in mass reduces the amplitude of any vibrations that might build up during washing or spinning.

Diving Board Oscillations

Jumping up and down on a diving board at its natural frequency of oscillation causes large amplitude oscillations to build up as a result of resonance.

Electrical Resonance – Tuning a Radio or Television

Tuning a radio or television involves selecting a particular transmission frequency from all the incoming frequencies received by the aerial by matching the natural frequency of the detecting circuit in the radio or TV with the desired incoming frequency.

This can be achieved by varying the capacitance of the detecting circuit. It will only produce a large oscillating voltage, i.e. resonance occurs, for the incoming frequency equal to its own natural frequency. The circuit will produce only small oscillating voltages for other frequencies.

Resonance in Musical Instruments

Musical instruments would be barely audible without resonance.

The air in a pipe or tube vibrates with a large amplitude only if caused to vibrate at one of its natural frequencies. The same applies to stringed and percussion instruments.

Stringed Instruments

All stringed instruments are periodically tuned to correct pitch by adjusting tension. The piano and harp contain a large number of stretched strings of different length, tension, and mass per unit length.

The violin has only four strings, whose length and hence frequency can be altered by fingering. Short thin strings under tension give a high pitch when plucked or bowed. The fundamental pitch of the speaking voice is controlled by muscular tension in the vocal cords.

Automatic Focusing of Cameras Using Ultrasonic Pulses

A miniature loudspeaker which also acts as a microphone emits ultrasonic pulses of one millisecond duration. The signal is reflected from the subject and, based on the time taken for it to return, the focus is automatically adjusted. If the time is short, for instance, the focus is adjusted for a close-up shot. Some cameras use infrared beams instead of ultrasound.

Ultrasound Spectacles for Visually Impaired People

In the early seventies a hand-held sonar blind guidance device was developed. An ultrasonic beam is emitted by the device which is reflected on encountering an obstacle. The reflected beam is picked up by a headset on spectacles worn by the visually impaired person and is indicated by a tone. The pitch of the tone indicates how far away the obstacle is. The transmitter and receiver can both be incorporated into the spectacles.

Ultrasound Imaging in Monitoring Unborn Babies and Detecting Tumours

The sound waves travel easily through the baby and are partially reflected by parts of it where a boundary of different density or structure is encountered. Different types of tissue have different reflective properties for ultrasound. When the reflection is from a moving part, for example blood, the reflection is Doppler shifted and, with the aid of a computer, images are built up. Hence a picture may be assembled, even of the beating of the baby's heart.

It is safer than X-rays which cause ionisation and can thus damage cells.

It allows continuous monitoring of, for instance the heart, or a baby, without any risk.

It can measure the depth of an object below the surface of the body from the time it takes for the echo to return. No indication of depth is given from an X-ray photograph.

It is sensitive to differences between soft tissues in the body which X-rays do not show up and can therefore help to detect the presence of tumours.

Killing Bacteria in Liquids

Ultrasound of high intensity can cause the destruction of living cells suspended in liquids. Ultrasound can reduce the bacteria count in milk to 8 per cm^3 – pasteurised milk has 30 000 per cm^3.

Detecting Faults in Metal Castings

The 3 pulses on the screen, Fig. 3.14, indicate the transmitted pulse X, the pulse reflected from the flaw Y, and the pulse reflected from the end of the piece of material, Z. Any flaw in the internal structure is indicated by a change in the echo pattern displayed on a screen.

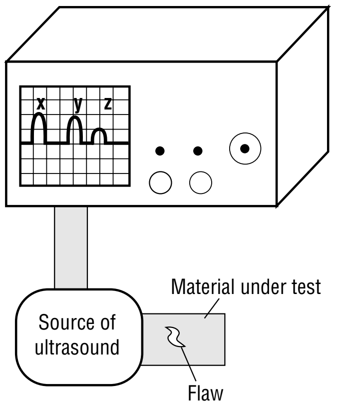


Fig. 3.14

Cleaning Teeth in Dentistry

Small particles are vibrated against the teeth at frequencies of 20–30 kHz.

Cleaning Medical Instruments

They can be thoroughly cleaned by immersion in a liquid irradiated by an ultrasonic beam of frequency of 20 kHz–1 MHz.

Ultrasound 'Bleepers' Used to Deter Household Vermin

High frequency sounds, inaudible to humans, are emitted which the mice and other undesirables find so unpleasant that they just move out. However the frequency should be altered regularly or they may become immune to it.

Dog whistles also utilise ultrasound frequencies inaudible to humans.

Ultrasound Intrusion Detector

A transducer sets up a pattern of ultrasonic waves in a room or in a car, producing a certain amplitude at a detector. An intruder will alter this pattern and hence alter the amplitude at the detector, thereby setting off the alarm.

Homogenising Milk

Ultrasound can break down fat globules into smaller sizes.

Opening Garage Doors

An ultrasound whistle on the car activates a microphone. Amplified current from the microphone opens the doors.

3.5 Worked Examples

Example 1

To combine two sound levels in decibels, e.g. normal conversation (level 60 dB) and a radio playing loudly (level 90 dB) it is necessary to convert to intensities.

Conversation

$$60\text{ dB} = 6\text{ B}$$

$$6 = \log \frac{I_c}{I_0}$$

$$10^6 = \frac{I_c}{I_0}$$

$$I_c = 10^6 \times I_0$$

$$= 10^6 \times 10^{-12}$$

$$= 10^{-6}\text{ W m}^{-2}$$

Radio

$$90\text{ dB} = 9\text{ B}$$

$$9 = \log_{10} \frac{I_R}{I_0}$$

$$10^9 = \frac{I_R}{I_0}$$

$$I_R = 10^9 \times I_0$$

$$= 10^9 \times 10^{-12}$$

$$= 10^{-3}\text{ W m}^{-2}$$

Total Intensity

$$I_T = I_R + I_c$$

$$= 10^{-6} + 10^{-3}$$

$$= 1.001 \times 10^{-3}\text{ W m}^{-2}$$

Intensity Level

$$= 10 \log_{10} \frac{1.001 \times 10^{-3}}{10^{-12}}$$

$$= 10 \log_{10}(1.001 \text{ x } 10^9)$$

$$= 10 \log_{10}(1.001) + 90$$

$$= 0.004 + 90$$

$$= 90.004\text{ dB}$$

This is approximately 90 dB showing that the conversation is almost completely drowned out by the radio. However, it may still be heard due to variations in its intensity level.

Example 2

Two loudspeakers each produce sound at a certain point of intensity level 80 dB. What is the intensity level at the point if both speakers are producing sound simultaneously?

$$80\text{ dB} = 8\text{ B}$$

$$= \log_{10} \frac{I_s}{I_0}$$

$$10^8 = \frac{I_s}{I_0}$$

$$I_s = 10^8 \times I_0$$

Total Intensity

$$I_T = (10^8 \text{ x } I_0) + (10^8 \text{ x } I_0)$$

$$= 2 \times 10^8 \times I_0$$

Intensity Level

$$= 10 \log_{10} \frac{2 \times 10^8 \times I_0}{I_0}$$

$$= 10 \log_{10} 2 + \log_{10} 10^8$$

$$= 10(0.3 + 8)$$

$$= 83\text{ dB.}$$

CHAPTER 4 REFERENCE MATERIAL

4.1 Acoustics

Acoustics may be defined as the science of sound. A fundamental aspect of acoustics is the reflection and absorption of sound.

The shape of a room and the type of surfaces determine its acoustics, i.e. its sound reflecting and absorbing properties.

In large cathedrals or concert halls a sound may persist for a long time after it has been produced, due to multiple reflections from the walls, roof, floor and other objects present. This persistence of the sound after the source has ceased to produce it, with many echoes blending into a prolonged sound, is known as **reverberation**. At each reflection some sound is also absorbed, causing each successive echo to become weaker.

In St Paul's Cathedral in London the organ sound persists for about six seconds after the organist has ceased playing. In the same cathedral is the famous whispering gallery which is circular in shape. A person whispering quietly at one side can be heard clearly at the other side due to repeated reflection of the sound around the chamber.

If reverberation time is too long sound from one syllable can overlap that from the next syllable, causing speech to become confused and impossible to understand. If on the other hand it is too short, sounds are too weak and a room sounds 'lifeless', especially for music.

The Royal Albert Hall in London was designed with very little thought given to what would be heard or seen by an audience. Not only did large pillars obstruct the view but also the large curved surfaces, especially the dome, tended to focus echoes at certain places. It had to have a curtain across it at orchestral concerts. As sound could reach a person's ear directly and after multiple reflections some notes might be heard twice if the time between them was long enough to distinguish them.

Constructive or destructive interference could also occur between sounds received directly and after reflection thus cancelling out some sounds altogether and amplifying others.

The technical problems involved were first investigated only in 1906 by Sabine in America. He measured the reverberation time, which is the time for a sound of a specific intensity to become just inaudible (or to fall by 60 dB, i.e. to one millionth of its original value), in empty halls and churches, using a small organ pipe and then investigated the effects of cushions on seats and various wall and floor coverings to see which reverberation time was best. Soft materials move slightly when sound waves push or impinge on them and hence absorb some of the sound's energy.

He found that reverberation time T depended on

(i) the volume of the room V

(ii) its surface area A and

(iii) the absorptive power of the surfaces in the room, a, according to the formula

$$T = kV/Aa.$$

This formula was found very useful when designing new halls and research continues on the investigations begun by Sabine.

The ideal reverberation time for a room varies for speech, orchestras, and choirs. Since a lot of halls must be multipurpose, compromises must be made and the reverberation time is usually between one and two seconds. Broadcasting studios used for speech only, have zero reverberation time. Acoustically designed lecture rooms and concert halls like the Royal Festival Hall in London have plush cushions on the seats with similar sound absorptive properties to clothing so that the reverberation time is the same whether or not the hall is full. Walls, floors, and ceilings are covered with special sound absorbent materials and the walls are made up of many small flat surfaces inclined at various angles to distribute the sound evenly. The ceiling in the Point Depot in Dublin has soft black cloth draped across it to avoid unwanted reflections. Using present day sound-absorbent materials and changing seating arrangements to avoid spots with acoustic problems places like the Royal Albert Hall have been made more acoustically pleasing.

Sound insulation is especially important for people living in blocks of flats where the spaces between different floors are filled with materials which, being inelastic, tend to absorb rather than transmit sound.

In recording studios sound engineers vary the inclination and position of reflecting surfaces around playing instruments to change the sound quality.

4.2 The Ear

Characteristics of the Human Ear

1. The ear can distinguish frequencies which differ by less than 1% and can recognise very precisely different frequency ratios, as in musical scales.
2. The ear can analyse sounds into their component frequencies.
3. Hair cells in the inner ear have different lengths – hence they respond to different frequencies.
4. The frequency range of the ear varies by a factor of 1000 (20 Hz to 20 kHz) whereas the frequency range of the eye varies by a factor of 2. The loudness range varies by a factor of more than 10^{12} which is 1000 times more than the range of intensities which can be seen by the eye.
5. The ear filters out sounds we don't want to hear, concentrating on, for example, one instrument in an orchestra or one person's voice in a noisy room.
6. The ear is extremely sensitive to pressure variations and can detect pressure changes as low as 2×10^{-5} Pa.
7. The combination of the ear and the brain allows us to memorise the quality of sounds so that we can recognise different voices and musical instruments from their particular combination of harmonics. Computers can still not match these skills in voice recognition. (See Voice Prints.)

Essentially the ear converts weak mechanical vibrations of the air into electrical signals that are sent to the brain.

Having two ears enables us to judge the direction a sound is coming from, i.e. the binaural effect.

Hearing Loss

Hearing loss can be due to a number of factors, including the following.

1. Ageing.
2. Malfunctioning in the processing of information by the brain.
3. A broken eardrum – caused by a blow to the ear or very loud noises too close to the ear, e.g. fireworks going off very close to the ear.
4. A blow on the head can damage the inner ear.
5. Exposure to very loud noises can also cause damage to the individual hair cells in the inner ear. (See also Noise Pollution.) The louder the sound and the longer the exposure the greater the damage. Rock band players can have their upper frequency limit reduced to 10 kHz.
6. Young children can suffer temporary deafness due to a cold-type infection which causes the middle ear to become filled with fluid. Grommets are inserted into the ear to drain this fluid.

   Infections of the middle ear can, if left untreated, cause damage to the bones of the middle ear. Beethoven's deafness was due to osteosclerosis, i.e. bone hardening, of the stirrup bone in each of his middle ears. Since this bone was unable to move it could not transfer the vibrations of sound to the inner ear. Nowadays this bone can be replaced by an artificial one if it becomes diseased.

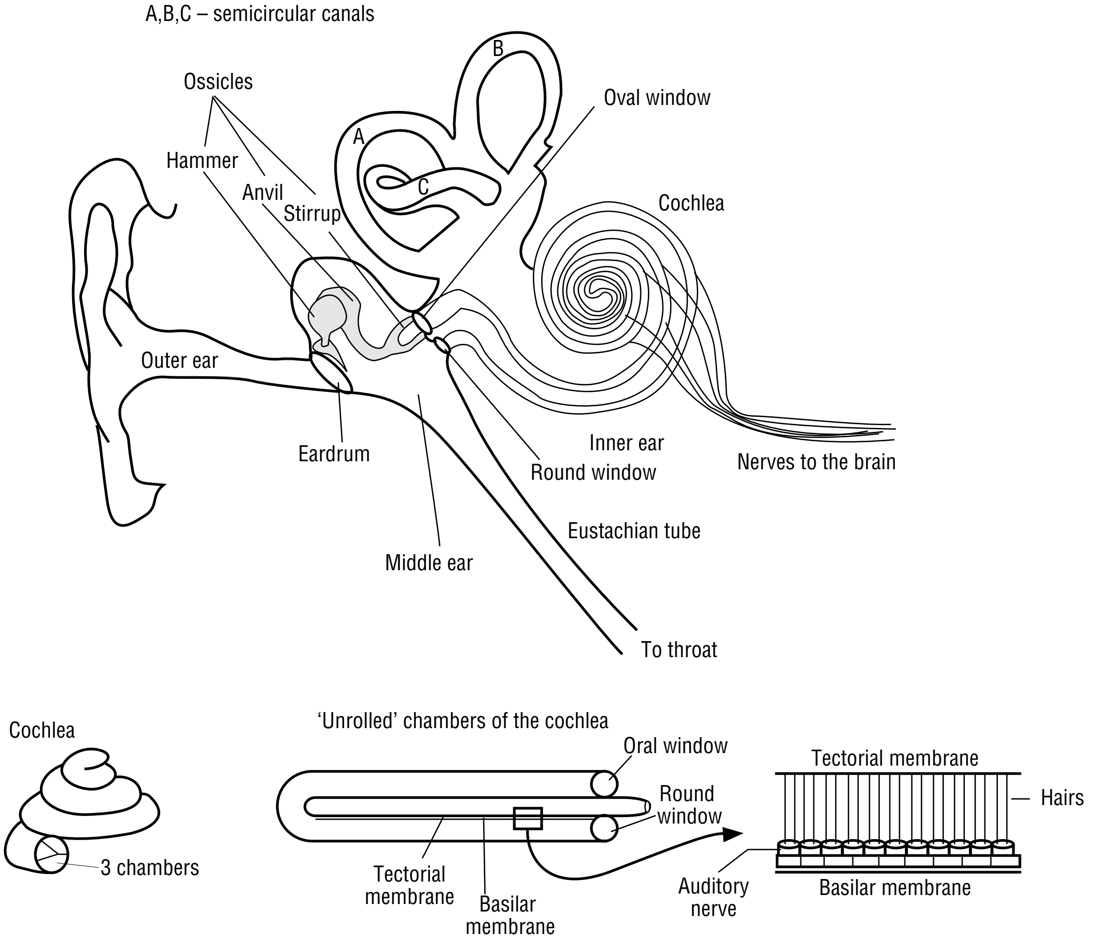


Fig. 4.1 *How we hear*

7. Nerve loss results in the cochlea not passing impulses to the brain. There is no cure at present for this type of hearing loss.

Tinnitus is a hearing defect where the sufferer is subjected to a constant ringing in the ears. Permanent tinnitus may be produced by a few hours' exposure to high noise levels, especially if the exposure is regularly repeated.

Hearing aids are miniature electrical amplifiers which are fitted inside or behind the ear to make sounds much louder before they pass into the ear canal.

Newly developed digital hearing aids can now enable people who have had total deafness in the inner ear to hear once again. A programmable speech processor about the size of a personal stereo analyses sounds and changes them into digital code which travels along a wire to a radio transmitter in an ear plug. A chip implanted behind the ear picks up the radio signal, decodes it and transmits it through a wire, through bone, to tiny silicon-platinum electrodes in the cochlea. Nerve cells in the cochlea are thus electrically stimulated and send messages to the brain which processes the information, thus enabling the person to hear.

Structure of the Ear

The ear consists of three sections, the outer ear, the middle ear, and the inner ear, separated from each other by skin membranes or 'windows' which transmit the sound from one section to the next.

The outer ear acts as a funnel collecting and directing the longitudinal sound waves to the first skin membrane, the eardrum (tympanic membrane), which is made to vibrate by the compressions and rarefactions of the sound wave. The amplitude of vibration may be as low as 0.1 mm for ordinary speech.

The vibrating eardrum causes the linked bones (ossicles) in the air-filled middle ear (the hammer, the anvil, and the stirrup) to vibrate. These bones act as a lever which means that the force of the vibrations is amplified (by more than 20 times) when transferred through these tiny bones to the oval window.

The oval window has a much smaller area than the eardrum (about 1:15) and since the force on it is distributed over a smaller area this increases the pressure and so increases the amplitude of vibration.

When sound passes from one medium to another some of the sound is reflected depending on a property called acoustic impedance. If the impedance of the two media is similar this avoids loss of sound energy by reflection. The design of the middle ear is such that there is good acoustic impedance matching when the sound passes from the air-filled middle ear to the fluid-filled inner ear.

The Eustachian tube connects the middle ear to the mouth and it allows pressure on both sides of the eardrum to be kept equal. Differences in pressure can cause pain and impair our hearing. This pain is experienced when one is subjected to a change in air pressure at take-off or before a plane lands or in hilly country. To obtain relief and equalise pressure in an aircraft the tube which is usually closed can be opened by yawning or chewing or swallowing. When you have a cold your hearing is impaired because fluid fills the Eustachian tube which means that the pressure across the eardrum cannot be equalised.

The vibrations of the oval window are transmitted to the fluid-filled inner ear, a cavity deep within the skull. It contains two organs, the semicircular canals, whose function has to do with balance and not hearing, and the cochlea which is shaped like a curled up horn, about the size of the little finger nail, and contains thousands of tiny hairs. It is here that the mechanical vibrations of the sound waves are converted to electrical signals (as in a microphone). The sound wave causing the fluid of the inner ear to vibrate may be made up of several different frequencies. As the pressure wave travels from the oval to the round window the basilar membrane is set vibrating and this vibration is transmitted to the hair cells which brush against the tectorial membrane, but particular hairs only vibrate if forced to vibrate at their own natural frequencies which is determined by their length. (See Resonance.) When a hair vibrates the nerve attached to it ‘fires’. A signal is thus sent to the brain which processes the information and so we hear the sound. The brain can recognise the frequency of the sound by the particular nerves which transmit the signal.

4.3 Graphic Equalisers

In the recording studio recordings may be made with emphasis at different parts of the frequency spectrum, with low frequencies perhaps attenuated and high frequencies strengthened. On playback when the final ‘mix’ is being produced this can be compensated for by using an equaliser which can alter the sound of any instrument by boosting or attenuating selected frequency bands coming from the tape. It requires considerable skill to know which frequencies to boost, and by how much, to make the output of a recording the same as the original or to get the desired effect.

Graphic equalisers may be added on to a domestic hi-fi system for increased control for listening pleasure.

4.4 Musical Scales and Standard Concert Pitch

Two different pianos might be in tune in so far as the correct frequency ratios might exist between the strings in each instrument but they might not be in tune with each other. This was the case especially with church organs in the eighteenth century where organs rarely had the same frequency for their middle C pipes. With the development of orchestral music it became more important to standardise musical concert pitch. It was only as recently as 1939 that standardisation, based on a frequency of 440 Hz for A, was agreed on by an international committee.

Musical scales evolved over the centuries and were different in different parts of the world, both in the number of intervals in the scale and the pitch of the notes. The Chinese use totally different sets of musical intervals to Europeans.

African music has much smaller intervals, with two or four times as many notes in their scales as in European scales.

Traditionally, European music has developed from a scale of eight notes based on natural ratios of frequencies where the eighth note, the octave, has twice the frequency of the first. This is called the diatonic scale. To make it possible to play the same melodies in other pitches or keys, it was found necessary at an early stage to add five other notes which, on the usual modern keyboard, are black. However, it then became clear that, owing to approximations, some keys were less perfectly in tune than others, and in the 18th century the frequencies were adjusted, or 'tempered', to make each scale sound equally good or bad. On this scale of 12 notes (each a semitone apart) the frequency ratio between any note and the previous one is $2^{1/12}$, i.e. 1.06.

Tuning forks found in laboratories are usually in scientific pitch which is not the same as the pitch of corresponding notes on an equally tempered scale and are thus unsuitable for tuning musical instruments.

Before the 17th century up to 12 different mode permutations of the diatonic scale were in use but for the last three centuries only two modes, i.e. major and minor have been in general use.

Major and minor scales have different sequences of tones (T) and semitones (S). In major scales the sequence is T-T-S-T-T-T-S, whereas in a minor scale the sequence is T-S-T-T-S-T-T.

In practice this form of the minor scale is rarely used in composition and it is usually altered in one of two ways to create a greater emphasis on particular pitches. In the harmonic minor scale the seventh note is raised a semitone and in the melodic minor the sixth and seventh notes are raised by one semitone in the ascending pattern and left unaltered in the descending pattern. Composers are constantly experimenting and the nature of scales can change radically over a number of centuries.

Claude Debussy and others in France and England used a whole-tone scale consisting of six whole tones. Arnold Schoenberg, an Austrian composer of the 20th century, used all 12 semitones within the octave, i.e. the chromatic scale. Scales using intervals smaller than a semitone, i.e. microtonal scales, have been used in contemporary music and special instruments can be used to produce the required pitches. We must remember that our scales are only conventions adopted for pitch tuning and that singers, trombone and stringed instrument players can easily alter the pitches they produce which may not necessarily match the pitches of standard scales.

Generally, the simplest scales are found in old music and the music of non-literate cultures whose creators were not even aware of the theoretical concept of scales, while more complex scales occur in more developed cultures.

4.5 Sound Transducers

Devices which convert sound waves into other forms of oscillation, or vice versa, are sound transducers.

Microphone

A microphone converts sound waves into electrical signals.

Loudspeaker

A loudspeaker converts electrical signals into sound waves.

Recording Head of a Tape Recorder

The recording head converts an amplified electrical signal from a microphone into a varying magnetic field. The current from the microphone is fed into the coil of the electromagnet which magnetises the tape moving past it; a series of small permanent magnets is formed in the oxide layer of the tape. On playback the varying magnetisation of the tape moving past the recording/playback head induces small varying currents in the coil. This electrical signal is then amplified and converted back to sound in the loudspeaker.

Note. If tape speed varies significantly the frequency of sound emitted on playback will differ noticeably from the recorded frequency giving rise to 'wow' and 'flutter'. 'Wow' is a slow variation in the music's pitch. 'Flutter' is a faster variation in pitch than 'wow' but with similar causes.

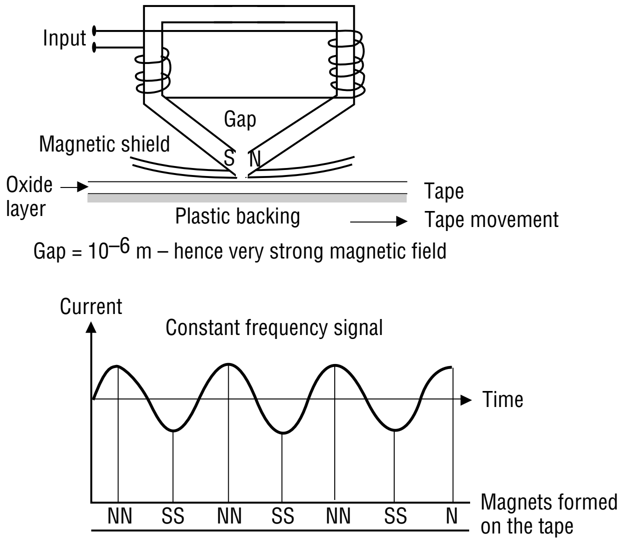


Fig. 4.2 *Recording and playback head of a tape recorder*

Stylus and Cartridge of a Record Player

Vibrations due to the stylus following the curvy nature of the grooves on the record cause magnets attached to the stylus to move inside the cartridge coils, inducing a small varying voltage in these coils (only a few millivolts). This voltage is amplified and produces a varying electric current in a loudspeaker where the signal is converted back to sound. Unfortunately anything in the way in the groove makes the stylus move and so scratches, fingermarks and dirt cause clicking noises to be produced.

On old gramophones the stylus or needle was attached to the end of a large horn which made the tiny vibrations in the tip of the needle loud enough to hear.

Stereo sound was introduced in the fifties. During recording different patterns are cut into either side of a groove. The left and right movements of the stylus occur simultaneously so that each channel produces different signals all the time so that instruments for instance can be heard spatially separated from each other. However it is not possible for this analogue groove on a record to truly represent all the sounds, say from an orchestra.

Analogues are the various alternative forms in which sound waves exist with variations of amplitude and frequency exactly matching the original sound wave. Examples include the following.

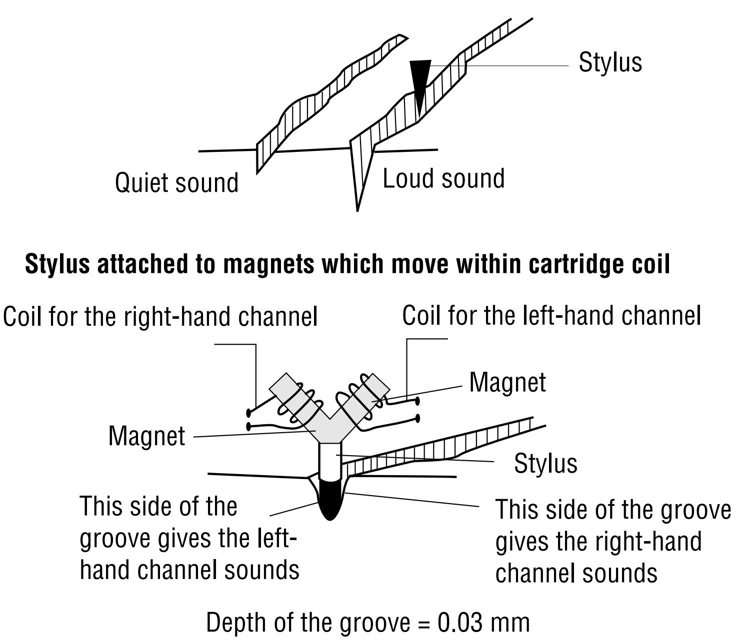


Fig. 4.3 *Grooves in a record*

1. The bumps in the grooves of a record.
2. The current flowing in a microphone.
3. The variation in the magnetisation of a tape.
4. The current flowing in a loudspeaker.
5. ‘Flats’ and ‘pits’ of equal width but different length in the surface of compact discs (originally developed by Phillips and Sony in 1982) represent the 0s and 1s of binary code into which the original signal has been converted after its voltage levels have been sampled over 40,000 times per second.

   The pits are burned into the disc by a laser.

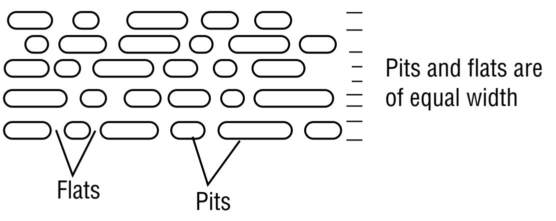


Fig. 4.4 *Surface of the CD*

CD Player

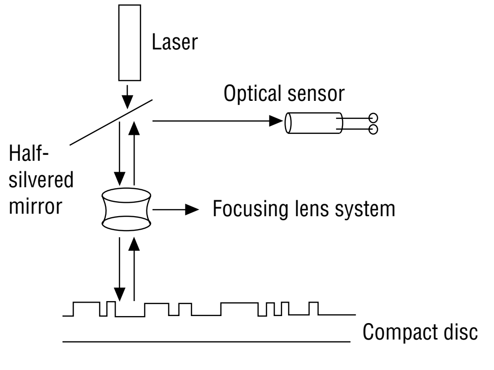


Fig. 4.5 *CD Player*

A very narrow beam of light from a low intensity laser passes through a half-silvered mirror and is focused onto the surface of the compact disc. Since a ‘track’ on a CD consists of a series of bumps which are called ‘pits’ and flats called ‘lands’ the light from the laser is either reflected straight back from the ‘lands’ or is not reflected straight back from the ‘pits’. No light represents a ‘0’ and the presence of light represents a ‘1’. Thus the pattern of ‘pits’ and ‘lands’ is converted into a series of 0s and 1s, the binary code for the original sound. These digital pulses are then decoded and converted back into sound.

The CD player has a number of advantages compared with the record player.

The laser performs the function of the stylus in the record player without damaging the disc.

The sound is reproduced with less than 1% of the distortion produced by an ordinary cassette player.

Since it is a digital system, encoding sounds as a series of ‘0’ or ‘1’ states, it will reproduce the sounds exactly as long as these two states are distinguishable.

‘Hiss’, i.e. random noise, is eliminated.

More music can be stored on a CD than on a vinyl record of the same size.

Since the CD player has a computer controlling it and a quartz crystal clock keeping the whole system accurately timed the motor does not vary in speed and the distortion which is due to variation in motor speed (‘wow’ and ‘flutter’) is thus removed.

The pickup of a record player can pick up all sorts of vibrations but since the pickup of a CD player is a light beam which is simply on or off it cannot pick up these vibrations. Other forms of laser disc are video-disc, which is larger in diameter than a CD and can be used to view motion pictures, and the CD-ROM which looks the same as a CD but contains computer programs and data bases and can store over 100 000 pages of information.

An erasable CD is being developed to compete with Digital Audio Tape (which became available in the late 80s). A laser is used to read or write onto a magnetic coating and the signal is encoded onto the disc by varying the magnetisation. The tape can be demagnetised using the laser also, thus erasing the original data and allowing the disc to be reused.

These tapes, which are only about the size of a credit card, produce the same high quality sound as compact discs but can be erased and used again. They record 14 μm tracks diagonally across the 4 mm wide tape. Three of these tracks are less than the width of a human hair. The tape is coated with small crystals of pure iron whose varying magnetisation records the sound in digital form. They can record over 2 hours of sound which is much more than a compact disc. By running the tape in LP mode (i.e. half standard play speed) recording time can be made twice as long, giving 4 hours of high quality digital recording and playback from a 120 minute tape. A new Pioneer DAT player has doubled the sampling rate to 96 kHz, surpassing the CD sampling rate of 44.1 kHz.

4.6 Speed of Sound

Speed of sound, v, in a gas of density ρ, is,

$$v = \sqrt{\frac{\gamma P}{\rho}} \qquad \text{Formula (i)}$$

where P = pressure

and γ = a constant

Therefore if pressure and density are multiplied by the same factor no change will occur in speed. If pressure of a fixed mass of gas at constant temperature is doubled then volume will be halved according to Boyle's Law and density will be doubled therefore speed will be unchanged. The speed of sound is independent of pressure changes at constant temperature.

Therefore even though speed of sound varies between gases of different density it doesn't necessarily change for one particular gas as its density changes.

The Effect of Temperature

The speed of sound in a gas increases with increasing temperature. The speed of sound in air is 331 m s^{-1} at 0 °C and increases by 0.6 m s^{-1} for every 1 degree Celsius rise in temperature. At room temperature its speed is about 340 m s^{-1}. As the temperature rises the gas expands and if the pressure remains constant then density decreases and as it occurs without a corresponding change in pressure it can be deduced from formula (i) above that speed increases.

Speed of sound in a gas is proportional to the square root of the absolute temperature of the gas. To calculate the speed of sound at different temperatures:

$$\frac{v_1}{v_2} = \sqrt{\frac{T_1}{T_2}}$$

The Effect of Humidity on the Speed of Sound

The density of moist air is less than that of dry air as water vapour is less dense than nitrogen and oxygen. Also γ (the ratio of the principal molar heat capacities), the constant in formula (i), is slightly less for moist air than for dry air. However, the density effect predominates and as a result the speed of sound increases with humidity.

CHAPTER 5

MULTIPURPOSE EQUIPMENT

5.1 Ripple Tank

A ripple tank may be used to show properties of water waves which are also properties of waves in general, i.e. reflection, refraction, interference, diffraction.

Guidelines on Setting Up

1. Take care not to scratch the surface of the tank. Components should be stored in a separate tray.
2. The tank must be free from dust and grease. Use a small amount of detergent and a clean cloth to wash it out before use, especially the beaches. Rinse thoroughly to remove all traces of detergent. Components used in the tank must also be cleaned thoroughly.
3. Fill to a depth of 5–10 mm with clean water so that the water reaches a short distance up the sloped beaches, which should be wet. Use a cloth to wet the beaches to reduce unwanted reflections from the sides.
4. The tank may be mounted on adjustable legs and placed on the floor. The tank must be level. Mount the lamp (typically 12 V, 24 W or 36 W) on a support rod at the end of the tank using a bosshead so that its helically coiled filament is parallel to the waves being observed (if they are straight waves) and connect it to the 12 V power supply. Look at the two reflections of the lamp, one from the water surface and one from the glass surface of the tank. The two images should coincide if it is level. Alternatively a spirit level may be used. Adjust the legs or use wedges until it is level.

   Viewing of wave patterns in this case may be by looking along the plane of the water or better still by looking at the shadows of the waves cast onto a sheet of white paper on the floor below the tank.

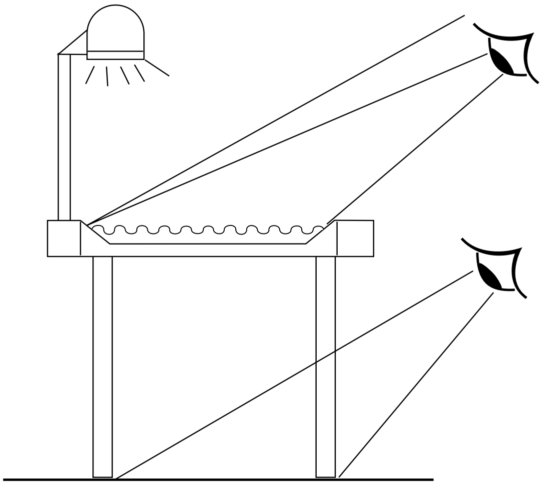


Fig. 5.1 *Viewing wave patterns*

   Alternatively, the lamp may be placed underneath the tank, and the pattern viewed on the ceiling.

5. An alternative and better approach for class demonstration is to mount the ripple tank on the overhead projector, producing the results on a vertical screen. Some models can be used either way while some are too awkward for use on the OHP.

   The water surface should be focused onto the screen. This can be achieved by putting a small plastic ruler in the tank and focusing on its graduations. Slight adjustments may then be made to account for the depth of the water although the water is so shallow that this is hardly necessary.

The waves may be made to appear stationary using the stroboscope. The number of waves occupying the length of the ruler may then be counted on the viewing screen and the wavelength may be calculated.

If using a Xenon tube flashing stroboscope reduce the ambient light. (Set to about 25–30 flashes per second and adjust the speed of the ripple motor until the waves become steady.)

Caution. Flashing lights at certain frequencies and viewing repetitive wave motions could possibly trigger epileptic fits.

(The ripple and support bar may be left off until pulse work has been completed.)

Generation of Waves

Single pulses are short wave trains which are more effective for illustrating reflection than continuous wavetrains where observations are confused by the reflected waves overlapping the incident waves and the resulting interference pattern becoming more dominant than the reflections.

To obtain a circular wave pulse allow a single drop of water to fall on the water surface from a water dropper. A wave in the form of an expanding ring travels to the edges where it is reflected but since the edges are sloped to minimise reflection the reflected wave is much smaller in amplitude.

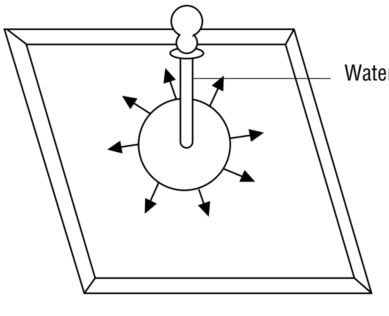


Fig. 5.2 *Circular wave pulse*

Note. When using pulse waves for reflection or refraction it should be emphasised that the wave travels in a direction perpendicular to the wavefront.

To obtain a straight wave pulse place the wooden roller in the tank and rock it by hand gently but quickly, forwards or backwards, being careful not to rock the entire tank.

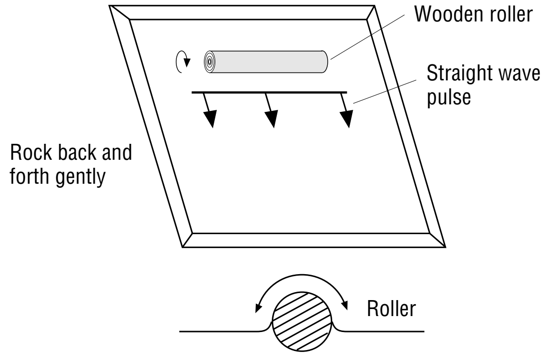


Fig. 5.3 *A straight wave pulse*

Waves

Continuous straight waves may be generated by fixing the ripple bar to the horizontal support suspended by elastic bands, Fig. 5.4. The ripple bar is moved up and down continuously by the vibrations of a small electric motor having an off-centre (eccentric) metal disc attached to its rotating axle. It should just touch the water surface along the whole of its length and its surface should be wetted thoroughly and evenly on all sides. In fact the forces of adhesion should make the water cling to the underside of the bar, Fig. 5.5.

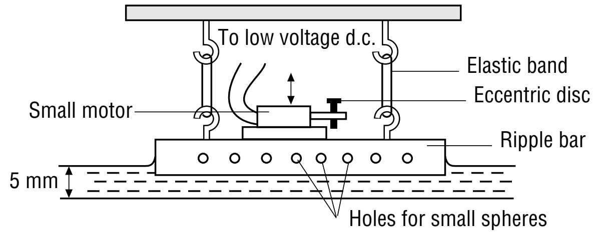


Fig. 5.4 *Ripple bar for generating continuous waves (front view)*

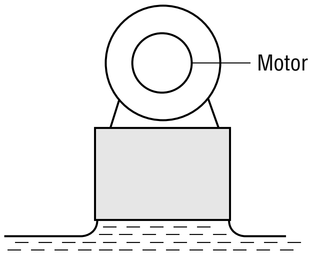


Fig. 5.5 *Forces of adhesion making water cling to the underside of the ripple bar (side view)*

The power pack usually powers both the lamp and the motor with a rheostat incorporated to control the speed of the motor and hence the frequency of the waves. The clarity of the wave pattern is affected by the speed of the motor and is usually best at low speeds.

Continuous circular waves may be generated by raising the ripple bar and fitting a small sphere to the beam so that its widest part is just below the surface of the water. The speed of the motor may be adjusted to vary the wave frequency. One or two spheres may be used (2 for interference) and their separation varied by using the other holes in the beam.

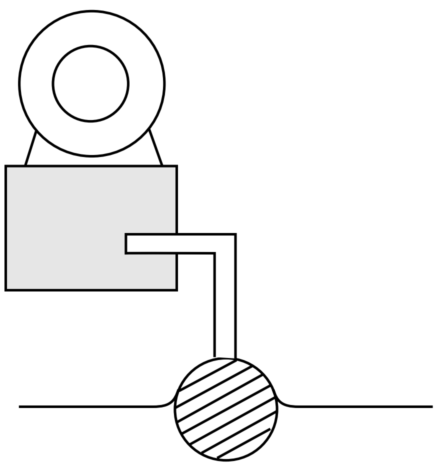


Fig. 5.6 *Ripple bar raised and a sphere fitted to it for the production of continuous circular waves*

5.2 The Cathode Ray Oscilloscope

The oscilloscope displays an input voltage on the Y-axis as a function of time on the X-axis.

The scale on the Y-axis can be altered using the **Volt/div** control while the scale on the X-axis can be altered using the **Time-div** control.

Guidelines on Setting Up

1. Switch on using the **Power/intensity** control set about mid-way. (They may be two separate switches.) The intensity should not be too high as it could burn the screen phosphor.
2. Set **Focus** to about mid-way.
3. If it's a dual-trace CRO then switch the **vertical mode** control to **Channel 1** (or **A**). Other modes provided are **Channel 2 only, Dual,** and **Add,** the latter of which produces a single trace representing the algebraic sum of the signals to the two channels.

   The two traces are produced from a single electron beam either by 'chopping' from one to the other or by 'sweeping' complete traces of one and then the other alternately. Use **Alternate** at high frequencies and **Chop** at low frequency.
4. Set **Sweep Variable** and **Volt/div Variable** to Cal (i.e. if you want to make use of the Time/div and Volts/div calibrations to make measurements).
5. Set the **Time/div** control to 0.5 ms/div.
6. Set **Trigger mode (Coupling)** to **Auto** and **Source** switch to **Channel 1**.
7. Adjust **Channel 1 Vertical Position** control (*y-shift*) and the **Horizontal Position** control (*x-shift*) until a trace is centred on the screen.
8. Adjust **Intensity** to a minimum for comfortable viewing, to protect the screen and to obtain the sharpest trace for accurate measurements. Adjust **Focus** to get the sharpest trace.
9. If the trace is inclined to the horizontal due to stray magnetic fields the **Trace Rotation** control can be used to realign it.
10. The **Timebase** is the sawtooth voltage which causes the spot to travel across the screen from left to right at a steady speed which can be varied by the **Time/div** and **Var** controls.

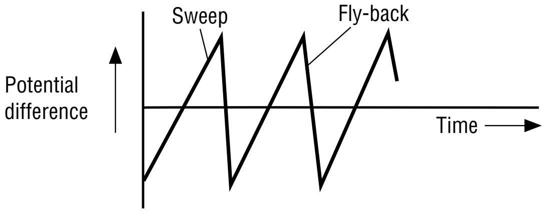


Fig. 5.7 Sawtooth potential difference generated by the time-base

Vary the **Time/div** setting to vary the sweep speed of the spot across the screen. Change the **Time/div** control down to 0.2 s/div – the spot taking 2 s to traverse the screen. The **Variable Sweep speed** control is continuously variable and allows for intermediate values between the **Time/div** settings.

To look at Wave Forms of Signals from a Signal Generator

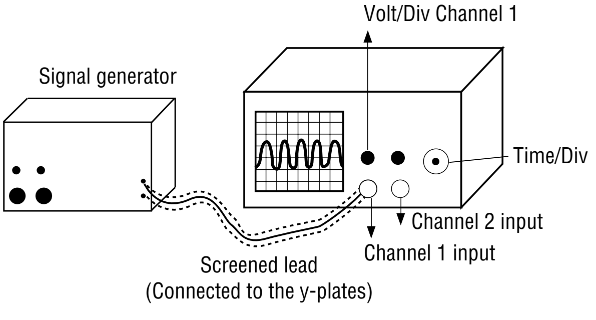


Fig. 5.8 *Displaying waveforms on the CRO*

1. Connect the signal generator, switched to 'sine wave' to the **Y-input, Channel 1**, of the oscilloscope using a coaxial cable. The signal generator delivers a voltage which varies with time.

   (In the coaxial cable the central wire carrying the signal is shielded from stray electric fields by the outer cylindrical sheath of braided wires. The central wire is connected to the red (output) terminal of the signal generator and is electrically insulated from the outer sheath which is connected to the black (earth) terminal of the signal generator and to earth via the oscilloscope.)
2. Set the **AC-GND-DC** controls to **AC.**
3. Set the frequency of the signal generator to 10 Hz, and the amplitude to about 1.5 V.

   Set the oscilloscope **Time/div** to 0.1 s/div and the **Volt/div to** 0.5 volt/div.

   As the spot travels across the screen it is now deflected sinusoidally in the Y direction. Spread it out by decreasing the **Time/div**.
4. Speed it up by changing the signal generator frequency to 1 kHz and the oscilloscope setting to 0.2 ms/div and 1 volt/div. The display may not be stable. Several waveforms may appear on the screen all starting at a different point on the waveform.

   The first diagram below shows 3 different waveforms all starting at different points. The oscilloscope is not **Triggered.**

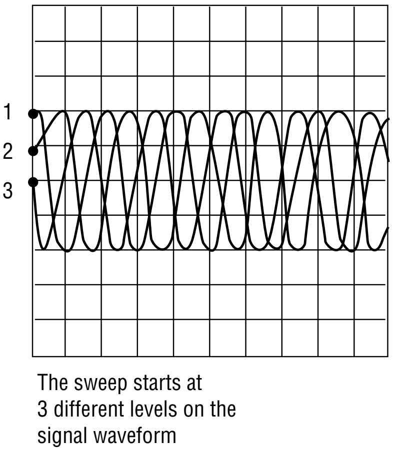


Fig. 5.9 *Un-triggered oscilloscope*

The next diagram shows a stable trace, i.e. each horizontal sweep starts at the same point on the waveform.

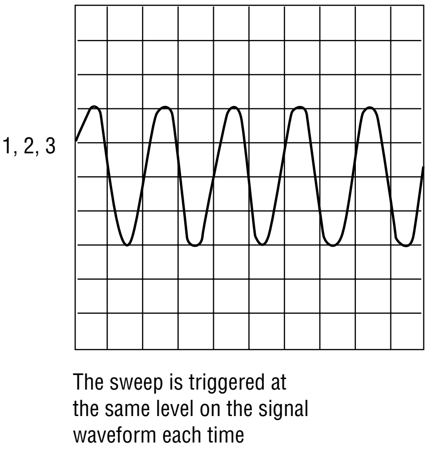


Fig. 5.10 *Triggered oscilloscope*

(**Explanation.** Part of the input signal is fed to a trigger circuit, which gives a pulse which starts the sweep across the screen every time at precisely the same instant in the signal's cycle.

The particular point in the signal's cycle at which the sweep (of the time base sawtooth voltage) across the screen is triggered is determined by the **Trig Level**, i.e. the sweep is begun when the input signal reaches a particular level.

By adjusting the **Trig level** the trace can be stabilised.

The point at which the waveform starts to sweep across the screen can be changed slightly before the trace becomes unstable again.

The timebase may be triggered on the positive-, or negative-going slope of the incoming signal.

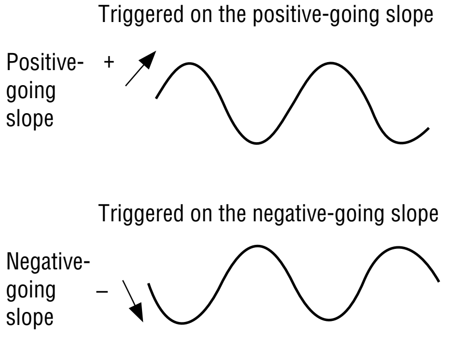


Fig. 5.11

Changing the **Time/div** control and opening up the display does not affect the **Trig Level**.

Changing the **Volt/div** control, to decrease the waveform amplitude will, at a certain point, cause the trace again to become unstable.

5. Vary the amplitude of the waveform using the **Volt/div** control and vary the wavelength using the **Time/div** control. The frequency of the signal may be varied using the frequency control on the signal generator.

The number of divisions occupied by a complete cycle x time per division = Period. Frequency = 1/Period.

AC-GND-DC

With a.c.-coupling (switch in **AC** position) any d.c. or steady component of the signal is blocked by a capacitor. It is useful if one wants to examine an a.c. ripple superimposed on a large d.c. voltage where it would be impossible to increase the Y sensitivity without the signal going off the screen.

It is not necessary to always set the switch to AC when displaying a.c. voltages. The DC setting (direct coupling) is suitable for most applications.

The **GND** setting switches off the input signal and is used to set or check the zero voltage level of each channel.

Trig Source

If it is required to trigger the signal on Channel 1 then part of the input signal from Channel 1 should be fed to the trigger circuit. Therefore the **Trig Source** control is set to Channel 1.

Likewise part of the signal from Channel 2 is used to trigger the signal on Channel 2 and the Source control is then set to Channel 2.

Trig Mode

This is normally set to **Auto** but may also be set to **Norm**.

In **Auto** mode the signal is displayed whether the oscilloscope is triggered or not.

In **Norm** mode the signal is only displayed when the oscilloscope is triggered.

If an untriggered signal is annoying, switching to **Norm** removes it from view instantly.

If no Trace Appears

Check that the X and Y position controls are in their mid-positions.

Set **Trig Mode** to AUTO.

Turn down the Y-sensitivities and adjust the **Trig Level**.

Increase Intensity slightly.

5.3 Microwave Apparatus

This apparatus may be used to demonstrate all of the properties of waves. The transmitter produces electromagnetic radiation of wavelength 2.8 cm. A horn attached to the transmitter reduces diffraction while a similar horn on the receiver improves the collection of the microwaves. In the receiver the microwaves act on an aerial, producing oscillating electric currents. These are then rectified and registered on a microammeter.

Microwaves exhibit reflection, refraction, diffraction and interference. They also exhibit polarisation, showing that they are transverse waves. As their wavelength is about fifty thousand times greater than the wavelength of light the properties of waves are easier to demonstrate with microwaves than with light.

Reflection

Microwaves obey the laws of reflection, Fig. 5.12.

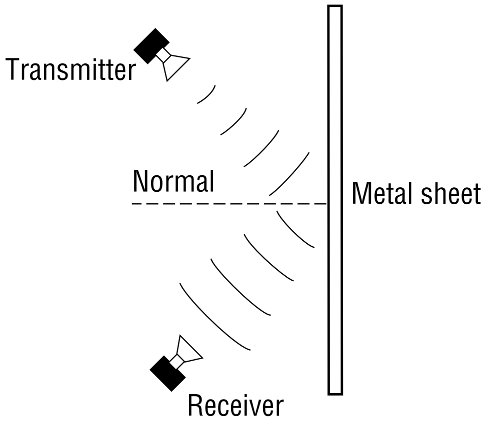


Fig. 5.12 *Reflection at a plane surface*

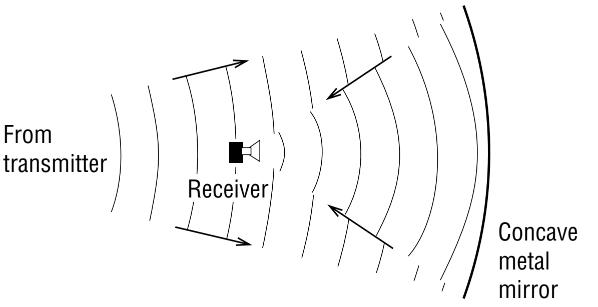


Fig. 5.13 *Focusing by a concave mirror*

Microwaves are focused by a concave surface, Fig. 5.13.

Refraction

Microwaves may be refracted using a hollow perspex prism filled with paraffin, Fig. 5.14.

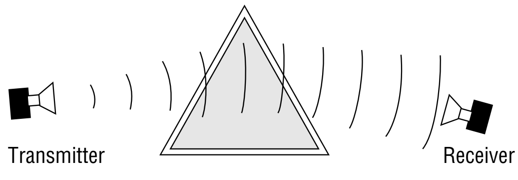


Fig. 5.14 *Refraction by a prism*

A hollow perspex lens filled with paraffin may be used to focus the microwaves in the same way that a glass lens focuses light, Fig. 5.15.

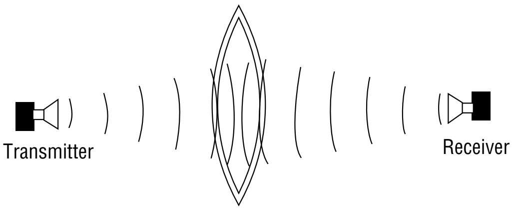
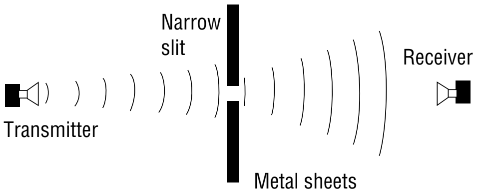


Fig. 5.15 *Focusing by a converging lens*

Diffraction

Two metal sheets may be used to create a slit to demonstrate diffraction. By adjusting the width of the slit the relationship between slit size and the amount of diffraction can be demonstrated, Fig. 5.16.



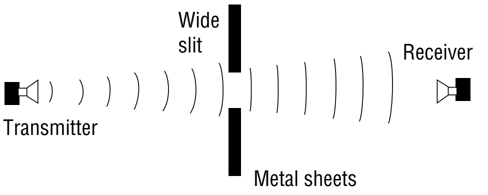


Fig. 5.16 *Diffraction at a slit*

Interference

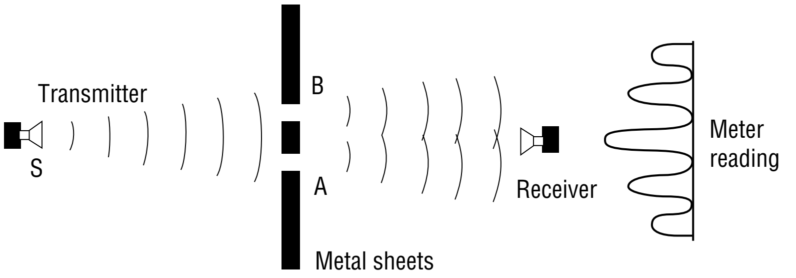


Fig. 5.17 *Interference using two slits*

Interference may be demonstrated by using sheets to make 'Young's Slits', Fig. 5.17.

Polarisation

Polarisation may be demonstrated in two ways.

1. Rotate the receiver relative to the transmitter. When the receiver is on its side, i.e. turned through 90°, the meter should read zero. When the receiver is inverted, i.e. turned through 180°, the meter should return to maximum value.

2. A wire grid may be used as a polarising sheet. The emitted microwaves are already polarised so only one grid is required. Position the grid with its plane parallel to the face of the transmitter so that the meter reading is a maximum. When the grid is rotated through 90° the meter reading should reduce to zero. When the grid is rotated through a further 90°, i.e. a total of 180°, the meter reading should return to a maximum. This shows that the microwaves are polarised.

Note

Every emitter of electromagnetic radiation emits polarised radiation. For example, in the case of light each atom emits polarised light. However, in any given source there are so many atoms, each emitting in its own plane of polarisation, that all possible planes are covered and no particular direction is preferred. Light from an ordinary lamp is therefore unpolarised, Fig. 5.18.

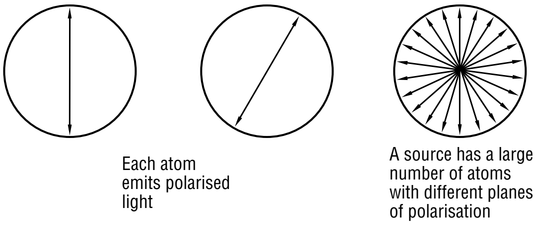


Fig. 5.18 *Unpolarised light*

APPENDIX 1

3cm. Microwave Apparatus

This apparatus may be used to demonstrate all the properties of waves. The transmitter produces electromagnetic radiation of wavelength 3 cm. The horn on the transmitter reduces diffraction just as a simple megaphone can be used to reduce diffraction of the voice.

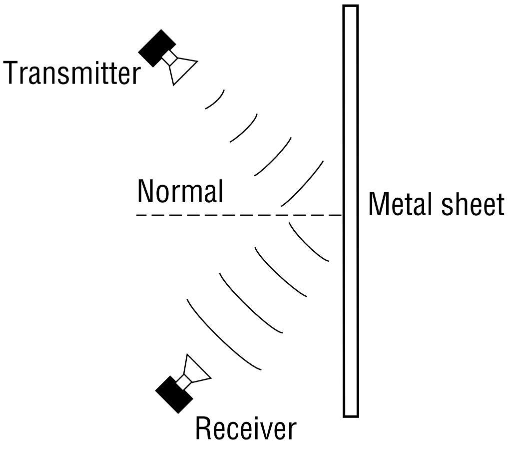
The receiver also has a horn to assist in the collection of the electromagnetic signal. This signal acts on the aerial causing oscillating electric fields which are rectified and register as a current on a microammeter.

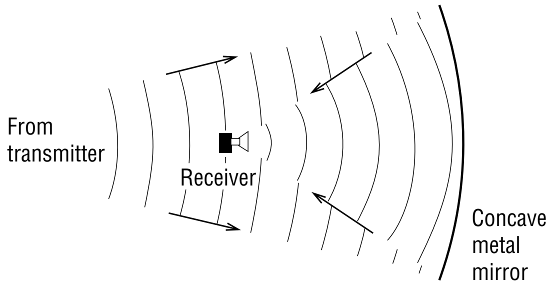
Microwaves obey all the ordinary laws of waves, reflection, refraction, diffraction, interference and polarisation.

As the wavelength is about half a million times greater than the wavelength of light the experiments are much easier to demonstrate using microwaves.

Reflection

Microwaves obey the laws of reflection at a metal surface.

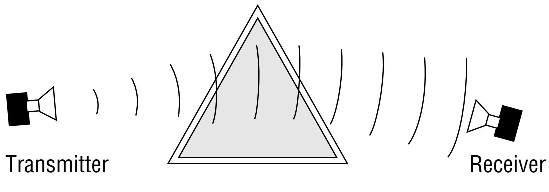




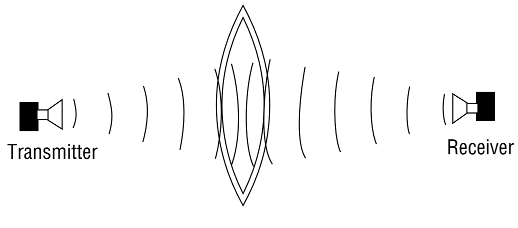
Microwaves are focused by a concave metal mirror.

Refraction

Microwaves are refracted by a hollow perspex prism filled with paraffin.

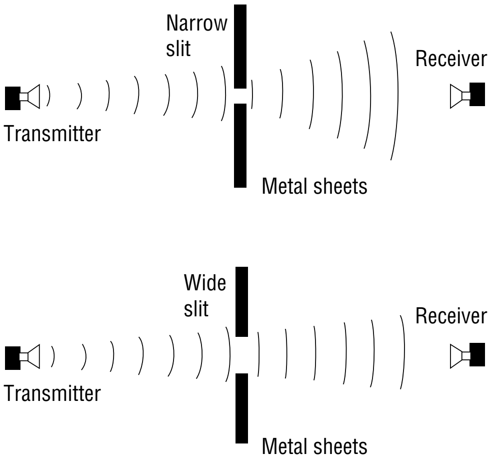


A hallow perspex lens filled with paraffin will focus the microwaves just as a glass lens will focus light.



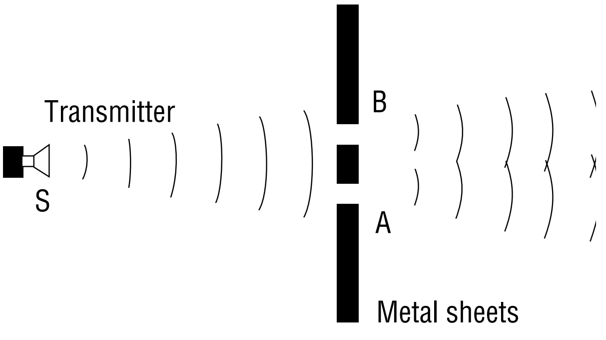
Diffraction

Two metal barriers will create a slit for the diffraction of microwaves. By adjusting the size of the slit the relationship between slit size and the amount of diffraction can be demonstrated.



Interference

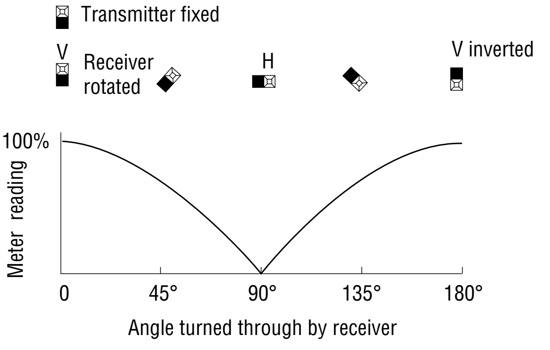
This can be demonstrated by using metal plates to make “Young’s Slits”.



Polarisation

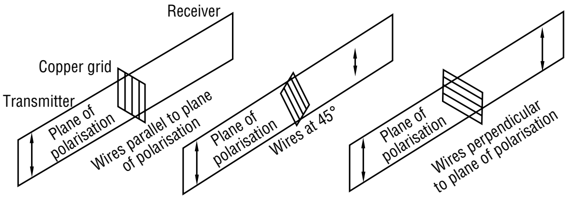
This can be demonstrated in o ways.

1. Rotate the receiver relative to the transmitter. When the receiver is on its side the meter should read zero but when it is inverted the meter should return to maximum value.



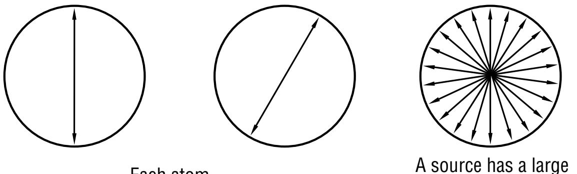
2. Use of a wire grid as a polarising sheet.

   The emitted microwave is a polarised wave. As the grid is rotated the meter reading should reduce to zero when the grid has rotated through 90° and then increase again to max value when the grid is inverted. This proves that the microwave is polarised.



Note

Every emitter emits polarised electromagnetic radiation. In the case of light each atom emits polarised light but there are so many atoms each emitting in its own polarised plane wave that every possible plane is covered and all directions will be equally favoured, light is therefore unpolarised.



Each atom emits polarised light

A source has a large number of atoms with different planes of polarisation

MODULE 3

Electricity

CONTENTS

CHAPTER 1 # INTRODUCTION

1.1 Teaching Approach

The main intention in writing this unit has been to cover all the material on Electricity required by teachers in preparing students for the Leaving Certificate Physics examination.

The material is dealt with in a sequence which follows very closely that outlined in the official Physics syllabus. To avoid repetitive statements and parts of questions (e.g. Define, etc.), the reader is referred to a standard textbook on occasions. To stimulate students' interest and to serve as an aid for discussion or revision, the unit contains a number of 'Do you know' items in Chapter 1. The text illustrates points of contact with practical situations in everyday life. A number of worked examples are interspersed throughout the unit, each of which has been chosen to illustrate and to bring down to earth some physical concept.

Practical work must play a major part in any course in electricity if the understanding, appreciation and investigative skills of the student are to be developed. Where possible, challenging personal tests or projects should be set for the student, e.g. using a digital multimeter to identify the components contained in a sealed (black) box. The components, whose terminals only are accessible, might be a resistor, a capacitor, a piece of wire, a diode and also, maybe, a blank to add to the confusion!

The appendices contain references to material, some of which is not specifically mentioned on the syllabus. This is an attempt to anticipate some of the questions frequently directed at teachers of Senior Cycle Physics. For example, with the advent and development of technology, many schools utilise a 3-phase supply to power the workshop lathes, etc., hence the reference to three-phase electricity in Appendix E.

The teaching of **Important Concepts** en bloc (see p. 4) might serve as a worthwhile precursor to a more in-depth study of electricity. Certainly, if a determined effort were made by students to grapple with these concepts initially, the subsequent study might prove to be easier and all the more rewarding. Using Important Concepts as an introduction to Electricity should equip the student with the necessary understanding and the basic knowledge required. This should provide the foundation for greater appreciation and enjoyment of the subject.

Having dealt with Important Concepts in class, it might then be worth considering devoting a few classes to some practical and experimental aspects, e.g. the understanding and the use of a digital multimeter and maybe some basic demonstrations with the cathode ray oscilloscope. Simple procedures, e.g. checking connecting leads for continuity, testing diodes and transistors, determining resistance values, all help to generate goodwill and to arouse interest.

Students often experience difficulty with units and learning definitions by rote. They should be encouraged to focus on the relevant formula in order to obtain the required units. Some have difficulty using the negative index notation, i.e. $m\ s^{-1}$ rather than m/s. Again, by the appropriate substitution of *unity* for some of the quantities in a formula, a suitable definition will be obtained.

Special attention should be devoted to graphical work (on graph paper!). In particular, the significance of the slope of a graph ought to be fully comprehended. Students often fail to recognise that a required answer might be the inverse of the slope in question.

1.2 Background

Over 2500 years ago, it was recorded that a piece of amber (a translucent fossil resin), when rubbed, would attract small particles of matter, e.g. straw. From this simple experimental fact has developed the whole science of electrostatics, that is, the properties of electric charge at rest. We know now that electric charge consists of two kinds which are called positive and negative; *like charges repel each other but opposite, or unlike, charges attract each other*.

Many important devices which are used in science, industry and the home, are directly based on electrostatic principles. Examples of such devices include ultra-high-speed particle accelerators, air cleaners and many other types which are used to minimise the effects of static electricity in the paper, printing and textile industries. Electrical charges may be stored temporarily in capacitors, then removed and replaced in a very short period of time, thus giving rise to a variety of applications in television, radio, radar and communications industries.

1.3 Do You Know

- Industrial plants need to take safety precautions to prevent the build up of static electricity in order to avoid dangerous sparks where flammable liquids are being used.
- Floor tiles in an operating theatre should be made of a conducting material to avoid build-up of static electricity since some anaesthetics form an explosive mixture with air.
- Passengers disembarking from aeroplanes which may have become charged during flight could receive shocks when touching the handrails of the steps used to assist them from the aircraft. To overcome this problem, special electrically conducting rubber is used in the tyres of aircraft (and in some floor coverings). This helps to earth the charged object and so prevents the risk of electric shock.
- Vinyl records, i.e. LPs, can become charged while being played and this leads to the accumulation of dust on the disc and stylus.
- Some types of integrated circuits (ICs) are susceptible to damage from high (static) voltages. Such ICs are therefore supplied in anti-static or conductive carriers.
- When spray painting, especially into inaccessible areas, less paint is wasted if the object to be painted is charged.
- Smoke pollution can be reduced by placing charged objects (scrubbers) inside a chimney stack.
- A recent development in the control of bacteria, algae and viruses in water involves the use of electrostatics. Water is pumped through tanks containing electrodes made of silver/copper alloy which release positive charges (ions) into the water. These positive ions are attracted to negatively charged bacteria cells. They break through the cell wall and neutralise the cell's reproductive ability. This water disinfecting system is used in air-conditioning cooling towers, swimming pools, hospitals, and in sea shrimper vessels, where it is being used to maintain clean water for shellfish, so that they do not cause food poisoning when eaten.
- Electrostatic plotters produce drawings through the use of a row of styli across the width of the paper. As the paper rotates on a drum, an electrostatic charge is created on the paper by the signal to the styli. The paper then passes through a developer and the drawing appears. The electrostatic printer image is made up of a series of very small dots, resulting in a high quality output.
- Laser printers and photocopiers also depend on electrostatics for their operation.
- An ingenious high-tech scheme that could help to protect the earth's ozone layer is being proposed by an American plasma physicist, Alfred Wong. The plan involves the use of huge 500 metre-long curtains of zinc or aluminium carried on platforms up to 40 km above the earth's polar regions. The curtains would deliver electrostatic charges to neutralise harmful chlorofluorocarbon (CFC) molecules. The curtains would later be brought back to earth for cleaning. At the heart of the plan lies the fact that if chlorine, fluorine and bromine (which are the most damaging elements in CFCs) are given an extra negative charge, they no longer break down the ozone.

If the global emission of ozone-destroying chemicals were stopped now, it is suggested that it would still take 50–100 years for the earth to recover unaided. Some scientists have predicted that an annual 10% decrease in the ozone layer would lead to a 26% increase in the incidence of skin cancer, 150,000 new cases of blindness due to cataracts, and a weakening of the body's immune system.

Some scientists view the success of Professor Wong's experiments as vital to the earth's future environmental health. However, most would consider his scheme to be somewhat fanciful.

- A crackling sound is heard when a person removes clothing made of nylon or other synthetic material in dry weather. This is due to sparks from the material which has become charged (maybe as it is being removed). The sparks may be visible if the clothing is removed in the dark.
- A person sliding off a car seat may become charged and consequently experience an electric shock when contact is made with the door handle.
- Washed and dried hair follows the comb due to the hair and the comb becoming oppositely charged.
- Rubbing inflated balloons against clothing can cause the balloons subsequently to stick to walls or ceilings.
- As a car wheel rotates a static charge is built up between the axle and the wheel bearing, since they are insulated by a thin film of grease. The bouncing of the wheel causes contact between the two surfaces, and the static charge is intermittently dissipated. This might produce interference (noise or crackling) on a car radio at certain speeds and so static collector springs are sometimes positioned under the axle dust cap to counter this action by establishing a constant electrical path between the axle and the bearing.
- During a thunderstorm in 1749, the American physicist, Benjamin Franklin (1706–1790), flew a kite with a pointed metal top attached to a silk insulating thread with a metal key attached at the lower end. When the thread got wet, it became a conductor and sparks jumped when Franklin brought his knuckles near the key. This was the first time that it was shown that lightning is basically an electric spark.

Franklin also designed the lightning conductor. This is a thick copper strip running up the outside of a tall building. The upper end of the strip terminates in one or more sharp spikes above the highest point of the building. The lower end is connected to a metal plate buried in moist earth. The lightning conductor protects a building from being damaged by lightning in a number of ways.

During a thunderstorm, the value of the electric field intensity in the air can be very high near a pointed lightning conductor. If the value is high enough, ions, which are drawn towards the conductor, will receive such large accelerations that, by collision with air molecules, they will produce vast additional numbers of ions. Therefore the air is made much more conducting and this facilitates a flow of current between the air and the ground. Thus, charged clouds become neutralised and lightning strikes are prevented. (Glow or corona discharges from sharp points during thunderstorms are often observed.) Alternatively, in the event of the cloud suddenly discharging, the lightning strike will be conducted through the copper strip, thus protecting the building from possible catastrophic consequences.

Raised umbrellas and golf clubs are not to be recommended during electrical storms for obvious reasons. Benjamin Franklin was very fortunate to survive his kite-flying experiment.

On high voltage electrical equipment, pointed or roughly-cut surfaces should be avoided to minimise the possibility of point discharge, e.g. blobs of solder should be rounded.

Since there is no electric field inside a hollow conductor the inside of a motor car is a relatively safe place to be during an electrical storm.

For the same reason engineers work from inside special metal cages when working on high voltage power transmission lines.

Co-axial cable, Fig. 1.1, is used to link the aerial to a television. This prevents external electric fields from causing interference with the actual TV signal. The braiding acts as a hollow conductor.

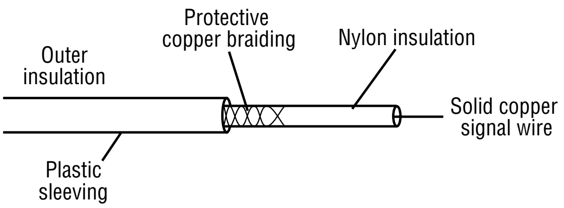


Fig. 1.1 *Co-axial cable*

1.4 Important Concepts

Positive and Negative Charge

A glass rod, when rubbed with silk will attract a plastic rod which has been rubbed with fur. Benjamin Franklin named the kind of electric charge that appears on the glass positive, and the kind that appears on the plastic he named negative.

Electric effects are not limited to the above materials. Most substances, when rubbed with another under suitable conditions will become charged to some extent. Since like charges repel and unlike charges attract the unknown charge on a body may be identified by comparing it with a glass rod that had been rubbed with silk or a plastic rod that had been rubbed with fur.

Electrification by Contact

The effect of charges moving from one body to another is shown below.

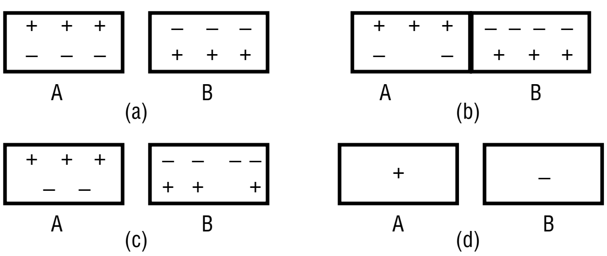


Fig. 1.2

In Fig. 1.2(a), two neutral bodies are represented, i.e. in each body, there are equal numbers of positive and negative charges (or there are positive and negative charges of equal strength).

In Fig. 1.2(b), the two bodies have been placed in contact with each other and some of the negative charge from A has moved onto B. There is now an excess of negative charges on B and a deficiency of negative charges on A.

In Fig. 1.2(c), the two bodies have been separated, with the unequal charges remaining on each one. The body marked A has an excess of positive charges, i.e. there exists a net positive charge. Similarly on B there is a net negative charge. Rather than talking about excesses and deficiencies of charges, we simply say that A is positively charged and that B is negatively charged. We indicate the net charge on a body as shown in Fig. 1.2(d).

Similarly, of course, if negative charge had moved from B to A, body A would have become negatively charged and body B would have become positively charged.

Electrification by rubbing is really a process of electrification by contact. If a glass rod is briefly touched to a piece of silk, negative charges leave the glass and attach themselves to the silk. Rubbing the glass with silk is a more effective way of electrifying it than merely touching it with silk, because more parts of the surfaces are brought into contact in this way. Thus, in electrification, the primary mechanism is simply one of contact.

Electric Fields

Consider the region surrounding a positively charged sphere, Fig. 1.3. No matter where we place an imaginary test charge of magnitude Q, it experiences a force. The magnitude of this force depends on its proximity to the charged sphere, i.e. $F_3 > F_2 > F_1$.

We can therefore draw a line at any point to represent, in magnitude and direction, the force acting on the test charge at that point. The infinite collection of such lines, i.e. vectors, is called a vector field or simply a field. The direction of the field is defined as the direction of the force on a positive test charge.

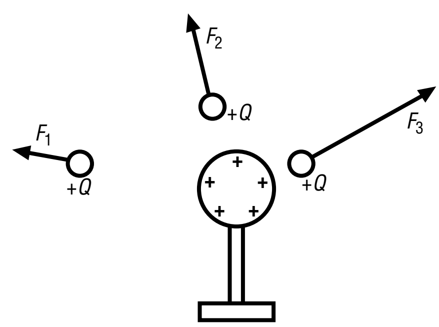


Fig. 1.3

The strength or intensity, E, of the electric field at a point is defined as the force on a test charge at that point divided by the magnitude of the test charge.

Thus, $E = F/Q$. E is a vector quantity since F is a vector quantity and Q is a scalar quantity. E is in the same direction as F.

Flux

The word flux is derived from the Latin word fluere, i.e. to flow. Flux is a property of all vector fields which obey the inverse square law.

By way of introduction to the concept of electric flux, first consider a cylinder lowered into water which is flowing with constant velocity, v. Let the cross-section A of the cylinder be at right angles to the direction of flow of water, Fig. 1.4.

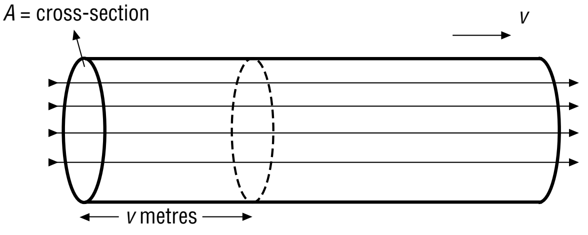


Fig. 1.4

Let the mass flux (Φ_m) represent the mass of water passing through the surface A per unit time. In one second, all the water that flows through the surface A will be contained in the volume ($A \times v$), (see Fig. 1.4).

Since mass = density $\times$ volume, the mass of water which flows through the surface A in one second = $\rho(vA)$, where ρ = the density of water.

Therefore, the mass flux, $\Psi_m = \rho vA$.

Electric Field Flux

Fig. 1.5 shows two large circular metal plates arranged close together. The plates are parallel to each other and one carries a positive charge on its inside surface while the other carries an equal negative charge.

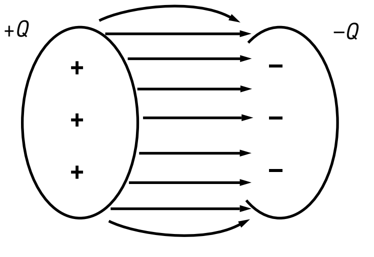


Fig. 1.5

Let each plate have a cross-sectional area of A. Ignoring edge effects, a uniform electric field exists in the space between the plates, i.e. the electric field intensity E is constant. Field lines drawn between the plates give the direction of E at any particular point. The closer together the field lines are in a particular region the greater is the intensity of the field in that region. The electric field flux, Ψ, passing through the area A which is perpendicular to E is defined as the product of the area A and the field intensity E, i.e. $\Psi = EA$.

Potential Difference

Water tends to flow downhill because of the force of gravity. From an energy point of view, we might say that water flows downhill because its gravitational potential energy is less in the lower position. A heavy object, when subjected to no other force than its own weight, will always seek the lowest possible level, where its gravitational potential energy is least.

Similarly, positive electric charge tends to flow from points of higher potential to points of lower potential, i.e. "downhill".

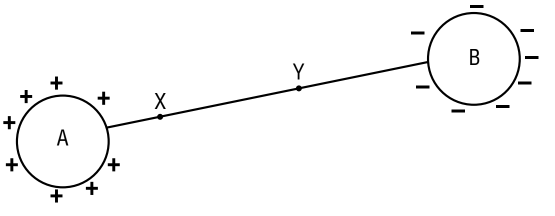


Fig. 1.6

In Fig. 1.6, two insulated metal spheres, A and B, are charged as shown. Since work must be done on a unit positive charge to bring it from point Y to point X, clearly therefore, X is at a higher potential than Y, i.e. there exists a potential difference, or voltage, between X and Y.

Conductors and Insulators

Conductors are materials which contain large quantities of 'free' charges, i.e. charges which are relatively free to move within the material.

Insulators are materials which contain hardly any free charges.

In modern times, we are accustomed to the idea of an atom as comprising negative charges (electrons) and positive charges. Rutherford's scattering experiment showed the existence of a positively charged nucleus. Millikan's oil-drop experiment showed that charges on bodies are multiples of a unit charge, the charge of one electron.

In a solid body, the nuclei are fixed in position except for some slight vibrations that are the manifestation of thermal energy. The amplitude of these vibrations increases at higher temperatures. However, unless the solid melts, nuclei rarely move from their positions in the crystal structure. In certain substances, on the other hand, some of the orbital electrons may be rather easily detached from the electron cloud surrounding a nucleus. Research has now shown that the phenomenon of electrical conduction is quite complex and that all substances may be classified in a sequence ranging from good conductors to poor conductors (i.e. good insulators).

In liquids and gases, both positive and negative charges move. In vacuum tubes, negative charges (i.e. electrons) alone usually constitute the current. In solid conductors, the positive charges are generally held firmly in place by atomic forces and the electrons can move. Since conduction through solids is by far the most frequently encountered case, current normally consists of moving electrons or negative charges.

In metals, electrons can move relatively easily and so metals are generally good conductors. Silver is an excellent conductor but because it is so expensive, it is normally only used for electrical contacts in high-grade electrical relays. Copper is the most commonly used conductor.

In some insulating materials, e.g. argon gas, there are hardly any free electrons and the electrons in the atoms are tightly bound so that electric forces can only move the electrons a tiny distance in relation to the nucleus of an atom. If the potential difference acting on a layer of insulation is increased its atoms become 'strained', Fig. 1.7, but no appreciably number of electrons will break away from the atoms to form an electron flow since, even for very high potential differences, the electric field would not be strong enough to overcome the electrostatic attraction between the electrons and the nucleus. However, at a certain potential difference an avalanche effect occurs. There will then be a rapid flow of electrons through the insulation resulting in great heat, high temperature and a possible fire.

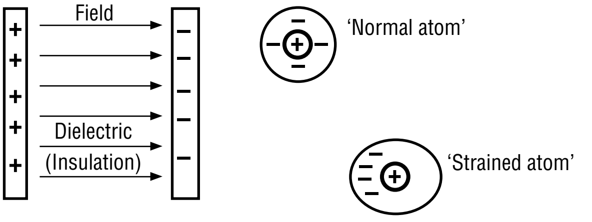


Fig. 1.7

The potential difference at which the insulation breaks down is called the breakdown potential and the potential difference per unit thickness of the insulating material at breakdown is called the dielectric strength of the material. The breakdown potential of a material is affected by many factors, e.g. thickness, moisture content, whether the applied potential is alternating rather than direct, the frequency of an alternating potential, etc.

One important advantage of using gases and liquids as insulation is that, in the event of breakdown and the passage of sparks, these insulators immediately restore themselves as soon as the excessive potential difference is removed.

Conductors and insulators differ immensely in their ability to conduct electric charge. This is quite striking when we compare copper, which is a good conductor, with glass, which is a good insulator. If we apply to a copper conductor a potential difference that will cause the flow of 3×10^{12} electrons per second, the same potential difference applied to a piece of glass of the same dimensions and at the same temperature will cause the flow of only *one* electron per second!

Examples of Insulating Material

Gaseous argon, helium, and neon are used in some types of thermionic devices.

Liquid mineral oils are used for filling housings of some types of high-voltage apparatus, e.g. transformers.

Distilled water is classed as a good insulator but is not of practical use as it is difficult to obtain water which does not contain some dissolved minerals.

Soft solids, e.g. paper, asbestos, plastics, rubber, are used for conductor coverings, for tubing sleeves and as spacers between layers of metal.

Hard solids are used for structural supports, e.g. glass, ceramics, mica.

Current

When the ends of a copper wire are connected to a battery an electric field will be set up at every point within the wire. Consequently there will be a flow of charge, i.e. an electric current is established. The rate of flow of charge is called current. If a net charge Q passes through any cross section of the wire in time t, the current I (assumed constant) is given by $I = Q/t$

By convention, the direction in which an electric current is established is taken to be that in which a positive charge at that point of the circuit would move. Thus, an electric current in a circuit is said to flow from the positive terminal of a battery to the negative terminal.

The Unit of Charge

The unit of charge is called the coulomb after Charles Augustin Coulomb (1736–1806), a French scientist.

1 coulomb = 1 ampere second,

i.e. 1 C = 1 A s

Since the charge on an electron equals 1.6×10^{-19} C, it follows that approximately 6.25×10^{18} electron charges make up one coulomb. A current of 1 A in a wire means, therefore, that about 6.25×10^{18} electrons pass a point on the wire every second.

Resistance

There is a flow of charge through a conductor when a potential difference is applied across it. The current which is established in the conductor depends on a number of factors in addition to the potential difference across it. The most important of these is the temperature, but such variables as the elastic strain to which the conductor is subjected, the illumination of its surface, or indeed almost any of its physical conditions, may also affect the current. The German physicist, Georg Simon Ohm (1787–1854) found however, that if all of these were kept constant, the current in a given conductor is proportional to the potential difference across it, i.e.

potential difference/ current = constant.

The ratio of potential difference to current is known as the resistance, R, of the conductor. (See p. 31 for more details.) For a conductor which obeys Ohm's law R is constant.

Combinations of Resistors

Many circuits contain networks of resistors and it is important to be able to calculate the combined effective resistance of various arrangements. By equivalent or effective resistance we mean that single resistance which, if substituted for a combination of resistances in a circuit, would leave the current unchanged in that circuit.

Resistors in Series

In this arrangement the same current I flows through all the resistances, R_1, R_2 and R_3, Fig. 1.8(a). The combined effective resistance, R_s, is given by:

$$R_s = R_1 + R_2 + R_3$$

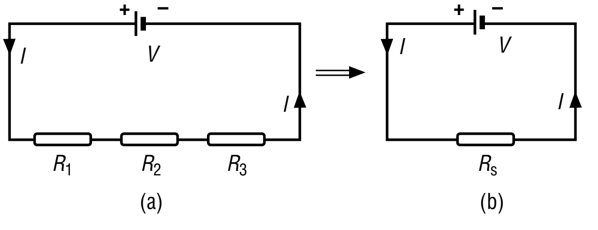


Fig. 1.8 *Resistors in series*

Because R_s is effectively the same as the combination, the current that flows in Fig. 1.8(b) will be of exactly the same magnitude as that in Fig. 1.8(a), (assuming that the same battery is used in both circuits).

For resistors in series:

– the total potential difference equals the sum of the individual potential differences;

– the same current flows through all the resistors;

- the individual potential differences are proportional to the individual resistances;
- the total resistance is always greater than the greatest individual resistance;
- the total resistance equals the sum of the individual resistances.

Resistors in Parallel

In this arrangement, the same potential difference (i.e. V) exists across all the resistances, R_1, R_2 and R_3. Students have difficulty accepting this particular fact. It is advisable, therefore, for teachers to devote some time to demonstrations involving combinations of resistors or bulbs, especially parallel networks.

The combined effective resistance, R_p, in this case, is given by:

$$\frac{1}{R_p} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}$$

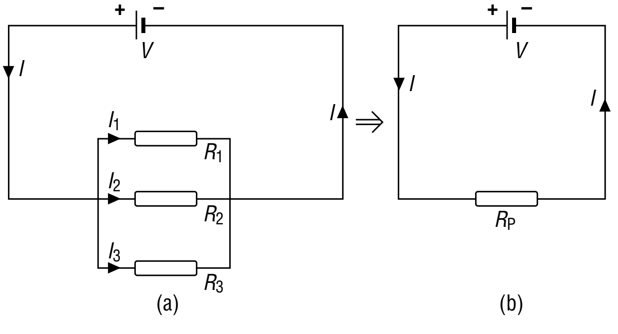


Fig. 1.9 *Resistors in parallel*

Since R_p in Fig. 1.9(b) is effectively the same as the resistor combination in Fig. 1.9(a), the magnitudes of the currents, I, in both circuits will be equal (again assuming that the same battery is used in each case).

For resistors in parallel:

- the potential difference is the same across each resistor;
- the total current equals the sum of the individual currents, i.e. $I = I_1 + I_2 + I_3$;
- the individual currents are inversely proportional to the individual resistances;
- the effective resistance is always *less than* the least individual resistance.

Internal Resistance

The chemical solution between the electrodes of a cell has a certain resistance. This, together with the resistance of the electrodes, is referred to as the internal resistance, r, of the cell. When a current is maintained by a cell, the current flows through its internal resistance as well as through the external resistance, R, in the circuit. Thus:

$E = I(R + r)$, where E is the emf of the cell (see p. 18).

Example

What is the maximum current that may be obtained from a cell of emf 1.5 V and internal resistance 0.5 Ω?

Solution

The maximum current is obtained by directly joining the two terminals of the cell together, i.e. a short-circuit (not to be recommended!).

$$I_{max} = \frac{E}{r} = \frac{1.5}{0.5}$$

$$= 3 \text{ A}$$

The internal resistance limits the current to 3 A.

Example

A voltmeter of resistance 1000 Ω is connected across the terminals of a well-used dry cell whose emf is 1.50 V and whose internal resistance is 60 Ω. What is the reading on the voltmeter?

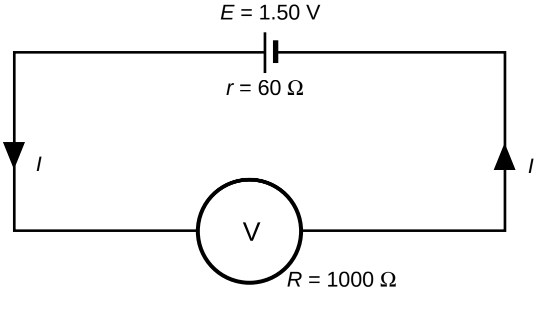


Fig. 1.10

For the entire circuit:

$$I = \frac{\text{total } V}{\text{total } R}$$

$$= \frac{1.50}{1060}$$

$$= 0.00142 \text{ A}$$

For the voltmeter:

$$V = I\,R$$

$$= 0.00142 \times 1000$$

$$= 1.42 \text{ V.}$$

Hence, the reading on the voltmeter is 1.42 V, which is substantially less than the actual emf value of 1.50 V. The discrepancy arises because of the potential difference across the internal resistance of the cell which is caused by the current flowing through the cell and the voltmeter. Since the size of this current depends on the impedance (resistance) of the voltmeter, only very high impedance voltmeters should be used for accurate measurement of potential difference and emf. This latter point is very relevant in the mandatory student experiment – measurement of the internal resistance of a source of emf. (Refer to the description of this experiment in the available textbooks.)

Semiconductors

Semiconductors belong to a class of materials which in the pure state and at room temperature have few free charges carriers. Consequently, semiconductors are materials with resistivity values (see p. 36) between those of conductors and those of insulators. Heating, or the addition of impurities can, however, change conditions very greatly, and we can engineer different electrical characteristics for semiconductor materials.

Both germanium and silicon, and indeed most solids, are crystalline, i.e. their structure conforms to a regular pattern. Note also that in germanium (Ge) and silicon (Si) atoms, the outer shell in each case consists of four electrons. These valence electrons form covalent bonds with the valence electrons of adjacent atoms.

Doping

By introducing into the crystal lattice structure a material that has a different number of valence electrons to the semiconductor, it is possible to alter the structure so that more charges are free to move. This process is known as doping. For example, if an antimony atom with five valence electrons is introduced into a germanium crystal lattice, four of these valence electrons will form covalent bonds with the valence electrons of the four adjacent atoms of germanium, but the fifth will be free of any such bond and will provide a free negative charge carrier, Fig. 1.11. The diagram of the germanium crystal shows the free electron donated to the material by the donor impurity atom, antimony (Sb).

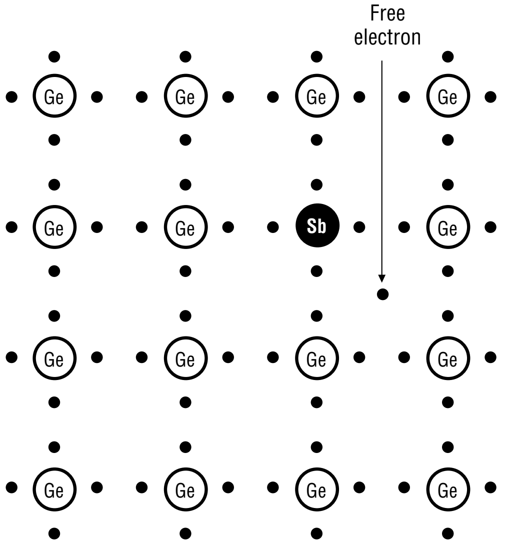


Fig. 1.11 *Germanium doped with antimony*

Other atoms with five valence electrons that may be used in this way to create an excess of free electrons in germanium or silicon are arsenic and phosphorus. An impurity that provides free electrons is called a donor impurity – it donates electrons to act as charge carriers. Since the majority charge carriers are negative germanium or silicon doped with donor atoms is therefore known as n-type germanium or silicon.

A preponderance of mobile positive holes is achieved by introducing into the semiconductor crystal lattice structure a substance whose atoms have three valence electrons, e.g. indium, boron or aluminium. For example, the indium atom has only three valence electrons and so three covalent bonds will initially be formed with the adjacent Ge atoms. To complete the crystal symmetry, however, a fourth covalent bond will be created by the indium atom capturing an electron from a nearby atom. In this way, positive holes in the semiconductor crystal lattice are created. Fig. 1.12 shows the hole created by the indium impurity atom accepting an electron to complete its covalent bonds with adjacent Ge atoms. Trivalent impurity atoms are called acceptors. In this case, the doped germanium or silicon is called p-type.

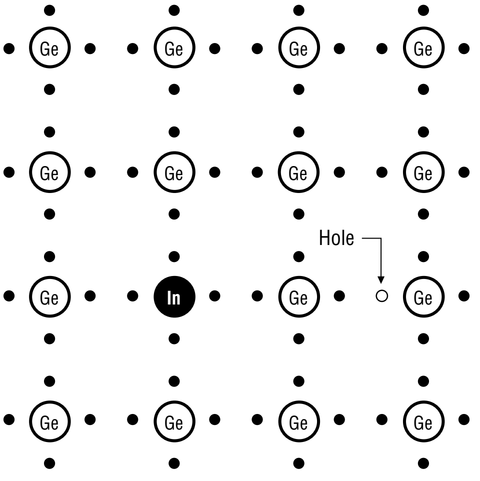


Fig. 1.12 *Germanium doped with indium*

Fig. 1.12 shows the hole created by the impurity indium (In) atom accepting an electron from a neighbouring atom to complete its covalent bonds with the surrounding Ge atoms.

Commonly Used Terms

1. Semiconductor material that has not been doped is sometimes referred to as intrinsic semiconductor material. Current flow in such material is known as intrinsic conduction. Equal numbers of electrons and positive holes are involved in intrinsic conduction.
2. Current flow in doped semiconductor material is called extrinsic conduction or impurity conduction. Unequal numbers of electrons and positive holes are involved in this instance.
3. In n-type material, there will be a greater number of free electrons than free holes. Thus, electrons are the majority carriers in n-type material and holes are the minority carriers. The converse is true in p-type material.

Note

The semiconductor family of devices is extensive and includes diodes, transistors, thermistors, photoresistors, thyristors, diacs, triacs, etc. A number of these devices are discussed in detail later.

1.5 Experimental Approach

Experiment 1.1

Demonstrating Forces Between Charges

Apparatus

Van de Graaff generator, (see Appendix F, p. 75), pith balls or light polystyrene spheres.

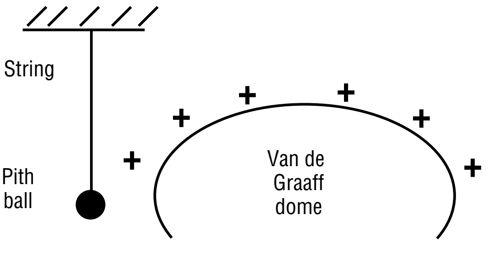


Fig. 1.13

Method

Switch on the Van de Graaff generator.

Bring a pith ball, suspended from a long insulating nylon thread, towards the dome, Fig. 1.13.

Observation

The pith ball is initially attracted and touches the dome. This is due to electrostatic induction (see below). The pith ball is subsequently repelled because it picks up some positive charge from the positively charged dome.

Conclusion

Like charges repel and unlike charges attract.

Note

1. If two pith balls are brought in contact with the charged dome of the Van de Graaff, they will subsequently repel each other if freely suspended, Fig. 1.14.

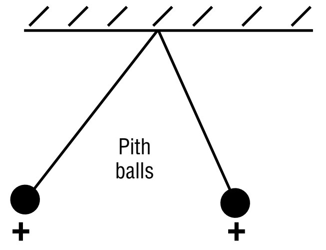


Fig. 1.14

2. If an earthed sphere is placed close to the high voltage dome of the Van de Graaff, a suspended pith ball will bounce back and forth between the large dome and the earthed sphere. This demonstrates that like charges repel, unlike charges attract. The earthed sphere has a negative charge induced on the side nearest the dome. The pith ball is first attracted to the dome where it acquires a small positive charge. It is then repelled by the dome and attracted by the sphere. When it touches the sphere it loses the positive charge and gains a small negative charge. It is now repelled by the sphere and attracted by the dome. The pith ball in fact transfers charge between the dome and the earthed sphere.

Electrification by Induction

A charged object, e.g. a plastic rod, can induce charges in another which is placed close to it if the latter is a conductor, Fig. 1.15. Some charges move along the conductor under the influence of the electric field caused by the original charge. This results in the conductor having equal and opposite charges at its ends. Consequently, the conductor experiences an attractive force towards the charged object since the unlike charges are closer together than the like ones.

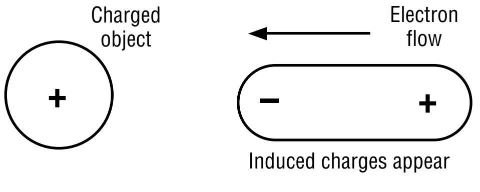


Fig. 1.15

Experiment 1.2

To Give a Conductor a Permanent Positive Charge by Induction

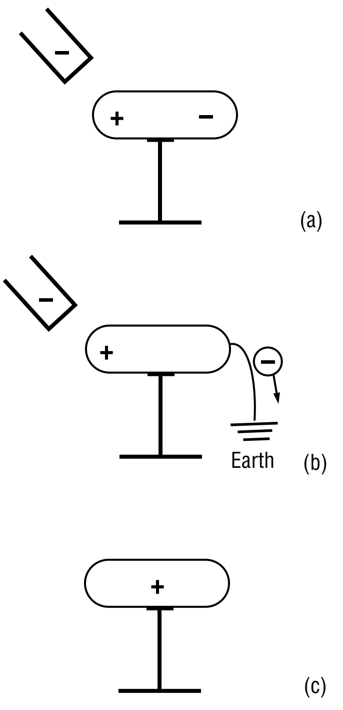


Fig. 1.16

Bring a negatively charged rod near a conductor which is mounted on an insulating stand, Fig. 1.16(a). Separation of charge occurs in the conductor.

Earth the conductor by touching it momentarily, allowing the free negative charges to flow to earth, Fig. 1.16(b).

Remove the charged rod. The conductor is left with a net positive charge, Fig. 1.16(c), by induction. The remaining charge on the conductor is redistributed.

While the actual number of electrons transferred during any charging process is very large the fraction of the charged body's electrons involved is minute. For example, a copper sphere of diameter 6 cm, when charged positively to a potential of one million volts, loses only 1 in 10^{11} of its electrons!

The Electrophorus

This is a rather old-fashioned electrostatic generator based on an earthing-induction technique for charging a conductor. It was invented by Alessandro Volta about the year 1800. The modern electrophorus consists of a metal disc with an insulating handle and a separate polythene base. Sufficient charge may be produced to give an audible and visible spark.

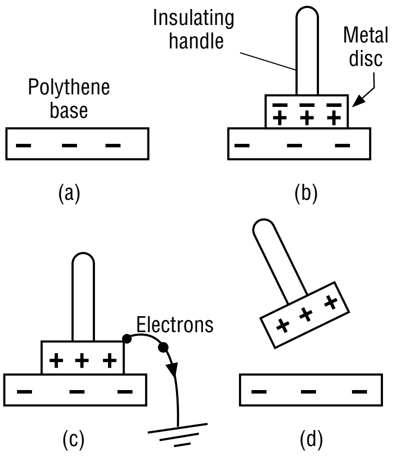


Fig. 1.17

To charge an electrophorus disc proceed as follows.

Rub the polythene base with fur (synthetic fabrics may also be used although they give the opposite charge). The base becomes negatively charged, Fig. 1.17(a).

Holding the insulated handle, place the disc on the base. Electrostatic induction occurs, Fig. 1.17(b).

Briefly touch (earth) the disc, Fig. 1.17(c), to provide an escape route for the free negative charge.

Having removed the earth connection, the positively charged disc may now be removed, Fig. 1.17(d). The disc may now be used for charging a gold leaf electroscope or any other conductor.

It should be noted that the negative charge in the base does not move onto the positively charged disc and neutralise it. This is because the metal disc and the polythene base, while apparently smooth, make contact in relatively few places, Fig. 1.18, and, since polythene is an insulator, charge cannot flow to the points of contact.

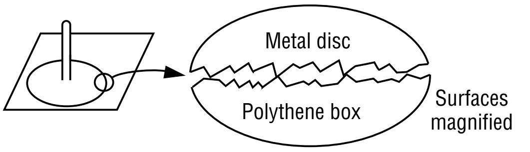


Fig. 1.18

Experiment 1.3

To Investigate the Distribution of Charge on a Charged Conductor

Apparatus

Electroscope and can, conductors of various shapes (see Fig 1.21), proof-plane.

A proof-plane, Fig. 1.19, is a useful device for transferring samples of charge from the surface of a conductor.

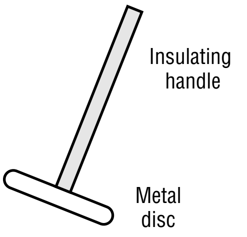


Fig. 1.19

(**Note:** the use of a can on the cap of the electroscope ensures that practically the entire charge collected by the proof plane is transferred to the electroscope.)

Method

1. Charge one of the conductors, using the method of induction or conduction.
2. Using the proof-plane, transfer a sample charge from the surface of the conductor, to the can (Fig. 1.20).
3. The amount of charge on the proof plane is a measure of the density of charge on the surface from which it was charged.
4. Note the angle of divergence of the leaves. This is a measure of the charge on the proof plane.
5. Earth the can briefly and repeat the test for sample charges taken from other areas of the conductor. Take samples from the inside and the outside if possible.

6. Repeat the procedure for different shapes of conductor.

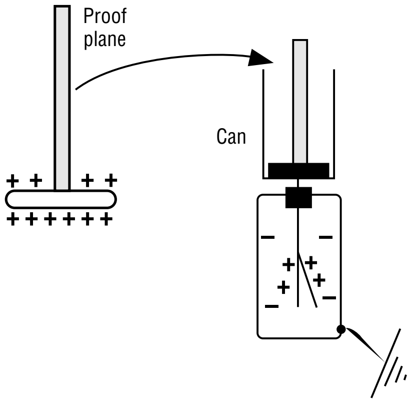


Fig. 1.20

Results

The results obtained would indicate the distribution of charge as shown in Fig. 1.21, (a), (b), (c) and (d).

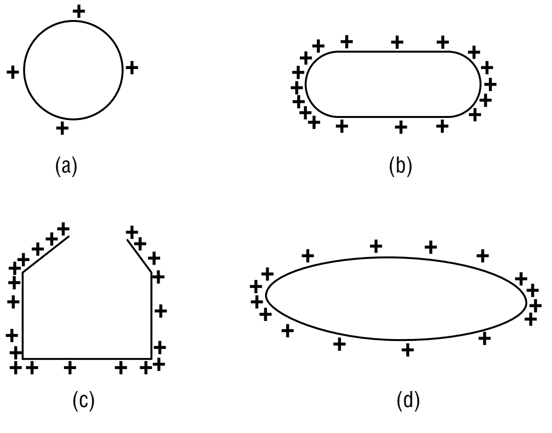


Fig. 1.21

Conclusions

1. All charge on an empty hollow conductor is distributed on the outside surface.
2. There is an even distribution of charge on the outside of a spherical conductor.
3. There is a concentration of charge on the most curved outside parts of an irregularly shaped conductor.

Note: moisture can easily ruin an electrostatics experiment, so it is wise to use a small heater, e.g. a hair dryer, to ensure that all materials to be used are dry.

Point Discharge

The very high concentration of charge at sharply curved parts of a charged conductor has many useful applications. The strong electric field in the region of a sharp point can lead to ionisation of the air in the immediate vicinity. Ions with charges of opposite sign to that of the pointed object, will be attracted to it and neutralise its charge. Ions of similar sign will be strongly repelled, possibly giving rise to a noticeable electrostatic wind, Fig. 1.22(a).

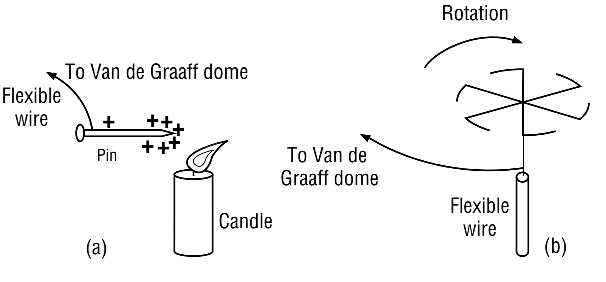


Fig. 1.22

The wind effect may also be shown with Hamilton's mill, Fig. 1.22(b), a freely pivoted framework of pointed metal rods which is seated on the dome of a Van de Graaff generator. The ions in the air experience a force away from the charged pointed rods while the rods experience an equal and opposite force, causing the mill to rotate.

Experiment 1.4

To Determine Whether a Substance is a Conductor or an Insulator

Apparatus

Gold leaf electroscope, a dry wooden peg, metal rod, plastic rod, etc.

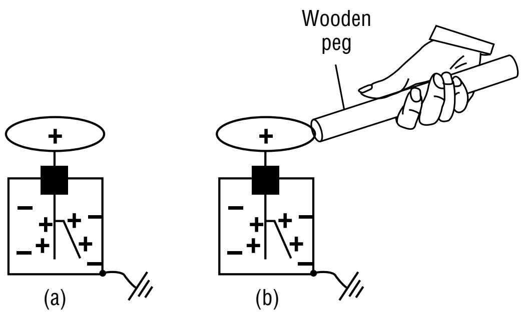


Fig. 1.23

Method

Place a charge on the gold leaf electroscope using the method of induction (see the electrophorus) or conduction. Note that the leaves diverge, Fig. 1.23(a).

Touch the cap of the gold leaf electroscope with the wooden peg, Fig. 1.23(b), and observe that the divergence of the leaves does not change, indicating that wood is an insulator (the wood must be very dry).

Repeat the procedure with a metal rod. The leaves will converge, indicating that the charge on the leaves has passed through the metal rod to earth via the person holding the rod. Therefore, the metal rod is a conductor, as is the person holding it.

Repeat for different materials.

Experiment 1.5 (Teacher Demonstration)

To Demonstrate that the Divergence of the Leaves of an Electroscope Increases as the Voltage is Increased

Apparatus

Gold leaf electroscope, EHT power supply (2 kV, d.c., variable). (The EHT supplies used in schools are perfectly safe. However, they should be treated with caution as it is possible to get a substantial shock from them.)

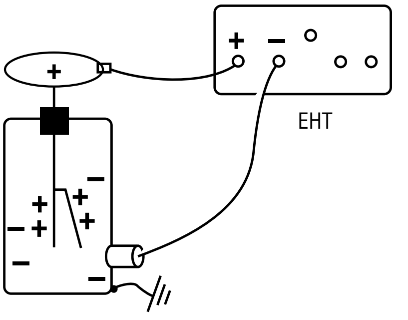


Fig. 1.24

Method

Connect the case of the gold leaf electroscope to the neutral of the EHT supply, Fig. 1.24, using a standard connecting lead. Join the positive of the EHT to the cap of the gold leaf electroscope (a crocodile clip may be required, or the cap may be removed and a standard 4 mm plug may fit in the socket). Switch on and slowly increase the EHT up to 2 kV.

Observe the divergence of the leaves.

Result

The divergence of the leaves increases as the value of the EHT voltage increases, though not linearly. Thus, the degree of divergence is an indication of the potential difference between the leaves and the case.

CHAPTER 2 STATIC ELECTRICITY

2.1 Electric Fields

Coulomb's Law

There is a striking similarity between Coulomb's law of force between two point charges and Newton's law for gravitational forces. Some scientists believe that the similarity is more than just coincidental but have been unable to identify the precise connection, if one does exist.

Coulomb's law may be expressed as

$$F = \frac{kQ_1Q_2}{d^2}$$

where d is the distance between the charges Q_1 and Q_2 and k is a constant.

A value for k may be found indirectly by an analysis of experiments that are in turn based on Coulomb's law. Such experiments give the value of $k = 9.0 \times 10^9$ N m^2 C^{-2}.

Coulomb's law is often written in the form

$$F = \frac{Q_1Q_2}{4\pi\varepsilon d^2}$$

where ε is the permittivity of the medium. ε is equal to $\varepsilon_0\varepsilon_r$, where ε_0 is the permittivity of free space (i.e. vacuum) and ε_r is the relative permittivity of the medium. The value of ε_0 is found, experimentally and from electromagnetic theory, to be 8.9 x 10^{-12} C^2 N^{-1} m^{-2}. (The more commonly used unit is the farad per metre, F m^{-1}.) The relative permittivity of air is almost exactly equal to one and so the permittivity of air is usually taken to be equal to ε_0.

Example 1

What is the force of attraction between a positive charge of 2 μC and a negative charge of 5 μC, Fig. 2.1, if they are separated by 30 cm of air?

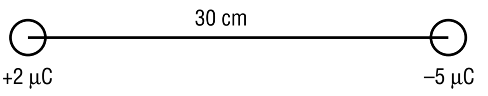


Fig. 2.1

Solution

$$F = \frac{Q_1Q_2}{4\pi\varepsilon d^2}$$

$$= \frac{2 \times 10^{-6} \times 5 \times 10^{-6}}{4 \times 3.14 \times 8.9 \times 10^{-12} \times (0.3)^2}$$

$$= 1 \text{ N.}$$

Example 2

Three charges, each of +100 μC, are equally spaced along a straight line, successive charges being 3 m apart, Fig. 2.2. Calculate (i) the resultant force acting on one of the end charges, (ii) the resultant force on the central charge.

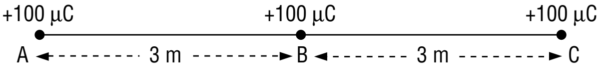


Fig 2.2

Solution

(i) The force on the end charge at point C is due to the charges at points A and B.

$$F = \frac{kQ_1Q_2}{d^2}$$

The force (F_1) due to the charge at the point B is given by

$$F_1 = \frac{(9.0 \times 10^9) \times (100 \times 10^{-6}) \times (100 \times 10^{-6})}{3^2}$$

$$= 10 \text{ N}$$

The force (F_2) due to the charge at the point A is given by

$$F_2 = \frac{(9.0 \times 10^9) \times (100 \times 10^{-6}) \times (100 \times 10^{-6})}{6^2}$$

$$= 2.5 \text{ N}$$

Therefore, the total force on the end charge (at C) is

$$F = 10 + 2.5$$
$$= 12.5 \text{ N.}$$

This force acts in a direction away from the centre.

(ii) The net force on the central charge is zero. This is so because the charge at the point B is subjected to two forces which are equal in magnitude and opposite in direction.

Example 3

Equal charges of +20 μC are placed at the four corners of a square of side 0.5 m, Fig. 2.3. Calculate the magnitude and the direction of the resultant force on one of the charges.

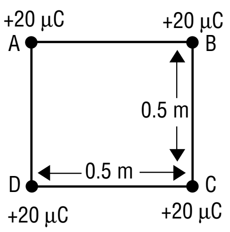


Fig. 2.3

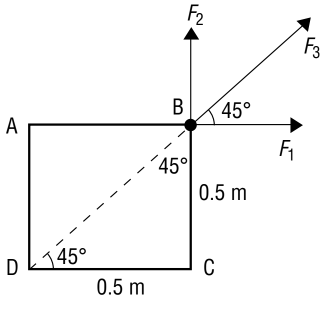


Fig. 2.4

Solution

Since BDC forms an isosceles triangle,

$|DB| = 0.5 \times \sqrt{2}$ m

Let F_1, F_2 and F_3 represent the forces due to the charges at points A, C and D respectively, that act on the charge placed at B, Fig. 2.4.

$$F = \frac{kQ_1Q_2}{d^2}$$

$$|F_1| = \frac{(9.0 \times 10^9) \times (20 \times 10^{-6}) \times (20 \times 10^{-6})}{(0.5)^2}$$

$$= 14.4 \text{ N}$$

Clearly, $|F_2| = |F_1| = 14.4$ N

$$|F_3| = \frac{(9.0 \times 10^9) \times (20 \times 10^{-6}) \times (20 \times 10^{-6})}{(0.5 \times \sqrt{2})^2}$$

$$= 7.2 \text{ N}$$

To obtain the resultant force on the charge at B, we add the three forces vectorially.

In Fig. 2.5, the resultant (R) of F_1 and F_2 is given by

$$|R| = \sqrt{(14.4)^2 + (14.4)^2}$$
$$= 20.4 \text{ N}$$

So, the total resultant force, F, on the charge at B is

$$F = (F_1 + F_2) + F_3$$
$$= 20.4 + 7.2$$
$$= 27.6 \text{ N, outward along the diagonal.}$$

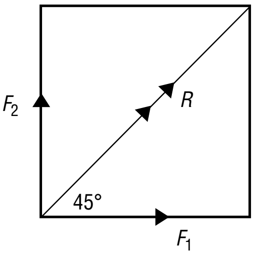


Fig. 2.5

Example 4

Calculate the force of attraction between two 1 C charges separated by a distance of 1 m.

Solution

$$F = \frac{kQ_1Q_2}{d^2}$$

$$= \frac{9 \times 10^9 \times 1 \times 1}{1}$$

$$= 9 \times 10^9 \text{ N.}$$

To emphasise the size of this force, assume that the mass of the average person equals 50 kg. Thus the weight of the average person is

$$W = mg$$
$$= 50 \times 9.8$$
$$= 490 \text{ N.}$$

Therefore, a force of 9×10^9 N is equivalent to the weight of $9 \times 10^9/490 = 1.8 \times 10^7$ people, i.e. 18 million people!

It is clear from this calculation that 1 C constitutes a very large charge indeed. The charges produced electrostatically are usually of the order of 10^{-6} C or smaller. However, charges in thunder clouds have been estimated to be several coulombs.

Electric Fields

The behaviour of electric charges closely resembles that of magnetic poles. Just as there are magnetic fields around magnetised objects, so also, electric effects are experienced in the regions of space round charged objects.

To observe the flux patterns of electric fields, grass seed or semolina granules are placed on an insulating liquid like olive oil (see Experiment 2.1 p. 21). Electric flux patterns develop which may not be very clear because surface tension forces interfere with the electric forces trying to align the particles. Some flux patterns between electrodes of different shapes are shown in Fig. 2.6.

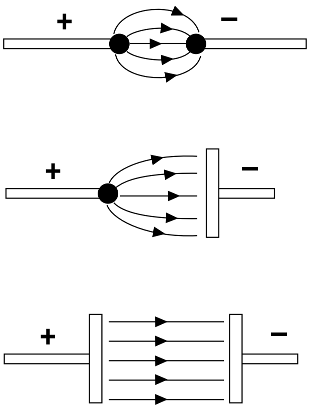


Fig. 2.6

The direction associated with electric field lines is chosen arbitrarily, namely, that the lines originate on positive charges and terminate on negative charges. This means that the direction of a field line at any point shows the direction of the resultant force which would act on a small positive test charge placed in the field at that point.

Another descriptive property of the field lines is their density; the field is strongest where the lines are closest together.

Electric Field Intensity

The intensity, or strength, E, of the electric field at a point is defined as the force per unit positive charge acting at that point. For example, if a charge of +1 C at point P, Fig. 2.7, experienced a force, F, of 800 N due to the presence of the charged sphere, we would say that the electric field intensity at P equals 800 N C^{-1}.

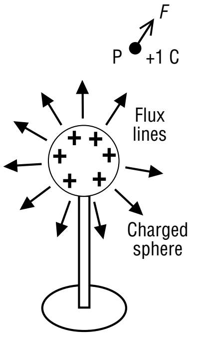


Fig. 2.7

In general, then $E = F/Q$, i.e. the intensity of the electric field equals the force on a test charge divided by the magnitude of the test charge.

Example

Calculate the magnitude and direction of the electric field intensity at a point where an electron experiences an upward force of 3.2×10^{-16} N, Fig. 2.8.

Solution

$$E = \frac{F}{Q}$$

$$= \frac{3.2 \times 10^{-16}\text{ N}}{1.6 \times 10^{-19}\text{ C}}$$

$$= 2 \times 10^{3}\text{ N C}^{-1}$$

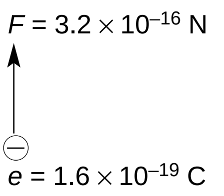


Fig. 2.8

The direction of E is downwards, since by definition, the direction of E is the direction in which a positive charge would move.

2.2 Potential Difference

The term voltage is often used instead of potential difference, partly because it is shorter to write but mainly because it tells us what unit is to be used. It puts a numerical dimension on the difference in electrical state between two points similar to the way in which contour lines on a map indicate the difference in height between two places. Potential difference is a scalar quantity.

The potential difference between two points is defined as the work done in moving unit charge from one point to the other. This may be expressed as $V = W/Q$, i.e. potential difference = work done per unit charge.

The Volt

A potential difference of one volt exists between two points if 1 J of work is required to transfer 1 C of charge from one point to the other, i.e. 1 V = 1 J C^{-1}.

Zero Potential

When heights and depths are measured for surveying purposes, the values given are related to an entity (i.e. zero) called mean sea level. When it comes to electrical potential, our zero for measuring purposes is usually the earth. Since the earth is such a huge body, its electrical state is not noticeably altered by the addition or removal of a few million electrons. This is similar to the concept of adding or removing a few bucketsful of water from the sea – the sea level does not alter.

When a conductor is earthed, i.e. connected to the earth by a conductor, it means that its potential is brought to zero, a value it holds as long as the earth connection is not broken.

Electromotive Force (EMF)

A source (or seat) of emf is a device in which chemical, mechanical or some other form of energy is converted into electrical energy, e.g. a battery, an electric generator, etc. In a cell, for example, chemical energy is converted into electrical potential energy. This causes charge to move in an external circuit whenever the opportunity arises. The emf of a particular cell is decided simply by its chemistry and does not depend on its size. Emf is measured in volts.

The potential difference value across the terminals of a cell on open circuit (i.e. no current flowing) equals the emf of the cell.

Definition

The emf in a closed circuit is the total work done in moving unit charge once around the complete circuit.

Example 1

How much work is required to move a charge of +3 C from a point X to a point Y if the potential of X relative to the earth is + 100 V and the potential of Y relative to the earth is + 118 V, Fig. 2.9?

Solution

The potential difference between X and Y is

$$V = 118 - 100$$
$$= 18 \text{ V}$$

Since,

$$W = Q\,V$$
$$W = 3 \text{ C} \times 18 \text{ V}$$
$$= 54 \text{ C V}$$
$$= 54 \text{ J.}$$

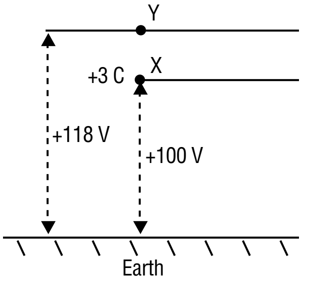


Fig. 2.9

Example 2

Calculate the magnitude and the direction of the electric field intensity at the centre of a square of side 2 m if charges of + 6 μC each are situated at three of the corners of the square, Fig. 2.10(a).

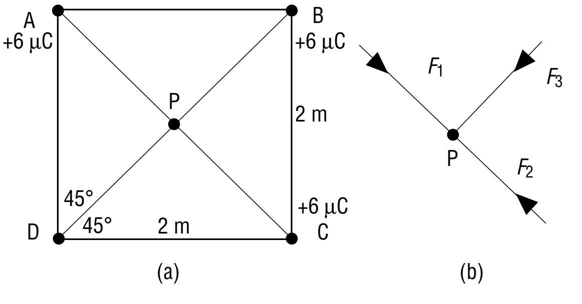


Fig. 2.10

Solution

Let the charges be at points A, B, and C. We require the electric field intensity at P, i.e. E_p. Imagine a +1 C charge at point P. The resultant force that acts on this unit charge is, by definition, E_p. The force F_1 due to the charge at A is equal in magnitude and opposite in direction to F_2, the force due to the charge at C, Fig. 2.10(b). Therefore, the resultant force at P is due solely to F_3, which is due to the charge at B.

$$F = \frac{Q_1 Q_2}{4\pi\varepsilon d^2}$$

$$F_3 = \frac{(6 \times 10^{-6}) \times 1}{4 \times \pi \times 8.9 \times 10^{-12} \times (\sqrt{2})^2}$$

$$= 2.7 \text{ x } 10^4 \text{ N}$$

Therefore, $E_p = 2.7 \times 10^4$ N C^{-1}, towards D.

2.3 Capacitance

Capacitance

When we add a litre of water to a number of different sized containers, Fig. 2.11, the level to which the water rises depends on the area of the base of the container. The greatest depth of water occurs in the narrowest container.

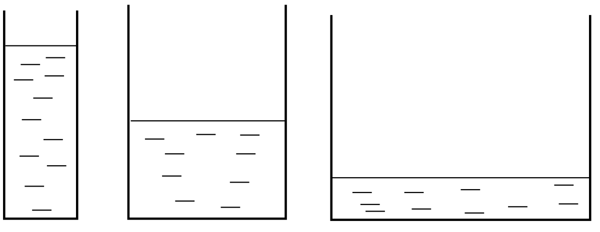


Fig. 2.11

Similarly in electricity, giving the same quantity of charge to different objects will result in their reaching different potentials depending on their electrical 'size', i.e. their capacitance. The conductor with the largest capacitance will be raised to the smallest potential. The largest capacitors can hold the greatest amount of charge before their potentials become so large that breakdown (resulting in a spark) occurs.

The capacitance of a body may thus be defined as the ratio of charge to potential or, in symbols, $C = Q/V$. The unit of capacitance is the farad, F. From the definition of capacitance 1 F = 1 C V^{-1}.

The farad is an inconvenient unit for practical use. Most everyday electronic devices contain capacitors with capacitance values stated in pF up to μF.

$$\begin{aligned} 1 \text{ F} &= 10^3 \text{ millifarad (mF)} \\ &= 10^6 \text{ microfarad } (\mu\text{F}) \\ &= 10^9 \text{ nanofarad (nF)} \\ &= 10^{12} \text{ picofarad (pF)} \end{aligned}$$

Capacitors

A capacitor consists essentially of two conducting plates (or sets of plates) separated by a layer of insulating material known as the dielectric. It is a device which is capable of storing electrical energy. The dielectric may be air, mica, ceramic, tantalum, aluminium oxide, etc.

Note. A large charged capacitor can be dangerous. After experimental work, it is safe practice to discharge a large capacitor by touching both of its terminals simultaneously with an insulated tongs, Fig. 2.12, or simply with the ends of a length of insulated wire.

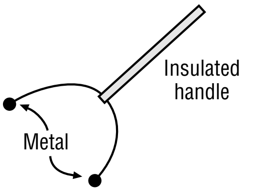


Fig. 2.12

Uses of capacitors include the following.

1. Smoothing the output of power supplies.
2. Frequency filtering. A capacitor has a reactance (i.e. an effective resistance to a.c.) which varies with frequency. By combining a coil and a capacitor it is possible to restrict current at certain frequencies and to pass current at other frequencies. These types of circuit are called filters because they filter out or pass only certain frequencies. Filters are widely used in radios and televisions.
3. Tuning circuits. A current in a circuit containing a coil and a capacitor will oscillate with a certain frequency (its natural frequency) which depends, amongst other things, on the capacitance of the capacitor. By adjusting the capacitance the circuit can be made to resonate with an incoming radio wave of a particular frequency (see module on Waves, p. 52).
4. Storing large quantities of charge, e.g. in flash guns for photography.
5. Blocking d.c. but passing (i.e. transmitting) a.c.

Correct polarity must be observed when connecting an electrolytic capacitor to a d.c. supply. If incorrectly connected, the internal aluminium oxide film could undergo breakdown and large currents would flow. The casing could then burst, spraying both the user and the rest of the circuit with corrosive material. Electrolytic capacitors used to smooth power supplies are particularly at risk in this respect because very large currents would be involved in the event of an internal short-circuit. Electrolytic capacitors should not be used in a.c. circuits.

In rolled capacitors, two long, thin sheets of metal foil with an insulating sheet between them are rolled up for compactness just like a swiss-roll! Lead-out wires are then attached to the metal foil sheets for connection to an external circuit.

Factors affecting capacitance

The capacitance of a capacitor depends on its construction, viz.

- the sizes and shapes of the surface areas of the plates;
- the separation of the plates, which is often the same as the thickness of the dielectric between them;
- the type of dielectric material.

Textbooks describe the experimental procedure for demonstrating the relationship between the capacitance and the various factors.

Parallel Plate Capacitor

For a parallel plate capacitor, i.e. a pair of parallel plates with a dielectric filling the space between them, the capacitance is given by

$$C = \frac{\varepsilon A}{d}$$

where C = capacitance

A = the common area of the plates

d = the distance between the plates

ε = permittivity of the dielectric.

Example 1

It is desired to construct a fixed capacitor having a capacitance of 600 pF, using mica as a dielectric. If the available metal plates are 2 cm × 1 cm and the mica has a thickness of 0.05 mm, find the number of dielectric layers required. The relative permittivity, ε_r, of mica is 5.6. Note: $\varepsilon_{mica} = \varepsilon_r\varepsilon_0$.

Solution

$$C = (n-1)\frac{\varepsilon A}{d} = \frac{(n-1)\varepsilon_r\,\varepsilon_0 A}{d}$$

$$n - 1 = \frac{C\,d}{\varepsilon_r\,\varepsilon_0 A}$$

$$= \frac{600\times10^{-12}\times0.05\times10^{-3}}{5.6\times8.9\times10^{-12}\times(2\times10^{-2}\times1\times10^{-2})}$$

$$= 3$$

Therefore, three layers of dielectric are required.

Energy Stored

The energy, W, stored in a capacitor is given by $W = \frac{1}{2}CV^2 = Q^2/2C$. A capacitor stores the electrical energy it receives from whatever charges it, e.g. a battery. If a lamp or electric motor is connected across it, current flows and the capacitor loses all of its charge, i.e. it is discharged. Some other form of energy is produced, e.g. internal energy, mechanical energy, etc.

Example 2

A 50 μF capacitor is charged to a potential difference of 300 V. How much electrical potential energy is stored in the capacitor?

Solution

$$W = \tfrac{1}{2}CV^2$$

$$= \tfrac{1}{2}(50\times10^{-6})\times(300)^2$$

$$= 2.25\text{ J.}$$

Example 3

A photographer's flashbulb unit contains a capacitor which is charged by a current of 30 mA for 20 seconds. The capacitor is then discharged quickly through a xenon-filled tube which requires 100 J of energy for a successful discharge. Calculate the value of the capacitance necessary.

Solution

Each flash uses a charge, Q, given by

$$Q = I\,t$$
$$= (30 \times 10^{-3}) \times 20$$
$$= 0.6 \text{ C}$$
$$W = \tfrac{1}{2} C V^2$$
$$= \frac{Q^2}{2C}$$
$$100 = \frac{(0.6)^2}{2C}$$
$$C = \frac{(0.6)^2}{200}$$
$$= 1.8 \times 10^{-3} \text{ F}$$
$$= 1800\ \mu\text{F}.$$

Variable Capacitors

Variable capacitors consist of two sets of parallel metal plates. One set is fixed and the other moves on a spindle within the fixed set. The plates are separated by a dielectric, usually air. When the spindle is rotated, the common area, A, between the plates is changed which causes the capacitance to vary.

When an alternating voltage is applied to a circuit containing a capacitor and a coil in series, the current in the circuit is a maximum (i.e. electrical resonance occurs) when the frequency of the applied voltage equals the natural frequency of the circuit.

In the aerial circuit of a radio receiver, Fig. 2.13, radio signals from different stations induce voltages of various frequencies in the aerial. This causes radio frequency (r.f.) currents to flow in the coil, L. If C is adjusted (tuned) so that the resonant frequency of the LC circuit equals the frequency of the required station, a large voltage at that frequency is developed across C. This voltage is then applied to the next stage of the receiver.

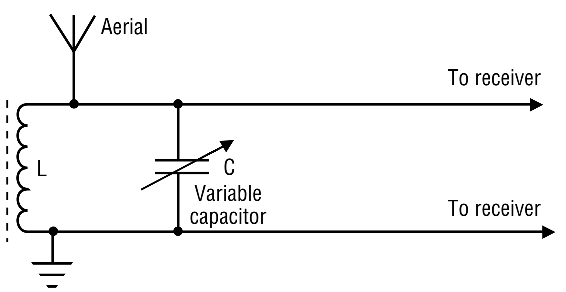


Fig. 2.13 *Tuning circuit*

Variable capacitors based on semiconductors are also available. In these the dielectric is the depletion layer in a reverse-biased p – n junction (see p. 44). The width of the dielectric, and hence the capacitance, is varied by varying the voltage applied to the semiconductor, the larger the voltage the wider the depletion layer and hence the smaller the capacitance.

2.4 Experimental Approach

Experiment 2.1

To Investigate the Electric Field Pattern Between a Positive and a Negative Rod

Apparatus

Petri-dish, olive oil, semolina, copper nails, EHT supply. (A Van de Graaff generator may be used as a source of EHT.)

Method

Half fill a petri-dish with olive oil and sprinkle semolina on the surface.

Arrange the two copper nails as shown in the circuit, Fig. 2.14(a).

Switch on the EHT and note the movement of the semolina as the particles are rearranged. Sketch the pattern observed.

Repeat the experiment using two copper plates as electrodes.

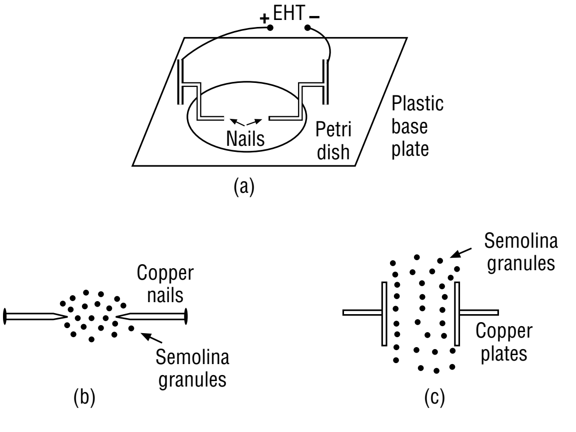


Fig. 2.14 *Demonstrating electric field patterns*

Observation

The expected field patterns would be as shown in Fig. 2.14(b) and 2.14(c).

Note. This experiment may be done very effectively on an overhead projector.

Experiment 2.2

To Examine the Charging and the Discharging of a Capacitor

Apparatus

9 V d.c. supply, 470 Ω resistor, 2-way switch, electrolytic capacitor (1000 μF, 9 V), 6 V lamp, connecting wires. (The 470 Ω resistor prevents too large a current flowing during the charging stage.)

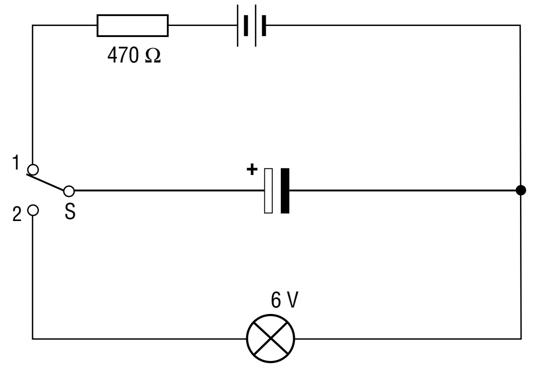


Fig. 2.15 *Charging and discharging a capacitor*

Method

Set up the apparatus as shown in Fig. 2.15, making sure that the electrolytic capacitor is correctly connected.

Move the switch S to position 2 after a few seconds and note the effect on the lamp.

Repeat the experiment using capacitors of different capacitances.

Observation

The lamp lights brightly at first but then fades rapidly (see the graph in Fig. 2.17(b) below).

Conclusion

With the switch in position 1, the capacitor charges quickly. In position 2, all the stored energy is released and converted into heat and light energy in the lamp.

Note. If the voltage across a capacitor is monitored during the charging and the discharging processes, Fig. 2.16, the graphs obtained will be similar to those shown in Fig. 2.17(a) and 2.17(b).

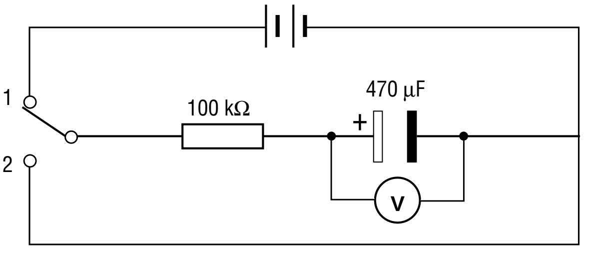


Fig. 2.16 *Charging and discharging a capacitor through a resistor*

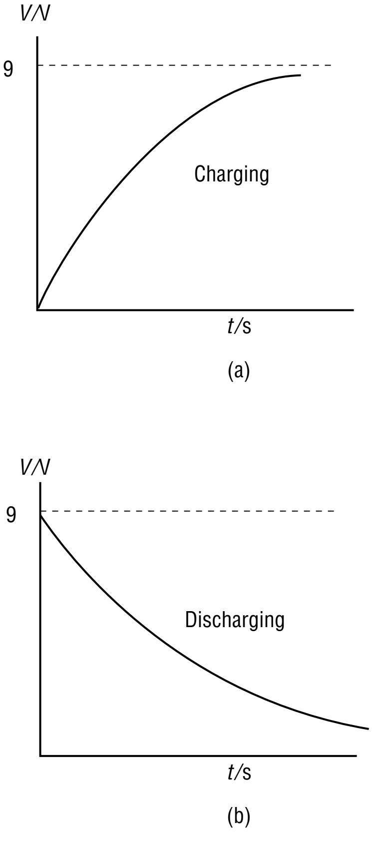


Fig. 2.17

Experiment 2.3

To Investigate the Effect of a Capacitor on the Passage of d.c. and a.c.

Apparatus

12 V supply (d.c. and a.c.), 10 μF capacitor, bulb (6 V, 60 mA).

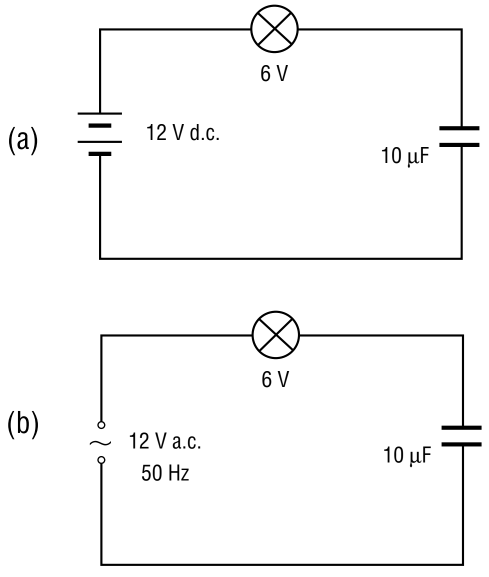


Fig. 2.18

Note. A non-electrolytic capacitor should be used above, since electrolytic capacitors are not intended for use in a.c. circuits.

Method

Set up the circuit using a d.c. supply, Fig. 2.18(a), and switch on the power.

Observation

The lamp does not light.

Replace the d.c. with the a.c. supply, Fig. 2.18(b), and switch on.

Observation

The lamp lights.

Conclusion

A capacitor blocks d.c. but allows a.c. to flow in a circuit.

When d.c. is used there is only a brief current flow (insufficient to light the lamp) as the capacitor is charged. When a.c. is used, charge flows backwards and forwards in the connecting wires between the a.c. supply and the capacitor plates. This process is repeated fifty times a second on a 50 Hz supply and a current would be recorded on an a.c. ammeter, i.e. the to-and-fro movement of charge makes it seem that a current flows through the capacitor.

CHAPTER 3

CURRENT ELECTRICITY

3.1 Current Electricity

The rate of flow of charge through a conductor is called current. For charge to flow in an electrical circuit, a potential difference or an electromotive force (emf) must exist in the circuit. A basic electrical circuit consists of a source of voltage, called the supply; the load, which has a resistance and which dissipates heat; and conducting wire which connects the supply to the load, Fig. 3.1. Switches are used to interrupt or to restore the flow of current. Fuses are placed in circuits to protect against overheating and the risk of fire which would result from the very large current which might flow if the circuit were faulty.

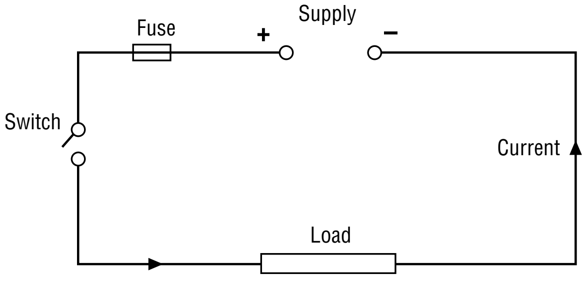


Fig. 3.1

Common Circuit Problems

Two common circuit faults which may arise are open circuit (o/c) and short circuit (s/c). An o/c fault is due to a break in the conducting path such that current can no longer flow through the circuit, Fig. 3.2, e.g. breakage in a wire.

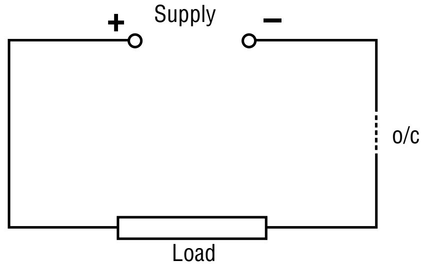


Fig. 3.2 *Open circuit fault*

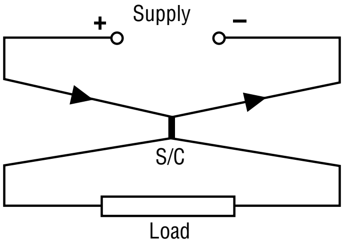


Fig. 3.3 *Short circuit fault*

A s/c fault is one in which an unwanted contact is made such that the current flow takes the wrong path or paths, Fig. 3.3. A s/c fault may result in blown fuses, overheating, damage to components, etc.

Production of EMF

In an electrical circuit, if any other form of energy is converted into electrical energy, an electromotive force or emf is said to exist in that circuit. An emf may be generated in a number of ways.

1. Chemical

Chemical energy is transformed into electrical energy by the immersion of two dissimilar conducting materials in an electrolyte, e.g. as in the simple cell.

2. Thermoelectric

Heat energy is transformed into electrical energy by heating one of the junctions formed by joining two wires made from dissimilar metals, e.g. in the thermocouple, Fig. 3.4.

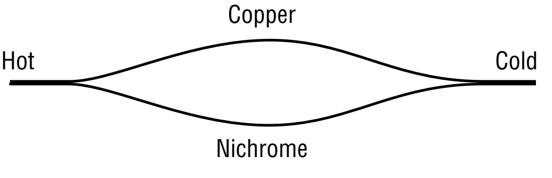


Fig. 3.4 *Thermocouple*

3. Electromagnetic

The energy stored in a magnetic field may be combined with some kinetic energy, e.g. a moving conductor, such that some of the kinetic energy is transformed into electrical energy, e.g. plunging a magnet into a coil of wire generates an emf in the coil. This is the most practical method of obtaining electrical energy and is used in all dynamo-electric machinery.

3.2 Sources of Electric Current

Primary and secondary cells are now examined. (A detailed study of power supply units is provided in Appendix A.) In a cell, chemical energy is converted into electrical energy. A collection of cells, usually connected in series, is called a battery.

A primary cell is capable of producing current as soon as its constituent chemicals are put together. Once the chemical reaction is finished, the cell is exhausted and cannot be regenerated.

A secondary cell generally needs to be charged before it can be used. Its chemical reactions take place in one direction during charging and in the opposite direction during use, i.e. during discharging.

Primary Cells

The Italian scientist Alessandro Volta (1745–1827) developed (c. 1800) the first primary cell based on earlier experiments performed by another Italian, Galvani, on the nerves from frogs legs; hence words like galvanometer and galvanise! A typical primary cell, consisting of two dissimilar metals immersed in a container of electrolyte is shown in Fig. 3.5.

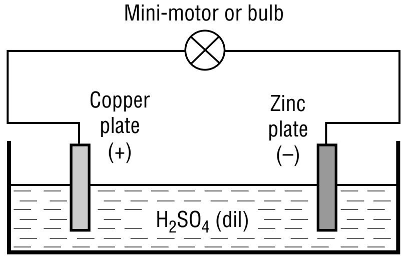


Fig. 3.5 *Simple cell*

In this circuit, chemical energy is converted into electrical energy which, in turn, operates the mini-motor (1.5 V type) or the small bulb (1.5 V type).

The disadvantages of this simple cell soon become apparent. Bubbles of hydrogen quickly form on the copper electrode (polarisation) and prevent the cell from operating. Another defect of the cell concerns the tendency of the zinc, which is seldom 100% pure, to dissolve in the acid even when no external circuit is connected. This problem may be overcome by making an alloy of the zinc with mercury (called an amalgam) by rubbing mercury over the surface of the zinc.

Note. When demonstrating the simple cell with the arrangement shown in Fig. 3.5 note the following points.

1. Always clean the copper and zinc electrodes with emery/glass paper before immersion in the electrolyte.
2. If a mini-motor is used, a piece of coloured cardboard glued onto the spindle indicates clearly the rotation of the motor.
3. Place a voltmeter across the electrodes of the cell. Note the drop in the reading as current flows in the circuit due to polarisation of the cell.

Most primary cells are of the zinc/carbon (Leclanché) type, Fig. 3.6. The ammonium chloride paste gradually dissolves the zinc. The subsequent chemical action provides the energy from which the voltage is obtained.

The manganese dioxide mixture oxidises hydrogen gas which would otherwise gather on the carbon electrode (polarisation) and would prevent current from flowing.

The Leclanché cell has a reasonable shelf life and yields a fairly steady voltage throughout its working life.

Towards the end of the useful life of a cell or battery, the value of its internal resistance rises. This causes the voltage across the terminals of the cell to drop below its normal value when current flows through the cell. A voltage check with this cell removed from its circuit will show a normal voltage rating, but the cell should nevertheless be replaced.

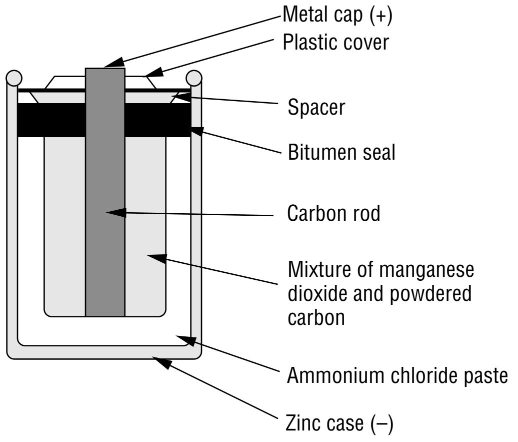


Fig. 3.6 *Leclanché cell*

The only useful check on the state of a battery is a comparison of the voltage reading on-load (i.e. with normal current flowing through it) with the known on-load voltage of a fresh cell.

There are various other types of primary cell, both of the wet and dry types, but since the wet type is not easily transportable, the dry type is normally preferred, e.g. the Mallory or mercury type. This cell consists essentially of a (negative) zinc electrode, a (positive) electrode of mercuric oxide and an electrolyte consisting of a saturated solution of potassium hydroxide (caustic potash) and zinc oxide. The emf of the mercury cell is lower than that of the Leclanché cell (approximately 1.35 V as opposed to 1.5 V). However, it possesses the advantages that it has a much longer shelf life, provides a more constant voltage, (see Fig. 3.7), has a self-depolarising action and, for a given volume, the electrical capacity is greater.

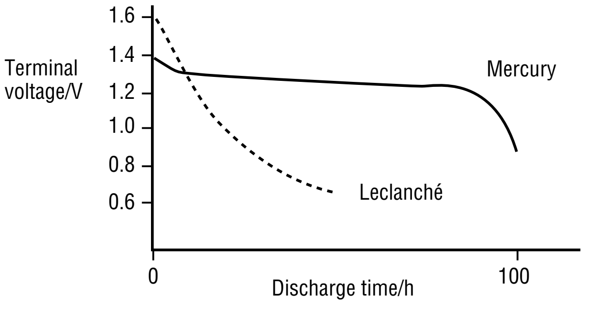


Fig. 3.7 *Change in terminal voltage with discharge time*

The small dimensions of the Mallory cell make it suitable for use in hearing aids and watches where space is severely restricted. The cells tend to be relatively expensive.

Secondary Cells

The lead-acid cell (used in car batteries), the nickel-iron-alkaline (NiFe) cell (used for powering some electric cars) and the nickel-cadmium (NiCd) rechargeable cell are all classed as secondary cells. Their common feature is that, unlike primary cells, they may be recharged after having been discharged.

The Lead-Acid Cell

If two clean lead plates are placed in a beaker of dilute sulphuric acid, a voltmeter placed across the plates will read zero since both electrodes are made of the same material, Fig. 3.8(a). However, if current from another source is passed through the cell for a few minutes and then stopped, the voltmeter does indicate the presence of an emf. A small bulb placed across the lead plates will light up for a short time indicating that some of the energy passed via the charging current had been stored, Fig. 3.8(b). The current used to charge the cell causes changes in the appearance of the lead plates. The positive plate shows a brownish deposit of lead dioxide and the other lead plate acquires a fresh layer of spongy grey lead.

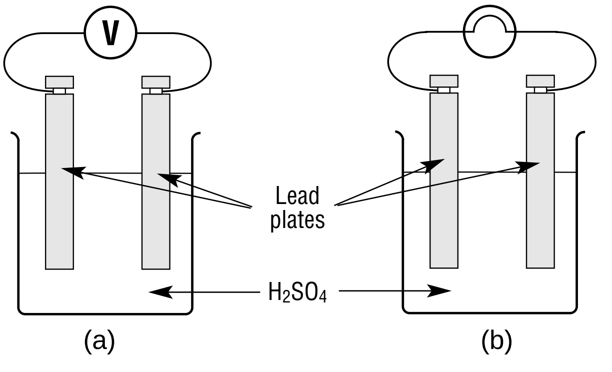


Fig. 3.8 *Lead-acid cell, (a) before charging and (b) after charging*

The emf of a fully charged lead-acid cell (accumulator) is just over 2 V but this drops quickly to 2 V when current is drawn. The emf drops slowly during use and before it reaches approximately 1.85 V, it should be recharged or it may be damaged. On charging, the voltage rises fairly rapidly to about 2 V and then rises more slowly to about 2.2 V. Further charging then causes the voltage to rise rapidly to about 2.6 V, when the accumulator is fully charged. Towards the end of the charging process, hydrogen and oxygen are liberated (known as gassing) and, for safety, no naked flame or spark should be used in the vicinity of the cell.

Charging an Accumulator

Accumulators (i.e. secondary cells) may be charged using a standard battery charger or a power pack capable of delivering a current of at least 5 A. It is essential that the positive terminal of the charging supply be connected to the positive side of the accumulator so that current passes into the accumulator in the opposite direction to when the latter delivers current. Manufacturers normally recommend a maximum charging current of about 4 A. A suitable rheostat or a variable voltage power supply is therefore required in the charging circuit, as the internal resistance of an accumulator is very low (e.g. 0.02 Ω). A typical charging circuit for a set of accumulators to be charged from a power pack through a rheostat is shown in Fig. 3.9.

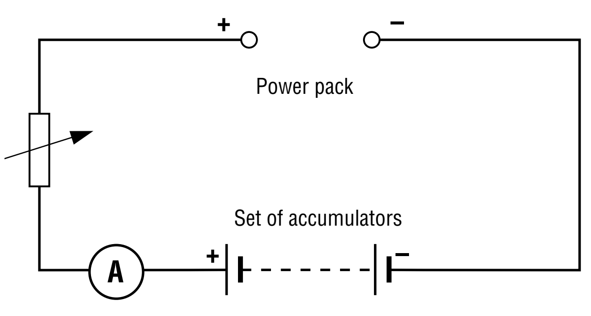


Fig. 3.9 *Recharging a set of accumulators*

Suppose there are ten cells in series, each of emf 2 V and that the power pack is set at 30 V. To obtain a charging current of 4 A, a value for the required series resistance is calculated as follows.

Total emf of battery of cells = 10×2 = 20 V

Supply voltage = 30 V.

Resultant emf in the circuit = 30 – 20 = 10 V.

The series resistance required = voltage/current

= 10/4 = 2.5 Ω.

Since the current flowing is 4 A the power developed in the resistor is

$$P = RI^2 = 2.5 \times 4^2 = 40 \text{ W}.$$

The total power supplied by the power supply is

$$P = VI = 30 \times 4 = 120 \text{ W}.$$

The power delivered to the accumulators is therefore 120 – 40 = 80 W.

So, one third of the energy supplied appears as heat in the series resistor. Thus, a rheostat rated as, e.g. 6 Ω 10 A, could be suitably adjusted for use in the circuit of Fig. 3.9, to allow a current of 4 A to flow, whereas a rheostat rated as 30 Ω 2 A would tend to get dangerously hot during use.

Care of Accumulators

The lead-acid accumulator requires periodic maintenance to prevent deterioration. It should not be left for long periods in a discharged state because the lead sulphate on the plates changes to a crystalline form which makes subsequent essential chemical reactions inoperable (the cell is said to be sulphated). A monthly topping-up charge would prevent this. Regular inspection of the acid level is important and any water lost due to evaporation should be replaced by distilled water. The top of the cells should be kept clean and free from moisture (which could result in a loss of charge). Excessive currents should not be drawn from accumulators. In extreme cases, the heat generated inside the cell could cause buckling of the plates, resulting in the active ingredients flaking off the plates and the cell being irreparably damaged. The maximum current that could flow in the event of a short circuit is determined by the magnitude of the internal resistance, e.g. in the case of a 2 V cell with internal resistance 0.02 Ω, the maximum current is given by:

$$I = \frac{V}{r} = \frac{2}{0.02} = 100 \text{ A}.$$

Car Battery

A typical 12 V car battery consists of six 2 V cells in series. The starting motor in the car draws a very large current from the battery for a short time. When starting a car, it is advisable, therefore, to have all headlights, wipers, rear window heater, etc., switched off, as these accessories also draw a large current which may prevent the battery supplying enough current to start the car and could also result in damage to the battery.

When the car engine is running, the alternator, driven by the fan-belt, recharges the battery.

The Nickel-Cadmium Cell

In this cell the negative electrode consists of cadmium oxide pressed into a steel plate and the positive electrode consists of nickel oxide in a steel plate. The container is also made from steel and contains the electrolyte, potassium hydroxide, Fig. 3.10. The normal emf is about 1.2 V.

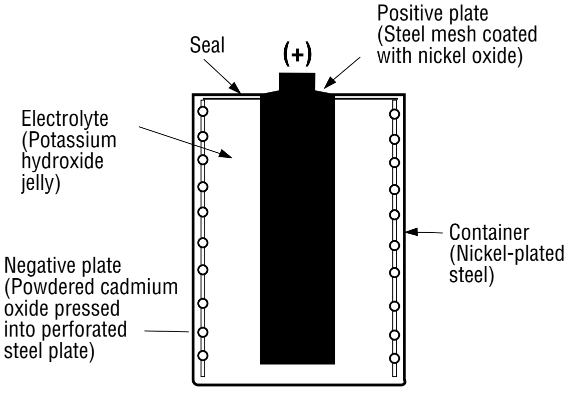


Fig. 3.10 *Nickel-cadmium cell*

During charging and discharging, the density of the electrolyte remains constant. The cells themselves are mechanically strong and do not require the same care as the lead-acid type to maintain them in good working order. When charging a nickel-cadmium cell a constant current must be used as otherwise a build-up of gas inside the cell could rupture the casing and cause injury. Only specially-designed chargers must therefore be used (see Appendix B). Under no circumstances should a nickel-cadmium cell ever be charged from a battery-charger of the type used for lead–acid cells, or from an ordinary mains power supply unit.

Some portable radios and torches contain a sealed type of nickel-cadmium cell which may be recharged as required by plugging into the a.c. mains supply using a special adaptor which steps down the voltage of the a.c. and rectifies it.

The Ni-Fe (Nickel-Iron) Accumulator

The emf of this cell is about 1.2 V, falling to 1.0 V when fully discharged. The positive plate contains nickel oxide and the negative plate contains finely divided iron. In both plates, the active material is held in pockets of finely perforated steel which results in much greater mechanical strength than in the lead-acid cell. The electrolyte is a solution of potassium hydroxide.

The NiFe accumulator is much lighter than the lead-acid cell and can withstand heavy charge or discharge currents without damage. These factors, coupled with its greater mechanical strength and freedom from problems of chemical pollution (e.g. sulphating), are leading to its increasing use for many applications, e.g. in some electric vehicles and as standby batteries for emergency lighting in large buildings.

The Ampere

‘The ampere is that steady current which, when flowing in two infinitely long, straight, parallel conductors of negligible circular cross-section, placed one metre apart in a vacuum produces a force between them of 2×10^{-7} newton per metre length of conductor.’

The ampere is one of the base units of the SI system. Its definition is derived from the (experimental) relationship for the force per unit length between two current-carrying wires, viz.

$$\frac{F}{l} \propto \frac{I_1 I_2}{d}$$

$$\frac{F}{l} = \frac{k I_1 I_2}{d}$$

The definition of the ampere makes $k = 2 \times 10^{-7}$. This value was chosen for historical reasons to make the SI ampere equal to the ampere that earlier research workers had defined.

The apparatus used in the standardising laboratories for fixing the value of the ampere is the current balance, one type of which is shown in Fig. 3.11. Ammeters used in laboratories are (indirectly) calibrated relative to a carefully maintained current balance.

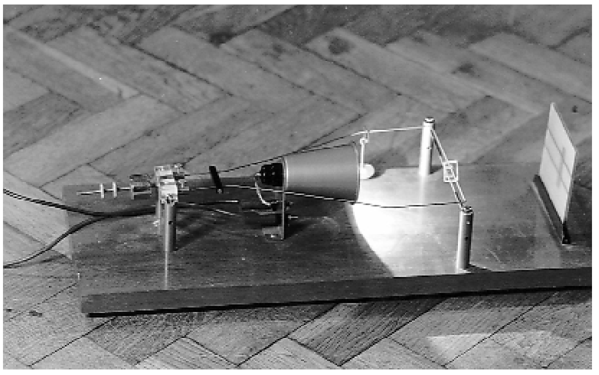


Fig. 3.11 *Current balance*

To demonstrate the force between current carrying conductors, two parallel strips of aluminium foil are arranged as shown, Fig. 3.12(a). When the current flows in opposite directions, the two lengths of foil repel each other. However, when current flows in the same direction through each strip, Fig. 3.12(b), an attractive force is experienced and the two strips jump towards each other.

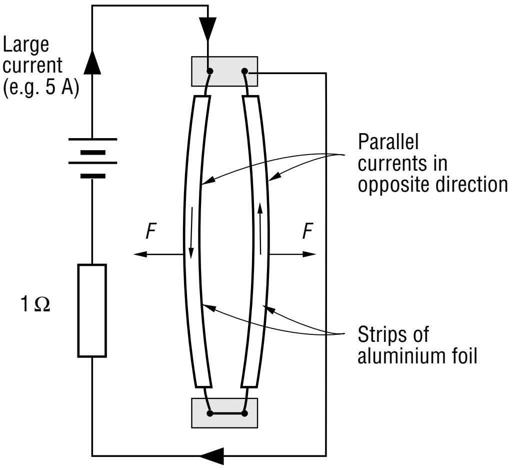


Fig. 3.12(a) *Repulsive force between currents*

Example

Two wires, each of length 50 cm, are arranged parallel to each other, 3 cm apart in air. What force is exerted by one on the other if a current of 10 A flows in each wire in opposite directions?

Solution

$$F = \frac{2 \times 10^{-7} I_1 I_2 l}{d}$$

$$= \frac{2 \times 10^{-7} \times 10 \times 10 \times 0.50}{.03}$$

$$= 3.3 \times 10^{-4} \text{ N}$$

The force is one of repulsion.

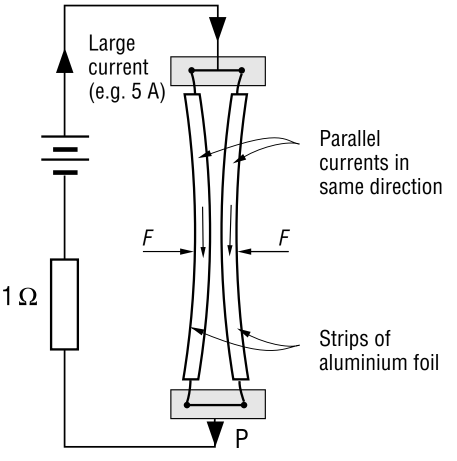


Fig. 3.12(b) *Attractive force between currents*

Demonstration

To Investigate How the Voltage Across a Battery is Affected When Current is Drawn from the Battery

Procedure

1. Set up the circuit shown in Fig. 3.13 using a new 9 V battery.

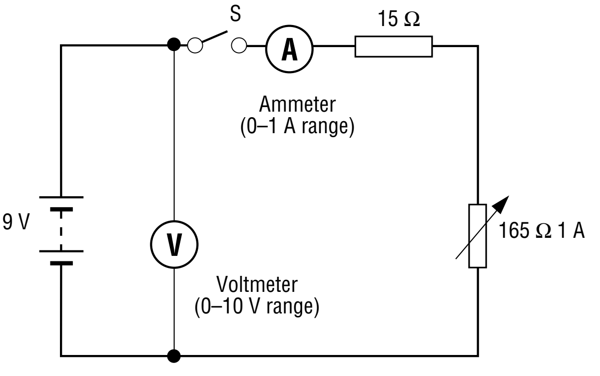


Fig. 3.13

2. Note the voltage across the battery when no current is being drawn (i.e. S open) from the battery. Record the value on the data table.
3. Close S and adjust the variable resistor until the ammeter registers 50 mA. Note V and again record the readings.
4. Take a series of readings at higher currents until voltage readings of less than 5 V are being recorded. (NB. To avoid draining the battery, ensure that for every reading, S remains closed for only as long as is needed to take the reading.)
5. Plot a graph of output voltage against the current, Fig. 3.14.
6. Calculate the slope $(V_2 - V_1)/(I_2 - I_1)$ by selecting two points on the graph. The magnitude of the slope equals the internal resistance of the battery.

Explanation

At any instant, the emf of the cell equals the sum of the voltage across the internal resistance and the voltage across the external circuit, i.e. $E = Ir + V$.

For the point (I_1, V_1) on the graph: $E = I_1r + V_1$

For the point (I_2, V_2) on the graph: $E = I_2r + V_2$

$\therefore\ I_2r + V_2 = I_1r + V_1$

$\Rightarrow V_2 - V_1 = r(I_1 - I_2)$

$$r = \frac{V_2 - V_1}{I_1 - I_2}$$

$$\Rightarrow \qquad r = \frac{-(V_2 - V_1)}{I_2 - I_1}$$

i.e. $\qquad r = -$ slope.

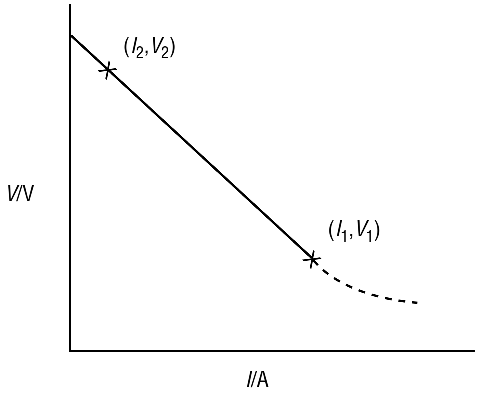


Fig. 3.14 *Variation in terminal voltage with current drawn*

Conclusion

The output voltage across the battery drops noticeably as increasing current is drawn. This is due to the increasing voltage drop across the internal resistance, r, of the battery as the current increases ($V = Ir$).

Towards the end of the useful life of a battery, the value of its internal resistance rises. As a result, the output voltage drops sharply as soon as current is drawn from such a cell.

3.3 Resistance

Temperature and Resistance

Temperature may be thought of as an indication of the ‘concentration’ of internal energy in a body. Thus, a large quantity of internal energy in a body of a given material and mass results in a higher temperature than when there is less internal energy in the same body. Internal energy is associated with the motion of the atoms and molecules of the substance. The molecules of all substances are in motion provided the temperature is above absolute zero. In solids, the motion is a vibrational one but in gases the molecules move about at high speeds.

As the temperature of a body rises, the amplitude of vibration of its atoms and molecules increases. This increase in the amplitude of vibration has many effects, two of which are now mentioned. Firstly, the greater amplitude of vibration impedes the progress of free electrons. Consequently, a rise in temperature means an increase in electrical resistance.

The second effect of the increased amplitude of vibration is to knock more electrons free from the atoms and thus to increase the supply of free electrons available for conduction. Since this is equivalent to less resistance, it follows that a rise in temperature also tends to lower the resistance of a body.

Whether the overall effect of the temperature rise is an increase or a decrease in resistance depends on the type of substance in which the temperature changes occur and on the temperature range in question. In all pure metals and in most metallic mixtures (i.e. alloys), there is an increase in resistivity (see p. 36) with rise in temperature, Fig. 3.15(a).

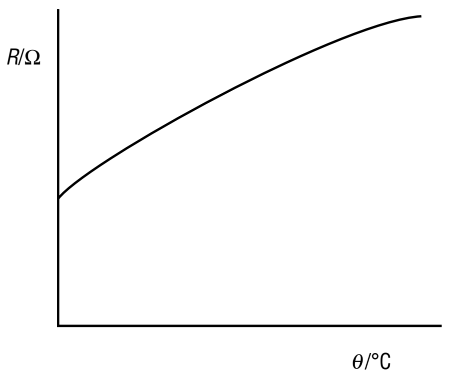


Fig. 3.15(a) *Increase in resistance with temperature, e.g. metal conductors*

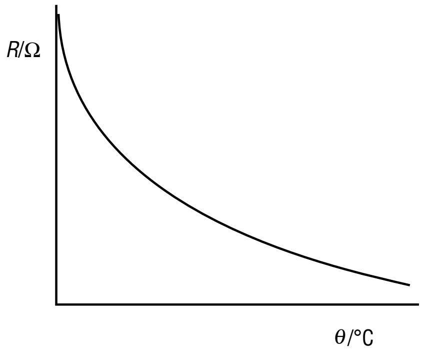


Fig. 3.15(b) *Decrease in resistance with temperature, e.g. semiconductors*

However, in some alloys, in insulators, in semiconductor materials, in carbon, and in liquids which have been made conducting by the addition of acids or salts (i.e. electrolytes), there is a decrease in resistivity with rise in temperature over some temperature range, Fig. 3.15(b).

Ohm's Law

Since the resistance of a conductor depends not only on its temperature but also on other physical conditions, e.g. strain, Ohm's law is stated as follows.

Under constant physical conditions, the potential difference across a conductor is proportional to the current flowing through it .

Not all conductors obey Ohm's law. Those which do, e.g. a length of nichrome wire, are known as ohmic devices or linear resistors.

Ohm's Law and Resistance

Methods for investigating the relationship between potential difference and current for various types of conductor are now examined.

Experiment 3.1

To Investigate How Current Varies with Potential Difference for a Length of Nichrome Wire

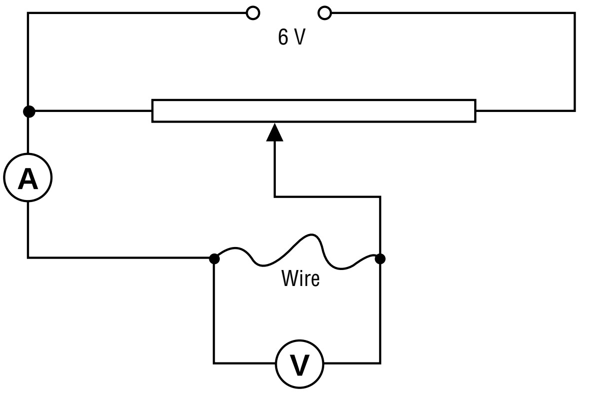


Fig. 3.16

Apparatus

As in Fig. 3.16. Power supply unit (6 V, d.c.), rheostat (e.g. 2 A, 33 Ω), ammeter (0–1 A) or milliammeter (0–500 mA), voltmeter (0–5 V) (or multimeters set to appropriate range), nichrome wire (approximately 50 cm of 28 swg wire; see Appendix L for s.w.g. i.e. standard wire gauge).

Procedure

As in text-books. In brief, adjust the potential divider (see p. 41) to obtain different values for V and hence for I. The following set of readings was obtained from such an experiment.

V/V	0.48	0.73	0.95	1.25	1.43	1.70	1.90	2.20
I/A	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45

(Note that the current was kept to quite small values to minimise temperature variations in the wire. With larger currents, the wire should be maintained at constant temperature by immersing it in a beaker of non-conducting liquid, e.g. distilled water.)

A graph was plotted of I vs. V as shown, Fig. 3.17.

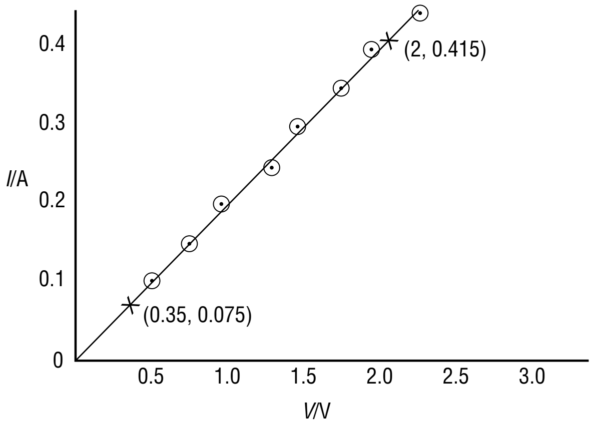


***Fig. 3.17** Graph of I vs V for a length of nichrome wire*

Result

A straight line through the origin was obtained.

$\Rightarrow I \propto V$

$\Rightarrow V/I$ = a constant, R, where R is the resistance of the wire.

$\Rightarrow$ Ohm's law is obeyed.

From the graph, the value of R may be determined.

$$\text{slope} = \frac{\text{change in current}}{\text{corresponding change in voltage}}$$

$$= \frac{\Delta I}{\Delta V}$$

$$= \frac{y_2 - y_1}{x_2 - x_1}$$

$$= \frac{0.415 - 0.075}{2.0 - 0.35}$$

$$= \frac{0.34}{1.65}$$

$$= 0.206$$

$$\therefore \quad R = \frac{\Delta V}{\Delta I} = \frac{1}{\text{slope}} = \frac{1}{0.206}$$

$$\therefore \quad R = 4.9\ \Omega.$$

A similar experiment may be performed using (a) a filament lamp, (b) acidulated water with platinum electrodes, (c) copper sulphate solution with copper electrodes, (d) a dry cell (e.g. 1.5 V), (e) a neon (gas) bulb (90 V type).

The resulting graphs from the above experiments are similar to those shown in Fig. 3.18(a) to (e).

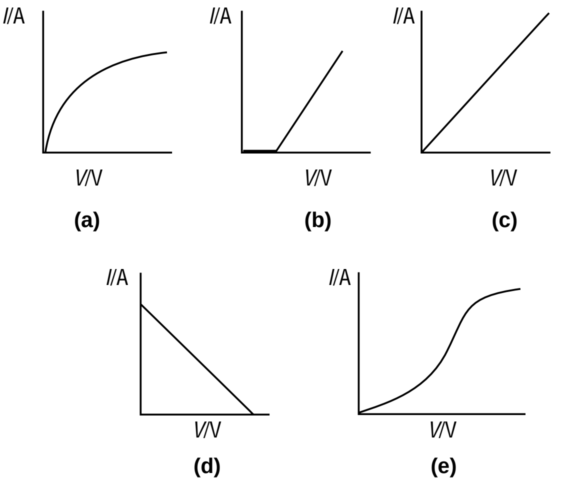


***Fig. 3.18** Variation of current with voltage for various devices*

Devices referred to in (a), (b), (d) and (e) above do not obey Ohm's law and are known as non-ohmic devices or non-linear resistors.

Conduction Through Gases

The passage of electricity through a gas (at low pressure) is called an electrical discharge. Gases that are commonly used in discharge tubes include neon – used in advertising signs, etc., sodium vapour – used in street lighting and laboratories, mercury vapour – used in street lighting and in fluorescent tubes. The phase tester used by electricians to check for the presence of a high voltage at a point in a circuit is simply a screwdriver containing a small neon lamp. When the metal end of the phase tester is pressed against a point at high potential and its metal cap earthed by the electrician's hand, the neon gas glows. (See Appendix I for neon oscillator demonstration.)

Electrolysis and Ohm's Law

The apparatus shown in Fig. 3.19 is recommended for investigating the relationship between potential difference and current for electrolytes. There is little risk of the distance between the electrodes being inadvertently altered during the course of an experiment.

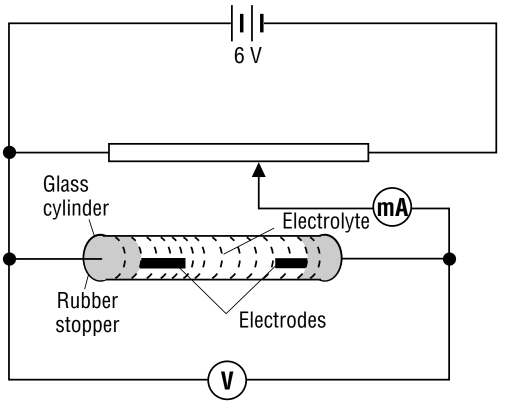


Fig. 3.19

If copper sulphate solution is electrolysed with platinum or carbon electrodes, copper is deposited on the cathode but the anode is not dissolved away; instead oxygen gas is evolved at it. Neither carbon nor platinum forms a sulphate and so each is said to be inactive (or insoluble) in the electrolysis of copper sulphate. Ohm's law is not obeyed when inactive electrodes are used. If a current is passed through an electrolyte using active (or soluble) electrodes, then Ohm's law is obeyed, e.g. copper sulphate solution with copper electrodes, Fig. 3.18(c). However, if a large current is used, the solution becomes non-uniform and the proportionality between *I* and *V* breaks down. Consequently, the best way to perform the experiment is to use a small d.c. current or to use an alternating current. As a result, no significant change in the uniformity of the electrolyte takes place.

Resistors

Definition: one ohm is the resistance of a conductor through which a current of 1 A flows when the potential difference across its ends is 1 V. (Since $R = V/I$, it follows that if $V = 1$ V and $I = 1$ A, then $R = 1$ V/A, i.e. 1 Ω.)

Resistance values of resistors range from very small, e.g. 0.01 Ω for a piece of copper wire, to very large values used in TV sets and other electronic equipment. A variable resistor such as a rheostat may have a resistance of 10 Ω whereas a rotary carbon track potentiometer used as a volume control in radio and TV equipment may have a total resistance of 1 kΩ. The nichrome element of a 1 kW electric fire has a resistance of about 60 Ω, whereas the tungsten filament of a 60 W lamp has a resistance of almost 1000 Ω. Wirewound (w/w) resistors are normally used to control relatively large currents whereas carbon and metal film resistors are used in low-current circuits.

The carbon track potentiometer contains a sector of a circle of resistive material (carbon) connected at each end to fixed terminals. A sliding contact is held against the resistive material by a spring leaf. As the position of the sliding contact is altered, so will different values of resistance be established between the sliding contact and each of the two fixed terminals.

Note. A preset or trimmer pot is a potentiometer whose resistance value is set (using a small screwdriver) when the circuit is initially tested and seldom, if ever, altered after that.

In most circuits it is unnecessary to have exact values of resistors and so manufacturers make only a certain number of preferred values having a particular tolerance, i.e. accuracy. Thus, a limited number of resistors provide maximum coverage with minimum overlapping. For example, the so-called E12 series involves resistors with a ±10% tolerance, which have the following twelve basic values: 1.0, 1.2, 1.5, 1.8, 2.2, 2.7, 3.3, 3.9, 4.7, 5.6, 6.8, 8.2. Resistors having these values are available in ohms, kilohms, megohms, and also with values 10 and 100 times greater.

Another series of resistors with a ±20% tolerance have preferred values 1.0, 1.5, 2.2, 3.3, 4.7, 6.8. The preferred values are chosen so that no resistor, whatever its actual resistance value, becomes a reject by reason only of its measured value. (See Appendix K for more data on resistors.)

Equivalent Resistance

When a number of resistors are connected together in a complex manner, their equivalent resistance can be worked out by treating the network section by section. The series or parallel formula for resistors may be applied to each section of the circuit in turn until the resistance of the whole combination is found.

Example

Find the equivalent resistance of the network shown in Fig. 3.20(a), given that the accumulator has an emf of 12 V and an internal resistance of 0.2 Ω. Calculate also, the current, *I*, that flows through the cell.

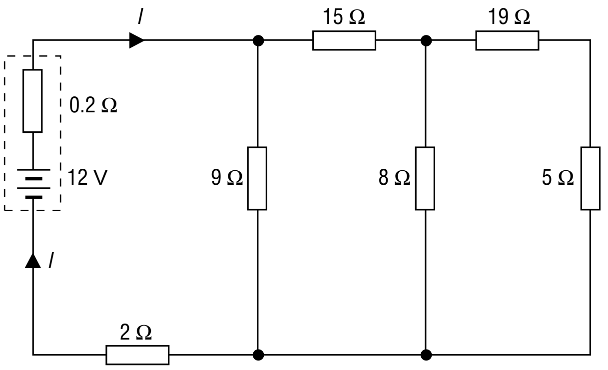


Fig. 3.20(a)

Solution

The 5 Ω and 19 Ω are in series, so their equivalent resistance is

$R_s = R_1 + R_2$

$= 5 + 19$

$= 24\ \Omega$

The network is now reduced to

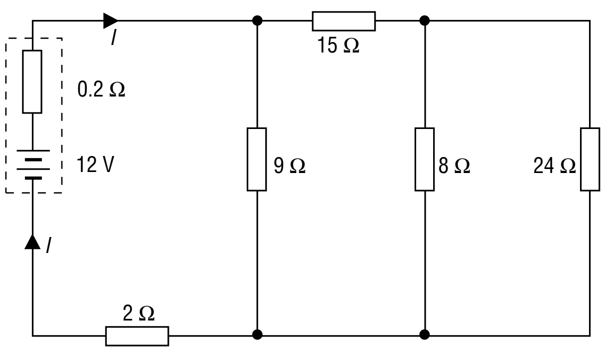


Fig. 3.20(b)

The 8 Ω and 24 Ω are in parallel, so their equivalent resistance is

$$\frac{1}{R_p} = \frac{1}{R_1} + \frac{1}{R_1}$$

$$= \frac{1}{8} + \frac{1}{24}$$

$$= \frac{4}{24} = \frac{1}{6}$$

$$\therefore R_p = 6\ \Omega.$$

The network is now reduced to

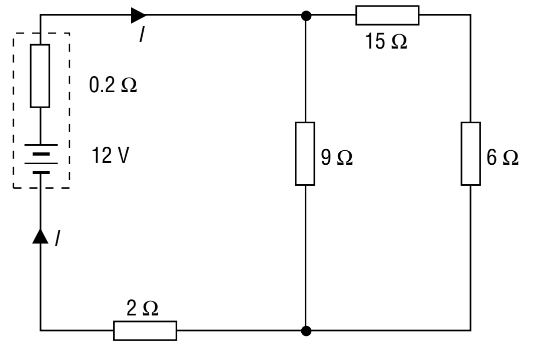


Fig. 3.20(c)

The 15 Ω and 6 Ω are in series.

$$R_s = 15 + 6 = 21\ \Omega.$$

The network is now reduced to

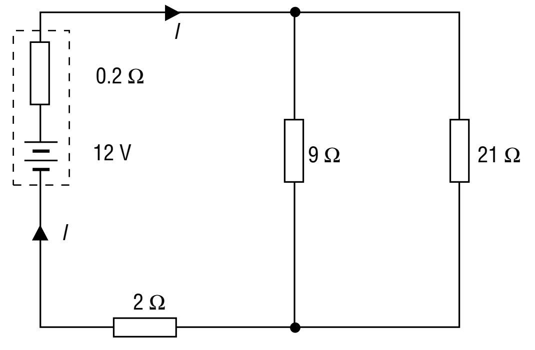


Fig. 3.20(d)

The 9 Ω and 21 Ω are in parallel.

$$\therefore \frac{1}{R_p} = \frac{1}{9} + \frac{1}{21} = \frac{10}{63}$$

$$R_p = \frac{63}{10}$$

$$= 6.3\ \Omega.$$

The network is now reduced to

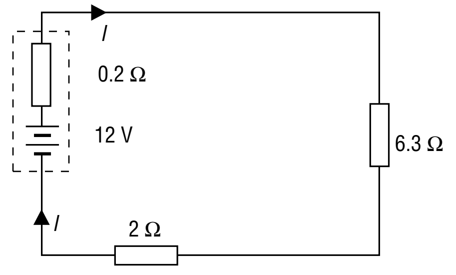


Fig. 3.20(e)

Since the three resistors, including the internal resistance of the cell, are in series the equivalent resistance is

$$R = 2 + 6.3 + 0.2 = 8.5\ \Omega$$

and so we may finally reduce the original network to

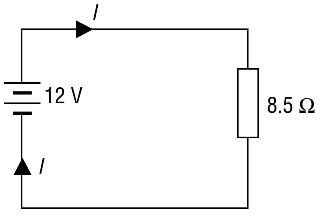


Fig. 3.20(f)

Clearly, $I = \dfrac{V}{R}$

$= \dfrac{12}{8.5}$

$= 1.41$ A

Answer

Equivalent resistance = 8.5 Ω.

Current through the cell = 1.4 A.

Note. Having obtained values for the equivalent resistance and the current in the network, any other values required may generally be obtained by referring to the appropriate circuit above. For example, if the voltage across the 9 Ω resistor is required, then Fig. 3.20(e) is the appropriate circuit to refer to.

For the 6.3 Ω resistor, $V = IR$

$= 1.41 \times 6.3$

$= 8.88$ V

The voltage across the 6.3 Ω resistor is 8.9 V. Therefore the voltage across the 9 Ω resistor is 8.9 V. (The 6.3 Ω is the equivalent resistance for the parallel branch in Fig. 3.20(d).)

Resistivity

The resistance, R, of a conductor at a given temperature depends on its length, l, and cross-sectional area, A, and on the material of which it is made. In fact, for a uniform conductor, it may be shown that

$$R \propto \frac{l}{A},$$

and so, $$R = \frac{\rho l}{A}$$

where ρ is a constant and is a property of the material of the conductor. It is called the resistivity of the material. It follows that resistivity is defined by the equation

$$\rho = RA/l$$

and the unit of ρ is the Ω m. From the defining equation it follows that resistivity may also be considered to be:

(i) the resistance of unit length of material of unit cross-sectional area, or

(ii) the resistance of a cube of material of side 1 m.

Resistivity Values

At room temperature silver has the smallest resistivity of any metal and the value for copper is not much greater. The resistivity of a pure metal is greatly increased by even small traces of impurity material and for this reason the copper used for electrical connecting wire is always electrolytically purified. Aluminium has a resistivity about twice that of copper but since its density is only about one third that of copper, an aluminium conductor can be much lighter and cheaper than a copper one. Some overhead power lines are made of aluminum and the lighter cables enable less expensive pylons to be used to support them. The cables contain a steel core to give them the necessary tensile strength to allow very long loops to be suspended between pylons; the aluminium strands are twisted around this core.

The range of resistivity values is enormous, as Table 3.1 indicates. Conductors have very low values of resistivity while the plastics (insulators) have very high values. Semiconductors have values between those of conductors and insulators.

	Material	ρ/Ω m
Conductors	Silver	1.6×10^{-8}
	Copper	1.7×10^{-8}
	Aluminium	2.7×10^{-8}
Semiconductors	Germanium	10^{-1}
	Silicon	10^{2}
Insulators	PVC	10^{12}
	Polythene	10^{14}

***Table 3.1** Resistivities at room temperature – values for semiconductors and insulators are approximate*

Uses of values

Once the resistivity of a material is known, lengths of it of known cross-sectional area can be used to make (i) standard resistors, i.e. resistors whose values are known to a high degree of accuracy, (ii) shunts which are used to convert moving-coil galvanometers to ammeters, etc., (iii) a particular resistance for an appliance.

Example

A coil of resistance 5 Ω is required to be made from a reel of manganin wire of diameter 0.40 mm. What length, l, of wire is required, given that the resistivity of manganin is 4.4×10^{-7} Ω m?

Solution

$R = 5\ \Omega$

$A = \pi r^2$

$= \pi (0.20 \times 10^{-3})^2\ \text{m}^2$

$\rho = 4.4 \times 10^{-7}\ \Omega\ \text{m}$

$$l = \frac{RA}{\rho}$$

$$= \frac{5 \times \pi (0.20 \times 10^{-3})^2}{4.4 \times 10^{-7}}$$

$= 1.43$ m

Thus, a length of 1.4 m of the wire can be wound into a coil of 5 Ω resistance.

Example

A steel wire of diameter 0.1 mm is shaped into a large circular ring of radius 20 cm. The terminals of a cell of emf 12 V and internal resistance 0.2 Ω, are connected to opposite points on the circumference of the ring. Calculate the current that flows in the circuit. (The resistivity of steel = 2×10^{-7} Ω m.)

Solution

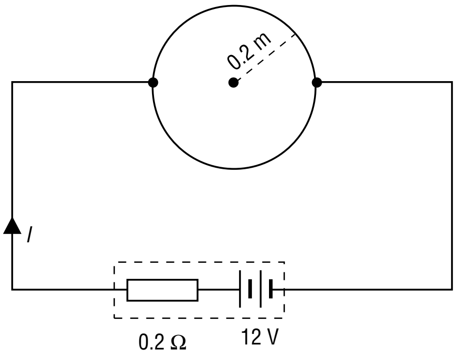


Fig. 3.21(a)

For the total length of steel wire:

$$R = \frac{\rho l}{A}$$

$$= \frac{2 \times 10^{-7} \times 2\pi \times (0.2)}{\pi \times (.05 \times 10^{-3})^2} \qquad (l = 2\pi r)$$

$= 32\ \Omega.$

Therefore, each half of the ring has a resistance of 16 Ω. Fig. 3.21(a) is equivalent to the parallel network shown in Fig. 3.21(b), which may be further reduced to Fig. 3.21(c) as follows.

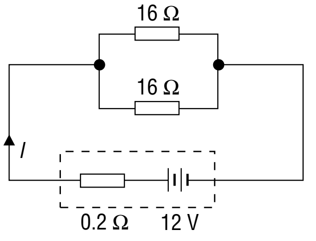


Fig. 3.21(b)

$$\frac{1}{R_p} = \frac{1}{R_1} + \frac{1}{R_2}$$

$$= \frac{1}{16} + \frac{1}{16}$$

$$= \frac{1}{8}$$

$R_p = 8\ \Omega$.

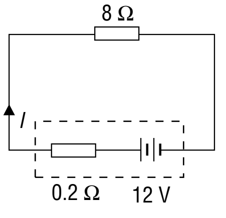


Fig. 3.21(c)

So, the total resistance in the circuit, Fig. 3.21(c), equals 8.2 Ω. Therefore, the current that flows in the circuit is given by

$$I = \frac{V}{R}$$

$$= \frac{12}{8.2}$$

$$= 1.46\text{ A}$$

The current is 1.5 A.

Experiment 3.2

To Measure the Resistivity of Manganin

Apparatus

Approximately 2 m of manganin 26 swg wire, ohmmeter, micrometer screw gauge, metre rule.

Procedure

1. Set the ohmmeter to the appropriate range, e.g. 200 Ω, and adjust the zero setting, if necessary.
2. Firmly fix the ends of the manganin wire in the terminals of the meter and measure the resistance, *R*, of the wire.
3. Measure the length, *l*, of the wire with a metre rule. Have the wire taut with no kinks in order to obtain an accurate value for *l*. *l* is the length of the manganin wire protruding from the terminals of the meter, i.e. allow 1 cm at each end of the wire for connecting to the terminals.
4. Measure the diameter of the manganin wire using a micrometer screw gauge. Check for zero error in the micrometer, Fig. 3.22. (There are 50 divisions on the barrel. One complete rotation equals 0.5 mm, so 1 division = 0.01 mm.) Ensure that the wire (with insulation removed) is held firmly in the jaws of the micrometer without over-tightening. (Tighten using the ratchet knob only, not the barrel.) Read the main scale and the circular scale on the micrometer. Taking any zero error into account, obtain a value for the diameter of the wire, Fig. 3.23. The diameter of the wire should be measured at several positions along its length and at each position, in two directions at right angles (in case the cross-section is not exactly circular). If the values obtained differ very little, their mean is taken as the effective diameter.

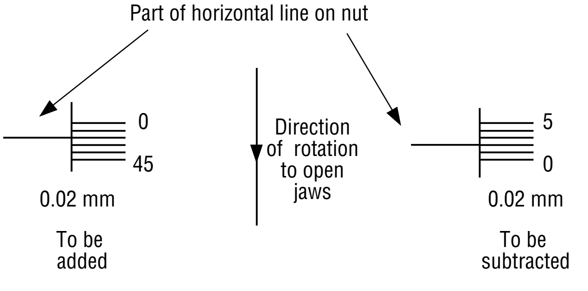


Fig. 3.22 *Finding zero error of micrometer scale*

5. Calculate the cross-sectional area, *A*, using $A = \pi r^2$, where *r* is the radius of the wire.

6. Substitute the measured values of R, A and l into the formula $\rho = RA/l$ and obtain a value for the resistivity of the manganin.

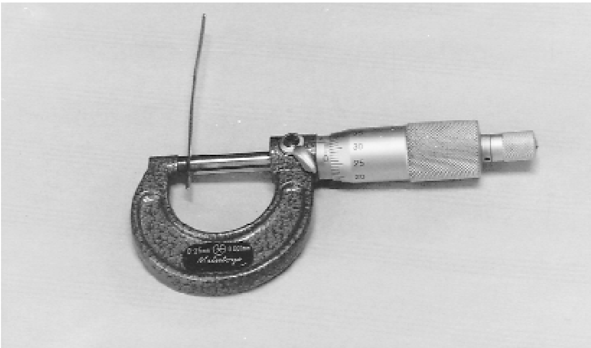


Fig. 3.23 *Measuring diameter of wire*

Example

A wire of resistance 5 Ω is drawn out to four times its original length. What is the resistance of the new length of wire, assuming that the diameter of the wire remains uniform?

Solution

Let D = diameter of the original length of wire.

d = diameter of the new length of wire.

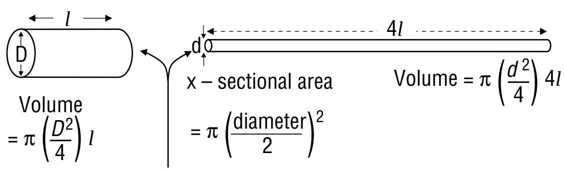


Fig. 3.24

Since the total volume of the wire remains constant

$$\frac{\pi D^2 l}{4} = \frac{\pi d^2 (4\,l)}{4}$$

$$D^2 = 4d^2$$

$$d = \frac{D}{2}$$

Apply $\rho = R\,A/l$

For the original length

$$\rho = \frac{5\pi D^2}{4l} \quad \text{(i)}$$

For the new length

$$\rho = \frac{R\pi d^2}{4(l)4}$$

$$= \frac{\pi R d^2}{16\,l}$$

$$= \frac{\pi R D^2}{16(l)4}$$

$$\therefore \quad \rho = \frac{\pi R D^2}{64l}$$

But, from Eqn (i)

$$\rho = \frac{5\pi D^2}{4l}$$

$$\therefore \quad R = \frac{5 \times 64}{4}$$

$$= 80\ \Omega.$$

The Wheatstone Bridge

A Wheatstone bridge circuit provides us with a method for measuring resistance accurately. It consists of a battery, a galvanometer and a network of four resistors, three of which are known and the fourth is to be determined, Fig. 3.25.

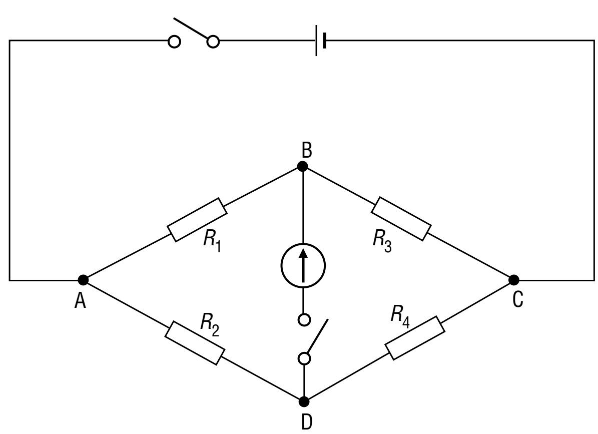


Fig 3.25 *Wheatstone bridge*

The combination is called a bridge because the galvanometer circuit is ‘bridged’ across the two parallel branches ABC and ADC. By adjusting the values of the resistances, the current in the galvanometer is made zero, as indicated by zero deflection, when both switches in Fig. 3.25 are closed. The bridge is then said to be balanced, i.e. the points B and D are at the same potential. For a balanced bridge it may be shown that

$$\frac{R_1}{R_2} = \frac{R_3}{R_4}.$$

R_1 and R_2 are often called the ratio arms of the bridge because their resistances determine the ratio of R_3 to R_4.

A factor limiting the accuracy of the Wheatstone bridge is the presence of contact resistances at the points where the resistors are joined in the circuit. It is difficult to ensure that a screwed-down or soldered connection has a resistance of less than 0.002 Ω, and the circuit unavoidably includes several such resistances. When the resistance to be found is less than 1 Ω, the errors due to connection resistances of this type are likely to be relatively large and a modified bridge method should be used.

The Wheatstone bridge forms the basis for many measurement and control devices used in industry. Any quantity such as temperature, humidity, strain, displacement, liquid level in a tank, etc., which results in change in the value of a resistance can be measured with a Wheatstone bridge circuit.

The well-known manner in which resistance varies with temperature is the principle behind resistance thermometers. A thermistor (or a coil of fine platinum wire contained in a porcelain tube, etc.) is connected in the R_4 arm of the Wheatstone bridge. It is then placed in the region whose temperature is to be measured. If the temperature of the thermistor varies, its resistance changes and the out-of-balance current in the bridge circuit is registered on the (calibrated) galvanometer and read as degrees Celsius. Alternatively, the out-of-balance current could be used to activate an electromagnetic relay which could in turn shut off power to an overheating device, e.g. an electric motor. Thus, the Wheatstone bridge may be used to measure (or control) temperature and also as a fail-safe device.

A method of examining the strains in aircraft surfaces and the surfaces of roads has been developed which utilizes the change of resistance of a wire with strain. The apparatus is called an electric strain gauge and consists of a fine eureka or nichrome wire in a Wheatstone bridge circuit. The wire is glued to the surface of an aeroplane, for example, and as the surface is strained, the metal wire is extended or compressed, resulting in a change in its resistance. This change bears a definite relationship to the strain of the aeroplane surface. By attaching wires to different parts of the surface, the strains at many locations may be measured simultaneously when an aircraft is being tested.

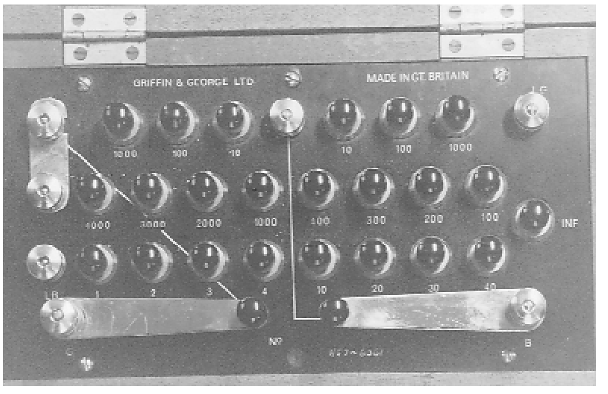


Fig. 3.26 *Post-office box*

Notes

1. In the Wheatstone bridge circuit, the galvanometer is used only as a null indicator, to show when the current is zero.
2. The Wheatstone bridge was popularised by Sir Charles Wheatstone (1802–1875), who used the bridge to locate breaks and short circuits in telegraph lines. An old-fashioned type of bridge with plugs rather than switches for changing resistance values, is called the Post-office box, Fig. 3.26.

Example

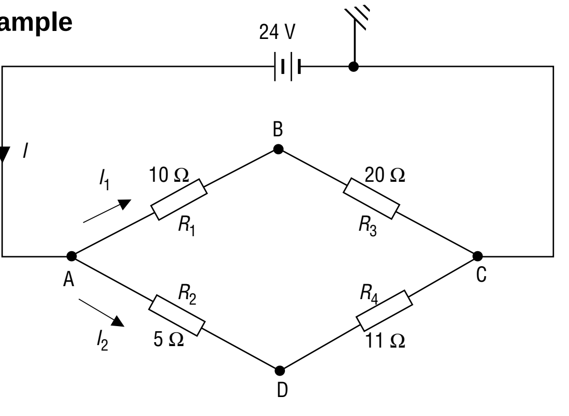


Fig. 3.27

A Wheatstone bridge circuit that is almost balanced is shown in Fig. 3.27. The negative side of the battery is earthed.

(a) Calculate the current that flows through each branch of the bridge.

(b) Calculate the potential of point B and of point D.

(c) If a galvanometer is connected between B and D, in what direction does the current flow through it?

(d) For what value of the resistance, R_4, would the bridge be in balance ?

Solution

(a) For the branch ABC, the effective resistance is $R = 10 + 20 = 30\ \Omega$.

The current, I_1, through this branch is given by

$$I_1 = \frac{V}{R} = \frac{24}{30} = 0.8\text{ A.}$$

For the branch ADC, the effective resistance is $R = 5 + 11 = 16\ \Omega$.

The current, I_2, through this branch is given by

$$I_2 = \frac{V}{R} = \frac{24}{16} = 1.5\text{ A.}$$

(b) The potential drop across R_3 is given by

$$V = I_1R_3 = 0.8 \times 20 = 16\text{ V}$$

Since point C is at zero potential, i.e. earthed, the potential at point B is 16 V. Similarly, for R_4

$$V = I_2R_4 = 1.5 \times 11 = 16.5\text{ V}$$

i.e. the potential at point D is 16.5 volts.

(c) Since point D is at a higher potential (16.5 V) than point B (16 V), current will flow from D to B through the galvanometer.

(d) For the bridge to be balanced

$$\frac{R_1}{R_2} = \frac{R_3}{R_4}$$

$$R_4 = \frac{R_2R_3}{R_1} = \frac{5 \times 20}{10} = 10\ \Omega.$$

Potential Divider

A potential divider may consist of a variable resistor or two (or more) fixed resistors in series. It divides the input voltage applied across it so that the output voltage it supplies is a certain fraction of the input voltage. For example, an arrangement that might be used for comparing the readings of two voltmeters is shown in Fig. 3.28, (see also, the potentiometer, page 43).

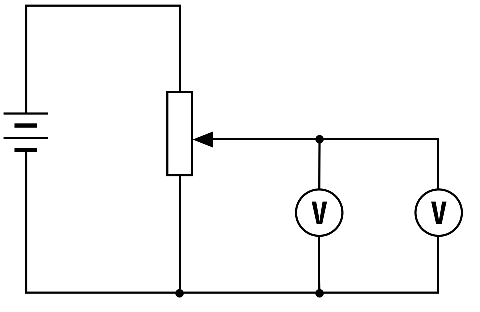


Fig. 3.28 *Comparing readings on two voltmeters*

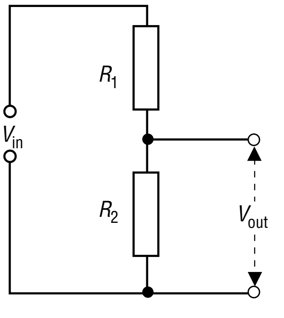


Fig. 3.29 *Potential divider circuit*

A general type of potential divider circuit is shown in Fig. 3.29. R_1 and R_2 form the potential divider to the input voltage, V_{in}. The output voltage, V_{out}, is taken off across R_2 and can be shown to be given by

$$V_{out} = \frac{V_{in}R_2}{R_1 + R_2}$$

Example

If V_{in} = 9 V, R_1 = 30 kΩ, and R_2 = 20 kΩ, then

$$V_{out} = \frac{9(20 \times 10^3)}{(30 \times 10^3) + (20 \times 10^3)}$$

$$= \frac{9 \times 20}{50}$$

$$= 3.6 \text{ V.}$$

As a general example in potential divider computation, consider the requirements of the circuit as illustrated in Fig. 3.30.

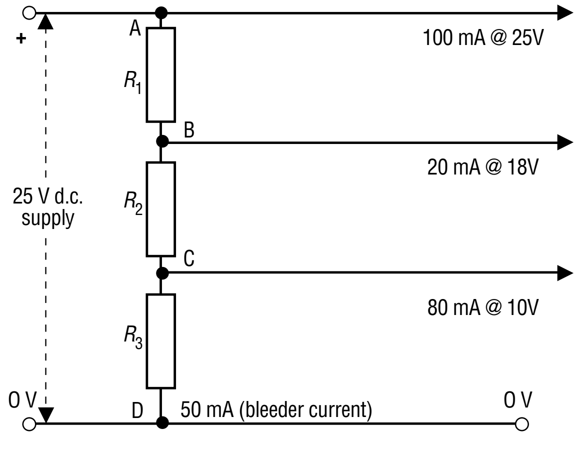


Fig. 3.30

It is required to provide for a load current of 100 mA at 25 V, 20 mA at 18 V and 80 mA at 10 V. What are the required resistance values of R_1, R_2 and R_3? What should be the power rating of each resistor?

Solution

Begin by examining R_3. From the circuit, it is seen that 50 mA flows through R_3 and the potential difference across it is 10 V. Note that the current through R_3 flows all the time the power supply is switched on and not just when current is being drawn by the three loads. This current through R_3 is known as the bleeder current and before a potential divider chain can be designed, some value for the bleeder current must be chosen. Generally, the value chosen for this is approximately 25 % of the total current (i.e. 200 mA in this case) taken from the tapping points but it may be chosen as low as 5%. The higher its value, the more the power wastage in the chain; the lower its value, the more prone to variation are the potential differences at the tapping points in the event of fluctuations in the supply voltage. In this circuit, 50 mA is *chosen* as the bleeder current.

$$R_3 = \frac{V}{I}$$

$$= \frac{10}{50 \times 10^{-3}}$$

$$= 200\ \Omega.$$

R_2 carries the 50 mA bleeder current and the 80 mA drawn at tap C, i.e. a total current of 130 mA.

$$R_2 = \frac{V}{I},$$

$$= \frac{8}{130 \times 10^{-3}} \quad \text{(p.d. between B and C = 8 V)}$$

$$= 62\ \Omega.$$

Similarly, R_1 carries 20 + 80 + 50 = 150 mA.

$$R_1 = \frac{V}{I}$$

$$= \frac{7}{150 \times 10^{-3}} \quad \text{(p.d. between A and B = 7 V)}$$

$$= 47\ \Omega.$$

Thus, R_1 = 47 Ω, R_2 = 62 Ω and R_3 = 200 Ω.

Note: these values apply only for the specific voltages *and* currents given in Fig. 3.30.

The power rating (i.e. the power dissipated) of each resistor may be found as follows.

$P = IV$

For R_1 $\quad P = 150 \times 10^{-3} \times 7$

$= 1.1$ W.

For R_2 $\quad P = 130 \times 10^{-3} \times 8$

$= 1.0$ W.

For R_3 $\quad P = 50 \times 10^{-3} \times 10$

$= 0.50$ W.

Since it is customary to allow for a safety factor of 100 %, the power rating of the resistors should be approximately double the value for the dissipation. Consequently, R_1 and R_2 should be 2 W resistors and R_3 should be 1 W.

The Potentiometer

A slide wire potentiometer may be used as a variable potential divider.

A uniform fall of potential occurs along a wire of uniform cross-sectional area when a steady current flows through it. The potential difference, V, between any two points is proportional to the length, l, of wire between those two points. This may be verified experimentally using the arrangement shown in Fig. 3.31.

Note that a resistance may be used in series with a galvanometer as a means of measuring potential difference. This serves as a useful teaching point, i.e. that a galvanometer and resistor (referred to as a multiplier in this instance) combination functions as a voltmeter.

Experiment 3.3

To Verify That Potential Difference is Proportional to Length for a Uniform Wire

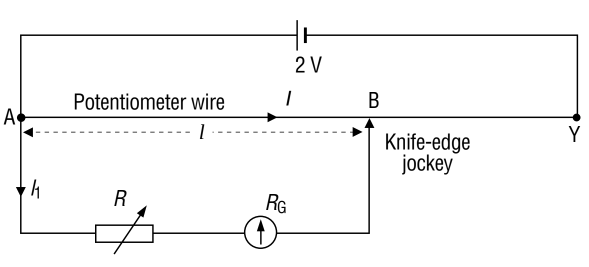


Fig. 3.31

Procedure

1. The apparatus is set up as shown in Fig. 3.31 and the value of R is adjusted so that the galvanometer gives almost a full-scale deflection (f.s.d.) when the complete wire is under test, i.e. the knife-edge jockey is at Y. This is to ensure a good distribution of points in the subsequent graph.
2. The potentiometer wire is then tapped at 10 cm intervals and the corresponding deflections on the galvanometer noted and tabulated.
3. The battery connections are reversed and the deflection readings checked.
4. A graph is plotted of length against deflection (the average of the two deflections is taken in each case).

Result

A straight line passing through the origin is obtained, showing that the deflection, θ, is proportional to the length of wire, i.e. $\theta \propto l$.

Conclusion

From Fig. 3.31 above it is seen that

$$V_{AB} = I_1(R + R_G),$$

i.e. $V_{AB} \propto I_1$ since $(R + R_G)$ remains constant.

I_1 is the current that flows through the galvanometer. But, $\theta \propto I_1$ for the moving coil galvanometer and so:

$$V_{AB} \propto \theta$$

From the (expected!) results of the experiment

$$\theta \propto l$$

$$\Rightarrow V_{AB} \propto l,$$

$\therefore V \propto l$ for the potentiometer wire.

(The resistance and the galvanometer combination in Fig. 3.31 may be replaced with an ordinary moving-coil voltmeter and the experiment repeated.)

3.4 Semiconductors

The operation of semiconductor devices depends on the effects that occur when p-type and n-type semiconductor material are in close contact. This is achieved by taking a single crystal of silicon (or germanium) and doping separate but adjacent layers of it with suitable impurities. The junction between the p-type and the n-type layers is referred to as the *p–n junction* and this is the key to some very important aspects of semiconductor theory. Devices such as diodes, transistors, silicon-controlled rectifiers, etc., all contain one or more p–n junctions.

The P–N Junction

At the junction of p-type and n-type material electrons diffuse across from the n-type region to the p-type region. Similarly, there occurs a diffusion of positive holes from the p-type to the n-type region. Many of these migrating electrons combine with holes on the journey through the junction and so effectively cancel out. Others, however, travel a little further before recombining. (The time for which a charge remains *'free'* is thought to be on average less than 10^{-6} seconds.) As a result, the n-type material near the junction becomes positively charged and the p-type material becomes negatively charged, Fig. 3.32. Due to the loss of mobile charge carriers on either side of the junction, the region is referred to as the *depletion layer*.

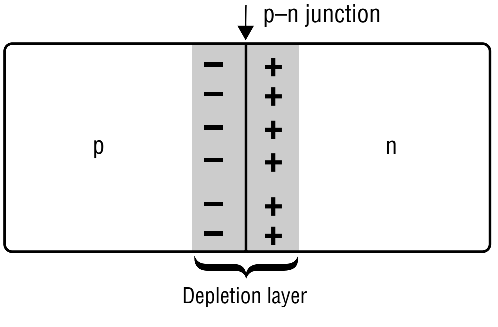


Fig. 3.32 *p–n junction*

Eventually a stage is reached when the charges built up across the junction are sufficient to repel further positive holes or electrons from crossing over. The value of the *potential barrier* arising from the charges remaining on each side of the junction depends on the crystal material and is called the *junction voltage*. It has a value of 0.6 V in the case of silicon.

Reverse Biased P–N Junction

If an external d.c. supply is connected across a p–n junction with its positive terminal joined to the n-type side and its negative terminal to the p-type side, it reinforces the potential barrier. As a consequence, holes in the p-region and electrons in the n-region are attracted towards the d.c. supply terminals. Thus the width of the depletion layer is increased, Fig. 3.33(a). This is called reverse biasing the junction – the word 'biasing' may be interpreted as 'influencing'. The resistance of the junction becomes very high.

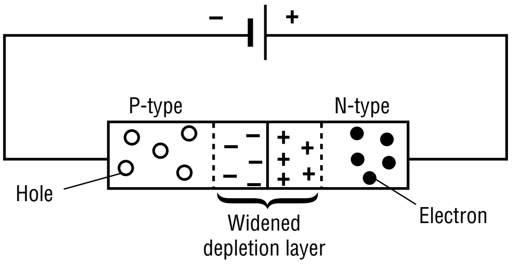


Fig. 3.33(a) *Reverse biased p–n junction*

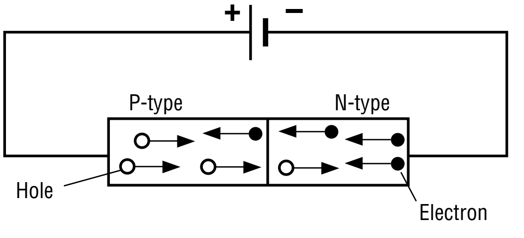


Fig. 3.33(b) *Forward biased p–n junction*

In Fig. 3.33(b) the d.c. supply reduces the width of the depletion layer. A significant current will flow if the battery voltage exceeds the junction voltage (0.6 V for silicon). The junction is said to be **forward biased**, i.e. the p-type side is connected to the positive terminal of the battery and the n-type side to the negative terminal. The resistance of this junction becomes very low.

P–N Junction Characteristics

The electrical characteristics of a typical p–n junction (silicon) are shown in Fig. 3.34.

When the junction is forward biased (i.e. forward voltage, V_F), the forward current, I_F, is small until

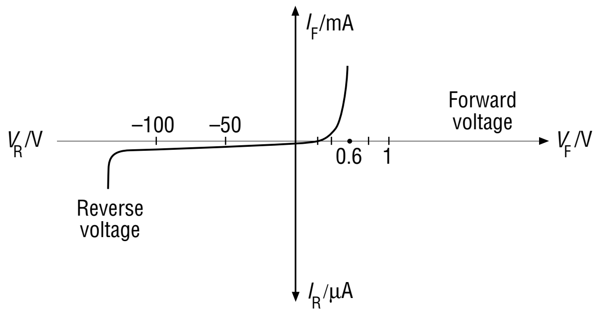


Fig. 3.34 *p–n junction characteristics*

V_F reaches 0.6 V. Beyond this value, a very small increase in V_F causes a large increase in I_F. The junction will be destroyed by excessive I_F due to high power dissipation at the junction.

When the junction is reverse biased, the reverse current, I_R, is negligible for low values of V_R (note the different scales on the negative axes of the graph). However, an excessive reverse voltage will cause junction breakdown. The diode will now conduct but it will be permanently damaged.

Note

A p–n junction diode (Fig. 3.35) consists of a p–n junction with one connection (called the anode, A) to the p-type material and another (called the cathode, K) to the n-type. Manufacturers produce many such diodes with different characteristics and ratings.

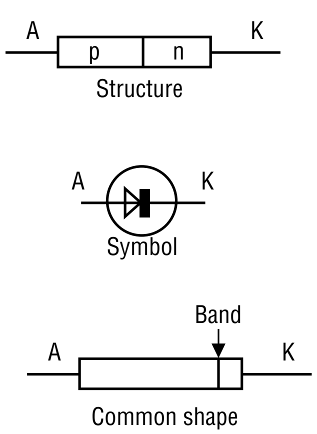


Fig. 3.35 *Junction diode*

One commonly used diode is the one coded IN4001. For this diode, the recommended maximum forward current, I_F = 1 A. The maximum reverse voltage, V_{RRM} = 50 V.

Uses

Because diodes only conduct in one direction, i.e. when forward biased, they are used:

(i) as rectifiers to change a.c. to d.c. in power supplies (see appendix A);

(ii) to prevent damage to a circuit by an inadvertently connected reversed voltage supply.

Some specialised diodes are manufactured for high-speed switching and many other applications.

Note. See Appendix A, page 70 for data on the Zener diode.

Demonstration 3.1

Forward and Reverse Biased Diodes

1. Set up the circuit shown in Fig. 3.36. Bread-board would be suitable for construction purposes.

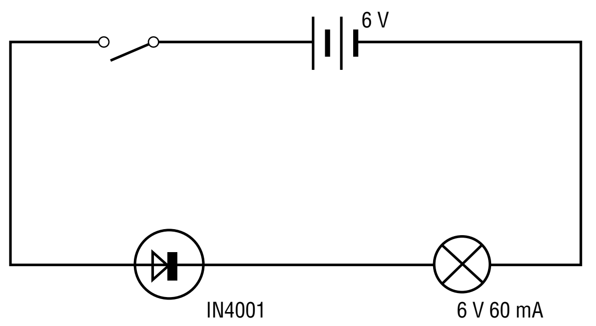


Fig. 3.36

1. Close the switch S and note that the lamp lights.
2. Reverse the diode (reverse bias) in its mountings and observe that the lamp does not light.

Conclusion

A diode conducts current when it is forward biased but not when it is reverse biased.

Demonstration 3.2

A Light Dimmer Switch

1. Set up the circuit shown in Fig. 3.37.

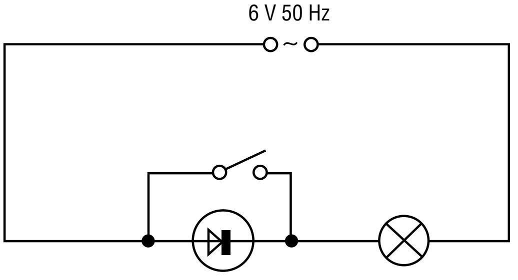


Fig. 3.37

1. Switch on the power and note that the lamp lights.
2. Close the switch and note that the lamp glows more brightly.

Explanation

With the switch closed, alternating current flows through the lamp which then lights brightly. However, when the switch is open, the lamp is not as bright because the diode only conducts one way (i.e. only when forward biased). As a result, only half-wave (pulsating) d.c. passes through the lamp. Thus, a diode in parallel with a switch, could be used as a very basic light dimmer.

Similarly, if the lamp in Fig. 3.37 above is replaced with a soldering iron, the iron will be 'on simmer' with the switch open but will quickly reach operating temperature when the switch is closed. Since a soldering iron is only used intermittently during circuit construction, this measure protects the iron from overheating and prolongs the life of the heating element.

The Light-Emitting Diode (LED)

Diodes which are designed to give a light signal are called light-emitting diodes. An LED (Fig 3.38) consists of a junction diode made from semiconducting material such as gallium arsenide or indium phosphide. Light is emitted from the junction when it is forward biased and the brightness is approximately proportional to the forward current. The colour of the light depends on the material used in the construction of the junction. Red, green and yellow LEDs are currently available.

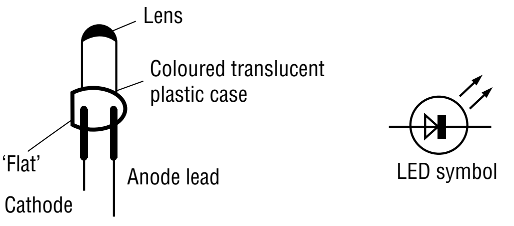


Fig. 3.38 *Light-emitting diode*

An LED does not light when reverse biased and the maximum permitted reverse voltages are low – typically 5 V. When an LED is forward biased a series resistor must always be used to limit the forward current to approximately 10 mA. The voltage drop across a conducting LED is approximately 1.7 V.

Example

Calculate the required value of the protective resistance, R, in Fig. 3.39. Assume a voltage drop of 1.7 V across the LED and that the forward current should not exceed 10 mA, i.e. 0.01 A.

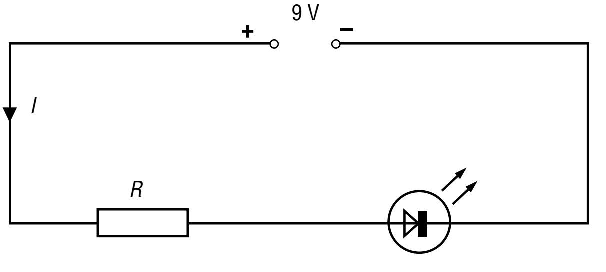


Fig. 3.39

Solution

$$9 = V_R + V_{LED}$$

$$= V_R + 1.7$$

$$\Rightarrow \quad V_R = 7.3 \text{ V}$$

$$V_R = IR$$

$$\Rightarrow \quad R = \frac{V_R}{I}$$

$$= \frac{7.3}{0.01}$$

$$= 730\ \Omega.$$

Uses

LEDs are widely used as indicator lamps in electronic systems to show whether an output is 'high' or 'low'. Many electronic devices show their 'output' by means of seven-segment LED displays. Each segment is an LED and, depending on which segments have a p.d. across them, the display lights up the numbers 0 to 9.

Advantages

The advantages of LEDs are: reliability, long life, low cost, small size, small operating current, and high switching speed.

Demonstration 3.3

A Simple Polarity Tester

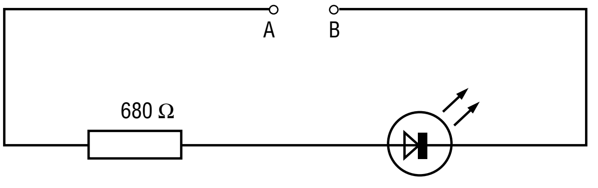


Fig. 3.40

If the LED lights in Fig. 3.40 when a battery is placed across A and B, then terminal A is the positive terminal of the battery. If the LED does not light, then A is the negative terminal. Note that the maximum reverse voltage which should be applied to an LED is 5 V.

Demonstration 3.4

A Simple Continuity Tester

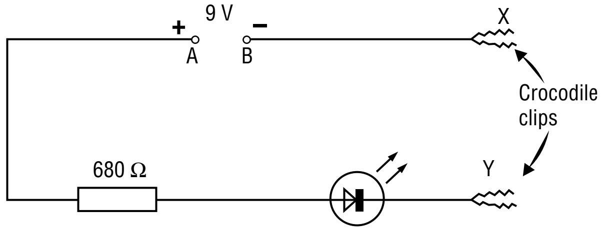


Fig. 3.41

The circuit in Fig. 3.41 may be used to check for blown fuses or lamps, etc. Simply grip both ends of a fuse with the crocodile clips at X and Y. If the LED does not light, the circuit lacks continuity and so therefore, the fuse wire must be broken. (Always test the circuit initially by short-circuiting X and Y.)

The Light-Dependent Resistor (LDR)

The influence of light on many semiconductor materials results in the creation of free holes and electrons. Light, in other words, has a similar effect to that of heat. For this reason, most semiconductor devices are provided with a light-proof coating or case. Alternatively, the effect can be put to good use in devices intended to react to changes in the intensity of illumination, such as phototransistors and light sensitive cells.

An LDR is a device whose resistance varies from a high value in the dark (somewhere in the range 500 kΩ to 10 MΩ) to a lower value in bright light (in the range 1 kΩ to 100 kΩ). The symbol for an LDR is given in Fig. 3.42.

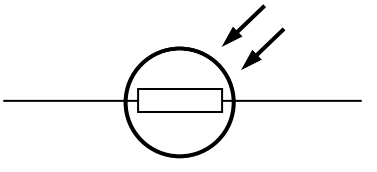


Fig. 3.42 *Light-dependent resistor*

The resistance changes non-linearly with the light intensity and the response of an LDR is considered to be somewhat sluggish, i.e. it might take 0.1 seconds to respond to a pulse of light! An LDR functions as an input transducer, i.e. a non-electrical input (light) results in an electrical output. A commonly encountered LDR is one identified by the code ORP12 which is also referred to as a cadmium sulphide (photoconductive) cell. Materials (other than CdS) used in the construction of LDRs include cadmium selenide and lead sulphide. They vary in their response to different wavelengths and are widely used as flame detectors in central heating boilers and furnace control systems because of their sensitivity to red light (and also to infrared radiation).

The response of an LDR to different wavelengths and light intensities may be investigated by directing a beam of light through colour filters onto the surface of an LDR. The resistance may be measured using an ohmmeter set to the appropriate range.

Thermistors

Carbon type resistors have a small temperature coefficient of resistance (see Appendix C). Thermistors are resistors made from semiconducting material which has a very large temperature coefficient of resistance. The resistance of a p.t.c. (positive temperature coefficient) thermistor increases appreciably as the temperature of the material increases. The converse holds for an n.t.c. (negative temperature coefficient) thermistor. The symbol for a thermistor is shown in Fig. 3.43.

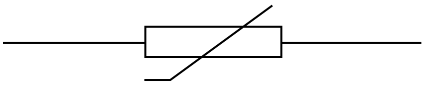


Fig. 3.43 *Thermistor symbol*

The applications of thermistors revolve around temperature measurement and control. They may be used for current regulation, stabilising oscillators and for providing temperature compensation for any changes in the resistance of the windings of motors and generators, etc.

The response of a thermistor to temperature changes may be demonstrated by connecting it to an ohmmeter. The thermistor may be heated with a match or cooled using a 'freezing' spray.

The Transistor

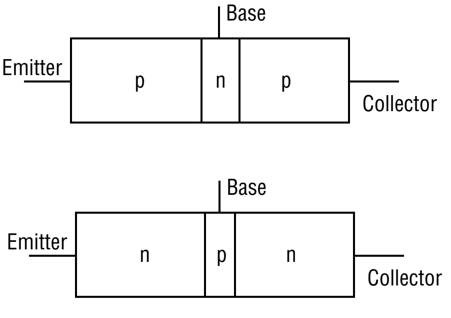


Fig. 3.44 *Transistor structures*

Transistors are three-terminal semiconductor devices originally made from germanium but the vast majority are now made from silicon. They are used for amplifying and for switching, e.g. in alarms and oscillators. They basically consist of two p–n junctions formed very close together in a three-layer arrangement, resulting in either a pnp or npn configuration, as shown in Fig. 3.44.

In the construction of a transistor, all three layers are 'grown' in one crystal. The central layer (the base) is very thin (a few μm) and lightly doped. The two outer layers (the emitter and the collector) are more heavily doped. Note that the two junctions are back to back: in the case of the pnp transistor, a p–n junction is followed by an n–p junction. The device can thus be considered as two semiconductor diodes connected back to back. This type of transistor is known as a bipolar transistor.

The symbols used in circuit diagrams to represent transistors are shown in Fig. 3.45.

Fig. 3.45 *Transistor symbols*

We now study the operation of some circuits involving npn transistors. With a pnp device, the operation is similar except that the biasing arrangements are of the opposite polarity.

The Transistor as a Current Amplifier

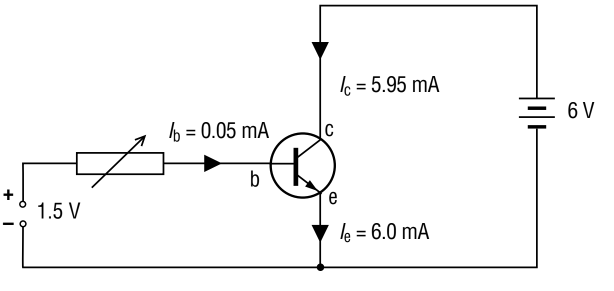


Fig. 3.46 *Current amplifier*

Bipolar transistors may be connected in a number of different ways in circuits. When the emitter is common to both the input and the output circuits, then the transistor is said to be connected in what is called the common-emitter mode (Fig. 3.46). This configuration gives a high power gain and is used more often than any other. The common-base and common-collector modes are used when certain other characteristics are required.

The principle of a transistor is that current (I_c) flows between the collector and emitter terminals only when a current (I_b) is flowing between the base and the emitter terminals, i.e. the bipolar transistor is a **current-operated device**.

I_c is much greater than I_b and so the transistor functions as a current amplifier. The ratio of these two currents is called the d.c. current gain (h_{FE}), i.e.

$h_{FE} = \frac{I_c}{I_b}$ where I_c and I_b are d.c. quantities.

Depending on the type of transistor, current gain values may vary from around 25, as in the case of power transistors, to over 1000 for high-frequency amplifier types. For the bipolar transistor in Fig. 3.46 above h_{FE} = 5.95/0.05 = 119. (Note also that $I_e = I_b + I_c$.)

Demonstration 3.5

To Show Current Amplification by a Transistor

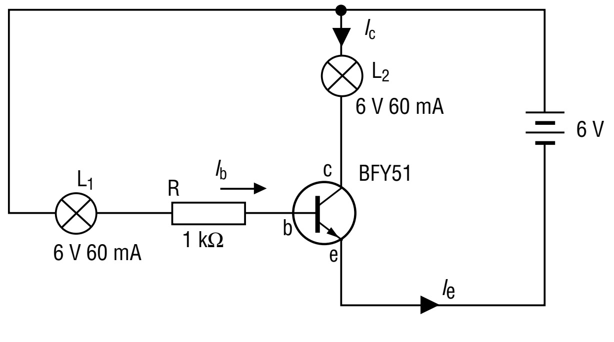


Fig. 3.47

Procedure

1. Set up (on breadboard) the circuit as shown in Fig. 3.47. Milliammeters may be included in the circuit, one (100 mA f.s.d.) in series with L_2 and another (1 mA f.s.d.) in series with L_1. Note that the base-emitter junction is forward biased and that the collector-base junction is reverse biased.
2. Switch on the power and note that lamp L_2 lights up but not L_1. This indicates that the collector current, I_c, is much greater than the base current, I_b (the milliammeter readings will clearly confirm this).
3. Remove L_1 so that the base circuit is incomplete and so I_b becomes zero. Note that L_2 goes out, indicating that I_c becomes (more or less) zero.

Conclusion

The small base current, I_b, flows through the lamp L_1 and the resistor R and switches on the transistor. This results in the flow of a relatively large collector current. (R acts as a protective resistor and prevents an excessive base current which would result in damage to, in particular, the fragile base region of the transistor.)

Note

The transistor BFY51 is an npn medium current transistor. The maximum average value of collector current (I_c max) the transistor can pass without damage is 1 A. The maximum total power rating (P_{tot}) is 800 mW.

The case or encapsulation of a transistor may be of metal (to aid in the dissipation of heat) or plastic. A wide variety of shapes and sizes are available. The outline for the BFY51 is shown in Fig. 3.48.

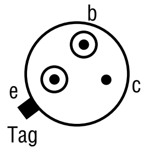


Fig. 3.48 *BFY51 viewed from underneath*

Viewed from underneath, the emitter, base, collector, are arranged clockwise from the tag. Some power transistors have only two separate leads – the metal case of the transistor serves as the third lead, e.g. the collector.

The Transistor as a Switch

Transistors have many advantages over other electrically operated switches such as relays, etc. They are small, reliable, cheap, have no moving parts, an almost indefinite life and they can switch on and off millions of times a second. Alarm circuits may be constructed in which some appropriate transducer, e.g. a thermistor or LDR, controls the operation of a transistor as a switch.

Demonstration 3.6

A Light-Operated Switch

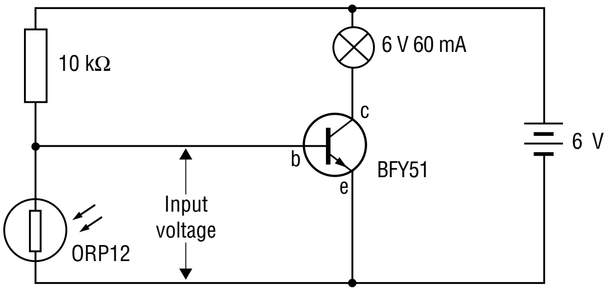


Fig. 3.49 *Light-operated switch*

1. Set up the circuit as shown in Fig. 3.49. A 6 V buzzer may be used instead of the 6 V lamp.
2. With the power switched on, and the LDR illuminated by a lamp or daylight, slowly move your hand closer to the LDR, thereby reducing the light intensity incident on its surface.
3. Note that the lamp gradually brightens and is fully on when no light reaches the LDR.
4. Slowly remove your hand and note that the light from the lamp gradually fades.
5. Replace the 10 kΩ resistor by a variable resistor and repeat the above procedure. It will be found that the light intensity level at which the lamp comes on can now be altered by moving the sliding contact.
6. Interchange the resistor and the LDR. The lamp will now be on in the light and off in the dark. This circuit could form the basis of an intruder alarm.

Explanation

The resistor and the LDR form a potential divider across the supply. The input applied between the base and the emitter of the BFY51 is the voltage across the LDR. This voltage depends on the resistance of the LDR.

In bright light, the resistance of the LDR is low (e.g. ~ 1 kΩ) compared with that of the resistor (10 kΩ). As a result, most of the supply voltage is dropped across the resistor and the input to the transistor, i.e. the voltage across the LDR, is then too small to turn on the transistor (~ 0.6 V is required for silicon).

In the dark, the LDR has a high resistance (~ 10 MΩ) and so more of the supply voltage is dropped across it and less across the resistor. The voltage across the LDR is then sufficient to switch on the transistor and produce a collector current large enough to light the lamp.

Demonstration 3.7

A Temperature-Operated Alarm

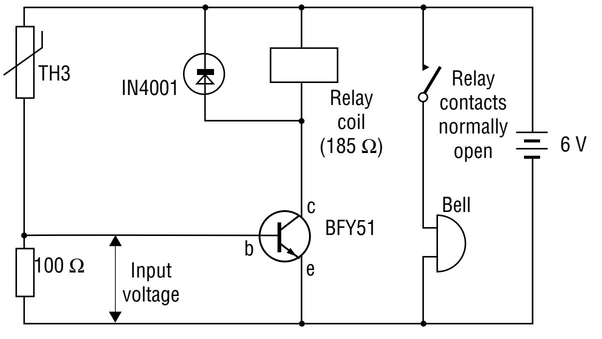


Fig. 3.50 *Temperature-operated alarm*

1. Set up the circuit as shown in Fig. 3.50. The thermistor (n.t.c. type, see p. 48) and the resistor form a voltage divider across the supply. The input to the transistor is the voltage across the resistor.
2. Place a burning match under the thermistor. After a few seconds, the 6 V electric bell will ring.
3. Spray the thermistor with a 'freezing' compound and the bell will cease ringing.
4. Interchange the thermistor and resistor and note the result.
5. Use different values of the resistor (e.g. use a variable resistor) and note the effect.

Explanation

When the thermistor is heated, its resistance drops, resulting in a smaller fraction of the supply voltage being dropped across it. Hence, the voltage drop across the resistor is increased and when this exceeds ~ 0.6 V, the transistor switches on and a collector current flows through the coil of the electromagnetic relay. As a result, the 'normally open' contacts close and the large current that the bell requires is obtained *directly* from the supply.

Note

(i) The diode short-circuits any (large) currents induced in the coil of the relay when the relay is switched off. Without the diode these currents would damage the transistor.

(ii) If the bell was placed in series with the transistor, the current flowing through the transistor would not be sufficient to operate the bell – hence the use of the relay.

The Transistor as a Voltage Amplifier

The function of an amplifier is to produce an output signal which has the same frequency and wave-shape as the input but which has a greater amplitude. When the output signal has a greater voltage amplitude than the input signal, the amplifier is said to have voltage gain, A, which is defined as the ratio of output signal voltage to input signal voltage, i.e.

$$A = \frac{V_{out}}{V_{in}}$$

When the output signal has a greater current amplitude than the input signal, the amplifier is said to have current gain, which is defined as the ratio of output signal current to input signal current.

When a bi-polar transistor is used for amplification, it may be used in the common-emitter mode, the common-collector mode or the common-base mode. The common-emitter amplifier is the type most often encountered, as it has what is generally the most useful set of characteristics. These include low input resistance, high output resistance and high values of voltage and current gain.

The normal function of a transistor with the base-emitter junction forward biased and the base-collector junction reverse biased is as a current amplifier. Voltage amplification is achieved by connecting a 'load' resistance R_L between the collector lead and the supply voltage, Fig. 3.51.

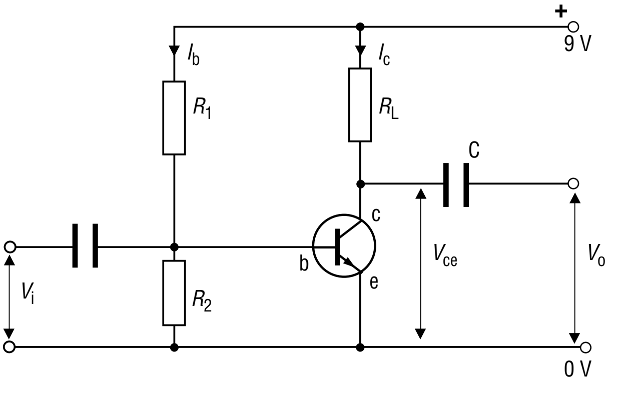


Fig. 3.51 *Voltage amplifier*

R_1 and R_2 form a potential divider which provides forward bias for the base-emitter junction. R_1 and R_2 are normally chosen so that the collector-emitter voltage (V_{ce}) is about half the power supply voltage, i.e. $V_{ce} \approx {}^1/_2 V_{supply}$. The small alternating voltage to be amplified, i.e. the input voltage V_i, is applied to the base-emitter circuit. This causes small changes in the base current which produce large changes in the collector current flowing through the load resistor. In practice, positive and negative swings of a few *millivolts* in V_i can result in a fall or rise of several *volts* in the voltage across R_L and therefore in the collector-emitter voltage as well. V_{ce} is therefore a varying direct voltage which may be regarded as an alternating voltage superimposed on a steady direct voltage. Only the alternating part is required and the capacitor C blocks the d.c. part but allows the a.c. part, i.e. the output, V_0, to pass on to the next stage.

It is important to note that the output is 180° out of phase with the input, i.e. when the input is at its maximum positive value, the output is at its maximum negative value, i.e. the amplifier functions as a *voltage inverter*.

(See Appendix G for an experiment to measure the voltage gain of an amplifier.)

Integrated Circuits (ICs)

The incredible advances made in the field of electronics have provided us with multi-media PCs, heart pacemakers, robots for industrial control, etc. Before the advent of microelectronics, individual components were made separately and then wired together to form the required (and often bulky) circuit.

With the development of the first integrated circuits (chips) in the early 1960s miniaturisation was on the way. An IC is a complete electronic circuit containing transistors and perhaps diodes, capacitors and resistors.

All of the components are formed at the same time on a chip of silicon, typically about 5 mm square and no more than 0.5 mm thick. The first ICs consisted of fairly simple circuits with fewer than 100 components per chip. They were classed as small-scale integrated (SSI) circuits. Today, very-large-scale (VLSI) ones consist of complex circuits with anything up to several million components.

Compared with circuits built from separate components, ICs are very much lighter, smaller, cheaper and more reliable. Unfortunately, their small size limits the power and voltage (~30 V maximum) they can handle.

Computers for Making Computers

As a consequence of the development of microelectronics, modern computers can perform a much wider variety of jobs than their predecessors. They are also faster, cheaper, more reliable and smaller. Paradoxically, silicon chip technology not only enables us to make much ‘larger’ computers because so much complex circuitry is packed into so little space, but it has also opened the way to the mass production of much smaller computers.

A microprocessor is a miniature version of the Central Processing Unit (CPU) of a digital computer and may be considered to be the simplest computer (of sorts) available. Due to the level of complexity and specialisation required, microprocessors can now only be designed by using computers which themselves probably contain microprocessors.

CHAPTER 4

EFFECTS OF AN ELECTRIC CURRENT

4.1 Heating Effect

Electric current may be detected by some definite effect to which it gives rise. One of these effects is the heating effect. Two others are the chemical and magnetic effects which will be discussed later.

Heat is produced in current-carrying conductors, resulting in an increase in temperature of the conducting material. The heating is as a result of the collisions between the moving free electrons and the relatively stationary atoms of the conductor material. As a result, heating increases rapidly with increase of current flow, since a greater rate of flow results in more collisions.

The heating effect of an electric current has many practical applications, e.g. in radiant electric fires, domestic irons, immersion heaters, etc.

Joule's Law

The rate at which thermal energy is produced in a resistor is proportional to the square of the current flowing through it, if the resistance is constant.

Apparatus

In the mandatory student experiment to verify Joule's law a heating coil is placed in a calorimeter containing water and is connected to a low voltage d.c. power supply. The rise in temperature of the water is determined for different values of the current, the mass of water being the same in each case.

Let m_1 and m_2 = mass of water and of calorimeter, respectively; c_1 and c_2 = specific heat capacity of water and of material of calorimeter respectively; Δt = time for which the current, I, flows; $\Delta\theta$ = rise in temperature. The energy supplied is

$$\Delta E = (m_1c_1\Delta\theta) + (m_2c_2\Delta\theta)$$

$$= (m_1c_1 + m_2c_2)\Delta\theta$$

By definition, power, P = energy converted per second.

$$P = \frac{\Delta E}{\Delta t} = (m_1c_1 + m_2c_2)\Delta\theta/\Delta t$$

If m_1, m_2, c_1, c_2, and Δt are constant, then $P \propto \Delta\theta$. Consequently, if a graph of $\Delta\theta$ vs. I^2 results in a straight line through the origin, Fig. 4.1, it may be concluded that $\Delta\theta \propto I^2$.

But, $P \propto \Delta\theta$

$\therefore \quad P \propto I^2$

Joule's law is verified.

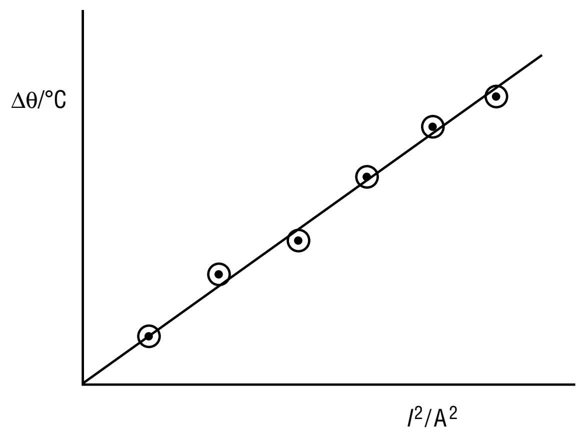


Fig. 4.1 *Variation of $\Delta\theta$ with I^2*

Example

An electric kettle draws a current of 10 A when connected to the 230 V mains supply. Calculate (a) the power of the kettle, (b) the energy produced in 5 minutes, (c) the rise in temperature if all the energy produced in 5 minutes is used to heat 2 kg of water. (Specific heat capacity of water = 4200 J kg^{-1} K^{-1}.)

Solution

(a) $P = IV$

$= 10 \times 230$

$= 2300$ W

$= 2.3$ kW.

(b) Energy produced in 5 minutes = Pt

$= 2300 \times 5 \times 60$

$= 690000$ J

$= 690$ kJ.

(c) Energy produced = energy gained by water

$690000 = mc\Delta\theta$

$= 2 \times 4200 \times \Delta\theta$

$$\Delta\theta = \frac{690000}{2 \times 4200}$$

$= 82.1$ K

$\therefore$ Rise in temperature of the water = 82 K.

Experiment 4.1

To Determine the Power of a Bulb Using the Formula $P = IV$

Apparatus

A low-voltage variable power supply unit capable of delivering at least 2 A is required. If a fixed voltage power supply unit is used, then a rheostat (e.g. 5 A, 25 Ω) must be placed in series with the bulb in order to vary the current. Ammeter (0–5 A), voltmeter (e.g. multimeter set to the 20 V range), and a variety of low voltage bulbs are also required.

Method

Set up the circuit shown in Fig. 4.2 using a bulb which is marked 12 V, 21 W. (A 21 W bulb converts electrical energy to heat and light at the rate of 21 joules per second.)

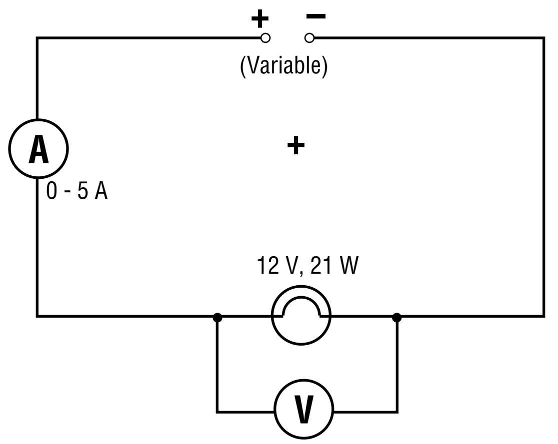


Fig. 4.2

Adjust the power supply unit until the voltmeter reads 12 V and then take the ammeter reading. Calculate the power, using the formula $P = IV$. The value obtained should be in close agreement with the markings on the bulb. Repeat the experiment using low voltage bulbs of different power ratings.

Experiment 4.2

To Determine the Fusing Current for a Wire Conductor

Apparatus

The wire conductor may be a thin strand of copper wire, a commercial fuse or fuse wire. The wire may be held in place using a terminal block, crocodile clips or a fuse holder. Low voltage power supply unit capable of supplying 5 A, ammeter (0–5 A), rheostat (5 A, 25 Ω).

Method

Set up the circuit as shown in Fig. 4.3 with the rheostat set to maximum resistance. Slowly decrease the resistance, allowing the circuit current to increase until the wire gets red hot and melts, thus breaking the circuit.

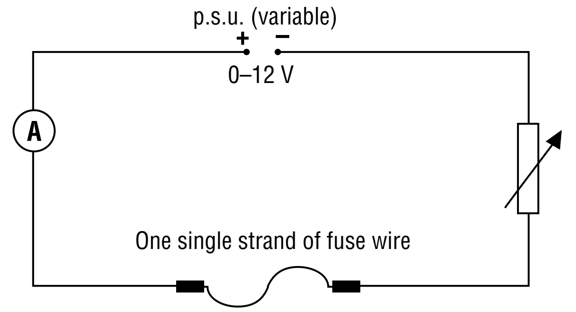


Fig. 4.3

Repeat the experiment three times and record the maximum ammeter reading each time. Take the average of these as the value of the fusing current. Investigate the fusing current values of different types of wire using the same procedure.

EHT and the Transmission of Power

Transmission of power at EHT (extra high tension) involves less energy loss than at low voltages. Because of the fact that transformers can easily and efficiently step up or down a.c., alternating voltages are commonly used for power transmission.

The circuits in Fig. 4.4(a) and 4.4(b) may be constructed to demonstrate the efficiency of high voltage transmission using transformers. The connecting leads may be of nichrome resistance wire to simulate many kilometres of cable.

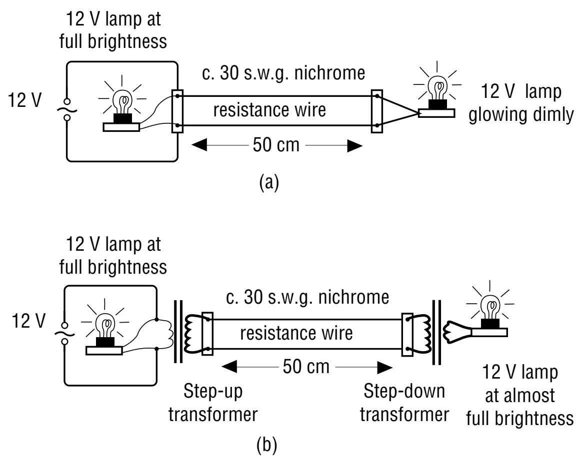


Fig. 4.4 *Model transmission system*

When power is transmitted through many kilometres of cable, the resistance, R, of the cable becomes significant. The power loss in the cable due to heating is given by I^2R. Therefore, for a given cable resistance, the smaller the current, the smaller the power loss. If the power, P, is delivered at a voltage, V, then the current through the cable is given by $I = P/V$. The power loss = $I^2R = (P/V)^2R$. Therefore, the larger the voltage, V, the smaller the power loss and so electrical power is best transmitted at high voltages.

In a power station the electricity may be generated at up to about 20 kV and is then stepped up to 220 kV or 400 kV for transmission over the long-distance cables. This is stepped down to 110 kV for supplying an area and a few large industries. Local transformers further reduce the voltage to 230 V for homes and offices. In addition to the 110 kV lines there are a number of 38 kV and 10 kV (currently being upgraded to 20 kV) lines, mainly for supplying industry.

4.2 Chemical Effect

The passage of electric currents through liquids causes heating just as it does in solids, but more importantly, chemical activity may occur in the liquids around the electrodes. Bubbles of gas are formed, deposits of metal may be seen and changes of colour may occur, depending on what liquids and what electrodes are used. In 1834, Michael Faraday began to investigate the behaviour of liquid conductors, such as acid and salt solutions.

Except for the case of molten metals (e.g. mercury at room temperature), the passage of an electric current through a liquid causes chemical changes; the process is known as electrolysis. Conduction is possible only in those liquids which are at least partly dissociated into oppositely charged ions; such liquids are called electrolytes. Solutions of many inorganic chemical compounds (e.g. common salt, sulphuric acid, etc.) are examples of this type of liquid. There are however, many substances, e.g. sugar, which dissolve without breaking up into ions. Such solutions do not conduct electric current and are called non-electrolytes.

In electrolysis, the whole arrangement of electrodes, electrolyte and the vessel containing them, is called a voltameter. The precise chemical changes that take place depend on the nature of the electrolyte and also on the material of the electrodes. Table 4.1 gives some of the facts about the commonest cases of electrolysis.

Electrolyte	Electrodes		Action at anode	Action at cathode
	anode	cathode		
copper sulphate solution	copper	copper	copper dissolves into electrolyte	copper is deposited from electrolyte
copper sulphate solution	platinum or carbon	platinum or carbon	oxygen liberated	copper is deposited from electrolyte
acidulated water	platinum or carbon	platinum or carbon	oxygen liberated	hydrogen liberated
fused sodium chloride	graphite	iron	chlorine liberated	sodium deposited
fused aluminium ore	carbon	carbon	oxygen liberated	aluminium deposited

Table 4.1

Conduction in Electrolytes

Conduction in electrolytes involves the movement of both positive and negative ions. One of the simplest examples is the electrolysis of fused sodium chloride as used in the extraction of metallic sodium. The molten electrolyte contains only sodium ions (Na^+) and chloride ions (Cl^-), apart from small quantities of impurities, Fig. 4.5.

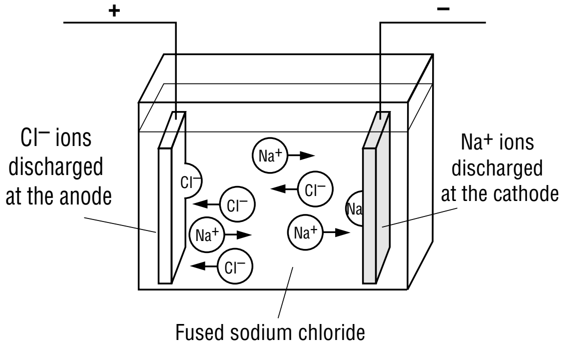


***Fig. 4.5** Electrolysis of sodium chloride*

At the cathode, the Na^+ ion gains an electron to become a neutral atom of sodium.

$$Na^+ + e^- \longrightarrow Na$$

At the anode, the Cl^- ion yields up its extra electron and becomes a neutral atom of chlorine.

$$Cl^- \longrightarrow Cl + e^-$$

The net result is a movement of one electronic charge into and out of the electrolyte for each pair of ions discharged. Metallic sodium is deposited at the cathode and chlorine gas bubbles off at the anode.

In the case of the copper voltameter, Fig. 4.6, which involves copper electrodes in copper sulphate solution, the net effect is that copper is dissolved off the anode and deposited on the cathode with the electrolyte remaining unchanged.

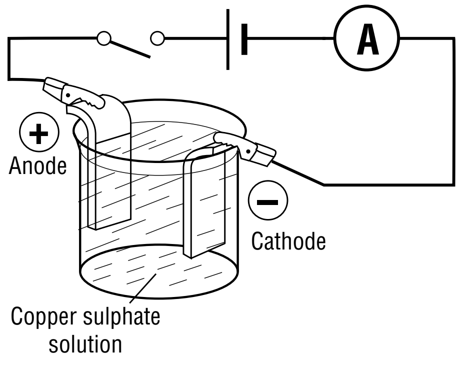


***Fig. 4.6** Electrolysis of copper sulphate*

The water voltameter (platinum electrodes in acidulated water) yields oxygen and hydrogen from its anode and cathode respectively in the ratio by volume of 1:2, Fig. 4.7. These gases come in effect from the water alone, the role of the acid being to make the water a conductor.

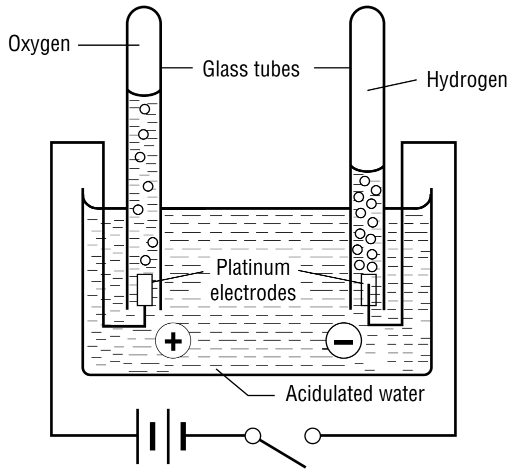


Fig. 4.7 *Electrolysis of water*

Uses of Electrolysis

1. The process of extracting pure metals from their ores is an electrolytic one in many instances, the ore, molten or in solution, being the electrolyte in a large-scale voltameter. Sodium, aluminium, calcium, magnesium and potassium are obtained commercially in this way. In the case of copper, entire lakes are sometimes used in Canada as electrolytic cells. Copper from the damp impure ore ends up on the pure copper cathodes – the cathodes might increase by hundreds of times their original mass in two weeks. The enlarged copper cathodes are then replaced by thin ones and the process repeated.

2. Electroplating is another common application of electrolysis. The metal to be plated is made the cathode of a voltameter with the appropriate solution as electrolyte and a metallic anode. Cutlery is often plated with silver, while steel (for cars and bicycles) can easily be plated with chromium. Even plastics can be made to accept a coat of metal if they are first made conducting by a layer of graphite. Techniques have also been developed to electroplate silver and gold onto glass.

3. Electrolysis plays an important role in the manufacture of electrolytic capacitors. The capacitor is made by the electrolysis of ammonium borate, for example, between aluminium electrodes. A very thin film of aluminium oxide is formed on the anode plate. The oxide layer is an insulating medium or dielectric and is so thin that the capacitance of the capacitor may be extremely high.

Demonstration 4.1

Colour Separation

1. Place a crystal of potassium permanganate in the centre of a damp filter paper, resting on a glass plate, e.g. a clock-glass.

2. Using crocodile clips at the connection points, apply 200 V d.c. to the filter paper, Fig. 4.8.

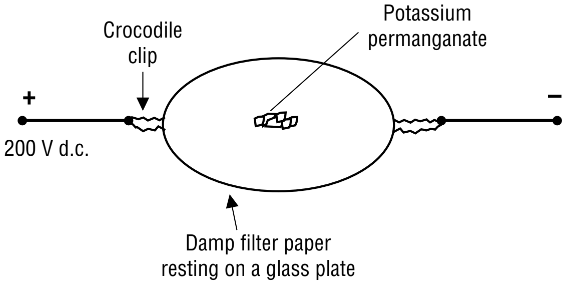


Fig. 4.8

Observation

There is a movement of colour across the paper towards the anode.

3. Reverse the polarity and observe the colour movement again.

4. Repeat the experiment using a.c. No separation of colour results because the polarity changes with a frequency of 50 Hz.

5. Perform the experiment using other coloured crystals.

Demonstration 4.2
Copperplating an Object

1. Replace the copper cathode of a copper voltameter, Fig. 4.9, with a clean metal object, e.g. a key.

2. Pass a small current for several minutes until the object becomes plated with copper. If too large a current is used, the fresh copper will not adhere firmly to the key.

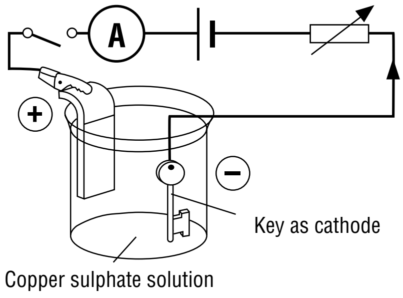


Fig. 4.9 *Electroplating*

Demonstration 4.3
Electric Writing

1. Soak filter paper or blotting paper in a solution of potassium iodide to which a small quantity of starch has been added.
2. Place the damp paper on a flat piece of metal and connect a 12 V d.c. supply as shown, Fig. 4.10.
3. Carefully move the wire over the paper.

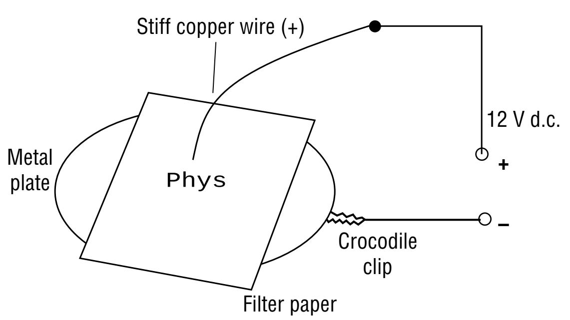


Fig. 4.10

Observation

The (positive) wire leaves a trace due to the liberation of iodine by electrolysis.

4.3 Magnetic Effect

We say that a magnetic field exists in any region of space where a magnetic force is experienced. The direction of a magnetic field at any point is the direction of the force on a north pole at the point.

The magnetic effect of an electric current was discovered by Hans Christian Oersted in 1820. To demonstrate the effect, a plotting compass is placed close to a long straight wire which is carrying current. A deflection is noted. By reversing the direction of the current, the compass needle will be deflected in the opposite sense. For class demonstrations, the compass and wire may be placed on an overhead projector.

To obtain clear results, a large current is needed and the compass must be reasonably close to the wire so that the effect of the earth's magnetic field is negligible.

Experiment 4.3

To Examine the Magnetic Effect of a Current in a Long, Straight Wire

Apparatus

Accumulator or power supply unit (capable of delivering at least 4 A), rheostat (5 A), ammeter (0–5 A), switch, connecting wires, stiff board, plotting compass, 20 cm of straight copper wire. The copper wire may be obtained from a waste piece of electrical cable.

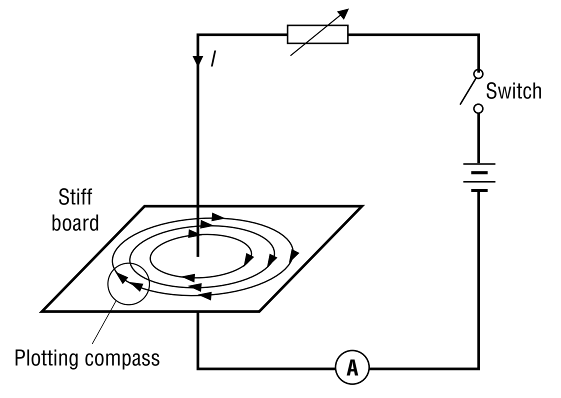


Fig. 4.11

Procedure

1. Set up the apparatus as shown in Fig. 4.11, with the board suitably supported horizontally and the wire passing vertically through its centre.
2. Switch on and adjust the rheostat to give an ammeter reading of 4 A.
3. Place the compass at any point on the board and mark the position of the head and of the tail of the compass, having tapped it gently to ensure it is not sticking.

4. Move the compass and mark the new position of the head when the tail is directly over the previous position of the head.
5. Proceed in this fashion until a complete loop has been traversed. Join the dots using a pencil.
6. With the compass placed at different distances from the wire, complete other loops.

Result

It will be seen that the magnetic lines of flux form circles, concentric with the wire.

Notes

(i) Instead of using a compass, the flux pattern may be observed by sprinkling iron filings on the board close to the wire.

(ii) By replacing the long straight conductor in Fig. 4.11 above with a circular loop or a solenoid, etc., other flux patterns may be observed, Fig. 4.12 and Fig. 4.13.

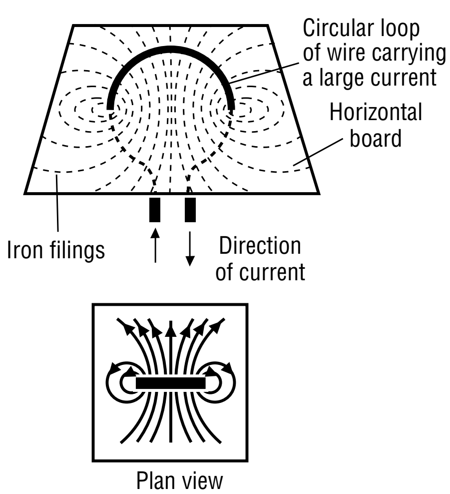


Fig. 4.12 *Flux patterns through a circular loop carrying an electric current*

Applications

The magnetic effect of an electric current has many applications, e.g. in electromagnets, electromagnetic relays, loudspeakers, electric motors, ammeters, voltmeters, etc.

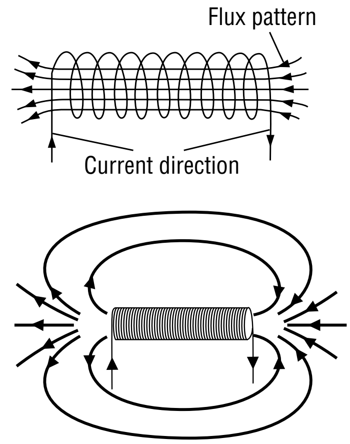


Fig. 4.13 *Flux patterns inside and outside a solenoid carrying an electric current*

Experiment 4.4

To Demagnetise a Magnet Using the Magnetic Effect

Apparatus

Power supply (12 V a.c.), rheostat (5 A), ammeter (0–5 A a.c.), solenoid, bar magnet and connecting wires.

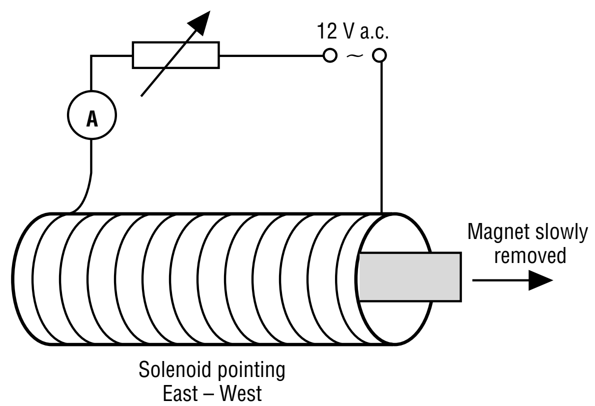


Fig. 4.14 *Demagnetising a magnet*

Procedure

Arrange the circuit as shown in Fig. 4.14 with the axis of the solenoid pointing east-west.

Place the bar magnet inside the solenoid and switch on the a.c. current.

Adjust the rheostat until approximately 3 A is registered on the ammeter. Slowly reduce the current to zero.

Result

The bar magnet will be found to be demagnetised.

Explanation

The alternating current is one which reverses direction every 1/100 second and hence reverses the magnetism in the bar material one hundred times per second. Reducing the current reduces the magnetism which soon decreases to a very small value and then disappears.

Direction

To obtain the direction of the magnetic field due to a straight wire carrying current, we may use one of the following rules.

(i) Maxwell's Corkscrew Rule

If a right-handed corkscrew is turned so that its point travels along the direction of current flow, the direction of rotation of the corkscrew gives the direction of the magnetic field.

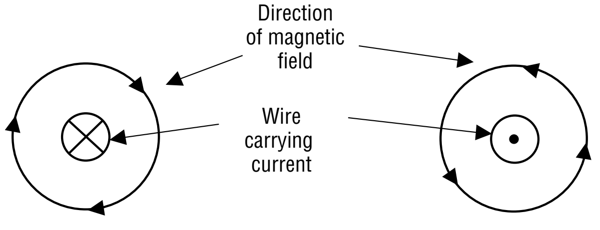


Fig. 4.15(a) *Fig. 4.15(b)*

In Fig. 4.15(a) the current is flowing into the paper. The direction is indicated by a cross which represents the tail of a retreating arrow. The associated magnetic field is in the clockwise direction. In Fig. 4.15(b) the current is flowing out of the paper – the dot in the centre of the wire represents the point of an approaching arrow. The magnetic field is anti-clockwise.

(ii) The Right Hand Grip Rule

Imagine grasping the wire with the right hand so that the thumb points in the direction of current flow. The direction of the curled fingers gives the direction of the field, Fig. 4.16.

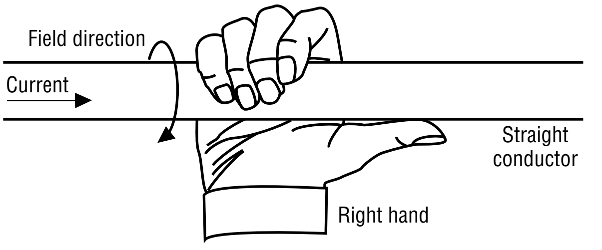


Fig. 4.16

4.4 Domestic Electricity

Electric Shock

Strict regulations exist in industry for the regular inspection and testing of electrical equipment. The onus is on the employer to ensure that all electrical installations and equipment are properly constructed, safely installed and correctly maintained. No such ongoing safety checks apply to the domestic situation and so the home can be a very unsafe place due to the risk of electric shock and fire.

Electric shock is caused by current flowing through the body, i.e. the person becomes part of a circuit. The effect may be harmless, painful or lethal depending on the magnitude of the current, the path it takes and the time for which it flows. For example, a current confined to the hand or arm is likely to have less of an impact than one which flows from one arm to the other, or from hand to feet, across the heart.

Current (mA)	Effect
up to 0.5	No effect
1	Barely detectable
5	Some pain – muscle spasm
10	Severe pain and muscle spasm
16	Cannot let go
20	Respiration ceases
100	Fibrillation of the heart
1000	Heart stops

Table 4.2 *Sample effects of a.c. current*

The data contained in Table 4.2 indicate that any current over 20 mA should be regarded as potentially lethal. It is worth noting also, that a shock which is sufficient to cause muscular contractions can throw a person with considerable force and possibly cause indirect injury. This reflex action may sometimes save life by clearing the person from the cause of the shock but, unfortunately, the same reflex action may sometimes cause the person to grasp the conductor more tightly. D.c. currents are generally less dangerous than comparable a.c. currents.

Body Resistance

In the event of an electric shock the current that flows through a body depends on the skin contact resistance of the body. The value of the resistance encountered is decided by the tightness of the grip, the area of contact and the presence of moisture (especially salty water or sweaty hands). A set of typical resistance values is given in Table 4.3.

Contact	**Resistance (kΩ)**
Dry hands, light contact	100
Dry hands, tight grip	20
Wet hands, tight grip	10
Hands wet with salty water	5

Table 4.3

The highest safe voltage is the highest voltage which will not drive a harmful current through a body when the skin contact resistance is as low as possible. If 5 mA is chosen as the largest acceptable current (Table 4.2) and 5 kΩ as the lowest likely skin contact resistance, (Table 4.3), then

$$V = IR$$

$$= (5 \times 10^{-3}) \times (5 \times 10^{3})$$

$$= 25 \text{ V}.$$

Remember, therefore, that 25 V is the highest safe voltage which should be applied between a person's hands unless the source has sufficient internal resistance to limit the current to 5 mA. (If the terminals of a 9 V battery are placed in direct contact with a person's tongue, an uncomfortable sensation (shock) is experienced if the battery is in good condition. A flat battery has no noticeable effect.)

Example 1

Holding the terminals of a 12 V car battery with wet salty hands (i.e. the extreme case, with R = 5 kΩ), results in

$$I = \frac{V}{R}$$

$$= \frac{12}{5000}$$

$$= 0.0024 \text{ A}$$

$$= 2.4 \text{ mA}.$$

This current of 2.4 mA, which flows through the body, could be felt but would be unlikely to cause any injury.

Example 2

Some portable power tools used on construction sites are operated from a 110 V transformer. If a worker with wet hands (R = 10 kΩ) experiences a shock, the current through the body is

$$I = \frac{V}{R}$$

$$= \frac{110}{10000}$$

$$= 0.011 \text{ A}$$

$$= 11 \text{ mA}.$$

This would be extremely painful but probably not lethal.

Example 3

The presence of moisture on the hands (R = 10 kΩ) is very likely to result in a lethal shock if contact is made with the 230 V mains.

$$I = \frac{V}{R}$$

$$= \frac{230}{10000}$$

$$= 0.023 \text{ A}$$

$$= 23 \text{ mA}.$$

Example 4

Slight dampness in the (insulating) material of a kite string could result in the string having a resistance of 10 MΩ. If the kite comes in contact with an overhead 220 kV power line, then

$$I = \frac{V}{R}$$

$$= \frac{2.2 \times 10^5}{10 \times 10^6}$$

$$= 0.022 \text{ A}$$

$$= 22 \text{ mA.}$$

This is a potentially lethal current.

Experiment 4.5

Measure the skin contact resistance of the body by firmly gripping the leads of a digital multimeter with (a) dry, (b) moist, (c) damp salty, hands.

Safety in the Home

Electrical wiring in the home involves the use of three wires, viz. the live or phase wire, the neutral wire (which is earthed by the ESB at the sub-station and at regular intervals along the distribution system) and the earth wire which is connected to the ground via a galvanised rod outside the house.

Fuses

Fuses are deliberate weak links in an electric circuit so that in the event of a fault developing and too large a current flowing, the fuse wire melts, causing an open circuit and thus preventing any major damage or fire. The following precautions must always be observed.

(i) The fuse used in a particular circuit must be of the correct rating. The fuse rating is the maximum current that the fuse can carry without melting.

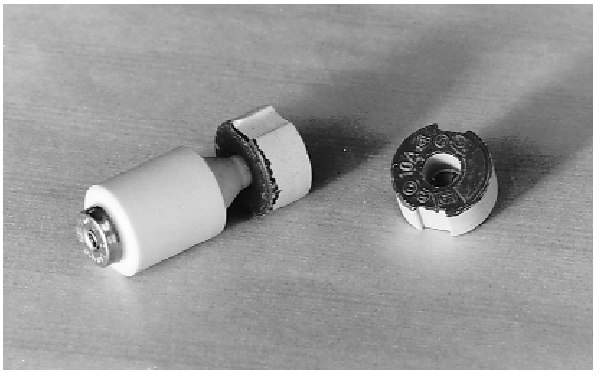


Fig 4.17

(ii) Lower rated fuses are generally narrower than higher rated ones. To ensure that only a fuse of the correct (or lower) rating is fitted in a particular circuit the socket in the fuse box contains a collar, Fig. 4.17, of the appropriate diameter. The collar also ensures that the fuse is properly seated in the socket. A loose-fitting fuse will result in heat being generated due to high resistance (at the poor contact point) and this could give rise to fire.

(iii) The fuse must be in the live side of the circuit for safety, Fig. 4.18.

If a fault develops in an electrical appliance, e.g. the live wire comes in contact with the metal chassis, a large current flows from the live wire, through the chassis to the earth wire and the fuse melts. Clearly, placing the fuse in the neutral side of the circuit would be totally ineffective in this case.

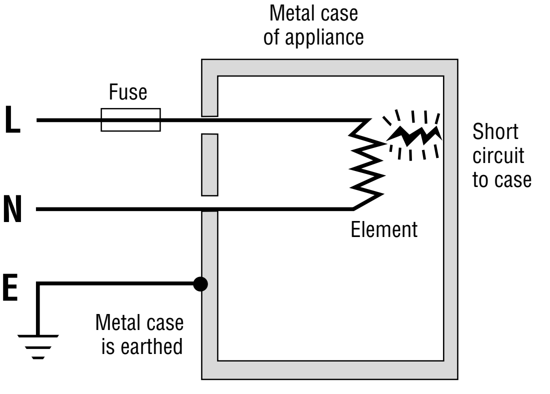


Fig. 4.18 *Fuse in live side of circuit*

Switches

Switches, like fuses, must also be placed in the live side of an electrical circuit. For safety, wall switches are not fitted in bathrooms but ceiling

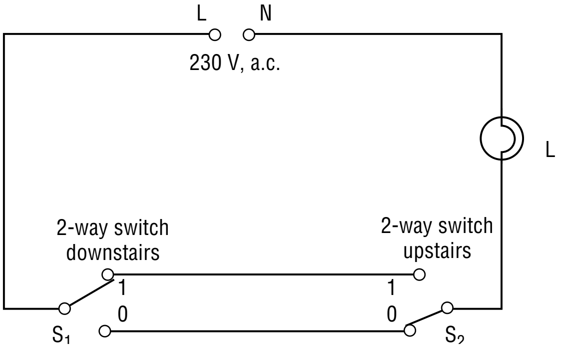


Fig 4.19(a) *2-way switches*

switches may be used. One-way switches are normally used to control room lighting whereas two-way switches are used to control a lamp from either of two positions, e.g. the top and the bottom of a stairs, Fig. 4.19(a). The results of the different positions of the switches, S_1 and S_2, may be summarised in the form of a truth table, Fig. 4.19(b).

S_1	S_2	L
1	1	1
1	0	0
0	1	0
0	0	1

Fig 4.19(b) *Truth table for 2-way switches*

Cable

Electrical cables are made in various sizes, with specified current ratings and it is important that the correct type of cable be used in each situation. The insulation around each wire is coloured to indicate the function of the wire – brown for live, blue for neutral and yellow/green for earth. (Some colour-blind people had difficulty identifying the live and earth wires in the old red-black-green colour system.)

Examples of cable sizes and the recommended current carrying capacity are given in Table 4.4.

Cable size (mm^2)	Current carrying capacity (amps)	Typical use
1.5	15	domestic lighting circuits
2.5	20	domestic power circuits
6	35	electric cooker
10	45	electric shower

Table 4.4

(Care should be taken with extension leads at all times. Such leads often comprise four 3-pin outlets protected by a 13 A fuse and may easily be inadvertently overloaded. Trailing leads may cause accidents; however, extension leads should be fully unwound when in use, to facilitate the dissipation of heat.)

Experiment 4.6

Use a micrometer screw gauge to calculate the cross-sectional area (= πr^2) of various samples (e.g. 1.5 mm^2, 2.5 mm^2, etc.) of electric cable.

Plug Tops

The 3-pin plug, which is used almost exclusively for connecting equipment and appliances to the mains supply, incorporates a number of safety features, e.g. earthing and a fuse link. The old 13 A three pin plug is being replaced by the new sleeved-pin plug for safety.

A plug should be wired so that (i) correct polarity is observed, i.e. brown = live, blue = neutral, green/yellow = earth; (ii) there are no loose strands of wire and all connections are tight; (iii) the earth wire is long enough so that if the cable is accidentally wrenched out of the plug, the live wire will pull free before the earth wire is dislodged, (iv) a fuse of the correct rating must be used at all times.

Earthing

A fuse protects an appliance and its cables in the event of an overload but an additional measure must be taken to safeguard a user. All electrical appliances, which have exposed metal parts, ought to be made safe to touch even if a fault develops inside them. This is achieved by earthing, i.e. by providing a wire which connects the metal parts to an object which is in good electrical contact with the earth such as a metal plate or rod, sunk deeply in damp soil. (Some electrical appliances are manufactured with an all-plastic outer case and do not require an earth connection.)

Furthermore, the ESB regulations stipulate that all exposed metal pipes be ‘bonded’ and connected to the main earth terminal. Bonding involves interlinking metal pipes using copper wire which is secured to the pipework using adjustable earthing clips. When all (exposed) metalwork is bonded and earthed, all parts are at the same potential (voltage) and are said to form an equipotential zone. A person touching such metalwork would not receive a shock because there exists zero potential difference, i.e. both person and metal are at the same (i.e. earth/ground) potential.

Ring (Main) Circuits

The standard domestic wiring technique involves taking a ring of 3-core cable, i.e. 2.5 mm^2, twin and earth, right around the house to supply all the power points, Fig. 4.20, starting at, and returning to, the distribution board. In a large house, several rings may be used to cater independently for lights upstairs and downstairs, power upstairs, power downstairs, etc.

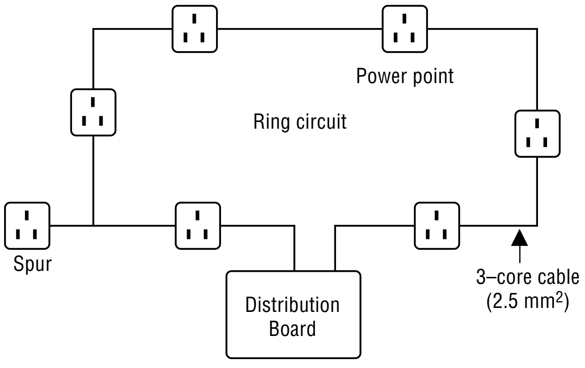


Fig. 4.20 *Ring main with spur*

Immersion heaters, electric cookers and showers require heavier cable and each has its own separate circuit. Note that in a ring circuit, there are two paths provided to each socket, i.e. clockwise and anticlockwise around the ring. Thus the current in any part of the ring circuit is only half the current to any appliance plugged into a socket. This minimises the risk of overheating in the cable and so prolongs its life and that of the surrounding insulation. In older houses, the distribution board consists of a main fuse, ESB meter, fuses for each ring circuit, etc., and a trip switch all of which are mounted in the one location inside the house.

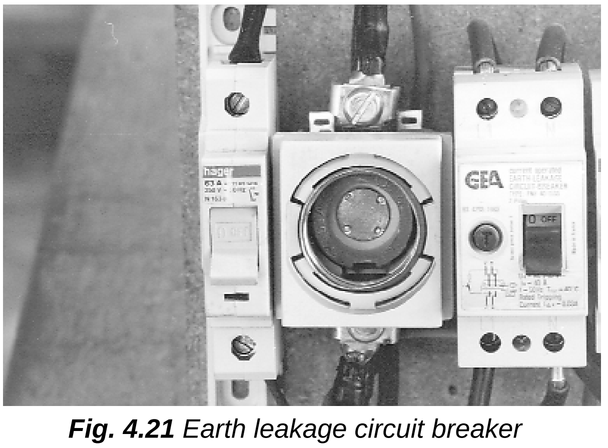


Fig. 4.21 *Earth leakage circuit breaker*

In modern houses, the main ESB fuse and meter are housed in a cabinet on the outside wall which facilitates the regular reading of the meter by the ESB authorities. Inside the house is mounted a modern domestic distribution unit, comprising main switch, miniature circuit breakers (MCBs) and an earth leakage circuit breaker (ELCB), Fig. 4.21.

The MCBs are a replacement for fuses and they are electromechanical devices which trip when an excess current flows. Once a fault has been identified and repaired, the circuit is restored by resetting the control button/switch; no rewiring or replacement of fuses is necessary. MCBs are primarily designed to protect the connected cable in an installation. They may be activated in a number of ways, depending on the excess current.

(i) **Thermal Operation**. Small overloads are detected by a carefully calibrated bimetallic element. A current greater than the rated current of the MCB will cause the bimetal element to get hot and to trip the MCB. The greater the overload current, the shorter the time required to trip the MCB.

(ii) **Magnetic Operation**. At a value of current between 4 times and 7 times the rated current, the thermal tripping operation is overridden by the magnetic tripping operation. Current flowing through a solenoid attracts an armature, causing the MCB to trip instantaneously, i.e. in less than 100 milliseconds.

At high values of overcurrent (i.e. short circuit current) a plunger in the solenoid is moved with sufficient force to physically separate contacts. The greater the short circuit current, the greater the force with which the plunger is moved and the faster the circuit is disconnected.

ELCBs are a means of protection similar to the old-style trip-switch. In the event of an earth fault, the unit trips in an operating time of 40 milliseconds (or less) at 30 mA. A test button is provided with the ELCB to enable the electrical and mechanical operations of the device to be checked. When the test button is pressed, an earth fault is simulated and the unit trips. Power is restored by simply resetting the control switch.

Residual Current Devices

Most modern installations now use residual current devices (RCDs). An RCD is a mechanical switching device which causes the opening of contacts when the residual current attains a given value under specified conditions. Such devices are checked for correct operation using a special transformer for safety which simulates fault conditions. When a test voltage is applied, the trip should operate instantaneously and cut off the power. The test button on RCDs may be used for periodic checking purposes by the consumer. RCDs have been previously called ELCBs. They detect earth leakage currents well below those detected by an overcurrent protective device such as an MCB or fuse. Their use in virtually all electrical circuits significantly reduces the risk of electrocution of humans and of animals and will cut off dangerous earth leakage currents which can initiate a fire and which are too small to be detected by an MCB or fuse. The majority of fires which occur as a result of faulty wiring are started by current flowing to earth. Fires can be started by fault currents of less than 0.1 A. The normal domestic overload protective device such as a fuse or MCB will not trip with such a small current. A correctly chosen RCD will detect this fault current and interrupt the supply, hence reducing the risk of a fire starting. A 30 mA RCD will provide a high degree of protection against electrocution in an accidental shock hazard situation. The current flowing through a human body could be between 80 mA and 240 mA, depending on the resistance of the body and the voltage across it. Typical specifications demand that an RCD operates within 50 ms at 240 mA and 150 ms at 80 mA.

Operation of an RCD

An RCD works on the current balance principle. The supply conductors to the load, i.e. phase and neutral, are passed through a toroid (a ring-wound solenoid) and form the primary windings of a current transformer. The secondary winding of the transformer is connected to the tripping relay, Fig. 4.22. In a normal circuit, the current in the phase conductor is equal to the current in the neutral and the algebraic sum of the currents is zero. If there is an insulation fault on the circuit and current flows to earth, then the phase and neutral currents will not balance and the algebraic sum of the currents is not zero. This imbalance is detected by the secondary winding of the current transformer. The RCD will trip and the supply to the load will be interrupted.

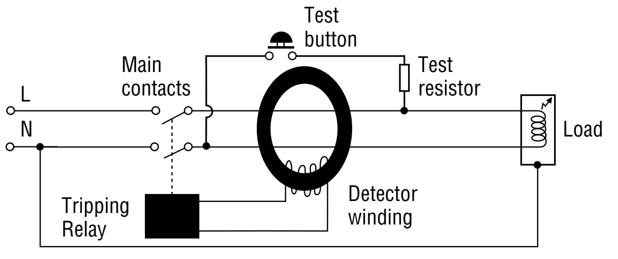


Fig. 4.22 *RCD circuit*

Example

Calculate the cost of using a 2.5 kW electric fire for 2 hours if the cost per unit is 8p. What fuse should be used in the plug connecting the fire to the 230 V mains?

Solution

The number of kilowatt-hours used = 2.5 kW $\times$ 2 h = 5 kW h,

i.e. 5 units. So, the total cost = 5 $\times$ 8 = 40 pence.

$$P = IV$$

$$2500 = I \times 230$$

$$I = \frac{2500}{230}$$

$$= 10.9 \text{ A}$$

The nearest higher rating to this requirement is a 13 A fuse.

Note. The kW h is a unit of energy. Since 1kW = 1000 watts = 1000 J s^{-1}, 1 kW h is equivalent to 1000 $\times$ 60 $\times$ 60 = 3.6 $\times$ 10^6 J, i.e. 1 kW h = 3.6 MJ.

Example

A bird perches on a power line that carries 160 A. Calculate the potential difference between the bird's feet if the resistance of the wire is 10 $\mu\Omega$ m^{-1} and the bird's feet are 4 cm apart?

Solution

The resistance of 4 cm of wire is

$R = 0.04 \times 10 \times 10^{-6}$

$= 4 \times 10^{-7}\ \Omega$

The p.d. across 4 cm of wire is thus

$V = IR$

$= 160 \times 4 \times 10^{-7}$

$= 6.4 \times 10^{-5}$ V

$= 64\ \mu$V.

APPENDIX A POWER PACKS

Power Packs

In the laboratory, power packs or power supply units (p.s.u.s) are often used instead of dry cells or accumulators. The p.s.u. lowers (or raises) the a.c. mains supply voltage to a suitable level and then converts this a.c. voltage into a d.c. one. Additional devices or circuits may be added to ensure that the d.c. output voltage remains steady even when the mains voltage varies or if the amount of current taken from the p.s.u. is altered. Initially, a.c. voltage from the mains is applied to the primary windings of a transformer, Fig. A.1. An a.c. voltage is then induced between the terminals of the secondary coil, the magnitude of which depends on the relative number of turns in the two windings.

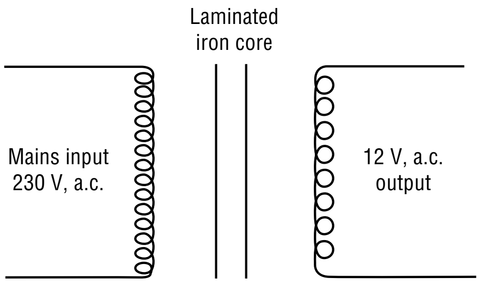


Fig. A.1 *Step-down transformer*

The 12 V a.c. output in Fig. A.1 results in the waveform shown in Fig. A.2.

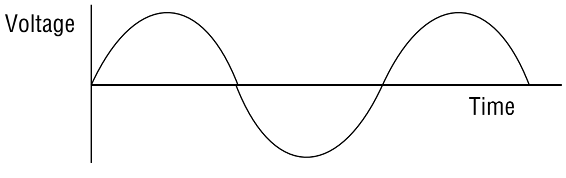


Fig. A.2 *A.c. output*

The next stage involves the use of a rectifier, or a series of rectifiers, which alter the wave shape of the a.c. into one of the wave-shapes shown below. If one diode is incorporated in a circuit as shown in Fig. A.3, then current flows only in one direction, resulting in half-wave rectification, Fig. A.4. Note that when using a cathode ray oscilloscope to demonstrate rectification the a.c./ d.c. switch should be set to the d.c. position.

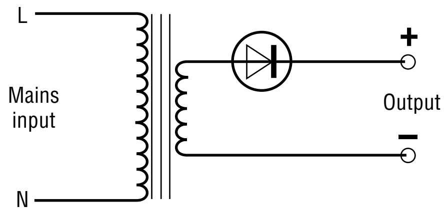


Fig. A.3 *Half-wave rectifier*

However, full-wave rectification is preferred for many applications and this may be obtained with the arrangement shown in Fig. A.5, using two diodes in a push-pull arrangement with a centre-tapped transformer. When point A is at a positive potential with respect to point B in the transformer, diode D_1 conducts and current flows clockwise through the load resistor, R. When point B goes positive with respect to point A, then diode D_2 conducts and again current flows clockwise through the load resistor. The output is then as shown in Fig. A.6.

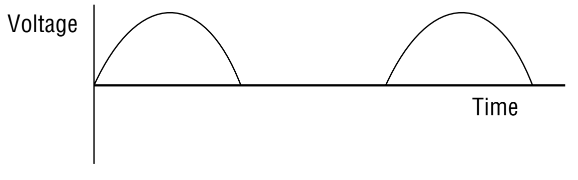


Fig. A.4 *Half-wave rectified output*

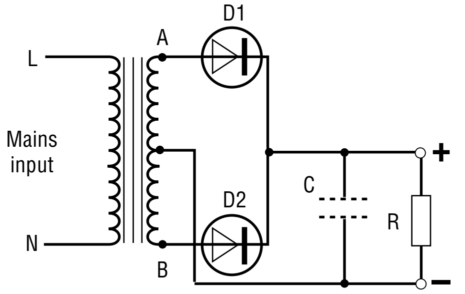


Fig. A.5 *Full-wave rectifier*

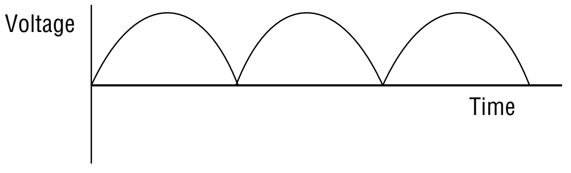


Fig. A.6 *Full-wave rectification*

By introducing a large (reservoir) capacitor smoother d.c. may be obtained from the circuit shown in Fig. A.5 provided the current drawn by R is small. The capacitor C is charged by the current from the diodes and supplies current to the circuit every time the voltage output from the diodes drops. Thus the waveform is smoothed, Fig. A.7.

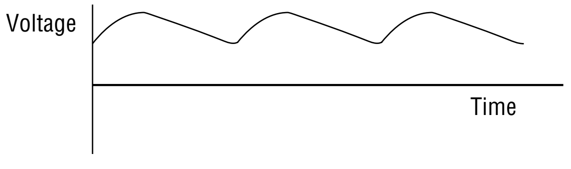


Fig. A.7 *Full-wave rectification with smoothing*

As an alternative to the circuit in Fig. A.5 above, four diodes may be used with a suitable transformer to obtain full-wave rectification, Fig. A.8. The four diodes may be obtained in the one package called a bridge rectifier.

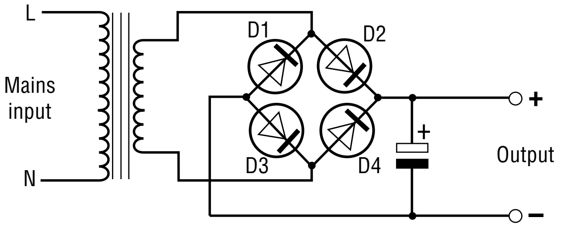


Fig. A.8 *Bridge rectifier circuit*

Regulation

Regulation is the term used to quantify the drop of output voltage which occurs when load current is drawn from the supply circuit. A perfectly regulated supply would have constant output voltage whatever the load current. In the circuits shown above a considerable ripple (voltage variation) becomes apparent when a large current is drawn from a p.s.u. Thus something additional is required to obtain good regulation, viz. a stabiliser circuit. This may be of the type shown in Fig. A.9 which uses a Zener diode (see p. 70). and is useful for small currents.

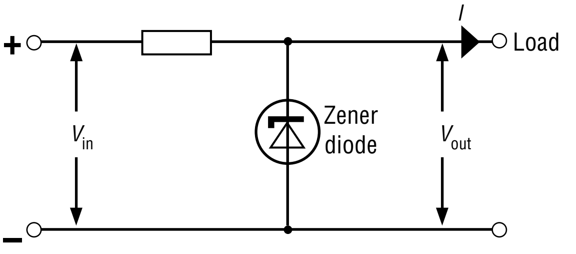


Fig. A.9 *Voltage stabiliser circuit*

A more advanced type of stabilised circuit is shown below in Fig. A.10. This involves the use of an integrated circuit (IC) called a voltage regulator.

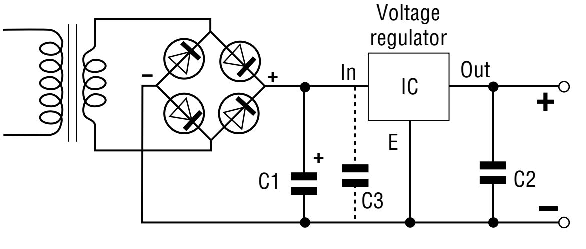


Fig. A.10 *Voltage stabiliser circuit*

Voltage stabilisation is achieved by feeding into the stabiliser circuit a voltage which is higher than the planned output voltage even at the worst possible combination of circumstances, i.e. low mains voltage and maximum load current. The stabiliser then controls the voltage difference between input and output so that the output voltage is steady.

A variety of p.s.u.s are available. Some have an automatic cut-out facility to protect the unit from harmful overloading. To reset such a unit, simply disconnect the offending circuit, push the overload switch back into position and the unit is again ready to deliver current.

P.S.U. Faults

P.s.u.s may exhibit a drop in output voltage or complete loss of output voltage, excessive ripple or very poor regulation. Complete failure can be caused by a blown fuse due to overloading or by an open-circuit (o/c) transformer winding. A short circuit (s/c) in the output circuit will also bring the output voltage to zero.

A drop in voltage, accompanied by a large ripple, can be caused by the failure of one of the diodes in a full-wave set, or by an o/c reservoir capacitor. Poor regulation can also be caused by an o/c reservoir capacitor, or by a diode developing unusually high resistance.

Anti-surge (AS) Fuses

A quick-blow fuse will melt if the current in the circuit rises above the value indicated by the fuse rating. Anti-surge fuses can carry currents much larger than their rated value, for short periods of time without breaking. P.s.u.s normally require anti-surge fuses, particularly if a very large smoothing capacitor is used, as the initial high charge current taken by the smoothing capacitor at switch-on can often be sufficient to activate quick-blow type fuses.

Smoothing Units

Some power supply units provide unsmoothed d.c. which renders them unsuitable for use in a variety of experiments (e.g. potentiometer/metre bridge experiments). Smoothing unit are available for use with such p.s.u.s, providing smoothed d.c. with less than 0.1% ripple on outputs of up to 3 A. At larger current outputs, the ripple will increase.

Demonstrating Rectification Using a Very Low Frequency Generator

A cathode ray oscilloscope was mentioned earlier as a means of demonstrating rectification. If an oscilloscope is not readily available, a very low frequency (VLF, c. 1 Hz) a.c. generator may be used with a centre-zero moving-coil d.c. meter to illustrate the concept of rectification.

With the arrangement in Fig. A.11 (no diode) the meter pointer will oscillate equally on both sides of zero, indicating an a.c. voltage across the 5 kΩ load resistor.

With the diode added, Fig. A.12, the pointer movement will be confined to one side of the zero mark, i.e. half-wave rectification.

With the four diodes (i.e. a bridge rectifier) in Fig. A.13, the pointer is again confined to movement on one side only of the zero mark but will oscillate at twice the frequency of the half-wave rectified circuit, i.e. full-wave rectification is obtained. A capacitor, C, (1000 μF) in parallel with the load will have a smoothing effect on the pointer movement in both half- and full-wave rectification.

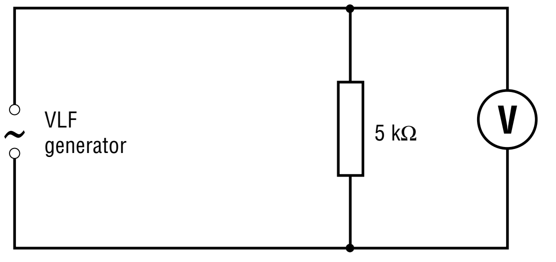


Fig. A.11

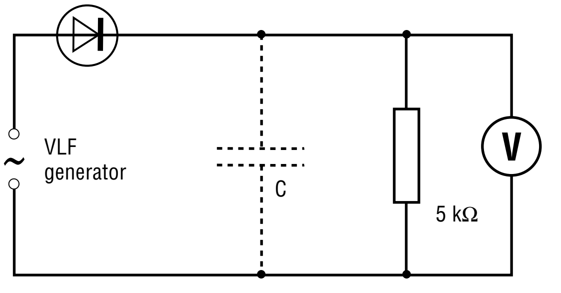


Fig. A.12 *Half-wave rectification (with smoothing)*

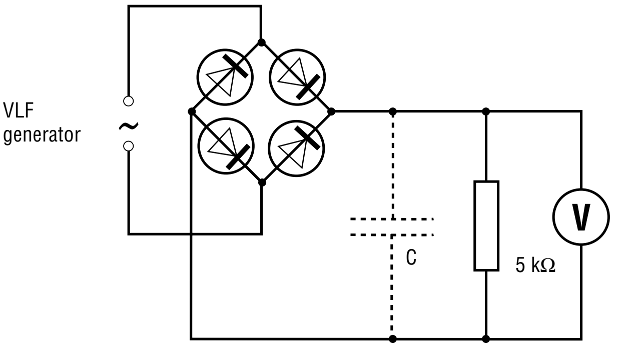


Fig. A.13 *Full-wave rectification (with smoothing)*

Zener Diode

When a large reverse voltage (e.g. 50 V) is applied to a p–n junction diode, the junction may break down and a large reverse current will flow (see p. 45). The precise voltage at which breakdown occurs (called the Zener breakdown voltage) depends on the degree of doping in the crystal and on the way in which the junction is constructed. Note that the diode is not damaged by the reverse current as long as it is kept low by a series resistor.

Some diodes are designed to breakdown at fixed and predictable voltages. Such diodes are called Zener diodes and are used for voltage stabilisation purposes because the reverse voltage across a conducting Zener diode remains (fairly) stable even when the current flow changes considerably, Fig. A.14. Devices of this type are available with Zener voltages at various standard voltages from 2.5 V upwards.

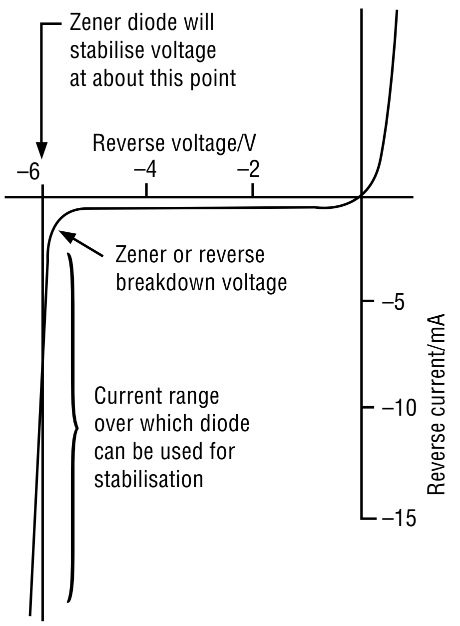


Fig. A.14 *Variation of current with voltage for a Zener diode*

Demonstration of Zener Diode Action

Arrange the circuit as shown in Fig. A.15 with the supply voltage (V_1) set at zero. Increase V_1 in 0.5 V steps up to 10 V. Record V_1 and the voltage across the Zener, V_0, at each stage. Tabulate the results and plot a graph of V_0 (y-axis) against V_1 (x-axis).

Observation

It will be found that V_0 reaches a maximum value (6.2 V) and then remains constant even though V_1 continues to increase.

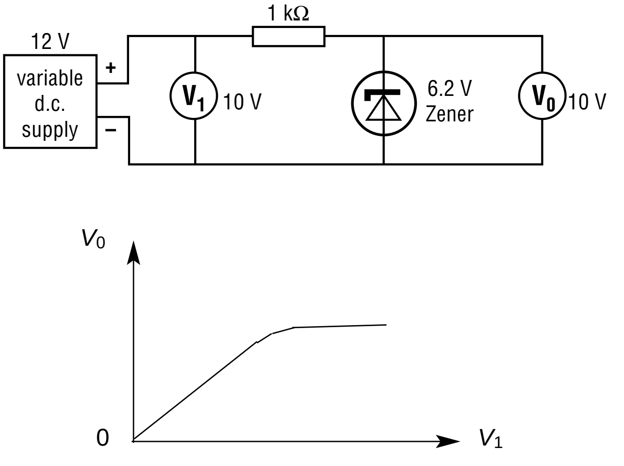


Fig. A.15

Principle of the Zener Diode

The principle of the Zener diode is based on the fact that the diode is used in the voltage range just beyond the reverse breakdown voltage, Fig. A.14. In this region, small changes in reverse voltage produce relatively large changes in the reverse current. It is recalled that in Fig. A.9, p. 68, the Zener diode was connected in the reverse sense across the load whose supply voltage it was to stabilise. Thus, if the load current were to increase (for any reason), causing a slight fall in the supply voltage, the reverse bias of the Zener is automatically decreased and the reverse current it passes, falls. This decrease compensates for the increase in the load current, i.e. the total current drawn from the supply remains constant. Since the supply voltage depends on the sum of the Zener and load currents (the variations in which tend to cancel out), the supply voltage remains constant.

Likewise, of course, if the supply voltage increases (for any reason), the Zener diode reverse bias is increased so that it passes a greater reverse current to bring the supply voltage down again, i.e. the supply voltage remains constant.

APPENDIX B

Ni-Cd BATTERY CHARGER

Ni-Cd Battery Charger

A basic constant-current circuit is illustrated in Fig. B.1. When a light-emitting diode is conducting, the voltage across it remains (reasonably) constant at approximately 1.6 V. Thus a LED may be used to maintain a constant voltage between the base and collector of a transistor. As a result, a constant current, whose size is controlled by R, will flow in the emitter circuit when Ni-Cd cells are connected into place. Once the supply voltage is sufficiently high to permit current to flow in the first instance, the charging current remains constant, irrespective of how many cells are connected in series, up to a certain limit determined by the p.s.u.

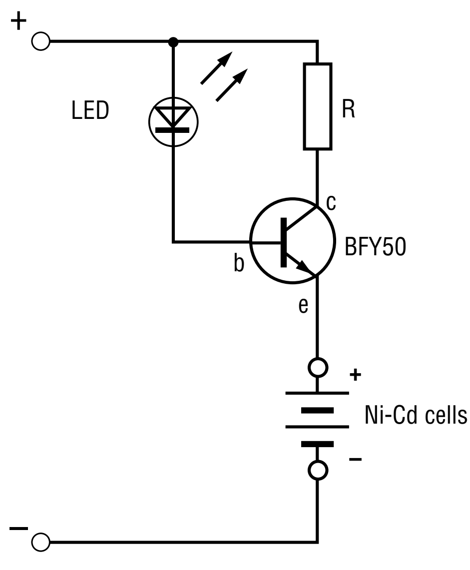


Fig. B.1

APPENDIX C

TEMPERATURE COEFFICIENT OF RESISTANCE

Temperature Coefficient of Resistance

The temperature coefficient of resistance, α, of a substance is the increase in resistance of a specimen of it per kelvin rise in temperature divided by its resistance at 0 °C, i.e.

$\alpha = (R_\theta - R_0)/R_0\Delta\theta$.

This may be rearranged to give

$R_\theta = R_0(1 + \alpha\Delta\theta)$

where R_θ = resistance at temperature θ, R_0 = resistance at 0 °C.

Example

A coil of copper has a resistance of 40 Ω at 10 °C. Calculate the resistance of the coil at 100 °C, given that the temperature coefficient of resistance of copper is 4.28×10^{-3} K^{-1}.

Solution

$$R_\theta = R_0(1 + \alpha\Delta\theta)$$

$$40 = R_0(1 + 10\alpha)$$

$$R_{100} = R_0(1 + 100\alpha)$$

$$\frac{40}{R_{100}} = \frac{1 + 10\alpha}{1 + 100\alpha}$$

$$R_{100} = \frac{40\,(1 + 100\alpha)}{(1 + 10\alpha)}$$

Since $\alpha = 4.28 \times 10^{-3}$ K^{-1},

$$R_{100} = \frac{40\,(1.428)}{1.0428} = 55\ \Omega.$$

Experiment C.1
Measurement of the Temperature Coefficient of Resistance

In the mandatory experiment to investigate the variation of the resistance of a metallic conductor with temperature, a graph similar to that in Fig. C.1 is obtained.

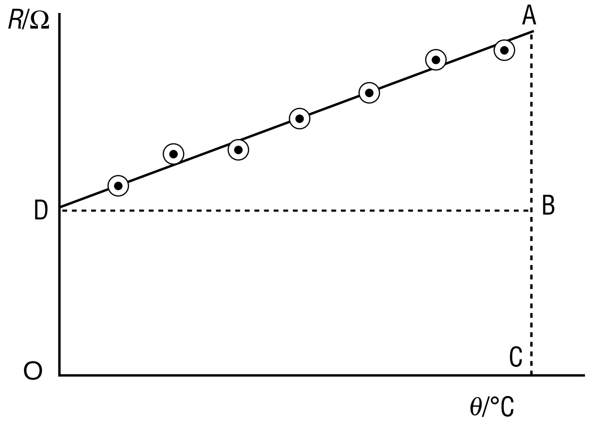


Fig. C.1

By definition,

$$\alpha = \frac{(R_\theta - R_0)}{R_0\Delta\theta}$$

$$= \frac{\text{increase in resistance}}{R_0 \times \text{(temp. rise)}}$$

$$= \frac{|AB|}{|OD| \times |DB|} \text{ from the graph, since } R_0 = |OD|$$

$$= \frac{\text{Slope}}{|OD|} \text{ since } \frac{|AB|}{|DB|} = \text{slope of the graph}$$

α = slope/intercept on R axis

Therefore, by calculating the slope of the graph in the experiment and measuring |OD|, i.e. R_0, the temperature coefficient may be calculated.

APPENDIX D

THE HOT-WIRE AMMETER

The Hot-Wire Ammeter

The hot-wire ammeter, which may be used for measuring a.c. as well as d.c., is based on the heating effect of an electric current. Instruments used for measuring alternating current must be constructed so that the pointer deflects the same way when the current flows through the instrument in either direction. In the hot-wire ammeter, Fig. D.1, heat is produced in the fine resistance wire AB when current flows through it.

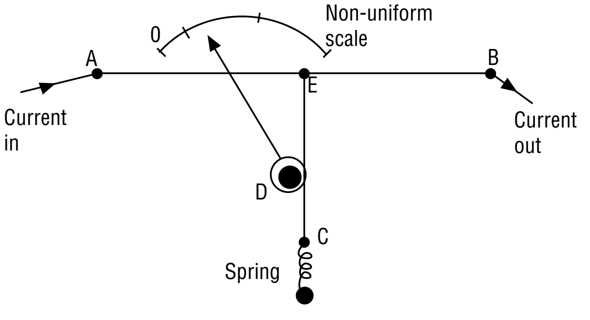


Fig. D.1

When the wire AB gets hot, it expands and sags. The slackness is taken up by another fine wire EC which passes round a pulley D and is kept taut by a spring. Any movement of the wire EC causes the pulley to rotate. Consequently, the pointer moves over the scale. The deflection, θ, of the pointer is approximately proportional to the average rate at which heat is produced (i.e. the power) in the wire AB,

$$\text{i.e. } \theta \propto P$$

But, by Joule's law, $P \propto I^2$

$$\therefore\ \theta \propto I^2$$

Thus, the deflection of the pointer is proportional to the square of the current and so the scale is not a uniform one, as indicated in Fig. D.1. In a.c. measurements the meter reads the r.m.s. current.

APPENDIX E

3-PHASE SUPPLY

3-Phase Supply

Economic considerations play a big part in the generation and transmission of electric power. In a power station, a generator produces an emf in three different coils arranged symmetrically around a rotating electromagnet. Consequently, each coil has an emf induced which is 120° out of phase with each of its neighbouring coils, Fig. E.1.

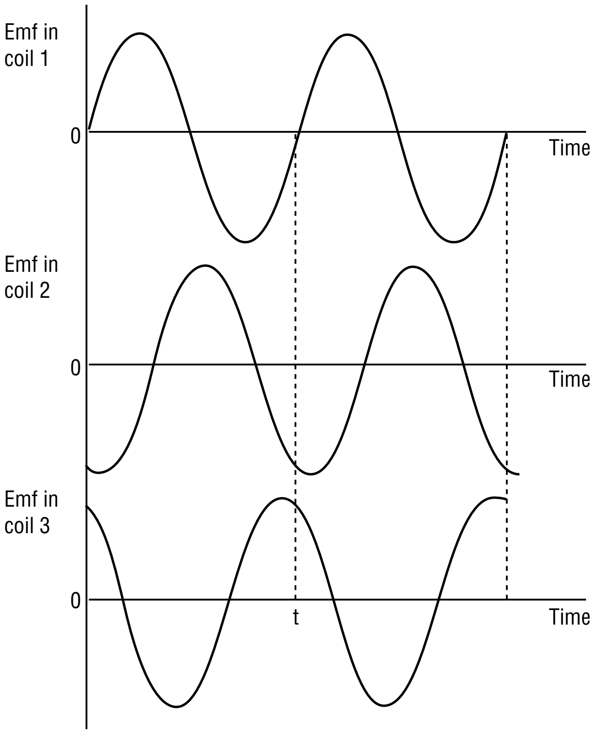


Fig. E.1

One end of each coil is connected to the metal frame, i.e. earth, of the generator. The other ends of the coils are the three outputs of the generator, each one having a varying voltage with respect to earth but 120° out of phase with each other. The voltages at any instant, may be written as:

$V_1 = A \sin \omega t$,

$V_2 = A \sin (\omega t - 120°)$ and

$V_3 = A \sin(\omega t - 240°)$.

These three voltages (i.e. 3-phases) are stepped up and down using transformers as required for distribution purposes by the ESB. Consumers of large quantities of electrical energy, e.g. industrial users, often use all three phases but only a single phase supply is needed for domestic use. Generally, adjacent houses take their supply from each of the three live cables in turn. As a result, each phase carries roughly the same load. Once the power has been distributed to the consumers, the three phases are then linked together, yielding a resultant current of (almost) zero, (see Fig. E.1 above). A thin (neutral) cable connected to earth, is all that is necessary to ensure a complete circuit. An inspection of HT pylons will (sometimes but not always) confirm the existence of either 4 or 7 sets of cables. The three live cables or pairs of them are attached to the pylon via large insulators. The remaining neutral cable is located at the very top of the pylon, linking neighbouring towers together. In brief, then, by using a three-phase system, it is possible to have three separate supplies requiring only 4 cables, one of which can be much thinner than the others. A single phase system would require six thick cables to supply the same quantity of electrical power. The three-phase system is thus more economically viable.

APPENDIX F

The VAN DE GRAAFF GENERATOR

The Van de Graaff Generator

In 1930, an American physicist, Robert J. Van de Graaff, designed a type of electrostatic generator capable of providing very high voltages, Fig. F.1. Such machines were originally used with high-voltage X-ray tubes and also for atom-splitting experiments. Much valuable information about the structure of atoms and nuclei was obtained from these experiments. (Textbooks describe the principle of the Van de Graaff generator.)

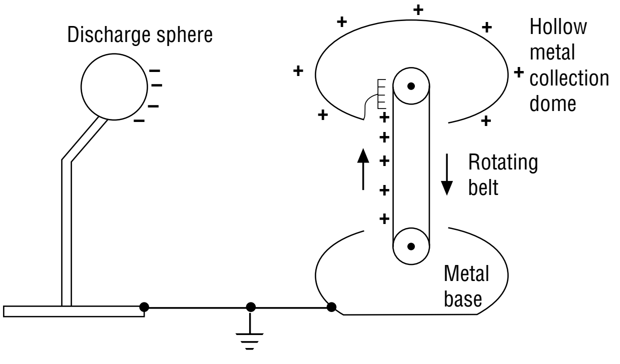


Fig. F.1

Dampness is frequently the root cause of problems encountered in static electricity experiments. This can be overcome by storing the apparatus in a dry place and by always drying the apparatus with a blow-heater or hair dryer before attempting any experiment. If drying fails to make the generator work properly, the rollers and the belt are more than likely dirty. The rollers are best cleaned with petroleum ether and cotton wool. The belt may be cleaned by immersing it in warm water containing a detergent followed by rinsing in (distilled) water and finally drying.

Demonstrations using the Van de Graaff generator

1. To show that a high voltage is generated.
2. To light a bunsen.
3. Force between charges.
4. Hamilton's mill.
5. Electric wind.
6. Length of spark.
7. Electric field patterns.
8. Electrostatic screening.

Some of the above have already been dealt with in the electrostatics section, i.e. nos. 3, 4, 5, 7.

Demonstration F.1

To Show that a High Voltage is Generated by the Van de Graaff Generator

Earth one terminal of a neon bulb (90 V type) and direct the other terminal towards the collecting dome. Switch on the generator and slowly move the bulb closer to the dome.

Result

The bulb will glow. This indicates that the dome is at a high potential. Note: a phase-tester, with its metal cap earthed, may be more conveniently used – do not touch the dome with the phase-tester as a shock may be experienced! Alternatively, three fluorescent striplights may be joined in series and one end earthed. The other end of the arrangement is touched to the dome. All the striplights will light.

Demonstration F.2

To Light a Bunsen Burner

Earth the metal frame of a bunsen using a flexible wire. Start the generator, turn on the gas and, holding the rubber tubing, point the burner towards the dome. The gas ignites. Note that sparks will pass to the flame after the gas ignites because of the high concentration of ions in the flame.

Demonstration F.3

To Investigate the Length of Spark Obtainable from the Generator

Place the earthed discharge sphere close to the collecting dome and observe the almost continuous sparking in the air, Fig. F.1. Slowly withdraw the discharge sphere. Note that the sparking frequency decreases but that the length and the strength of the spark increases. On a dry day with dust- free equipment, spark lengths of 25 cm may be obtained indicating the presence of tens of thousands of volts. A spectacular discharge is visible if the rounded end of an earthed metal rod is placed about 20 cm from the dome, with the generator operating in the dark.

Demonstration F.4

To Demonstrate Electrostatic Screening

The apparatus and the arrangement are similar to that used in Experiment 2.1 (p. 21). The anode used in this experiment is a point and the cathode is in the form of an almost-enclosed hollow square. The anode is connected to the dome and the cathode to the earthed base terminal of the generator. The field lines will form between the anode and the cathode but they will not penetrate the hollow square, i.e. there are no field lines inside. This is the principle of the Faraday cage.

Earthing the Generator

A convenient method of discharging the dome involves bringing an earthed pointed metal object towards the collecting sphere. Alternatively, simply touch the discharging sphere against the (earthed) collecting sphere. Always earth the dome after using the generator.

APPENDIX G

TO MEASURE THE VOLTAGE GAIN OF AN AMPLIFIER

Experiment

To Measure the Voltage Gain of an Amplifier

Apparatus

Cathode ray oscilloscope, signal generator, voltage amplifier.

Procedure

1. Construct the voltage amplifier on breadboard as shown in Fig. G.1. Using the resistance and capacitance values given in the diagram, the collector voltage should equal 4.5 V approximately, i.e. ≈ 1/2 (the supply voltage).
2. Set the sine-wave of the signal generator to 30 mV peak-to-peak at 400 Hz. Use the CRO to check the 30 mV p–p setting.
3. Connect the output of the signal generator to the input of the amplifier. Connect the output of the amplifier to the CRO.
4. Having made the appropriate adjustments to the CRO, measure the peak-to-peak amplitude of the signal from the output of the amplifier.
5. Calculate the voltage gain (A) from the formula:

$$A = \frac{\text{Output voltage p–p}}{\text{Input voltage p–p}}$$

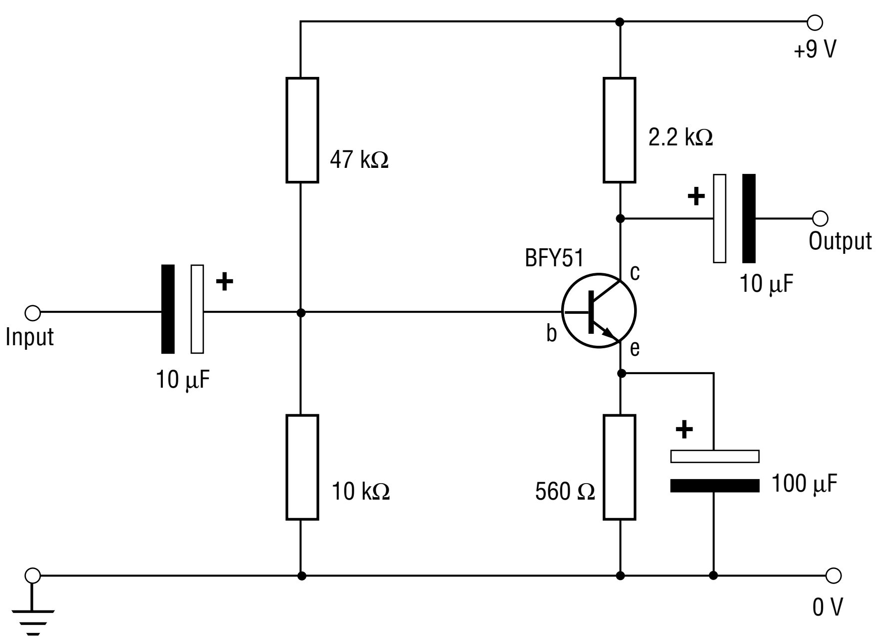


Fig. G.1

APPENDIX H

AN ELECTRONIC TIMING CIRCUIT

An Electronic Timing Circuit

This is a simple electronic timer which may be constructed on breadboard. When the circuit, Fig. H.1, is first set up, with the Time switch open, the red LED lights. When the Time switch is closed, the LED is extinguished and about 30 seconds later it glows again to indicate time elapsed. This timing circuit is similar to those used in Question Times and chess games, etc. To reset the timer the time switch must be opened and then closed again.

The time delay period depends on the capacitance of C. Increasing the capacitance of C extends the time period. The principle of the circuit is best understood by attaching a multimeter across C and closely monitoring the voltage before and after pressing the Time switch.

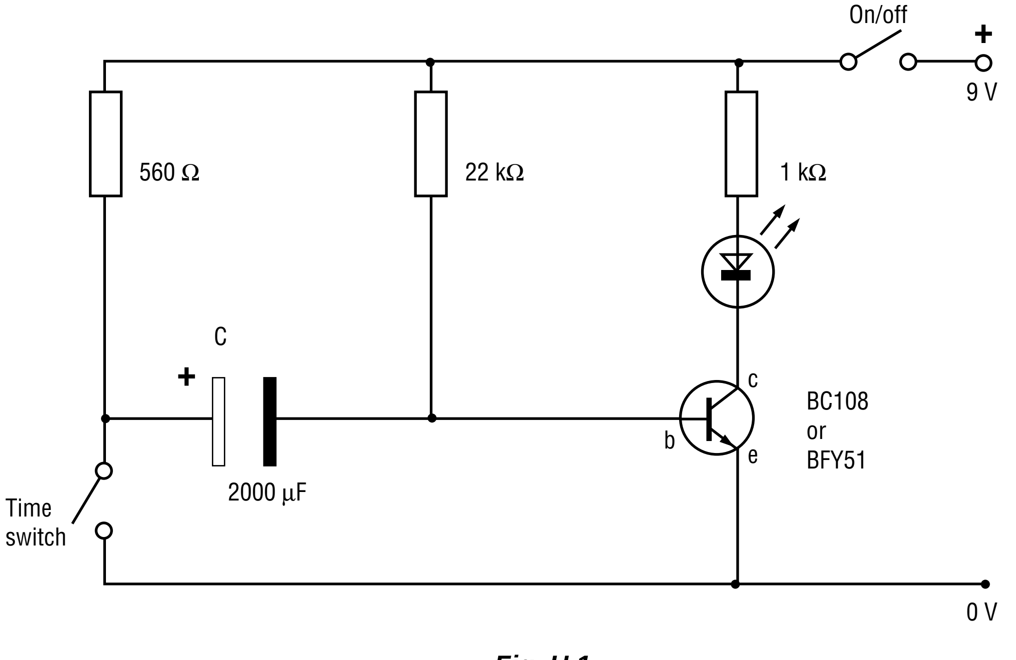


Fig. H.1

APPENDIX I NEON OSCILLATOR DEMONSTRATION

Neon Oscillator Demonstration

Apparatus

Neon lamp (90 V type), resistor e.g. 10 kΩ, capacitor, e.g. 4.7 μF, p.s.u. 100 V d.c.

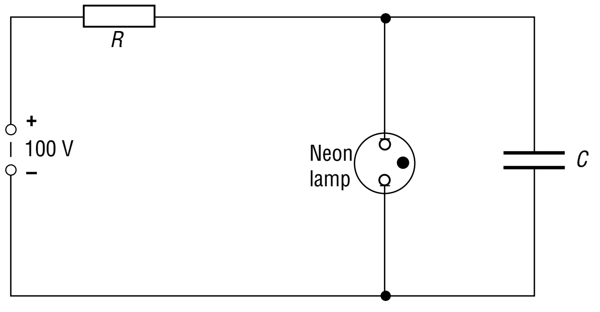


Fig. I.1

Procedure

1. Set up the circuit shown in Fig. I.1 on breadboard; ensure that all connecting wires are properly insulated.
2. Switch on the p.s.u. and observe the neon lamp flashing on and off.
3. Connect the y-plates of an oscilloscope (set to a.c.) across the capacitor. Have the y-gain fairly low and have the time-base switched on. Observe the approximately triangular waveform which is called a sawtooth wave-form, Fig. I.2.
4. Repeat for different values of R and C.

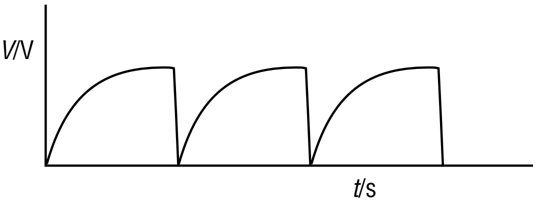


Fig. I.2

Explanation

The capacitor is slowly charged through the resistor. When the voltage reaches ≈ 90 V the neon gas conducts and the capacitor is rapidly discharged. The lamp is quenched and the cycle is then repeated.

Applications

This type of circuit, called a relaxation oscillator, is of use in a number of different situations:

(i) in automatic flashing displays (e.g. advertising);

(ii) an audible tone is heard if a loudspeaker is included in the circuit – selecting different combinations of R and C by means of push-button switches simulates a simple electronic organ;

(iii) it is basically similar to the time-base generator circuitry used in an oscilloscope for sweeping the spot across the screen.

APPENDIX J

DIGITAL MULTIMETERS (DMMs)

Digital Multimeters (DMMs)

Modern DMMs perform a wide variety of functions, e.g. resistance measurement up to 20 MΩ; audible continuity check; a.c./d.c. current measurement up to 10 A; overload protection on all ranges except 10 A, a.c./d.c.; a.c./d.c. voltage measurement; diode test function; transistor test function; capacitance measurement; frequency measurement; temperature measurement.

Some of these functions are discussed below with reference to a fairly standard style DMM. It is good practice to begin with the instrument switched to the highest range of the function to be measured. Always make good firm contact with the probes.

Most multimeters have four main terminals.

(i) The **COM** terminal is joined to the negative or black test lead for voltage, current, resistance, frequency and diode measurement.

(ii) The **Ω V** terminal is joined to the positive or red test lead for voltage, resistance, frequency and diode measurement.

(iii) The **A** terminal is connected to the positive lead for current measurement below 2 A.

(iv) The **10 A** terminal is joined to the positive lead for current measurement below 10 A.

The DMM as an Ohmmeter

This mode may be used to:

(i) verify the formulae for resistors in series and in parallel;

(ii) investigate how the resistance of an LDR varies as the incident light varies;

(iii) investigate how the resistance of a thermistor varies as its temperature changes;

(iv) measure body/skin resistance.

Students might enjoy taking a lie detector test with an ohmmeter. When a person is nervous, anxious or embarrassed, perhaps because he/she has told a lie, the amount of perspiration increases and so skin resistance goes down.

The audible continuity function (indicated on the meter by a wave symbol, i.e.) is particularly useful for checking short and open circuits, e.g. checking fuses and the continuity of some heating elements, etc. A built-in buzzer sounds if the resistance is less than 30 Ω.

The DMM as an Ammeter

Set the ac / dc selecting switch to either the ac or the dc position as appropriate. Connect the black test probe to the COM input terminal and the red test probe to the A terminal. The DMM will now function as an ammeter. For measuring current larger than 2 A, the red probe must be placed in the 10 A terminal. The 10 A range is unprotected and has a very low internal resistance. Consequently, the fuse will blow to protect the circuit if an excessive current flows.

The DMM as a Voltmeter

Set the ac/dc selecting switch to either ac or dc as appropriate. The black test probe is connected to COM and the red test probe to the ΩV input terminal. If the readout is positive when the meter is connected across a circuit, this means that the input at the red probe has a higher potential than the input at the black probe. The converse is true if the readout is negative. With an a.c. signal (i.e. voltage), the DMM measures an average value and displays it as the equivalent root-mean-square value for a sine wave. The most serious limiting factor in the use of a moving-coil (analogue) multimeter for voltage measurement is the rather low resistance the meter presents to the circuit across which it is measuring the voltage (see problem, page 8). Ideally, voltmeters should have an infinitely high resistance so that they do not draw any current from the circuit to which they are connected. An important measure of the loading effect of a voltmeter on a circuit is the ohm-per-volt (o.p.v.) rating of the meter. A poor quality multimeter may have an o.p.v. rating as low as 4000 ohm per volt when switched to the volts range, whereas an electronic (digital) multimeter is likely to have an o.p.v. rating of ten megohm per volt for all voltage ranges.

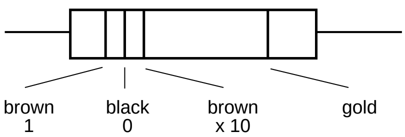
APPENDIX K COLOUR CODES OF RESISTORS

Each coloured band on a resistor represents a number.

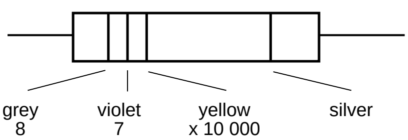
	Band 1	Band 2	Band 3
black	0	0	–
brown	1	1	0
red	2	2	00
orange	3	3	000
yellow	4	4	0000
green	5	5	00 000
blue	6	6	000 000
violet	7	7	0 000 000
grey	8	8	–
white	9	9	–

The fourth band represents the manufacturing tolerance – silver (±10%), gold (±5%), red (±2%), brown (±1%)

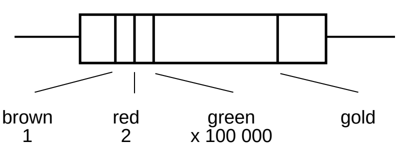
Examples



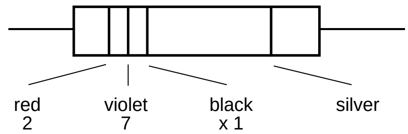
Resistance value 100 Ω
Tolerance ±5%



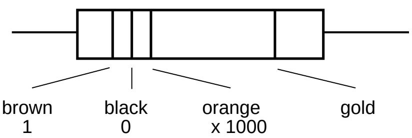
Resistance value 870 000 ohms (870 kΩ)
Tolerance ±10%



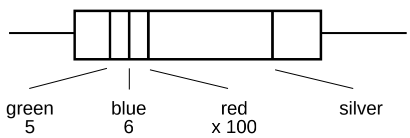
Resistance value 1 200 000 ohm (1.2 MΩ)
Tolerance ±5%



Resistance value 27 Ω
Tolerance ±10%



Resistance value 10 000 ohms (10 kΩ)
Tolerance ±5%



Resistance value 5600 Ω (5.6 kΩ)
Tolerance ±10%

To avoid many zeros, circuit values are often quoted using these prefixes:

mega (M)	1 000 000	(10^{6})
kilo (k)	1 000	(10^{3})
milli (m)	0.001	(10^{-3})
micro (μ)	0.000 001	(10^{-6})
nano (n)	0.000 000 001	(10^{-9})
pico (p)	0.000 000 000 001	(10^{-12})

These numbers are multiplying factors.

British Standards Resistor coding BS1852

This system will be found in use on recently produced resistors and is gradually replacing the colour coding system. It has the advantage of giving no problems to the colour blind. The letters R, K, M, represent $\times 1$, $\times 10^3$, $\times 10^6$, respectively. The position of the letter indicates the position of the decimal point.

Example

R47 = 0.47 ohm	100R = 100 ohm	1R0 = 1 ohm
1K0 = 1 000 ohm	4R7 = 4.7 ohm	10K = 10 kilohm
47R = 47 ohm	10M = 10 megohm	

Tolerances

F	±	1%
G	±	2%
J	±	5%
K	±	10%
M	±	20%

Examples

390RJ	390 ohm 5%
4M7M	4.7 megohm 20%
6K8F	6.8 kilohm 1%

APPENDIX L

COPPER WIRE SIZES

Size (standard wire gauge, s.w.g.)	**Diameter (mm)**	**Resistance (Ω m^{-1})**
8	4.064	0.0013
10	3.251	0.0021
12	2.642	0.0032
14	2.032	0.0054
16	1.626	0.0083
18	1.219	0.0148
20	0.914	0.0260
22	0.711	0.0435
24	0.559	0.0705
26	0.457	0.105
28	0.376	0.155
30	0.315	0.222
32	0.274	0.293
34	0.234	0.404
36	0.193	0.590
38	0.152	0.950
40	0.122	1.48
42	0.102	2.14
44	0.0813	3.34

Table L.1 *Copper wire sizes*

Note. The s.w.g. standard has been replaced with a millimeter standard (from 0.020 mm to 4.000 mm).

APPENDIX M ESB DECLARED SUPPLY VOLTAGE

ESB Declared Supply Voltage*

It has been decided that with effect from 1 January 1991 the nominal or declared voltage of low tension supplies to ESB customers will be 230 V (single-phase) and 400 V (three-phase). These will replace the values which have applied hitherto, 220 V and 380 V, respectively.

Electricity appliances are designed for a particular voltage, and they operate best when the voltage is at this nominal value. In general, if an appliance is operated at a voltage other than the nominal value, then either it performs less effectively, or its life is reduced. However, because variation of the voltage is unavoidable, standards are laid down which prescribe limits within which the voltage should be maintained, and within which both the performance and the life of electrical appliances should be acceptable.

There is an increasing trend towards standardisation of electricity supply voltages throughout the world. In Europe this trend has been given special force in pursuance of the single market. IEC (world-wide) and CENELEC (Europe)** have both adopted the following standard:

Nominal Voltage	**:**	**230 V (single-phase)**
	:	**400 V (three-phase)**

Under normal system conditions the voltage at the supply terminals should not differ from the nominal value by more than ±10%.

(This is the standard adopted for the type of supply system prevalent in Europe. A different standard applies to North-American-type systems.)

IEC and CENELEC have prescribed that the nominal voltage of existing systems should evolve towards the recommended standard in as short a period as possible, but not later than 2003. It is envisaged that after that date the tolerance range may be reduced, possibly to ±6%. During the period of transition towards the universal achievement of 230/400 V ±10%, they have recommended that countries whose nominal voltage is 220/380 V should bring their voltage within the range 230/400 V, +6%, –10%; and that countries whose nominal voltage is 240/415 V should bring their voltage within the range 230 V, +10%, –6%.

Many 220/380 V countries, including all such in the EC, have accepted this recommendation and are implementing it.

As long ago as 1979 ESB adopted the voltage limits of the above recommendation, i.e. an upper limit of 244 V (230 V +6%) and a lower limit of 207 V (230 V –10%), with corresponding values for three-phase supplies. However, no change was made at that time in the declared nominal voltage, which remained at 220 V.

As far as possible ESB voltages are maintained between the above limits. (All concerned, including our customers, should be aware that these limits apply at the point of supply: further variation of the voltage takes place in the electrical installation within the customer's premises.) As demand grows on the distribution networks, the variation of voltage tends to increase, so that reinforcement of the networks becomes necessary to ensure that the limits are not breached, particularly the lower limit of 207 V.

*Text of ESB circular issued October 1990
**IEC: International Electrotechnical Commission
CENELEC: European Committee for Electrotechnical Standardisation

ESB already has a substantial programme of system improvements of distribution networks, particularly in rural areas, to ensure that the standard is implemented.

Although most of the effort of controlling voltage is concentrated on ensuring that the lower limit is not breached, the pattern of electricity demand is such that, for most of the time and for most customers, the voltage tends to be nearer the upper limit of the range. In practice, the average ESB voltage is nearer to the new international norm of 230 V than to ESB's hitherto declared value of 220 V – the average is probably somewhat over 230 V for urban customers, and somewhat under 230 V for rural customers.

To an increasing extent, electrical appliances are being designed for the nominal voltage of 230 V. In fact, CENELEC Memorandum 14 recommends that "from 1 January 1993 at the latest, manufacturers mark equipment with the rated voltage 230 V or 400 V". 220 V appliances will become progressively obsolete.

In summary, the new nominal voltages of 230 V and 400 V are in accordance with international norms; they are close to the actual voltages received by most of our customers for most of the time; and they match the nominal voltage of an increasing proportion of the appliances being operated by our customers.

Finally, it should be noted that declaring the new nominal supply voltage will make no difference in regard to the voltages actually received by our customers. These derive from already established operating practices which are determined solely by the upper and lower limits, 244 V and 207 V respectively. There is no change in ESB operating voltages at low tension.

October 1990

MODULE 4

Mechanics I

CONTENTS

CHAPTER 1

LINEAR MOTION

1.1 Background

Units

Measurement is a very important component of physics (and indeed all science) and measurements cannot be made without units. The system of units used in physics today is called SI (SI stands for Système Internationale d'Unités - the International System of Units). This is based on seven base units, from which all other units are derived. These base units are given in Table 1.1.

Quantity	Unit	Symbol
Length	metre	m
Mass	kilogram	kg
Time	second	s
Temperature	kelvin	K
Electric current	ampere	A
Light intensity	candela	cd
Quantity of substance	mole	mol

Table 1.1 *SI base units*

This system evolved from the MKS system which was based on the metre, kilogram and second. Other systems previously in use were the CGS system, based on the centimetre, gram and second, and the imperial system, based on the foot, pound and second. None of the older systems is as satisfactory as the SI system, which includes nearly all units used in any branch of science. A few non-SI units are still in general use, e.g. centimetre, litre.

As has been stated, all other units are derived from the base units. For example, the unit of velocity is metre per second. Less obviously, the unit of force, the newton, is the kilogram metre per second per second, as the definition indicates.

Rules in the SI System for Writing Units and Their Symbols

1. The names of all units are written with lower case letters, not with initial capitals, despite many of them being derived from famous scientists' names, e.g. although the scientist's surname was Watt, the unit named after him is the watt.

2. In contrast, the symbols for units are given capital letters if the name of the unit is derived from the name of a person, but lower-case letters if this is not the case. Thus the symbol for the watt is W and for the ampere is A, but for the metre the symbol is m and for the second the symbol is s.

3 Multiples and sub-multiples of units are indicated by the standard prefixes, increasing and decreasing by factors of 10^3, Table 1.2.

Prefix	Symbol	Multiple
tera	T	10^{12}
giga	G	10^{9}
mega	M	10^{6}
kilo	k	10^{3}
milli	m	10^{-3}
micro	μ	10^{-6}
nano	n	10^{-9}
pico	p	10^{-12}
femto	f	10^{-15}
atto	a	10^{-18}

Table 1.2 *Prefixes for units*

Each of these prefixes can be used with any of the base units (except for kilogram, of course), e.g. kilonewton, nanosecond, attogram. The prefix and the unit make one word, with no spaces. Similarly the symbols for these units have no spaces between the letters, e.g. kN, ns, ag.

These are the only prefixes in the SI system; other prefixes, which used to be used in the metric system, are used in a small number of cases to create non-SI units, e.g. centi (10^{-2}) in centimetre, deci (10^{-1}) in decimetre (used in the definition of a litre: 1 litre = 1 dm^3; the litre is not an SI unit), and hecto (10^2) in hectare and hectopascal.

4. When two units are used to form a derived unit which has no separate name, e.g. metre per second or ohm metre, the symbols for the two units are written with a space between them, e.g. m s^{-1}, Ω m. This, of course, is to prevent confusion between a symbol with a prefix, e.g. ms (millisecond) and two separate symbols, e.g. m s (metre second). There should always be one space between the number and the unit as well, e.g. 12 m s^{-1}.

5. The recommended method of showing division of symbols is to use a negative index, e.g. metre per second is m s^{-1}. This makes it quite clear which quantities are divided by which, e.g. J kg^{-1} K^{-1}. As an alternative, the solidus (/) is permitted (but not recommended). If a solidus is used, is must be used once only in any set of symbols. Thus, J/kg K is permitted but not J/kg/K. As negative indices are used in all official examination papers and textbooks, and as they are used in all scientific literature, it is highly recommended that students be taught to use them instead of the solidus.

In connection with measurement, it should also be noted that international practice is now to write large numbers without commas to separate groups of digits, but with spaces between groups of three digits instead, e.g. 3 456 789 rather than 3,456,789. This is to avoid confusion with the continental practice which is to use a comma instead of a full stop as a decimal. A person from continental Europe would read 123,456 as being between 123 and 124 and not between 123 000 and 124 000.

The definitions of the three base units commonly encountered in mechanics are given below. The ideal is that these units can be independently measured at any place. This is the case for the metre and the second but not for the kilogram.

The metre is defined as the distance travelled by light in a vacuum during a time interval of 1/299 792 458 of a second.

The second is defined as the time interval occupied by 992 631 770 cycles of a certain radiation of a caesium-133 atom.

The kilogram is the mass of a piece of platinum-iridium alloy kept at the Bureau Internationale des Poids et Mésures at Sèvres, France.

Motion Under Gravity

In ancient times, the Greek philosopher **Aristotle** (384 BC - 322 BC) stated that heavier bodies fell faster than lighter ones, and this was accepted as true for many centuries. It was the Italian **Galileo** (1564 - 1642) who carried out experiments to show that this was not the case for most bodies (except those for which air resistance was significant). Tradition says that he carried out these experiments by dropping bodies of different mass from the Leaning Tower of Pisa, but this does not appear to have any basis in fact. While some of his work was carried out when he was lecturer in mathematics at the University of Pisa, most of it was done after he moved to the University of Padua. In 1604, he produced his proof that bodies falling under gravity do so with uniform acceleration.

For a body to fall with the acceleration due to gravity, the air resistance must be so low that it can be neglected. This is true for most bodies falling at slow speeds, except very light bodies with large air resistance such as a feather. A feather will fall very quickly in a vacuum. An apparatus, consisting of a feather in a long glass tube which can be evacuated by a vacuum pump, is available to demonstrate this.

At higher speeds, air resistance becomes larger for all bodies. As the speed increases the resistance also increases and eventually a point is reached where the resultant force on the body is zero. (The resultant force, $F = W - B - R$, where W, B, and R are the weight, the buoyancy force (the upthrust) and the resistance force, respectively.) When the resultant force is zero the acceleration is also zero and so the velocity is a maximum. This maximum velocity is known as the terminal velocity. The terminal velocity of a human falling freely (e.g. a sky-diver) depends on the attitude of the body as it falls; normally a sky-diver falls in a prone position, and the terminal velocity is then about 200 km h^{-1} (c. 50 m s^{-1}). If (s)he were to fall in a diving position (head-first), the terminal velocity would be much higher, as air resistance would be less.

1.2 Do You Know

Copies of the standard kilogram are kept in many countries.

The Irish copy is kept by the Metrology Laboratory in Glasnevin, Dublin. It is sent to Sèvres every 5 years to compare it with the standard kilogram; on its last trip it was found to have lost a few micrograms. The US has a copy of the standard kilogram called prototype kilogram no. 20. It has been taken to Sèvres twice to compare it with the primary standard. When it is moved from its vault, two people must always be present, one to carry the kilogram, and a second to catch it if the first person falls.

The metric system was invented in France after the French Revolution.

In 1791 the National Assembly of France appointed a commission of 12 outstanding men of science to revise the French system of weights and measures. They came up with a system which was so simple to use that it has become the international standard. They based their new unit of length, the metre, on the size of the earth; it was one ten-millionth of the distance from the equator to the North Pole along the meridian through Paris. The unit of mass, the gramme (the spelling has since been revised), was the mass of 1 cubic centimetre of water at 4 °C.

Prior to the invention of the metric system, the old measurements were based on measurements of parts of the body or other common objects. Many of these units were sufficient for approximate measurements but, when measurement became more exact, they would have caused much confusion, for one person’s inch, foot, mile, etc., would differ from another’s. It was necessary therefore for governments to pass laws setting out and standardising the measurements which were to be used for official and legally binding transactions. The British system of units is known as imperial measure, as referred to below. Each country or group of countries had its own units; one edition of the Encyclopaedia Britannica gives $3\frac{1}{2}$ pages to a listing of units used in different countries.

Measurements such as the ‘foot’ and the ‘hand’ (used for measuring horses) are obvious. The ‘inch’ was the distance between the first two knuckles on the second finger, the ‘span’ was the distance from the thumb to the little finger of an outstretched hand, the ‘cubit’ was the distance from the tips of the fingers to the elbow, and the ‘yard’ was the size of the king's waist! A 'fathom', used in nautical measurements, was the length of the outstretched arms, and thus about 6 feet. The 'perch' (also called the 'pole' or 'rod') was the length of a rod used for measuring land, usually $5^1/_2$ yards long (but an area of a square perch was also called a pole). A 'chain' was used in surveying; the standard chain was 22 yards long, divided into 100 'links', and 10 chains made 1 furlong. A 'furlong' (still used in horse-racing) was originally the length of a furrow in a standard (10 acre) field, and was later standardised as one-eighth of a mile, the same length as the Roman measure called a 'stadium'. The Romans invented the 'mile'; it was 1000 (double) paces - the name comes from the Latin mille, meaning 1000 (from which we also get the prefix milli). The Roman mile was about 1618 yards, but the mile was later defined in law (the statute mile) as 1760 yards. An Irish mile was longer, 2240 yards. The imperial distance table was:

12 inches	= 1 foot
3 feet	= 1 yard
$5^1/_2$ yards	= 1 perch, pole or rod
(or 22 yards	= 1 chain)
40 perches	= 1 furlong
(or 10 chains	= 1 furlong)
8 furlongs	= 1 (statute) mile

Two other non-SI units of length which are occasionally found are the 'micron', which is another name for a micrometre, and the 'angstrom' (symbol Å) which is used in crystallography; it is 10^{-10} metre or 100 pm, and is used to measure the distances between atoms in crystals. This unit was called after the Swedish physicist **Anders Jonas Ångström** (1814 - 1874).

Astronomical distances need larger units than these, for even the distance from the sun to the earth, 150 million km or 93 million miles, is a rather large figure when expressed in normal units of length. Astronomers use three different units of distance. The first, the 'astronomical unit' (AU), is the mean distance from sun to earth, 1.496×10^{11} m. The AU is used for measurements around the solar system, e.g. the planet Neptune is 30.06 AU from the sun. For larger distances, e.g. between stars and galaxies, the 'light-year' (ly) is commonly used; it is the distance light travels in one year, 9.46×10^{15}m. The nearest star, Proxima Centauri, is over 4 light-years away. The third unit, the 'parsec' (pc), is the distance at which one astronomical unit subtends an angle of one second of arc. One parsec is equal to 3.086×10^{16} m or 3.26 light-years.

Land area was measured in 'square perches' or 'poles' (see above); in 'roods', which were sometimes the same as square perches but later standardised as 40 square perches; and in acres. The 'acre' was originally the area of land which a team of oxen could plough in one day, later 4 roods or 4840 square yards. 640 acres made up a 'square mile'. The table of square measure was:

144 square inches	= 1 square foot
9 square feet	= 1 square yard
$30^1/_4$ square yards	= 1 square perch
40 square perches	= 1 rood
4 roods	= 1 acre
640 acres	= 1 square mile

Units of volume (or capacity) included the 'pint', 'quart', 'gallon' and 'bushel'. The English (imperial) gallon contains 4.546 litres, but the American (or Queen Anne) gallon, formerly used for measuring wine, is only 3.79 litres. A quart, as the name suggests, is a quarter of a gallon, while a pint is half a quart and a 'gill' or 'noggin' (sometimes a naggin) is a quarter of a pint - this measure was used for spirits in pubs. A 'peck' was 2 gallons and a bushel contained 8 gallons. All of these were used not only for the volume of liquids and for solids in small pieces (e.g. grain), but were also the names of vessels of different size. So the expression 'to hide one's light under a bushel' refers to a bushel-sized container. Of course, not all containers were exactly the same size, so the expression 'Put the milk in the small gallon' , while sounding peculiar to a scientist, would have a clear meaning for those familiar with the particular container. For medicines and other purposes, the 'minim' (about the size of a drop) and the 'fluid ounce' (a volume equal to that of 1 ounce of water at 62 °F - about 17 °C) were used. Wine bottles holding different volumes of wine were given names of ancient kings of Israel, Judah and Babylon; thus we can find bottle sizes called the 'jeroboam', the 'rehoboam' and the 'nebuchadnezzar' , while other names were given to barrels of different sizes; 'cask', 'firkin', 'kilderkin', 'hogshead', 'pipe' and 'tun'. The actual size of these varied with the liquid stored in them, e.g. a hogshead of liquor held 52.5 gallons, while a hogshead of other liquids held 100 to 140 gallons. The table of imperial capacity measures was:

60 minims	= 1 fluid dram
8 fluid drams	= 1 fluid ounce
5 fluid ounces	= 1 gill
4 gills	= 1 pint
2 pints	= 1 quart
4 quarts	= 1 gallon
2 gallons	= 1 peck
4 pecks	= 1 bushel
8 bushels	= 1 quarter

There were several different scales for measuring mass (called weight, of course). Troy weight (named from the French town of Troyes, where these weights were used at a fair) was used for measuring precious metals and precious stones, while the more common scale was called avoirdupois (clearly from the French 'to have weight' , but always pronounced in English fashion). There was also apothecaries' weight, used for measuring medicines, etc. The smallest unit of weight, the same in all three scales, was the 'grain', which was originally the weight of a grain of wheat taken from the middle of the ear. Troy units included the 'pennyweight' (the weight of a silver penny), the 'ounce' (derived from the Latin word, *uncia*) and the 'pound' (Latin *libra*, hence the abbreviation lb.). The troy pound contained only 12 ounces. Apothecaries' weight used units called the 'scruple', the 'drachm' and the ounce. In avoirdupois weight, the smallest unit was the 'dram' (formerly spelt drachm); the name comes originally from the Greek silver coin, the drachma. 16 drams made an ounce and 16 ounces a pound. The 'stone' varied from 7 to 24 pounds in different communities, but was later standardised as 14 lbs. Heavier weights were the 'quarter' (28 lbs or 1/4 cwt, hence the name), the 'hundredweight' (cwt), which in imperial measure was 112 lbs but in the USA is 100 pounds, and the 'ton', 2240 lbs in imperial measure but 2000 lbs in the USA (sometimes called a short ton). The name ton is derived from tun, a large container for holding wine, etc. By good fortune, the imperial ton is very close to 1000 kg (2205 lbs), the metric tonne. When the metric system was introduced into the building trade, many of the workers took to calling the (metric) tonne the 'tunny', to distinguish it from the old imperial ton. The table for avoirdupois weight was:

16 drams	= 1 ounce
16 ounces	= 1 pound
14 pounds	= 1 stone (or 28 pounds = 1 quarter)
8 stone	= 1 hundredweight (or 4 quarters = 1 hundredweight)
20 hundredweight	= 1 ton

The troy weight table was:

24 grains	= 1 pennyweight
20 pennyweight	= 1 ounce
12 ounces	= 1 pound

The apothecaries' weight table was:

20 grains	= 1 scruple
3 scruples	= 1 drachm
8 drachms	= 1 apothecaries' ounce

About the only other unit of mass referred to today is the 'shekel' . This was an old Jewish measure of weight (about 11.4 g) and also a coin used in Palestine.

When one compares the multiplicity of units of the old systems (of which only a part are listed above), and looks at the complicated tables which were used to relate the units, the simplicity of the SI system becomes apparent. Older teachers will remember learning tables as given above. Much time was spent in maths classes working out problems such as 'How much would it cost to surface 2 miles 6 furlongs 24 perches 3 yards of road at 3 shillings and sixpence per perch?' (of course, calculators were unknown in those days!)

What is time?

Everyone 'knows' what time is but most people would find it very difficult to define. **St Augustine** (354-430) is quoted as saying, 'What is time? If no one asks me, I know; if I wish to explain it to one who asks, I know not.' The American physicist, **Richard Feynman** (1918-1988) offered the following explanation: 'Time is what happens when nothing else happens'.

The first unit of time which was measured by ancient man was the year, because he needed to know when the season for sowing was near and also because many primitive societies worshipped the sun as a god. Prehistoric monuments like Newgrange were built with a particular orientation relative to the sun at a certain time. Newgrange is aligned with the rising sun on the winter solstice, while Stonehenge, in the south of England, marks both the summer and winter solstices. Many other prehistoric structures have been discovered which are aligned with the sun's position at midsummer, midwinter or the start of the agricultural year. There is no evidence that these structures were used to sub-divide the year into smaller units of time.

The Babylonians were probably the first people to use the month as a unit of time, basing it on the moon's phases. Their astronomers divided the year into 12 months of 30 days each, which meant that their calendar needed frequent correction. About 3000 BC, the Egyptians adopted a year of 365 days - they needed an accurate measurement of the year to forecast the flooding of the Nile. About 240 BC, the idea of a leap year day was suggested by one of the Ptolemies, rulers of Egypt. However, the introduction of this concept had to wait until Julius Caesar adopted it for the Roman Empire in 46 BC. This Julian calendar remained in use for over 1500 years, even though there was a gain of 1 day every 128 years, so that the seasons became more and more out of sequence with the calendar.

In 1582, **Pope Gregory XIII** (1502-1585) decreed that the leap year system should be revised, so that the years at the turn of the century, ending in '00', would not be leap years unless they were divisible by 400. The Gregorian calendar is accurate to 1 day in 3323 years. While most Roman Catholic countries adopted this new calendar immediately, and moved their dating forward by ten days, the Protestant countries, so soon after the Reformation, were not prepared to follow a command of the Pope, and it took many years before the Gregorian calendar was adopted all over Western Europe. England was the last country in the West to change to the new calendar; in 1752 eleven days were removed from the calendar, causing considerable unrest. It was not that people were confused as to what it meant, and were led to believe that their lives would be eleven days shorter, but that rents were due to be paid quarterly on fixed dates, and therefore the next payment would be expected eleven days earlier, and the landlords would receive rent for the non-existent days. It was not until 1918 that Russia adopted the Gregorian calendar, for the Orthodox church there had always resisted the change; it took the Communist revolution to change this. As a result of the change, 13 days were dropped from the calendar there, so that the October Revolution was celebrated on November 7 while the Communists remained in power. The same happens in this country, for the Battle of the Boyne, fought on 1 July 1690, is celebrated on 12 July - the difference is the 11 days lost in 1752. This is why the tax year ends on 5 April; before the eleven days' change it ended on 25 March, which was Lady Day, the end of the first quarter of the year.

While measuring the length of the year and the month were achieved early, measuring shorter intervals of time was more difficult. The Egyptians adopted a 24-hour day, but the lengths of the hours varied with the seasons, as there were always twelve hours in the day and twelve at night. The only methods of measuring time available then were the sundial and the water-clock, which could measure time by allowing water to flow through a narrow opening and measuring the volume flowing. (It is said that Roman lawyers were allowed a certain amount of water for their speeches, but sometimes the attendant who watched the flow was bribed to allow mud in, which would clog the outlet and so extend the time.)

When mechanical clocks were invented, it was necessary for the length of the hour to become constant. The first primitive mechanical clocks were made in China about 800 AD. The first clocks in Europe, about 1300 AD, were so inaccurate that they only had one hand – an hour hand; they could lose or gain 15 minutes per day. It was not until Galileo discovered that a pendulum could be used to control the speed of a clock that greater accuracy was possible. The story that Galileo discovered the constancy of the periodic time of a pendulum, despite changes in the amplitude of swing, by watching a lamp swinging in the cathedral at Padua while timing it with his pulse, is found in a contemporary biography, but some sources doubt the accuracy of the incident. It was not until near the end of his life that Galileo is actually recorded as suggesting that a pendulum could be used to control a clock - this was after he had gone blind. The first pendulum clock was built, not by Galileo, but by **Christiaan Huygens** (1629-1695) in Holland 14 years after Galileo's death; he followed this, some 20 years later, with the first watch controlled by a hairspring.

Clocks were further developed to improve their accuracy; one reason for this was the need for an accurate clock to use in measuring longitude on ships. This depends on knowing the time accurately; if a captain of a ship has a clock which is set to, e.g. Greenwich time, he can find his longitude by determining local time using the sun or the stars, and then comparing the two times. When a fleet of British warships, under Admiral Sir Cloudesley Shovel, was wrecked on the Isles of Scilly, losing 4 ships and 2000 men (including the admiral), all due to mistakes in calculating their longitude, the British Government, in 1714, offered a prize of £20 000 for the invention of a clock which would keep accurate time on board ship, and so allow the longitude to be calculated to within half a degree. **John Harrison** (1693-1776), the son of a Yorkshire carpenter, eventually won the prize; he built a total of 4 clocks. The first one, weighing over 30 kg, was submitted for the prize in 1734, and he was given £500 to develop the concept further. He continued to work on building better clocks for over 25 years. His fourth clock, the size of a large alarm clock, was tested in 1761. The ship carrying it headed for the island of Madeira. Nearing it, the ship's navigator calculated the longitude to be 13° 19' west, and predicted that the island would be sighted the next morning, which it was. Had the ship not had the clock on board, it would have missed the island altogether, 'the consequence of which', says a contemporary report, 'would have been inconvenient, as they were in want of beer'. While Harrison's clocks were successful, it was the Frenchman, **Pierre LeRoy**, who invented a different design on which later chronometers were based.

Modern clocks do not use pendulums or hairsprings. It was discovered that a suitably shaped quartz crystal will vibrate at a constant rate, and can be used to control the speed of an electric clock. The first quartz clock, built in 1928, achieved an accuracy of 1/1000 second per day - about ten times better than the best pendulum clock. For even more accurate timing, the atomic clock was developed; the second is now defined in terms of the vibrations of a certain wavelength of light emitted by a caesium atom. A caesium clock can measure time to an accuracy of one part in 10^{10}, or less than $^1/_4$ second in a lifetime.

Who needs such accurate timing? The electric power companies do, for they need to keep all their generators synchronised at the same frequency. Radio and TV studios need to keep their times synchronised to a high degree of accuracy also. Radar depends on measuring the length of time a radio signal takes to bounce off an object and return, and this is a very short time. Satellite location systems, such as the US Global Positioning System (GPS), depend on measuring the tiny time difference between radio signals arriving at the receiver from different GPS satellites. Researchers investigating the different sub-atomic particles need to be able to detect particles with lifetimes measured in tiny fractions of a second. Even the common home computer needs to be able to measure time to an amazing degree of accuracy, for a processor working at 200 MHz completes a single operation in 5 nanoseconds (or 5×10^{-9} s) - to say nothing of supercomputers, which are heading towards teraflop speeds - 10^{12} flops (<u>fl</u>oating point <u>op</u>erations per second), so measuring time in units of picoseconds (10^{-12} s).

The measurement of time to the degree of accuracy now possible causes perhaps more problems than any other unit. The trouble is that the length of the day is not constant, but varies slightly. The definition of the second given above is so accurate that the difference in the length of the day is measurable and it must be taken into account, e.g. for navigation. Apart from this seasonal effect, the length of the day is also increasing very slowly, mainly due to the effect of the moon's gravity. This, as we know, causes the tides, and the friction of the water slows the earth's rotation very slightly.

Atomic clocks, on which the definition of the second is based, are extremely accurate. In order to keep atomic time synchronised with solar time (time determined from the rotation of the earth) an additional second - a leap second - is inserted in atomic time as required.

How big and how small?

The range of measurements which man has succeeded in measuring is often difficult for our minds to conceive.

	Years	Seconds
Age of universe	1.5×10^{10}	5×10^{17}
Age of sun	1×10^{10}	3×10^{17}
Age of earth	5×10^{9}	1.6×10^{17}
First life on earth	2×10^{9}	6×10^{16}
Time since dinosaurs	1.6×10^{8}	5×10^{15}
Length of existence of Homo sapiens	1×10^{5}	3×10^{12}
Human lifetime	10^{2}	3×10^{9}
Lifetime of an average human cell	0.5	1×10^{7}
Time for light to travel from sun to earth		5×10^{2}
Time for light to travel around earth		1.3×10^{-1}
Time for 1 cycle of mains a.c.		2×10^{-2}
Time for 1 vibration of FM radio signal eg. RTE 2FM		1×10^{-8}
Time for light to travel across an atom		3×10^{-19}

Table 1.3 *Measurements of time*

	Light-years	Metres
Diameter of the universe (estimated)	10^{10}	10^{26}
Distance of the most distant visible object (quasar)	10^{9}	10^{25}
Diameter of the local group of galaxies	10^{7}	10^{23}
Distance of Andromeda galaxy (nearest major galaxy)	10^{6}	10^{22}
Diameter of Milky Way galaxy	10^{5}	10^{21}
Distance of star Canopus (second brightest star)	10^{3}	10^{19}
Distance of nearest stars	10	10^{17}
Farthest extent of solar system		10^{13}
Radius of earth's orbit		10^{11}
Diameter of moon's orbit		10^{9}
Diameter of the earth		10^{7}
Length of the River Liffey		10^{5}
Distance covered by a car in a minute		10^{3}
Diameter of a boxing ring		10^{1}
Diameter of a flower		10^{-1}
Smallest division on a ruler		10^{-3}
Grain of sand		10^{-5}
Size of a bacterium		10^{-7}
Diameter of a large atom		10^{-9}
Diameter of a nucleus		10^{-15}

Table 1.4 *Measurements of length*

Length of all blood vessels	6×10^{6} m	6000 km
Length of intestine	7 m	
Area of skin	2 m^2	
Area of gaseous exchange in lungs	50 m^2	
Volume of air filling lungs	5×10^{-3} m^3	5 litre
Average volume of air inhaled while resting	5×10^{-4} m^3	0.5 litre
Volume of blood in body	5×10^{-3} m^3	5 litre

Table 1.5 *Measurements related to human body*

1.3 Conceptual Approach

Displacement

Displacement is defined as 'the distance from a particular point in a given direction'. Two points must be made about this.

(a) The magnitude of the displacement is not the same as the distance, if the distance is not in a straight line. For example, if a body travels 5 m north and then 3 m south, its final displacement will be 2 m north, as it finishes up 2 m north of its starting point, but it will have covered a total distance of 8 m.

(b) Since the definition of displacement includes direction, it is clear that displacement is a vector quantity. Distance is a scalar quantity, as it is concerned only with length; the direction is not included.

Velocity

Velocity is defined as the rate of change of displacement with respect to time. The concept of 'rate of change' is often found difficult by students. It may be introduced by confining the discussion to cases of constant velocity and describing velocity as 'displacement divided by time taken'. Strictly speaking, this is of course the average velocity over the period of time in question. However, if the velocity is constant the instantaneous velocity at a particular instant is numerically equal to the average velocity. The concept of rate of change then follows by considering the average velocity over shorter and shorter times.

Velocity is a vector quantity. The corresponding scalar quantity is speed - the rate of change of distance with respect to time. While the distinction is sometimes ignored this is incorrect and confusing for students.

Acceleration

Acceleration is defined as the rate of change of velocity with respect to time. A similar approach may be taken to this definition as to the definition of velocity. There is an obvious progression from displacement to velocity to acceleration and this should be emphasised. The further progression to rate of change of acceleration, etc., may be referred to though these quantities are not encountered at Leaving Certificate level.

It should be emphasised that acceleration is the rate of *change* of velocity and not the rate of *increase* of velocity. Any change in velocity constitutes an acceleration. The change may be an increase (as in the everyday meaning of the word), a decrease (sometimes referred to as a deceleration or a retardation, although these terms are not defined and should be avoided), or a change in direction (as occurs in circular motion).

The unit of acceleration, the metre per second per second ($m\ s^{-2}$) can be a difficult one for students to understand. One approach to this unit is to begin with a problem such as, 'The velocity of a car increases from 5 m.p.h. to 25 m.p.h. in a particular direction in 5 seconds. What is its acceleration?' This leads to an answer of 4 m.p.h. per second in the same direction as the velocity, which is obvious to most, and avoids the confusion generated by 'per second per second'. Having then stressed the fact that the unit of time appears twice in the unit of acceleration, one may then proceed to change the unit of velocity to metres per second, and deduce the unit of acceleration as metres per second^2.

Momentum

The principle of conservation of momentum states that the total momentum in a closed system is constant. The principle can be demonstrated using Newton's laws of motion, as follows.

When two bodies A and B collide, A exerts a force on B, and B exerts an equal and opposite force on A (Newton's third law, see p. 26). Since force is proportional to rate of change of momentum (Newton's second law), the rate of change of momentum of A will be equal and opposite to the rate of change of momentum of B. The two forces act for the same length of time, therefore the total change in momentum of A will be equal and opposite to the total change of momentum of B. If the momenta of the two bodies are added together, their sum before the collision will be equal to their sum after the collision, i.e. momentum is conserved in the collision.

The principle only applies in a closed system, i.e. when no external force(s) act on the colliding bodies. An external force is one which is exerted on these bodies by a body other than the colliding bodies. Such an external force will cause a change in momentum of the body on which it acts and this will mean that the total momentum of the system will change, i.e. it will not be constant.

While momentum is conserved in collisions in closed systems, kinetic energy is not usually conserved in such collisions. The final kinetic energy will normally be less than the initial kinetic energy, as some of the energy will be converted to heat and/or other forms of energy. However, at a microscopic level kinetic energy is conserved, e.g. in collisions between molecules.

1.4 Applications

Newton's Cradle

A simple demonstration of the principle of conservation of momentum can be carried out using the executive toy called a Newton's cradle. This consists of five metal spheres, each suspended by two threads from a frame. If one sphere is lifted up and let swing back to hit the four stationary spheres, it will come to rest, and the last sphere will swing out on the other side to the same height. Since the last sphere rises to the same height as the first one fell from, its initial velocity must be the same as the final velocity of the first (in this demonstration it is assumed that conversion of kinetic energy to other forms is negligible). Since the masses of the two spheres are the same it follows that the momentum lost by the first must be equal to the momentum gained by the second. This can be repeated with two, three and four spheres, and each time the same number of spheres will swing to the same height on the other side, while the remainder will be brought to rest by the collision. In each case it is clearly demonstrated that momentum is conserved in the collision. Do not let the spheres continue to swing and collide, for gradually all will start to move in a chaotic manner, and the point of the demonstration will be lost.

In this demonstration it may be asked why, when one sphere is dropped with a certain speed, it is not the case that *two* others move off with *half* the speed, as this would still be consistent with the principle of conservation of momentum. The reason is that this would not result in the conservation of kinetic energy. For example, if the final speed of the first sphere is u and two others move off with initial speeds of $^1/_2u$, then the initial kinetic energy is $^1/_2mu^2$ and the final kinetic energy is $^1/_2m(^1/_4\,u^2 + {}^1/_4\,u^2)$, i.e. $^1/_4mu^2$.

Jet Aircraft and Spacecraft

The motion of jet aircraft and of spacecraft can be explained by the principle of conservation of momentum as an alternative to using Newton's third law. In the case of the jet aircraft, the engine takes in cold air, heats it and expels it, along with the exhaust gases, with a much greater velocity. Since the air and the exhaust gases are continuously gaining momentum backwards, and since momentum is conserved, the aircraft gains an equal momentum forwards, and so it accelerates. This continues until the forces opposing motion are equal to the force exerted on the engines by the expelled air and exhaust gases. A similar situation obtains in the movement of a ship (see Example 1, p. 28).

The essential difference between spacecraft and jet aircraft is that spacecraft must travel where there is no air, so all of the propulsion must come from expelled exhaust gases. Also because there is no air, spacecraft must rely on expelled gases for all changes in velocity - decreasing speed and changing direction as well as increasing speed.

Firing a Gun

When a gun is fired, the momentum gained by the bullet results in an equal and opposite momentum being gained by the gun; the momentum before the gun is fired is zero, so that the total momentum after firing must also be zero. Thus, as the bullet moves forward the gun moves back, i.e. recoils. In the case of a rifle, for example, this recoil makes it difficult to re-aim the gun quickly. In the ‘recoilless’ rifle the mechanism is arranged so that the gases generated by the firing of the gun are expelled in the opposite direction to the bullet. Thus most of the backwards momentum is accounted for by the momentum of the gases and the recoil of the gun is greatly reduced.

Other Collisions

Conservation of momentum also applies in the games of billiards, snooker and pool, when the balls collide. Calculations with these usually involve using two dimensions, as the line of action of the forces which the balls exert on each other will not usually be the same as the direction of travel of the moving ball, and the balls will head off in different directions.

Conservation of momentum does not usually apply in other sports, because forces act during the collision. For example, when a ball is hit by a tennis racquet, the hand of the player applies a force to the handle of the racquet, so that the principle does not apply, since the condition for conservation of momentum is that no external forces act. However, if the player were not in contact with the ground when striking the ball then the player, racquet and ball would constitute a closed system and the momentum gained by the ball in one direction would be equal to the momentum gained by the racquet and player in the opposite direction.

1.5 Worked Examples

Equations of Motion

The derivations of these equations can be found in all textbooks. It must be emphasised that the equations are only applicable in cases where the acceleration is constant. While $s = ut + \frac{1}{2}at^2$ is usually derived by a graphical method, taking s as the area under the graph of v against t, it may also be derived as follows.

$$s = \text{average velocity x time}$$

since acceleration is uniform,

$$\text{average velocity} = \tfrac{1}{2}(u + v)$$

$$s = \tfrac{1}{2}(u + v)t$$

but $\quad v = u + at$

$$s = \tfrac{1}{2}(u + u + at)t$$

$$s = \tfrac{1}{2}(2ut + at^2)$$

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

There are five quantities involved in these equations of motion, u, v, a, s and t. Each equation uses four of these quantities, so that there are altogether five equations, each one omitting one of the five quantities. Three of these are the standard equations:

$$v = u + at \text{ (not using } s)$$

$$s = ut + \tfrac{1}{2}at^2 \text{ (not using } v)$$

$$v^2 = u^2 + 2as \text{ (not using } t)$$

The other two are of less importance but they can be useful at times. They are:

$$s = \tfrac{1}{2}(u + v)t \text{ (not using } a)$$

$$s = vt - \tfrac{1}{2}at^2 \text{ (not using } u)$$

Students should be taught that, since each equation involves four quantities, three must be known in order to find the fourth. Then using a second equation, the fifth quantity may be found if necessary. To decide which equation is best to use in a particular problem, it is useful to list the five quantities at the start of the calculations, and to fill in the values of the three which are known, and to indicate (by a question mark) which is the one required. It should then be easy to select the equation which contains the four quantities involved and so calculate the unknown one. In listing the values for the quantities the unit should always be given. This is good practice and it also serves to draw attention to any quantity which is not in the appropriate SI unit.

Students should be made familiar with the terms used in questions which must be 'translated' to get some of these figures, e.g.

'starts from rest' $\Rightarrow$ $u = 0$

'comes to rest' $\Rightarrow$ $v = 0$

'is dropped' $\Rightarrow$ $u = 0$

'reaches its maximum height' $\Rightarrow$ $v = 0$

1. 10 seconds after coming around a corner, a racing car reaches a velocity in a particular direction of 70 m s^{-1}. If it accelerates at a uniform rate of 5 m s^{-2} in the same direction, at what speed did it come round the corner?

$u = ?$
$v = 70 \text{ m s}^{-1}$
$a = 5 \text{ m s}^{-2}$
$s = ?$
$t = 10 \text{ s}$

We require the equation which uses u, v, a and t, i.e. $v = u + at$

$$70 = u + 5 \times 10$$

$$70 = u + 50$$

$$u = 20 \text{ m s}^{-1}$$

Sometimes there is no standard equation which uses the four quantities involved in the question. In this case, there are two possibilities, (a) to find the fifth quantity first and use it to calculate the unknown, or (b) to use one of the other two equations, if known.

2. A truck accelerates from rest to a speed of 15 m s^{-1} in a distance of 60 m. How long does it take?

$u = 0$
$v = 15 \text{ m s}^{-1}$
$a = ?$
$s = 60 \text{ m}$
$t = ?$

None of the three standard equations uses u, v, s and t. If the standard equations only are known, we must calculate a first.

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

$$15^2 = 0 + 2 \times a \times 60$$

$$225 = 120\,a$$

$$a = 1.875 \text{ m s}^{-2}$$

$$v = u + at$$

$$15 = 0 + 1.875t$$

$$t = 8 \text{ s}$$

Alternatively, the equation $s = {}^1/_2\,(u + v)t$ can be used if it is known.

$$60 = {}^1/_2\,(15 + 0) \times t$$
$$60 = 7.5\,t$$
$$t = 8 \text{ s}$$

When a body is slowing down the acceleration is in the opposite direction to the velocity. Thus, if the velocity has a positive sign the acceleration must be given a negative value.

3. How far will a plane travel along the runway after landing at 60 m s^{-1}, if its brakes give it a uniform acceleration of magnitude 1.2 m s^{-2}?

$u = 60 \text{ m s}^{-1}$
$v = 0$
$a = -1.2 \text{ m s}^{-2}$
$s = ?$
$t = ?$

We require the equation which uses u, v, a and s, i.e. $v^2 = u^2 + 2as$

$$0 = 60^2 + 2 \times (-1.2) \times s$$

$$0 = 3600 - 2.4\,s$$

$$2.4\,s = 3600$$

$$s = 1500 \text{ m}$$

Students need to be warned that all distances should be in metres, times in seconds, and velocities in metres per second and accelerations in metres per second^2 for use in these equations.

4. A train starts from a station and after 1.0 minute has travelled a distance of 0.50 km. Calculate its acceleration (assumed to be uniform).

$$u = 0$$
$$v = ?$$
$$a = ?$$
$$s = 0.50 \text{ km} = 500 \text{ m}$$
$$t = 1.0 \text{ min} = 60 \text{ s}$$

We require the equation which uses *u, a, s* and *t*, i.e. $s = ut + {}^1\!/_2\, at^2$

$$500 = 0 + {}^1\!/_2 \times a \times 60^2$$
$$500 = 1800a$$
$$a = \frac{500}{1800}$$
$$= 0.28 \text{ m s}^{-2}$$

It is good experience to give questions which involve the use of scientific notation at this stage. Students need to be given practice in using scientific notation on their calculators.

5. An electron is moving at 3.0×10^6 m s^{-1}. If it has an acceleration of 5.0×10^{10} m s^{-2} in the opposite direction, how long will it take before it stops? How far will it go?

$$u = 3.0 \times 10^6 \text{ m s}^{-1}$$
$$v = 0$$
$$a = -5.0 \times 10^{10} \text{ m s}^{-2}$$
$$s = ?$$
$$t = ?$$

Since we know *u*, *v* and *a*, the first equation must contain these three quantities. The simplest is

$$v = u + at$$
$$0 = 3 \times 10^6 + (-5 \times 10^{10}) \times t$$
$$5 \times 10^{10}\, t = 3 \times 10^6$$
$$t = 6 \times 10^{-5} \text{ s}$$

This answer can be given as 6.0×10^{-5} s or as 60 μs (1 μs = 10^{-6} s)

$$s = ut + {}^1\!/_2\, at^2$$
$$s = 3 \times 10^6 \times 6 \times 10^{-5} + {}^1\!/_2 \times (-5 \times 10^{10}) \times (6 \times 10^{-5})^2$$
$$s = 90 \text{ m}$$

The most difficult calculation involving one equation only is when *t* is calculated from the equation $s = ut + {}^1\!/_2 at^2$ and *u* is not zero. The equation then gives a quadratic in *t*, which may or may not factorise. If it will not factorise, the usual formula will have to be used to find the roots. If one root is negative it may normally be disregarded, as negative times are not usually relevant, but if both roots are positive, they are both relevant answers. However, the physical interpretation of both roots should always be considered.

6. A particle initially travelling at 300 m s^{-1} is acted upon by a force in the opposite direction to its initial velocity which produces an acceleration of magnitude 200 m s^{-2}. After what time will it be 100 m from the point at which the force starts to act?

$$u = 300 \text{ m s}^{-1}$$
$$v = ?$$
$$a = -200 \text{ m s}^{-2}$$
$$s = 100 \text{ m}$$
$$t = ?$$

The equation using u, a, s and t is $s = ut + {}^1/_2\, at^2$

$$100 = 300t + {}^1/_2 \times (-200) \times t^2$$

$$100t^2 - 300t + 100 = 0$$

$$\text{or } t^2 - 3t + 1 = 0$$

By formula, $t = 0.38$ s or 2.62 s

It will pass the 100 m point 0.38 s after the force starts to act. If the force continues to act the particle will again pass the point, moving in the opposite direction, after 2.62 s, i.e. a further 2.24 s after it first passed the point.

Note
In all of these questions the acceleration is assumed to be uniform. In practice, uniform acceleration is a rarity; most acceleration varies with time, e.g. a runner or a car accelerates much faster at the start of the acceleration than when it has reached a certain speed. The graph of velocity against time for a runner or a car is not a straight line but a curve, much steeper at the start than at the end.

If figures are required for students to do calculations on acceleration, some motor magazines give lists of times for different cars to accelerate from 0 to 60 m.p.h. These could be used for the calculation of the average rates of acceleration, and might provide a link with everyday interests.

Motion Under Gravity

The main problem in these questions is to ensure that the direction of the acceleration due to gravity is correct.

If a body is thrown upwards, the initial velocity is usually taken as positive (although this is a purely arbitrary choice), which means that the upwards direction is positive. The acceleration (downward) must then be negative (-9.8 m s^{-2}).

1. A ball is thrown with an initial velocity of 20 m s^{-1} upwards. How long will it take to reach its highest point?

$u = +20$ m s^{-1}
$v = 0$
$a = -9.8$ m s^{-2}
$s = ?$
$t = ?$

$$v = u + at$$

$$0 = 20 + (-9.8)t$$

$$9.8t = 20$$

$$t = 2.0 \text{ s}$$

If the body is dropped or thrown downward, downwards is usually taken as positive, so that the acceleration is $+9.8$ m s^{-2}.

2. A person standing on the top of a building 50 m high throws a stone with a velocity of 15 m s^{-1} vertically downwards. How long will the stone take to reach the ground?

$u = +15$ m s^{-1}
$a = +9.8$ m s^{-2}
$s = +50$ m
$t = ?$

$$s = ut + {}^1/_2 at^2$$

$$50 = 15t + {}^1/_2 \times 9.8 \times t^2$$

$$4.9t^2 + 15t - 50 = 0$$

By formula, $t = 2.0$ s and -5.1 s

The time taken to reach the ground is 2.0 s.

Note
The minus answer is a mathematically correct result, although not valid for the question asked. If the stone had been thrown upwards from ground level 5.1 s is the time it would have taken for it to be at a height of 50 m with a downward velocity of 15 m s^{-1}.

One way in which students make calculations more difficult for themselves is that they do not realise that the motion of a body which is thrown upwards, rises to the highest point and falls again, can be treated as a single calculation. The acceleration remains constant during the whole period, while the fact that the direction of the velocity changes does not affect the calculation.

3. A stone is thrown vertically upwards from the top of a building. 10 s after it is thrown, it is at a height of 8 m above the ground and has a downward velocity of 50 m s^{-1}. Calculate the initial velocity of the stone and the height of the building from which it was thrown.

   Take upwards as positive.

   u = ?
   v = –50 m s^{-1}
   a = –9.8 m s^{-2}
   s = ?
   t = 10 s

$$v = u + at$$

$$-50 = u + (-9.8) \times 10$$

$$-50 = u - 98$$

$$u = 48 \text{ m s}^{-1}$$

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

$$s = 48 \times 10 + \tfrac{1}{2} \times (-9.8) \times 10^2$$

$$= 480 - 490$$

$$= -10 \text{ m}$$

The displacement is -10 m so the stone is 10 m below the point from which it was thrown. Since it is 8 m above the ground according to the question, the point from which it was thrown is 18 m above ground level.

It can be useful to know that

(a) the time to maximum height = the time down from maximum height (hence time to return to the starting point = twice time to maximum height);

(b) velocity at any height on the way up = – velocity at the same height on the way down.

4. A basketball player can jump off the floor to a height of 60 cm. For how long is she (a) off the floor, (b) at least 40 cm off the floor?

(a) Take upwards to be positive.

u = ?
v = 0
a = –9.8 m s^{-2}
s = +0.6 m
t = ?

There is no standard equation involving *v, a, s, t.*

First find u using

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

$$0 = u^2 + 2 \times (-9.8) \times 0.6$$

$$11.76 = u^2$$

$$u = 3.429 \text{ m s}^{-1}$$

$$v = u + at$$

$$0 = 3.43 + (-9.8)t$$

$$9.8t = 3.43$$

$$t = 0.35 \text{ s}$$

This is the time to the maximum height; time to return to starting point = 2 x 0.35 = 0.70 s

(b) The simplest method is to find the time to reach a height of 40 cm.

u = 3.43 m s^{-1}
a = –9.8 m s^{-2}
s = 0.4 m
t = ?

Note: as we are now finding the time to 40 cm, v is not 0.

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

$$0.4 = 3.43t + \tfrac{1}{2} \times (-9.8) \times t^2$$

$$4.9\, t^2 - 3.43t + 0.4 = 0$$

$$t = 0.15 \text{ s or } 0.55 \text{ s}$$

The player passes a height of 40 cm after 0.15 s on the way up and after 0.55 s on the way down again, so the time above 40 cm is 0.55 – 0.15 = 0.40 s.

Comparing the answers for (a) and (b) shows that she spent 57% of her time on the top 33% of her jump, giving rise to the expression ‘she hung in the air’.

Conservation of Momentum

For collision in one dimension the principle can be represented by the equation

$$m_1u_1 + m_2u_2 = m_1v_1 + m_2v_2$$

where m_1 and m_2 are the masses of the two bodies, u_1 and u_2 are their initial velocities and v_1 and v_2 are their final velocities.

1. A truck of mass 5.0 t is travelling along a straight level track at 6.0 m s^{-1}. It collides with a second truck of mass 7.0 t which is at rest. After the collision, the two trucks continue along the track together. Calculate their final speed.

$$\begin{aligned} m_1u_1 + m_2u_2 &= m_1v_1 + m_2v_2 \\ (5 \times 6) + (7 \times 0) &= (5 + 7)v \\ 30 &= 12v \\ v &= 2.5 \text{ m s}^{-1} \end{aligned}$$

Note:
It is not necessary to convert t to kg since the same unit occurs on both sides of the equation.

2. A gun of mass 3 kg fires a bullet of mass 5 g with a velocity in a certain direction of 600 m s^{-1}. With what velocity does the gun recoil?

$$\begin{aligned} m_1u_1 + m_2u_2 &= m_1v_1 + m_2v_2 \\ (3 \times 0) + (0.005 \times 0) &= 3v + (0.005 \times 600) \\ 0 &= 3v + 3 \\ v &= -1 \text{ m s}^{-1} \end{aligned}$$

(minus sign indicates that velocity of gun is in opposite direction to velocity of bullet)

3. A bullet of mass 10 g travelling at 400 m s^{-1} hits a block of wood of mass 5.0 kg and becomes embedded in it. With what velocity will the block move after the collision? If there is a frictional force of 20 N resisting the motion of the block, how far will it travel before it comes to rest?

$$\begin{aligned} m_1u_1 + m_2u_2 &= m_1v_1 + m_2v_2 \\ (10 \times 10^{-3} \times 400) + (5 \times 0) &= (10 \times 10^{-3} + 5)v \\ 4 &= 5.01v \\ v &= 0.80 \text{ m s}^{-1} \end{aligned}$$

$$\begin{aligned} F &= ma \\ 20 &= 5 \times a \\ a &= 4 \text{ m s}^{-2} \end{aligned}$$

$$\begin{aligned} v^2 &= u^2 + 2as \\ 0 &= 0.8^2 - (2 \times 4 \times s) \\ 8s &= 0.64 \\ s &= 0.080 \text{ m} \\ &= 8.0 \text{ cm} \end{aligned}$$

4. A truck of mass 15 t is travelling along a straight level track with a velocity of 5 m s^{-1} in a certain direction. It collides with a second truck of mass 30 t which is at rest. If the first truck rebounds after the collision with a velocity of 0.2 m s^{-1}, calculate the velocity of the second truck.

$$\begin{aligned} m_1u_1 + m_2u_2 &= m_1v_1 + m_2v_2 \\ (15 \times 5 + 30 \times 0) &= (15 \times -0.2) + 30v \\ 75 &= -3 + 30v \\ 78 &= 30v \\ v &= 2.4 \text{ m s}^{-1} \end{aligned}$$

(in same direction as first truck)

1.6 Student Experiments

There are four sets of apparatus which can be used in experiments to measure the velocity and acceleration of a body (e.g. a trolley) in the laboratory, viz. light gates and timer; ticker-tape timer; powder track timer; vibrating arm and brush or pen.

Details of the relevant experiments are to be found in all textbooks. Here, only problems which may arise in the experiments, and methods of improving the accuracy of the results, will be considered.

Measuring Acceleration Using Light Gates and Timers

The main inaccuracy in the measurements in this method is in measuring the time of passage of the card through the light gate, which may be very short if the body is travelling fast. Assuming the timer measures to 0.01 s, the time of passage should be at least 0.2 s to ensure the experimental error is less than 5%. Therefore the bodies should not be allowed to travel too fast. While increasing the length of the card which obstructs the light beam would increase the time, it would not normally be convenient to do so; one reason for this would be that a card longer than the trolley would be easily damaged, making its length inaccurate. If the timer measures to 0.001 s, the same problem does not arise until much higher speeds, which would not normally be attained in a laboratory experiment.

In measuring acceleration, the distance between the two light gates should be made as long as possible, subject, of course, to the acceleration being maintained over the distance. If the acceleration is produced by a falling mass pulling via a string and a pulley, the maximum distance of acceleration is the distance from the bottom of the mass to the floor, and the distance between the gates must be a good deal less than this, as the card must complete its passage through both gates while the acceleration continues. (This is another reason why a longer card would not be suitable.) Alternatively, if elastic cords are used to provide the acceleration, it becomes more difficult to maintain a constant force over a longer distance, and any increase in accuracy due to the longer distance is offset by variations in the force acting.

Measuring Acceleration Using the Ticker-Tape Timer

The quick way of calculating the acceleration is to measure the distance between two dots near the beginning and the distance between two dots near the end of the trace, and the number of spaces between the two measured distances, and from this to calculate, u, v and t and hence a, using the equation $v = u + at$.

Students may need to appreciate that the values of u and v are average velocities over the time interval, and that, in the calculation, we assume that the average velocity over the time interval is the actual velocity at the middle of the time interval. The time between the two velocity measurements should be measured from the middle of the first time interval, in which u is measured, to the middle of the second time interval, in which v is measured. In practice, it is simpler to measure the time by counting the number of spaces between the two measured spaces, adding 1, and multiplying by the time between the printing of each dot (0.02 s if the timer is driven by an a.c. power supply).

An alternative method of calculation is to measure the distance between the measured spaces. To this should be added half the sum of the two distances which were measured to find u and v. This gives s, the displacement, and a can be found using the equation $v^2 = u^2 + 2as$.

A more accurate method (though more time-consuming) is to calculate the acceleration graphically. To do this, each fifth dot on the tape is marked and the distances between the marked dots are measured. Since each distance measured corresponds to 0.1 s, the average velocity over each time interval can be found by multiplying the distance covered by 10. The velocities thus calculated are plotted against time, the times used being 0 s, 0.1 s, 0.2 s, etc. The graph will be a straight line (but not through the origin), and the slope is the acceleration. This method also has the advantage that it introduces students to the method of finding the slope of a graph. The calculations in this method could be done using a spreadsheet on a computer.

One problem which may occur in using a ticker-tape timer is that the dots printed on the tape may not be evenly spaced. This will have little effect if one is using the graphical method of finding the acceleration, but it can lead to appreciable errors if using the first method. To obtain an accurate answer in this situation, one could measure the distance over a number of spaces (e.g. 5) at both ends, and calculate the average velocity from this figure using the formula velocity = distance measured/(no. of spaces x 0.02 s). Care must be taken to ensure that the correct value of *t* or *s* is used in the calculation in this situation - see second paragraph in this section.

A second problem which may occur using a ticker-tape timer is that the striker may bounce, producing two dots close together instead of a single dot. In this case, students need to be instructed to ignore the second dot of each pair. Adjustment of the vibrating length of the bar of the timer may prevent this happening.

If the dots printed on the tape are faint and difficult to see, the carbon paper disc may need replacing (on timers which use a disc). If the supply voltage is low, this may also cause the dots to be faint; increasing the voltage should make the dots easier to see.

Measuring Acceleration Using the Powder Track Timer

Safety note

The powder track timer is connected to the 230 V mains supply through a large series resistor. The purpose of this resistor is to reduce the maximum current which can flow to a safe value. Under **no circumstances** should the resistor be removed or by-passed, as a lethal current would then be able to flow.

The calculations from this method are similar to the ticker-tape timer method, and either of the two alternative methods of calculation may be used. There is much less likelihood of irregularities in the distances measured using this method.

Vibrating Arm and Brush or Pen

This is the original method used by Fletcher, before any electronic devices were available. Today it is mainly of historical interest, although it is a perfectly viable method of measuring acceleration. A springy steel arm is clamped at one end and a brush (or possibly a pen) is attached to the other. A piece of paper is attached to the top of the trolley and the brush dipped in ink. A centre line is drawn on the piece of paper by pulling the trolley back under the brush while the arm is stationary. Then the arm is set vibrating and the trolley is released. The brush produces a wave-form trace on the paper. As the periodic time of the bar is constant, the distances covered by the trolley in equal intervals of time can be found by measuring the distances between the points where the wave-trace crosses the centre line.

The principal difficulty in carrying out the experiment is to ensure that a complete trace (or at least all the points where the trace cuts the centre line) is visible, without the brush or pen bearing too heavily on the paper and causing extra friction, so stopping the vibrations of the arm too quickly. The arm must vibrate in a horizontal plane and not at an angle to the horizontal.

The periodic time of the arm must be obtained before carrying out the experiment. The result can be recorded and perhaps marked on the apparatus, so that the measurement need not be carried out each time the apparatus is used. This measurement should be retaken if a different brush/pen is used.

The calculation is similar to that using the ticker-tape timer, except that the time for each measurement is the periodic time of the arm (or half the periodic time, if the distance is measured over half a wavelength instead of a whole wavelength).

As the periodic time of the arm is quite slow and the length of the paper quite short, there will not usually be a large enough number of measurements to use the graphical method unless the trolley is accelerating very slowly.

Measurement Of Acceleration Due To Gravity

There are a number of methods of carrying out this experiment.

1. Simple pendulum
2. Free fall methods
   (a) using centisecond timer
   (b) using ticker-tape timer.

Method 2(b) is carried out using a reasonably large mass (100 - 200 g) with a piece of ticker-tape attached to it by Sellotape. It is let fall and pulls the tape through the timer which is on the edge of a bench (or clamped vertically over the edge of the bench, if preferred). From the dots on the tape, the acceleration may be found in the usual way. If the mass is too small, the friction of the tape through the timer may slow it down significantly, resulting in a value of g which is too small.

Verification of the Principle of Conservation of Momentum

There are a number of possible ways to verify this principle but essentially there are two different methods:

1. ‘collision’ methods using two colliding trolleys;
2. ‘explosion’ methods using two trolleys with a spring between them.

Collision method

When using the collision method, the usual way to carry out the experiment is to leave one trolley at rest and allow a second trolley to collide with it in such a way that the two trolleys coalesce and move together after the collision. This allows the experiment to be carried out with the simplest timing measurements - one ticker-tape timer/sulphur track or two electronic timers, since only two velocities need be measured - the initial velocity of the moving trolley and the final velocity of the two trolleys moving together. Using any other method, three or more velocities must be measured, and the experiment becomes much more difficult.

There are a number of different sets of apparatus which can be used to carry out this method. The trolleys may be gliders running on a linear air track (which works on the hovercraft principle - an air pump blows air out through holes in the track, which lifts the gliders off the track and makes them hover with little resistance to motion), or trolleys on a sulphur track, or conventional trolleys running on a runway or laboratory bench. The timing may be carried out using light gates connected to electronic timer(s), a ticker-tape timer or the sulphur track.

When using a sulphur track or conventional trolleys, the runway needs to be sloped to counteract friction, as in the experiment to verify Newton’s 2nd law. The linear air track must be carefully levelled before the experiment, so that a glider remains at rest when placed on it. If the track is not level, the glider will start to move, as the resistance to motion is so small.

A number of ways can be used to make the trolleys or gliders coalesce: self-adhesive pads, Velcro strips, cork (or lump of Blu-Tak) and pin. The last is generally most satisfactory; the only difficulty is that the pin must not be crooked, as this would produce a sideways force as it went into the cork.

To time the gliders on the air track, two light gates are needed. The timers are set to measure the time during which the light is obstructed. The first glider (which is given a push to get it moving) carries the timing card, while the second (at rest at the start) does not carry a card and does not affect the light gate since, when it is moving, it is attached to the first glider. The position of the light gates is not critical, as the gliders move with (near enough to) constant velocity along the track; all that is required is that the timing card has passed completely through the first gate and has not reached the second gate at the instant of collision.

As an alternative to light gates, a ticker-tape timer can be used with an air track. The tape is attached to the first glider and the velocity before and after collision is found from the trace on the tape. The instant of collision is seen on the tape as the point at which the distance between the dots suddenly decreases; the velocity should be measured as close as possible to the moment of collision, as the timer will cause friction and slow the glider.

It is important that students remember to record the masses of the trolleys; it is a common error to omit reference to this measurement. The experiment should be repeated a number of times, using different masses of trolleys (the simplest method is to add standard masses to the trolleys used).

The 'explosion' method

The 'explosion' method uses a spring between two trolleys. The spring is most conveniently a piece of bent spring steel or brass rather than a coil spring; it is compressed and held in place with a thread or clip. The spring is attached to one of the trolleys and the second trolley is placed so that it touches the spring. The string is (preferably) burnt so as to release it without applying any force to the trolleys, as cutting it with a scissors or knife might do. The spring applies equal forces to the two trolleys and they move apart. The trolleys can be either conventional trolleys running on a level runway or laboratory bench, or gliders on an air track.

In the first case, a simple demonstration of the principle is to let the trolleys come to rest and measure the distance covered by each. Using trolleys of equal mass, the distances will be found to be equal, showing that the initial velocities are equal, and hence the sum of the momenta after the 'explosion' is zero, as it was before the collision. To

verify the principle, each trolley can be fitted with a card and the velocities measured with light gates; in this case, it is essential that the measurement of velocity is made as near to the starting point as possible, so that the effect of friction on the trolleys is as small as possible. If the trolleys are not free-running, the use of a shorter card than normal (perhaps 10 cm rather than 20 cm) might be more accurate, as it reduces the effect of the friction on the measurement. Two ticker-tape timers could also be used to find the velocities.

If gliders on an air track are used, each should carry a card and be made to pass through a light gate shortly after they start moving. The velocities measured can be assumed to be the velocities immediately after the 'explosion' and again the sum of the momenta will be found to be zero. In this case, the experiment can be repeated with gliders of different mass.

CHAPTER 2 FORCES

2.1 Background

Sir Isaac Newton (1642-1727)

Newton was born at Woolsthorpe, near Grantham, Lincolnshire, England. His father died before he was born, and when he was three years old, his mother remarried and left Isaac in the care of her mother. At school, he was a poor student until his spirit was aroused after a fight with another boy. After that he became the best student in the school. When he was 14, his stepfather died and his mother returned home. She took Isaac from school to run the farm, but he was more interested in mathematics and proved to be a poor farmer. His uncle was a member of Trinity College, Cambridge, and he advised that Isaac return to school and prepare for Cambridge.

Newton entered Trinity College, Cambridge in 1661 and took his BA degree in 1665, but later in that year an outbreak of plague led to the closure of the university, and Newton went home. He remained at Woolsthorpe for 18 months, and during that time he laid the foundations for his famous discoveries in mathematics and physics. In mathematics, he discovered the binomial theorem and calculus, while in physics, he experimented with light and colours, discovering dispersion, and worked out his law of gravitation, inspired, he claimed, by seeing the fall of an apple (that it fell on his head seems to be fiction). Later, he is reported to have said about that period, ‘I was in the prime of my age for invention, and minded mathematics and philosophy (i.e. science) more than at any time since’.

When he returned to Cambridge in 1667, Newton was made a fellow of the college. He did not publish any of his discoveries at that time. His professor of mathematics, Isaac Barrow, seeing Newton’s ability, resigned the chair of mathematics so that Newton could be appointed instead. Newton made a reflecting telescope, and when news of this reached the Royal Society in London, he was invited to come to demonstrate it. Instead, he sent a small telescope, and as a result was elected a Fellow of the Society. Next he sent a paper setting out his discoveries on dispersion and colour, and explaining why he had used a mirror instead of a lens in his telescope to avoid dispersion. Some of the members disagreed with his conclusions, particularly **Robert Hooke** (1635-1703), and as a result of the controversy, Newton developed his corpuscular theory of light, which conflicted with Hooke’s belief in a wave theory. In the controversy, Newton is reported to have said, ‘I was able to see further by standing on giants’ shoulders’, which was taken by some to be insulting to Hooke, who was physically very small. Newton also investigated mathematically what are now called Newton’s rings.

At this time, there were great discussions about planetary orbits. Hooke claimed that he had worked out a solution to the problem and **Sir Edmund Halley** (1656-1742), discoverer of Halley’s comet, offered a prize to Hooke or Newton if they could produce a satisfactory solution. Newton told Halley that he had already worked it out, but that he had lost the proof, so he worked it out again, and his treatise on the topic, called *De Motu*, was presented to the Society in December 1684.

In March 1686, Newton began work on his *Principia* (full title*, Philosophia Naturalis Principia Mathematica* - the mathematical principles of natural philosophy, or science), and it was published in summer 1687. Halley paid for its publication, but Hooke claimed that he had previously made some of the discoveries which Newton claimed as his own in his book. The book set the seal on Newton's reputation.

In 1689, Newton was elected member of parliament for the university, and he was re-elected in 1701, but he never took a prominent part in politics - he never spoke in parliament. In 1695 he was appointed to the post of Warden of the Royal Mint, and in 1699, Master of the Mint. In 1701, he resigned his professorship and moved to London, where he carried out his duties as master with great efficiency; he spent much time investigating the prevention of counterfeiting. In 1703, he was elected President of the Royal Society, and in 1705 he was knighted by Queen Anne. After a serious illness, he died in 1727 and was buried in Westminster Abbey, London, where his grave may be seen today.

Newton's fame was recorded by the poet Alexander Pope in this couplet,

> *Nature, and nature's laws, lay hid in night,*
> *God said, let Newton be, and all was light.*

History of Rocketry

Rockets were probably invented by the Chinese about the 13th century, firstly as fireworks, but later as weapons of war. They were possibly used against Mongol invaders in 1232 AD. The weapon was copied by the Tartars, the Arabs and finally by the Europeans before the end of the century. The English monk, **Roger Bacon** (c.1220 - c.1292) wrote down a formula for gunpowder in his writings about 1250, while the contemporary German monk, **Albertus Magnus** (c.1200-1280), also wrote formulas for powder charges for rockets. Rockets found some uses but were soon supplanted by guns. The term 'wildfire', still in use today, was originally a name for early rockets.

In the late 18th century, a prince of Mysore, in India, developed rockets as weapons, and used them effectively in battles against the British. This caused renewed interest in rockets in Britain, and the English inventor **Sir William Congreve** (1772-1828) developed them further, leading to a rocket corps being formed. Rockets were successfully used in the Napoleonic wars and the war with the USA in 1812. In the latter, their use was seen by the composer of the US National Anthem, Francis Scott Keys, and the anthem contains the words 'the rockets' red glare'. While improvement in guns made rockets obsolete as weapons, they were adapted to carry lifelines to ships in distress, and also to carry flares into the air.

The first suggestion that rockets could be used for space travel was made by a Russian mathematics teacher, **Konstantin Tsiolkovski** (1857-1935), in 1895. He also recognised that rockets using liquid fuels would be much more powerful than solid fuel ones. The American physicist, **Robert Goddard** (1882-1945), was the first to successfully fire liquid-fuelled rockets, in 1926, and he developed gyro-stabilisers for keeping his rockets on course. He, too, suggested the use of rockets in space. The German scientist, **Hermann Oberth** (1894-1989), not only suggested using liquid-fuelled rockets but also the concept of ion-propulsion. His writings led to experimentation on rockets in Germany which resulted in the production of the V-2 rocket, used against the French and British in 1944-45.

After the Second World War, rocket experts from Germany were captured by the US and Russia, and their work has led to the development of the guided missiles and space rockets which we know of today.

Multi-stage rockets were developed soon after the war; the first ones used captured V-2 rockets as the first stage. The idea of multi-stage rockets is that a large rocket (the first stage) is used to lift a small one (the second stage) into the air. When the first stage has used up all its fuel, the second stage is released from the top and its engine starts to work, sending it higher and faster than could be done with a single rocket. The Apollo mission, which carried men to the moon, used three-stage rockets, i.e. a very large rocket carried a smaller one which carried a still smaller one; this enabled the final stage to leave the earth's gravity and reach the moon.

2.2 Conceptual Approach

Newton's First Law

Before Newton, it was believed that there were two different types of matter, celestial matter (of which the moon, stars, etc. were made) and terrestrial matter. The difference was that celestial matter could keep moving at a steady speed without any force, while terrestrial matter needed a force to keep it moving, and would slow down and stop if no force acted. This was the way of explaining why the moon could keep moving around the earth, without any apparent force, while an object on earth could not keep moving without a force. Newton was the first to realise that the force of friction was what made the difference, providing a force which slowed down a moving object on the earth.

Newton's first law states that there can be no change in the velocity of a body unless a resultant external force acts on it. This law introduces the concept of inertia, the property of a body which enables it to resist changes in its velocity. The mass of a body can be defined as a measure of its inertia, a measure of how difficult it is to change its velocity. It is difficult to make a heavy body at rest start moving, but it is equally difficult to stop a heavy moving body. This can be seen in road accidents, where a heavily laden truck can demolish a wall of a house, for example, before it comes to rest. It is the inertia of a passenger in a car which makes the wearing of a seat belt necessary. If a car stops suddenly, as when it is involved in an accident, the passengers continue moving, due to their inertia, if they are not wearing seat belts, until they hit the part of the car in front of them - the steering wheel, the dashboard or the windscreen.

Many people have the idea that because bodies in space are weightless, they are easy to move - that anyone could do a Superman act by giving a spaceship in space a push, making it move away quickly. This is not the case; inertia does not change when gravity changes; the inertia of a body will be just as great when it is weightless. It is as difficult to accelerate a spaceship in space as it would be to accelerate it horizontally near the ground if the effects of friction with the air are ignored. (The term 'weightless' is usually a misnomer. Bodies in space are only truly weightless when the acceleration due to gravity is zero. This is rarely, if ever, absolutely the case; for practical purposes it may be taken as true if the distance from the nearest star or planet is great enough. The Mir space station, for example, is at an average height above the earth of approximately 380 km. At this height the acceleration due to gravity is 8.75 m s^{-2}. So the weight of a cosmonaut on Mir is only about 11% less than his/her weight on the surface of the earth. The cosmonaut appears weightless because Mir is travelling in a circular (approximately) orbit. See section on satellite motion in Module 7.)

Force

Force is defined as 'that which causes acceleration'. Force is a vector quantity and its unit is the newton (N).

A force can have a number of effects. The most obvious one is that it starts things moving; when a stationary ball is given a force as it is kicked, it moves. It also stops things which are moving; when a footballer traps a ball, he stops it by applying a force to it with his foot. It can change the speed of an object, either speeding it up or slowing it down. It can also change its direction, as when a player glances a moving ball and makes it change direction without changing its speed.

A force can change the shape of an object; this can involve stretching a rubber band or squeezing a ball, as well as changing the shape of a car when it collides with a wall.

Newton's Second Law

This law states that force is proportional to rate of change of momentum and is used to provide us with a way of measuring force. We can use the law to derive the equation $F = ma$ as follows.

$$F \propto \frac{d(mv)}{dt}$$

$$= k\frac{d(mv)}{dt}$$

$$= km\frac{d(v)}{dt} \quad \text{if } m \text{ is constant}$$

$$= kma$$

The unit of force is defined to make $k = 1$, i.e. 1 N is the force which gives 1 kg an acceleration of 1 m s^{-2}, so that the equation becomes $F = ma$.

Note
$F = ma$ is not a statement of Newton's second law. $F = ma$ is a special case of Newton's law, applying only when the mass is constant and the unit of force is defined as above.

Newton's second law illustrates two well-known facts.

1. The acceleration of a body depends on the force applied - greater force produces greater acceleration. A jet plane will accelerate faster if the pilot switches on reheat in the engine, making it produce a greater force.

2. The acceleration of a body depends on the mass of the body - greater mass results in less acceleration. A car will accelerate more slowly if it is towing a caravan than if it is not, since the mass to be accelerated is greater.

Newton's Third Law

Newton's third law states that when a body A exerts a force on a body B, B exerts an equal and opposite force on A. For example, when a book is placed on a table the book exerts a downward force on the table, and the table exerts an equal and opposite reaction force upwards on the book. (The table produces the reaction force by bending slightly under the weight - a good table would only bend by a tiny fraction of a millimetre. The bending can be seen more clearly if a person stands on a long plank or a trampoline, because the bending is much larger.) Note that, although the two forces referred to in Newton's third law are always equal in magnitude and opposite in direction, they never act on the same body.

Friction

Friction is a force which opposes relative motion between two bodies in contact. It is friction, for example, which is responsible for the fact that all objects on the earth need a constant force to keep them moving.

Friction results from the interaction between the surface atoms of the two surfaces in contact. No two surfaces are perfectly smooth and therefore the contact between them is not continuous, but occurs only at points here and there. At these points of contact cold welds occur and these welds must be broken before movement between the two surfaces can take place.

Limiting friction

When a heavy body is acted upon by a force which gradually increases, at first it does not move. This is because the frictional force balances and cancels out the force applied. As the force increases, the frictional force also increases, and it continues to cancel the applied force. This continues until the point at which the body starts to move, at which point the applied force is not completely cancelled out by the frictional force. It is clear that there must be a maximum value for the frictional force between a given pair of surfaces; this maximum value of the frictional force, which occurs when the body is just about to move, is called the limiting friction. This maximum value of the frictional force is found to be generally proportional to the force which the surface exerts on the body (the normal reaction); the constant of proportionality is called the coefficient of static friction, i.e. $F = \mu R$, where F is the maximum frictional force, μ is the coefficient of static friction and R is the normal reaction.

Static and dynamic friction

If a body is acted upon by a force, it can be seen that the magnitude of the force needed to start a body moving from rest is greater than the magnitude of the force needed to keep it moving at constant speed once it is moving. The two types of friction are called static friction and dynamic friction; the coefficient of static friction is greater than the coefficient of dynamic friction (defined as the ratio of the frictional force when the body is moving to the normal reaction) for a given pair of surfaces.

2.3 Experimental Approach

Inertia

A heavy mass (e.g. 500 g) is attached to the middle of a piece of thread and suspended from a support, so that the free end of the thread hangs down below the mass. Catch the lower thread and pull.

(a) If the lower thread is pulled slowly with a steadily increasing force, the upper thread will break, as it has to bear the weight of the mass plus the force applied, while the lower thread only has to bear the applied force.

(b) If the lower thread is given a quick, sharp pull, the lower thread will break. The inertia of the mass makes it resist the sudden change in its velocity, so the lower thread experiences a large force and breaks. The upper thread experiences very little extra force since, because of the inertia of the mass, the lower thread has broken before there is any significant extra force on the upper one.

Newton's Second Law

A jet of water is directed horizontally onto a plate which is free to turn, Fig. 2.1. The water exerts a force on the plate which is held in place by attaching a spring balance to it. The force on the balance is read. The force on the plate, F_p, can be calculated from the force on the spring balance, F_s, using moments.

$$F_p = F_s \times a/(a+b) \qquad \text{(see Fig. 2.1)}$$

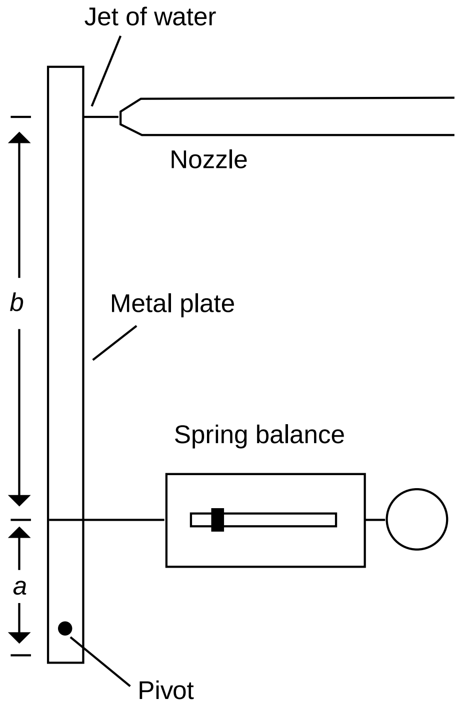


Fig. 2.1

To show that force equals rate of change of momentum, the momentum of the water per second must be calculated. The mass of the water per second is measured by collecting the water falling from the plate in a suitable container for a suitable length of time. The volume flowing per second, and hence the mass per second, can thus be calculated, knowing the density of water. The velocity of the water is measured by removing the plate and measuring the horizontal and vertical distances covered by the jet of water as it falls under gravity, Fig. 2.2.

Vertically: $h = \frac{1}{2}gt^2$

Horizontally: $d = vt$

Eliminating t: $v = d\sqrt{\dfrac{g}{2h}}$

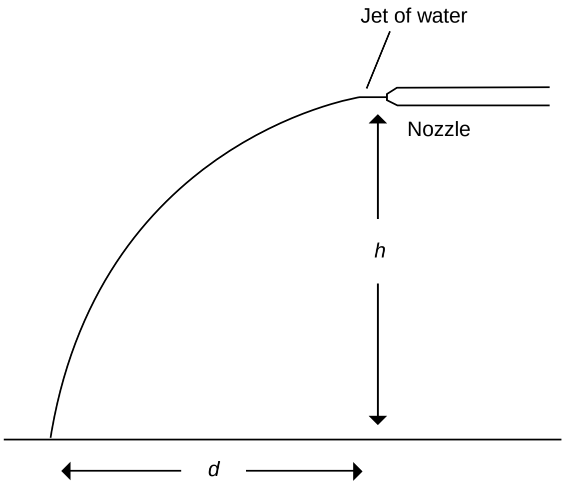


Fig. 2.2

The mass per second is multiplied by the velocity to find the rate of change of momentum; it should be found to be equal to the force measured above.

2.4 Applications

Inertia

A number of everyday effects may be explained in terms of the inertia of the bodies concerned.

Pendulum swinging

When a pendulum reaches the bottom of its swing, its inertia keeps it moving past that point to the other side of the swing. The same principle makes a weight on a spring keep bouncing up and down on each side of the rest position.

Car on icy road

A car slides off an icy road on the outside of the bend. The inertia of the car keeps it travelling in a straight line, and the forces between the wheels and the slippery road are not large enough to change that, so that the car continues in a straight line to the outside of the bend.

Tablecloth trick

For a party trick, the cloth can be pulled from a table laid for a meal, leaving the cups, saucers, etc. still on the table. If the cloth is pulled sharply enough, the inertia of the other items makes them remain at rest as the cloth moves.

Inertia reel seat belts

Inertia reel seat belts are used in cars. When the belt is pulled slowly, it unwinds easily, but if it is pulled suddenly, as would happen if the car were involved in an accident, it will not come because the inertia of the operating mechanism locks the reel.

Newton's Third Law

Firing a gun

When a gun is fired, the exploding cartridge exerts a force which accelerates the bullet or shot out of the barrel, and exerts a reaction force on the gun which forces it backward, giving the recoil or kick.

Helicopter rotors

In a helicopter, the engine exerts forces on the rotor to make it turn. The rotor exerts equal and opposite reaction forces on the helicopter, which try to make it turn in the opposite direction. The tail rotor of the helicopter provides a force to balance these reaction forces and stop the helicopter turning.

Rowing boat

When a person rows a boat, the oars push a mass of water backwards with each stroke. The force backwards on the water produces a reaction force on the boat which pushes it forward.

Ship's propellers

The propellers of the ship (or the paddle wheels of a paddle-steamer) push water backwards, and the water produces a reaction force on the ship pushing it forward.

Aeroplanes

The propellers of a propeller-driven plane push a large volume of air backwards, and the air produces a reaction force on the plane pushing it forward. The engines of a jet plane push a smaller volume of air backwards at higher velocity, and again the reaction force pushes the plane forward.

Rockets

The engines of a rocket push the exhaust gases out behind them, and the gas produces a reaction force on the rocket, pushing it forward. (Many students have the idea that the engines of a rocket push the gases against the earth to provide the forward motion, but this is not the case. A rocket works perfectly well in space where there is nothing for the gases to push against.)

Lawn sprinklers and hoses

The rotation of a lawn sprinkler is produced by the reaction forces at the bends in the pipes, Fig. 2.3. As the water travels around the bend its velocity in the direction BC is increased from zero to several metres per second. This change in velocity requires a force in the same direction (Newton's first law). This force is supplied by the side of the pipe which then experiences an equal and opposite force, causing the sprinkler to rotate. (At the same time the velocity of the water in the direction AB is reduced to zero, requiring a force in the direction BA.) A similar situation arises when water leaves any hose which is not straight; if the force of the water is strong enough, the reaction can make the hose move in the opposite direction. This is particularly the case with a fire hose, where the fireman must hold the hose very firmly to prevent it moving.

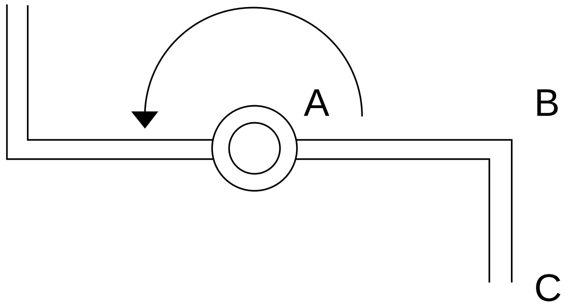


Fig. 2.3 *Lawn sprinkler*

In general, all forms of propulsion involve applications of Newton's third law. When we walk, we try to push the ground backward with our feet, and the reaction force pushes us forward. A car's wheels try to push the road backward, and the reaction force pushes the car forward. If the surface on which the car is standing is loose, the wheels succeed in pushing parts of the surface backwards and the reaction force is reduced, so that the car may not move at all.

Friction

Friction is a nuisance in some situations, but an advantage in others.

There are many everyday situations which depend on the presence of friction. We depend on the friction between our feet and the ground when we stand and walk - just consider how difficult it is to walk on very slippery ice, for example. A car depends on friction between the tyres and the road for both movement and steering - again, driving a car on ice, where the friction is small, is not easy. Our fingers depend on friction when we pick up objects between finger and thumb; it is said that the pattern of our fingerprints is useful in helping us grip objects using friction. All types of clamps and vices use friction to hold objects, not only in the laboratory but also in the workshop. Nails are held in wood by the frictional forces between them and the wood. When we sit down on a chair or lean against a table, we are depending on friction between the furniture and the floor to keep it in place. Friction is useful also in the action of sanding and rubbing objects to shape and polish them.

Friction is a nuisance in all types of machinery. When two parts of a machine slide against one another, friction tends to slow down the machine, converting mechanical energy into heat (internal) energy, and also to wear it out. To reduce friction in machinery, ball or roller bearings are used; in these, the sliding is largely replaced by rolling movement.

The normal method of reducing the effects of friction and wear in all machinery is to use lubricants such as oil and grease. These form a layer between the two surfaces and prevent them coming into contact. If a car engine is run without any oil in it, it will soon seize up; the heat produced as the metal surfaces rub against each other is sufficient to melt the metal and weld the surfaces together. Water can act as a lubricant; it does so between the tyres of a car and a wet road, especially if the tyres are bald. The tread of a tyre removes much of the water and so increases the grip available. Water is not a suitable lubricant in most other situations, for metals tend to corrode when in contact with water, but it is the lubricant for water pumps both in car engines and in central heating systems. In the body, synovial fluid acts as a lubricant in the joints.

2.5 Worked Examples

1. A ship is propelled through the water at a constant speed by the propellers pushing water backwards. Each second, the propellers give 500 kg water a velocity of 5 m s^{-1}. Calculate the force opposing the motion of the ship.

   Momentum given to water per second = 500 x 5

   = 2500 kg m s^{-1}

   Force applied to water = 2500 N

   (since force = rate of change of momentum)

   Force applied to ship = 2500 N

   (by Newton's third law)

   Since the ship is moving at constant velocity, there is no resultant force on it.

   Force opposing motion of ship = 2500 N

2. A machine gun is mounted on a truck. The mass of the truck and gun is 5 t. The gun fires 50 bullets per second with a velocity of 600 m s^{-1}. The mass of each bullet is 10 g. Calculate the force on the truck.

   Mass of a bullet = 10 g = 0.01 kg

   Momentum of a bullet = 0.01 x 600

   = 6 kg m s^{-1}

   Momentum of 50 bullets = 6 x 50

   = 300 kg m s^{-1}

   Change of momentum of truck per second = 300 kg m s^{-1}

   (since momentum is conserved)

   Force on truck = 300 N

   (since force = rate of change of momentum)

3. What is the acceleration of the truck in Example 2 (ignore mass of bullets)?

   F = 300 N
   m = 5000 kg

   $$F = ma$$
   $$300 = 5000a$$
   $$a = 0.06 \text{ m s}^{-2}$$

4. If the engine of a car of mass 1200 kg produces a force of 3000 N at the wheels, calculate the distance the car would travel from rest in 10 seconds, assuming frictional forces to be negligible.

   $$F = ma$$
   $$3000 = 1200a$$
   $$a = 2.5 \text{ m s}^{-2}$$

   $$s = ut + {}^{1}/_{2}\, at^{2}$$
   $$= (0 \times 10) + ({}^{1}/_{2} \times 2.5 \times 10^{2})$$
   $$= 125 \text{ m}$$

2.6 Student Experiments

Verification of Newton's Second Law

This experiment is an extension of the experiment to measure velocity and acceleration. There are two parts, each of which occupies as much time as any other complete experiment.

(a) The accelerating force is varied while the mass of the trolley is kept constant, and the acceleration is measured for each value of the force. A graph of force against acceleration should be a straight line through the origin, showing that $F \propto a$, when m is constant.

(b) The mass of the trolley is varied while the force is kept constant, and the acceleration is measured for each value of the mass. A graph of the reciprocal of the mass of the trolley against the acceleration should be a straight line through the origin, showing that $a \propto 1/m$ when F is constant.

Both of these results agree with the equation $F = ma$, from Newton's second law.

The principal precaution in this experiment is to slope the runway on which the trolley runs to compensate for friction, so that it moves at constant velocity with no force acting when given a small push. This should be checked using the timing device. (Once the correct slope has been found, the block used should be put away carefully, for it will probably do for future years, as it is likely that the slope needed will be the same each time as long as the same trolley is used.) If the track is not sloped correctly, friction will affect the results, and the graph of force vs. acceleration will not pass through the origin.

Sometimes it is stated in the instructions that, in order to keep the total mass constant, the masses used to produce the acceleration should be stored on the trolley, and then transferred one by one to the scale-pan as required to increase the accelerating force. This has a significant effect on the accuracy only if the mass of the trolley is small - less than about ten times the mass of the masses used, and such a mass in the pan would give a large acceleration. However, it is good practice to draw students' attention to the fact that the total mass must be kept constant when investigating the relationship between force and acceleration and to the fact that the total mass includes the mass of all moving parts of the system.

CHAPTER 3 MOMENTS

3.1 Conceptual Approach

Moments

When a force is applied to a body, it may tend to rotate about some axis. This turning effect is called the moment of the force. The magnitude of the moment depends on the magnitude of the force but also on the position of the line of action of the force relative to the axis about which the body is tending to turn. The moment, M, is defined as the product of the force and the perpendicular distance between the axis and the line of action of the force, i.e.

$$M = Fs \sin \theta$$

where s is the displacement of the line of action of the force, F, from the axis and θ is the angle which the displacement makes with the line of action. Moment is thus an example of the cross product of two vectors (see p. 60). The unit of moment is the newton metre (N m).

Couples

A couple consists of a system of forces which have a zero resultant but a non-zero moment. In most practical cases a couple consists of a pair of forces which are equal in magnitude, opposite in direction and not on the same straight line. Consider such a couple, where the magnitude of each force is F and the perpendicular distance between their lines of action is d, Fig. 3.1.

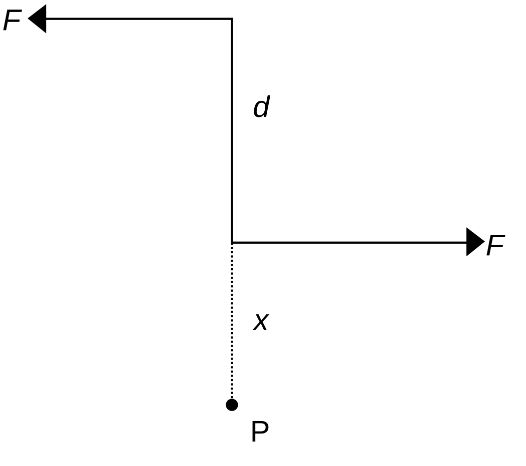


Fig 3.1 *A couple*

Taking moments about the point P as shown in the diagram:

Moment of top force (anticlockwise) $\quad M = F(d + x)$

Moment of bottom force (clockwise) $\quad M = Fx$

Total moment, (minus since they are in opposite directions)

$$M = F(d + x) - Fx$$
$$= Fd + Fx - Fx$$
$$= Fd$$

Since x does not appear in the final equation the moment of a couple is the same about all points in its plane.

The total moment of a couple is called a torque.

The torque of a couple about any point is the product of the magnitude of one of the forces and the perpendicular distance between their lines of action.

Levers

Archimedes said, 'Give me a place to stand and I will move the world'. He knew that using a lever, a small force applied to one end can produce a larger force at the other - and if the lever were long enough, the force could be large enough to move the world.

A lever is a rigid body which is free to turn about a fixed point (or axis) called the fulcrum.

The main purpose of a lever is to provide a greater moment than would otherwise be the case, by providing a means of extending the distance between the axis and the line of action of the force. A lever is an example of a simple machine, by which the force a person can exert can have its effect multiplied - for example, a spanner used to tighten a nut.

Every lever has a fulcrum, about which it turns. It also must have an effort, which is the applied force, and a load, which is the final force exerted. The relationship between the magnitude of the effort and the magnitude of the load is given by the law of the lever. In this case, the law may be stated as: effort multiplied by perpendicular distance from effort to fulcrum equals load multiplied by perpendicular distance of load from fulcrum.

The mechanical advantage of a lever is the ratio between the load and the effort; if the mechanical advantage is large, this means that the lever enables a small force applied as the effort to produce a large force as the load. The above statement of the law gives the following relationship.

$$\text{Mechanical advantage} = \frac{\text{load (L)}}{\text{effort (E)}}$$

Assuming 100% efficiency

$$\frac{L}{E} = \frac{\text{distance from fulcrum to E}}{\text{distance from fulcrum to L}}$$

A large mechanical advantage is produced by making the distance from fulcrum to effort much greater than that from fulcrum to load. For example, when a person uses a crowbar to lever apart two bodies, the fulcrum is at the bend in the crowbar, and the distance of the effort from the fulcrum is much greater than the distance of the load from the fulcrum. Typical figures for the distances would be 50 cm from fulcrum to effort and 5 cm from fulcrum to load, so that the mechanical advantage is 10, and an effort of 300 N applied to the crowbar produces a force at the load end of 3000 N. Levers with a large mechanical advantage are used to enable a person to apply a larger force than if he or she were to apply it directly to the load.

The downside of having a large mechanical advantage is that it requires a large movement of the effort to produce a small movement of the load. While having a large mechanical advantage increases the force acting, it cannot increase the work done. Since work is the product of force and displacement, if the force is increased by a certain factor then, assuming 100% efficiency, the distance moved must be decreased by the same factor. In the example of the crowbar given above, the force is multiplied by 10, so the distance moved is divided by 10, and a movement at the effort end of 10 cm produces a movement at the load end of only 1 cm. Sometimes a lever with a fractional mechanical advantage is used (see below); in this case, the force applied at the load is smaller than the effort, but the distance moved by the load is larger than the distance moved at the effort.

Three classes of lever

Levers are divided into three classes, according to the arrangement of the effort, load and fulcrum.

Class 1 levers have the fulcrum in the middle, with the effort at one end and the load at the other, Fig. 3.2.

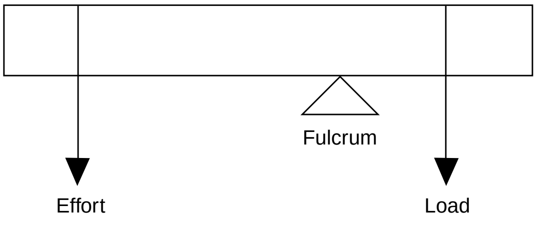


Fig. 3.2 *Class 1 lever*

Examples of class 1 levers include a metre stick when used to verify the law of the lever, a crowbar, a seesaw, and a screwdriver used to lever the lid off a tin of paint. Double class 1 levers include scissors, pincers, pliers, and shears. A lifting barrier which has a counterweight to help lift it (e.g. automatic railway level crossing barriers) is also a class 1 lever. The mechanical advantage of a class 1 lever can have any value, both greater than and less than 1.

Class 2 levers have the load in the middle, with the effort and the fulcrum at the ends, Fig. 3.3.

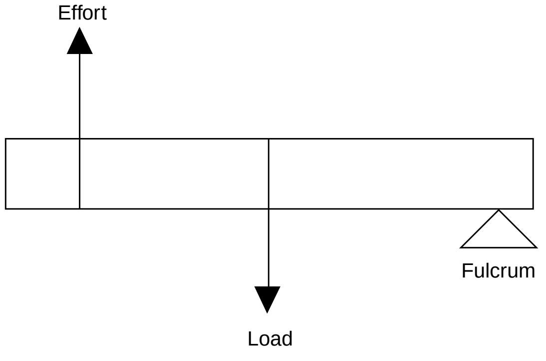


Fig. 3.3 *Class 2 lever*

Examples of class 2 levers include a wheelbarrow and a bottle-opener, and a nutcracker is a double class 2 lever. The mechanical advantage of a class 2 lever is always greater than 1 (assuming 100% efficiency), because the effort is always farther from the fulcrum than the load.

Class 3 levers have the effort in the middle, with the load and the fulcrum at the ends, Fig. 3.4.

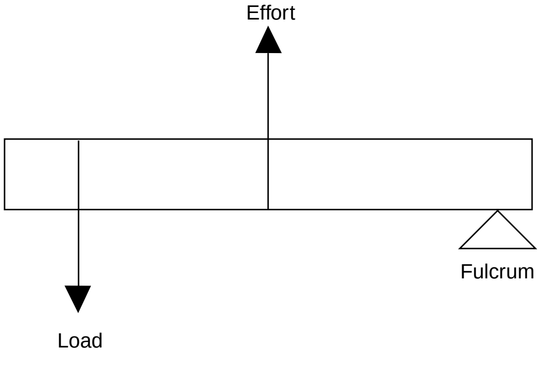


Fig. 3.4 *Class 3 lever*

Examples of class 3 levers include a fishing rod used to land a fish and the human forearm. Tweezers and tongs use a pair of class 3 levers. The mechanical advantage of a class 3 lever is always less than 1 (assuming 100% efficiency), because the load is always farther from the fulcrum than the effort.

Equilibrium

A body is said to be in equilibrium when its acceleration is zero. In many everyday examples, bodies which are in equilibrium are at rest.

There are two conditions for a body to be in equilibrium.

1. The (vector) sum of the external forces is zero.

2. The algebraic sum of the moments of the external forces about any axis is zero.

The second condition for equilibrium is sometimes called the principle of moments.

Law of the lever and principle of moments

The law of the lever is a special case of the principle of moments (the second condition for equilibrium). The law of the lever deals with a body (a lever) which is in equilibrium when free to rotate about one axis, usually referred to as the fulcrum.

If one applies the principle of moments to such a body, taking moments about the fulcrum, the moments acting are the clockwise and anticlockwise moments. From the principle of moments it follows that the algebraic sum of these moments must be zero, i.e. the sum of the clockwise moments equals the sum of the anticlockwise moments, since clockwise and anticlockwise moments have opposite signs. The law of the lever states: when a lever is in equilibrium the sum of the clockwise moments about the fulcrum is equal to the sum of the anticlockwise moments.

3.2 Experimental Approach

To Demonstrate that the Moment of a Force Depends on its Position

A metre stick is held at one end by a clamp which is held loosely in the boss-head of a retort stand so that the metre stick is free to turn about, say, the 10 cm mark. A spring balance is attached to the metre stick at various distances from the turning point (it will be sufficient to hook the end of the spring balance around the metre stick at the required point), and the force registered on the spring balance as it is used to turn the metre stick and lift up the free end is recorded. It will be seen that the force decreases as the distance from the turning point increases. If the force needed to turn the metre stick is too small to register properly on the spring balance, it can be increased by attaching a suitable mass to the end of the metre stick or by tightening the boss-head a little. If the spring balance is attached to the metre stick by a thread, it is possible to pull on the spring balance when it is at an angle to the metre stick, and thereby show that the moment depends on the perpendicular distance from the turning point to the line of action of the force. The experiment can be repeated with different masses on the end of the metre stick to vary the moment required to turn the metre stick.

To Demonstrate the Conditions of Equilibrium for a Body

Weigh a metre stick, then hang it from two spring balances, and hang a number of weights from the metre stick, as shown in Fig. 3.5. Ensure that the spring balances are hanging vertically. Note the magnitude of each weight and also the positions on the metre stick where all the threads are attached. Read the forces on the spring balances, and for accuracy lift up the metre stick, let it go again, and re-read the balances. The second readings should be identical with the first; if not, there is friction in the balances causing the readings to be inaccurate.

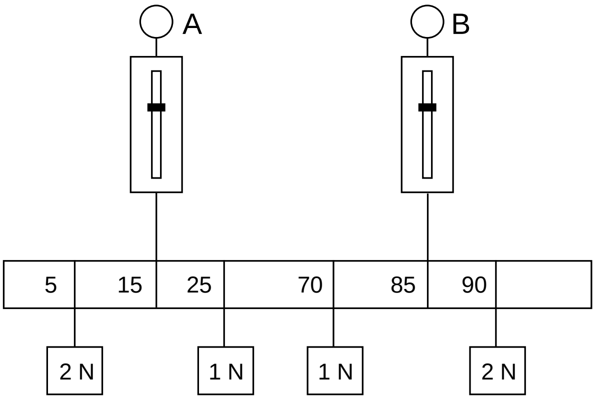


Fig 3.5

To show the first condition for equilibrium, find the total weight of the hanging weights, and add on the weight of the metre stick. Show that this total weight is equal to the total of the forces shown on the two spring balances. This would imply that when the forces acting upward (shown on the spring balances) are added to the downward forces of the weights (with appropriate signs - if upward forces are positive, downward forces must be negative), the total should be zero.

To show the second condition for equilibrium, take any point on the metre stick, and calculate the moments of all the forces acting on the metre stick about this point. The sum of the moments (again with appropriate signs - if clockwise moments are positive, anticlockwise moments must be negative) should be zero. Repeat the calculation for moments about other points on the metre stick.

Do not expect the answers to the calculations from this experiment to be exact, especially for the second one. A spring balance will not measure forces more accurately than ± 0.1 N normally, and an error of 0.1 N in the reading of the balances can produce quite a large difference in the moment produced by the force. It is more convincing for students if the clockwise and anticlockwise moments are worked out separately, so that the final answer is seen to be much smaller than either of the two moments used to calculate it (i.e. the anticlockwise moment is almost equal to the clockwise moment).

Example of calculations

(a) Given that the weight of the metre stick in the diagram (Fig. 3.5) is 0.5 N and the readings on spring balances A and B are 3.4 N and 3.1 N respectively, show that the first condition of equilibrium is obeyed.

(b) Assuming that its weight acts at the mid-point of the metre stick, show that the second condition of equilibrium is obeyed.

(a) Total upward force = 3.4 + 3.1
= 6.5 N

Total downward force = 2 + 1 + 1 + 2 + 0.5
= 6.5 N

Total upward force = total downward force;
first condition is obeyed.

(b) Taking moments about zero mark on metre stick (taking clockwise moments as positive)

Weight of metre stick is 0.5 N and acts through the 50 cm mark.

Sum of moments
= (2 x 5) - (3.4 x 15) + (1 x 25) +
(0.5 x 50) + (1 x 70) - (3.1 x 85) + (2 x 90)

= 10 - 51 + 25 + 25 + 70 - 263.5 + 180

= 310 - 314.5

= -4.5 N cm

Sum of moments is zero within the limits of experimental error; the second condition is satisfied for moments about a particular axis, i.e. an axis through the zero mark.

To show that the condition is satisfied for other axes, moments may also be taken about an axis through any other point on the metre stick.

Taking moments about an axis through the 15 cm mark (clockwise positive)

Sum of moments
= – (2 x (15 - 5)) + (3.4 x (15 - 15))
+ (1 x (25 - 15)) + (0.5 x (50 - 15))
+ (1 x (70 - 15)) - (3.1 x (85 - 15))
+ (2 x (90 - 15))

= -20 + 0 + 10 + 17.5 + 55 - 217 +150

= 232.5 - 237

= -4.5 N cm

Again, sum of moments is zero within experimental limits.

Using the Law of the Lever to Calculate the Weight of a Metre Stick

A metre stick is suspended by a length of thread from a retort stand and the position of the thread is moved until the metre stick balances. The point at which the thread is attached is noted; this is the centre of mass of the metre stick, i.e the point through which the weight acts.

The thread is then pushed away from the centre of mass, and a known weight, *W,* is attached to the shorter side, Fig. 3.6. The positions of weight and thread are adjusted until the metre stick again balances. These positions are noted, and the distances d_1 and d_2, as shown in the diagram, are measured.

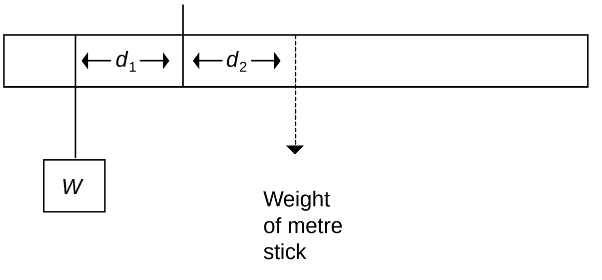


Fig. 3.6

Using the law of the lever,

$$W \times d_1 = \text{weight of metre stick} \times d_2$$

Hence the weight of the metre stick can be found.

3.3 Applications

Moments

When a door is to be closed, it is easy to close it by pushing at the handle side, for the distance between the point of application of the force and the axis of rotation is large, so a small force is multiplied by the large distance to give a moment to move it. If the door is pushed close to the hinge side, the distance between the point of application of the force and the axis of rotation is small, so to get a moment large enough to move the door, a large force must be applied. This can be observed when a passenger tries to open the front door of a car from inside. The car door is hinged at the front, for safety reasons, and the person usually tries to push it near the front, i.e. near the line of the hinges (the axis of rotation), because this is where the door handle is placed. Since the distance between the force and the axis of rotation is small, a large force must be applied to open the door. If the handle were placed farther to the rear of the door, the force needed to open the door would be smaller (but then the handle would not be easily visible to the passenger sitting in the seat).

Couples

Couples are commonly applied in everyday machines. To steer a bicycle, a couple is exerted on the ends of the handlebars - one end is pulled towards the rider, the other is pushed away. Similarly to turn the steering wheel of a car (or any other similar wheel), one side is pushed upwards and the other pulled downwards; again the two forces form a couple. In the operation of a simple electric motor or a moving-coil meter, the forces acting on the sides of the coil in the magnetic field form a couple which rotates the coil.

Levers

Levers are used in all sorts of everyday situations. Many tools are levers, e.g. crowbar, claw hammer (used to pull out a nail), spanner, socket wrench, wheel brace, pliers, pincers, wire cutters, shears, scissors. Levers are used in all sorts of machines, and in such everyday objects as door handles, bottle openers, can openers, as well as many parts of the human body.

To tighten a nut on a bolt, it is not usually enough to use our fingers, for the force we can apply is limited, being provided by the friction between the fingers and the side of the nut and, as well, the distance between the point of application of the force and the axis of rotation of the nut is small (half the width of the nut), so that the moment we can apply is very small. If we use a spanner to turn the nut the distance involved becomes much greater, being the length of the spanner, and the force applied can also be much greater, since it is not applied by friction, so that the moment applied to the nut becomes very much greater, and the nut can be made very tight.

If the wheel-nuts of a car are too tight to be loosened with a normal wheel brace, it is possible to arrange two metal bars (or strong pieces of wood), as shown in Fig. 3.7(a), to increase the distance between the force and the axis, and so increase the moment on the nut. Alternatively, if the wheel brace is a different shape, a strong metal tube or pipe placed over the end of it can be used to extend the distance from the axis to increase the moment, as shown in Fig. 3.7(b).

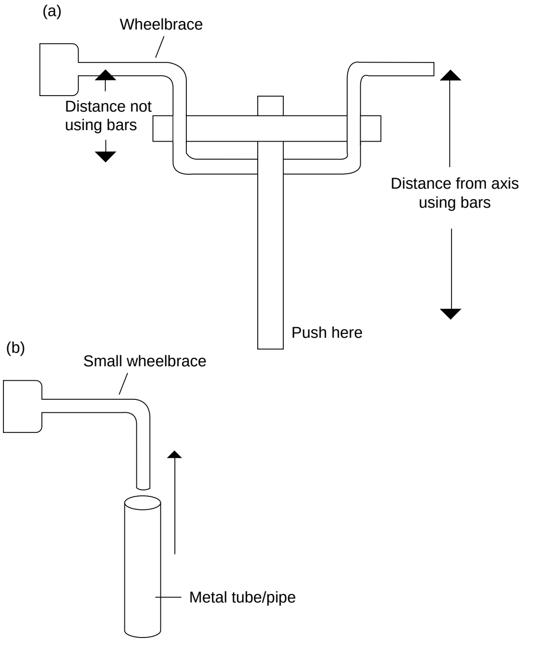


Fig. 3.7 *Increasing the moment of the force on a wheel brace*

If one is digging with a long-handled spade, one applies a large moment to the spade when one pulls back on the top of the handle, and the spade can exert a large force on the earth. If the handle of the spade is shorter, one cannot apply such a large moment, as the distance involved is less.

Centre of mass

When a body is supported at its centre of mass, it can be held up by a single force. This is used by a circus performer when he (or she) balances objects on the top of a long pole - it may be a stack of cups and saucers, or it may be the assistant, perhaps sitting in a chair. The centre of mass of the object(s) must be directly above the top of the pole. A tightrope walker must keep his centre of mass directly over the rope as he walks, so that his weight acts directly on the rope.

3.4 Worked Examples

1. A metre stick, of weight 0.8 N, is set up with a number of weights hanging from it, as shown in Fig. 3.8. Calculate the forces shown on the two spring balances supporting the metre stick. Assume that the centre of mass is at the 50 cm mark.

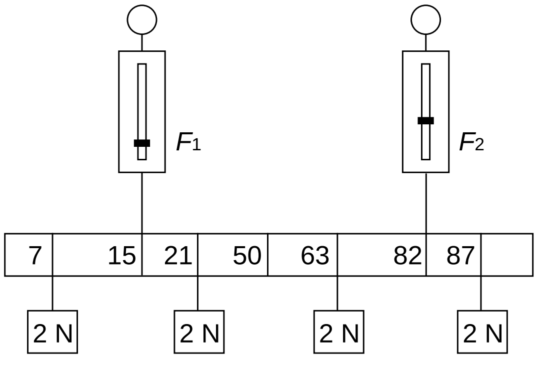


Fig. 3.8

Using the first condition of equilibrium:

upward force - downward force = 0

$$\begin{aligned} \text{upward force} &= \text{downward force} \\ F_1 + F_2 &= 2 + 2 + 0.8 + 2 + 2 \\ &= 8.8\text{ N} \qquad (1) \end{aligned}$$

Using the second condition of equilibrium:

sum of moments about any axis = 0

(The moment of a force about the axis through which its line of action passes is zero. So, if we take moments about the axis where one of the unknown forces acts, that unknown force will be eliminated from the equation, leaving only one unknown which can therefore be found easily.)

Taking moments about the 15 cm mark (anticlockwise positive):

$$\begin{aligned} (2 \times 8) - (2 \times 6) - (0.8 \times 35) - (2 \times 48) & \\ + (F_2 \times 67) - (2 \times 72) &= 0 \\ 16 - 12 - 28 - 96 + 67F_2 - 144 &= 0 \\ 67F_2 &= 264 \\ F_2 &= 3.94 \\ &= 3.9\text{ N} \end{aligned}$$

From (1) above,

$$\begin{aligned} F_1 + F_2 &= 8.8\text{ N} \\ F_1 + 3.94 &= 8.8 \\ F_1 &= 4.86 \\ &= 4.9\text{ N} \end{aligned}$$

The two forces exerted by the spring balances are 4.9 N on left, 3.9 N on right.

2. A bridge consists of a uniform girder of length 10 m and mass 20 t supported at its ends, A and B. Calculate the forces exerted on its supports when a truck of mass 10 t has its centre of mass 2 m from one end of the bridge.

$$\begin{aligned} \text{Mass of girder} &= 20\text{ t} \\ &= 20\,000\text{ kg} \\ \text{Weight of girder} &= mg \\ &= 196\,000\text{ N} \\ &= 196\text{ kN} \end{aligned}$$

Similarly, weight of truck $= 98$ kN

First condition of equilibrium

$$F_A + F_B - 98 - 196 = 0$$

$$F_A + F_B = 294 \text{ kN} \quad (1)$$

Second condition of equilibrium

Taking moments about end A of girder

$$(98 \times 2) + (196 \times 5) - (F_B \times 10) = 0$$

$$1176 = 10F_B$$

$$F_B = 117.6$$

$$= 118 \text{ kN}$$

Substituting in (1)

$$F_A + 117.6 = 294$$

$$F_A = 176.4$$

$$= 176 \text{ kN}$$

Forces on supports are 176 kN at A, 118 kN at B.

3. The diagram shows a crane lifting an object of weight 19.6 kN. The jib of the crane weighs 9.8 kN. Calculate the tension T in the cable supporting the jib.

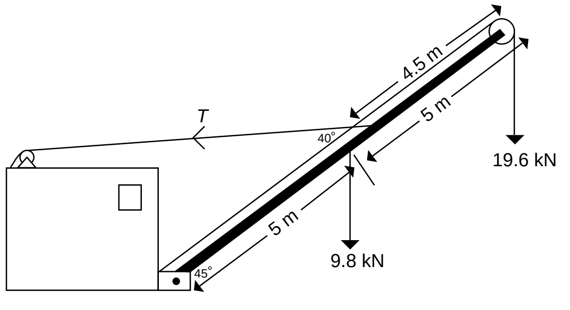


Fig. 3.9

There will be a reaction force at the bottom of the jib where the jib pivots on the base. The magnitude and direction of this force are unknown but, if we take moments about an axis through the pivot, this force does not enter into the calculations. We will also assume the tension in the cable running parallel to the jib has a negligible moment about the base, because the distance involved is very small.

Taking moments about the base of the jib

$$T \times 5.5 \sin 40^\circ = (19.6 \times 10 \cos 45^\circ) + (9.8 \times 5 \cos 45^\circ)$$

($5.5 \sin 40^\circ$ is the perpendicular distance from the pivot of the jib to the line of action of force T, $10 \cos 45^\circ$ is the perpendicular distance from the same point to the line of action of the load, and $5 \cos 45^\circ$ is the perpendicular distance from the same point to the line of action of the weight of the jib.)

$$3.535\, T = 138.6 + 34.6$$

$$3.535\, T = 173.2$$

$$T = 49 \text{ kN}$$

Note: the tension in this cable is $2^1/_2$ times the load the crane is lifting.

4. The diagram shows a tower crane as used on large building sites. The boom is attached to the tower. The length of the boom, measured from the tower, is 12 m and it has a weight of 150 kN, acting at its midpoint. The weight of the load is 80 kN, acting at 2 m from the end of the boom. The support cable is attached to a point 4 m from the end of the boom and makes an angle of 20° with the horizontal. Calculate the tension in this cable.

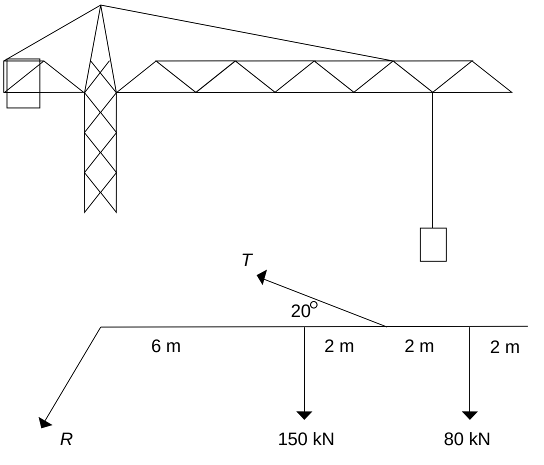


Fig. 3.10

The diagram shows the forces involved. The reaction at the end attached to the tower is in an unknown direction, but if moments are taken about an axis through that end, this force is eliminated from the calculations.

Taking moments about an axis through the end attached to the tower

$$(150 \times 6) + (80 \times 10) - (T \times 8 \sin 20°) = 0$$

$$T = 621 \text{ kN}$$

5. A metre stick is hanging at the 50 cm mark, Fig. 3.11. Weights of 2 N are hanging at the 22 cm mark, the 36 cm mark and at the 70 cm mark. Where should a 1 N weight be attached to the metre stick to make it balance?

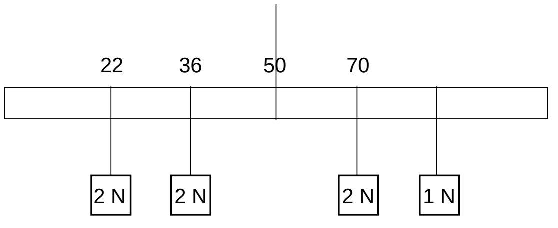


Fig. 3.11

Taking moments about the fulcrum

Moment of first weight

$= 2 \times (50 - 22)$

$= 56$ N cm anticlockwise

Moment of second weight

$= 2 \times (50 - 36)$

$= 28$ N cm anticlockwise

Moment of third weight

$= 2 \times (70 - 50)$

$= 40$ N cm clockwise

Total moment $= 44$ N cm anticlockwise
Moment required
to balance $= 44$ N cm clockwise

This is produced by a 1 N weight, so its distance from the fulcrum is 44 cm. The 1 N weight is 44 cm from the fulcrum on the clockwise side, i.e. at the 94 cm mark.

6. A steel girder of length 4 m has a weight of 1000 N and has its centre of mass at the midpoint. When it is supported at a distance of 1.5 m from one end, it balances horizontally when a man stands at a distance of 0.7 m from the same end. Calculate the weight of the man.

In this case the fulcrum is not at the centre of mass so the weight of the girder must be taken into account in the calculations.

Taking moments about the fulcrum

Clockwise moment $= 1000 \times 0.5$

$= 500$ N m

Anticlockwise moment $= W \times 0.8$

Since the lever balances,
clockwise moment = anticlockwise moment

$$W \times 0.8 = 500$$

$$W = 625 \text{ N}$$

7. What force does the biceps muscle apply to the radius bone of the forearm to raise a weight of 100 N in the palm of the hand?

The forearm is a class 3 lever. The fulcrum is the elbow, the effort is applied where the biceps muscle tendon is attached to the radius, and the load is the weight in the palm of the hand. Assuming that the distance from the elbow to the palm of the hand is 30 cm, and from the elbow to the attachment point of the tendon is 4 cm, we can calculate the effort required.

Moment produced
by weight $= 100$ N $\times 30$ cm

$= 3000$ N cm

Moment produced
by muscle $= F \times 4$ cm

$$4F = 3000$$

$$F = 750 \text{ N}$$

The force exerted by the muscle is much greater than the weight of the object being lifted.

CHAPTER 4

DENSITY AND PRESSURE

4.1 Background

Pascal

Blaise Pascal (1623-1662) was born in Clermont-Ferrand in central France. His father, a respected mathematician, moved to Paris in 1631 where he educated his son. The young Pascal proved to be a precocious mathematician and physicist, writing a number of mathematical theses. The Italian physicist **Evangelista Torricelli** (1608-1647) had measured the pressure of air for the first time and produced the first man-made vacuum in 1643. Torricelli stated that air pressure decreases with height, and Pascal decided to carry out an experiment to confirm this. As he lived in Paris, without mountains nearby, he wrote to his brother-in-law, Florin Périer, asking him to make a barometer and to compare the pressure at the bottom and at the top of the Puy-de-Dôme, a mountain about 1400 m high near Clermont-Ferrand. Périer found that the pressure was appreciably lower at the top of the mountain than at the bottom. Pascal obtained glass tubes up to 14 metres long from a glass works near where he lived and, using these, he made barometers filled with water and also with red wine. Using these barometers he was able to show that the height of liquid in a barometer varied inversely with the density of the liquid.

In 1654 Pascal experienced a religious conversion, and he wrote little on mathematics or science after that. Instead, he devoted himself to Christian apologetics. His main work is *Apologie de la Religion Chrétienne*, but many of his other writings were gathered together and published under the title *Pensée*s. He did not enjoy good health and died from cancer at the age of 39.

His name was given to the SI unit of pressure only in 1971; an alternative suggestion was that the unit be called the giorgi, after the Italian physicist **Giovanni Giorgi** (1871-1950), who devised the MKS system of measurement (see section on units).

Archimedes

Archimedes (c.287-212 BC) was a leading scientist of ancient times. He was born in Syracuse, in Sicily, the son of Pheidias, an astronomer. He studied at the school of mathematics and physics founded by **Euclid** (c.325 BC - ?) in Alexandria, and then returned to Syracuse, where he was closely associated with the king of Syracuse, Hiero II. He made many mechanical contrivances; the principal one is the Archimedean screw, said to have been invented for the purpose of removing water from the hold of a large ship belonging to the king. The Archimedean screw is still used today for raising water for irrigation, or grain, e.g. in a combine harvester. He also invented weapons of war which were used in the defence of Syracuse, which was besieged by the Romans; among them is said to have been a mirror which focused the sun's rays on ships to set them on fire.

His most famous exploit occurred when King Hiero gave a goldsmith a block of gold and asked him to made a crown from it. When the crown was complete, its mass was equal to the mass of the block of gold, but gossip spread that the goldsmith had kept some of the gold and substituted the same mass of cheaper silver for it. Archimedes was asked to investigate if the allegation was true. The story goes that when Archimedes was stepping into a bath and noticed the water overflowing, it struck him that he could solve the problem by comparing the volume of water displaced by the crown with the volume displaced by the same mass of pure gold. If the crown were pure gold, the volumes would be the same, but if the crown contained silver, the volume it would displace would be greater. He was so overjoyed that he jumped from the bath and, without dressing, ran from the bath-house to his home shouting, ‘Eureka, eureka’ (I have found it, I have found it). When he showed that the crown was in fact made from an alloy of gold and silver, the goldsmith who had made it admitted his guilt.

Archimedes is reported to have said, ‘Give me a place to stand, and I will move the world’. When the king asked him to illustrate this, he constructed a machine by which the king was able to move a large ship, fully laden.

Nine treatises written by Archimedes survive today, mostly containing propositions in geometry. He showed that the value of π was between $3^1/_7$ and $3^{10}/_{71}$ and, in some of his calculations, he arrived at a method of calculating areas which is equivalent to the modern method of integration.

Archimedes was killed when the city of Syracuse was captured by the Roman besiegers in 212 BC, despite the orders of the Roman general, who had commanded that Archimedes be spared. The general gave him an honourable burial and, some 150 years later, the Roman writer Cicero records finding the grave of Archimedes, much overgrown, near one of the city gates.

Robert Boyle

Robert Boyle (1627-1691) was born at Lismore Castle, Co. Waterford, the 14th child and 7th son of Richard Boyle, Earl of Cork. He learned to speak French and Latin at an early age and was sent to Eton College when only 8 years old. After leaving school and touring France, Italy and other parts, he devoted his life to study and scientific research. Reading in 1657 that the German physicist **Otto von Guericke** (1602-1686) had invented a vacuum pump, Boyle set himself the task, along with his friend, the English physicist **Robert Hooke** (1635-1703), of improving on von Guericke’s design. When he had completed his ‘pneumatical engine’, he began experiments on the properties of air. He published his results in 1660 under the title *New Experiments Physico-Mechanical Touching the Spring of the Air and Its Effects.* Boyle’s book attracted criticism, and it was while answering some of the criticisms that Boyle enunciated the law which bears his name; that the volume of a given mass of gas varies inversely as the pressure. (In continental Europe this law is attributed to the French physicist **Edmé Mariotte** (1620-1684) who published it in 1676.)

Apart from his experiments on air, Boyle studied chemistry, and published *The Sceptical Chymist* and other books on the subject. He advanced an atomic theory of matter and distinguished between mixtures and compounds. He also studied theology and learned Hebrew and Greek so as to read the Bible in its original languages. He spent large sums on translating the Bible and left a sum of money in his will to provide lectures for proving the Christian religion against ‘notorious infidels - atheists, theists, pagans, Jews and Mohammedans’.

Life Under Pressure

Atmospheric pressure at sea level is approximately 1×10^5 Pa, i.e. the atmosphere exerts a force of approximately 100 000 N on each metre^2 at sea level - this is equivalent to the weight of 10 tonnes on each metre^2. Thus the air pressure on a person’s chest, which has an area of about 0.25 m^2, is equivalent to having the weight of 2.5 tonnes pressing inwards on it. This causes no problems because there is an equal force inside the body, due to the air in our lungs, etc., pressing outwards, and the two forces balance. The same applies to animals living in the sea, even at great depths; the pressure outside is balanced by an equal pressure inside. When deep-sea creatures are raised to the surface they may explode, because the pressure inside stays the same while the pressure outside decreases. The same thing could happen to a person if the pressure around them was reduced quickly; if, for example, they were taken up to a great height in a plane or balloon. This is why aircraft have pressurised cabins and cockpits, so that while the plane is flying at 10 000 m, the pressure in the cockpit is only equivalent to what would be experienced at a height of, say, 3000 m. If, however, the cockpit were to become holed, particularly in a military plane flying at a very high altitude, the drop in pressure could be fatal, even if an oxygen mask were provided to supply the pilot with oxygen, for the internal pressure could make the body swell and perhaps explode. As well as that, the reduced pressure lowers the boiling point of water, and at a height of 20 000 m it becomes 37 °C, so that the blood would start to boil at body temperature at that height. To prevent this happening, pilots wear G-suits, which are made of non-stretch fabric, and which have tubes which contain air, attached along the seams. If the external pressure falls suddenly, the tubes expand because of the pressure of the air in them, thus tightening the suit and exerting a pressure on the pilot's body.

Flight

The first balloon to fly was made in July 1783 by two French brothers, **Jacques Étienne Montgolfier** (1745-1799) and **Joseph Michel Montgolfier** (1740-1810). The Montgolfier brothers, who were paper manufacturers from Lyons, made a bag over 10 m in diameter and filled it with hot air from a straw fire. Later in 1783, another Frenchman, **Jean Pilâtre de Rozier**, ascended in a Montgolfier balloon, so becoming the first person to fly. In November of that year, he and the **Marquis d'Arlandes** made the first free flight for 25 minutes over Paris.

The news of the Montgolfiers' balloons stimulated others to make balloons. The previous year, **Tiberius Cavallo** had made bubbles filled with hydrogen and had seen them rise to the ceiling. In August 1783 the French physicist, **Jacques Charles** (1746-1823) (after whom Charles's law is named) filled a small balloon with hydrogen and saw that it flew. In December of the same year he, accompanied by a **M. Robert**, flew 40 km and rose to over 600 m in a large hydrogen balloon.

The first crossing of the English Channel by balloon was achieved on 7 January 1785 by **Jeffries** and **Blanchard**, who flew from Dover to a forest 20 km from Calais. They only just made it to land at the end, having thrown out not only all their ballast but also part of their clothing to lighten the balloon. Pilâtre de Rozier, with **P.A. Romain**, attempted to cross from France to England using a hot-air cylinder attached below a hydrogen balloon. After half an hour in the air, the hydrogen caught fire and both men died as the balloon plunged to earth.

A major problem with balloons was that they were at the mercy of the wind direction, as they could not travel against the wind. Early experiments with large oars to 'row' the balloon in the desired direction failed, and it was not until mechanical propulsion was used that it was possible to steer a balloon in the desired direction. While there was a little success with steam engines, it required the development of the petrol engine before balloons could become airships, able to follow a set course despite the wind. Some of these were non-rigid, using the pressure of the gas in the balloons to support the structure, while others had a metal framework to give rigidity. The German inventor, **Count von Zeppelin** (1838-1917) was very successful with hydrogen-filled rigid airships or Zeppelins. These were used by the Germans in the First World War for observation, and between the wars they ran a regular transatlantic service from Berlin to New York.

The fact that hydrogen is highly flammable led to its eventual disuse for manned balloon flight, although it took almost 150 years for this to happen. In 1922, the US adopted a helium-only policy for manned balloons, but other countries, with no sources of helium, continued to use hydrogen. In 1937, the 'Hindenburg', a Zeppelin, caught fire as it arrived in New York and was completely destroyed, with the loss of 36 lives. Previous to that, a British airship, R101, had also crashed and caught fire in France on its maiden voyage to India, again with heavy loss of life. These accidents, coupled with the coming of the Second World War, led to the end of the use of hydrogen-filled balloons for manned flight. The US Navy used helium-filled non-rigid balloons ('blimps') during and after World War 2 for reconnaissance and anti-submarine activities. However, helium is expensive and this makes helium balloons too expensive for sport, which is today the preserve of hot-air balloons. Hydrogen, of course, is still used to fill meteorological balloons, which carry instruments into the air to make measurements at high levels in the atmosphere.

The first heavier-than-air flying machines were gliders, which were first constructed and flown by the English baronet, **Sir George Cayley** (1773-1857), between 1804 and 1853. His coachman was the (unwilling?) passenger in the first man-carrying glider in history in 1853. The German aeronautical pioneer, **Otto Lilienthal** (1848-1896) developed methods of controlling gliders in flight between 1891 and 1896 (he died when one of his craft crashed). Powered flight had to wait until the development of the internal combustion engine. It was the American brothers, **Wilbur Wright** (1867-1912) and **Orville Wright** (1871-1948) who used their bicycle-building experience to make a number of gliders, and then the first successful powered aeroplane. This aeroplane made six flights on 17 December 1903, at Kitty Hawk, North Carolina. The Wright brothers' success led to many others building and flying planes.

Milestones in aviation history include the first flight in Europe, made by Wilbur Wright in August 1908, the first flight across the English channel (**Louis Blériot** (1872-1936) on 25 July 1909), the first transatlantic flight (US Navy seaplane, May 1919, from Newfoundland to Portugal, via the Azores), and the first non-stop transatlantic flight (**Alcock** and **Brown,** June 1919, from Newfoundland to Clifden, Co. Galway). The first transatlantic flight from East to West was made by the Irish aviator, **Col. James Fitzmaurice** (1898-1965), on 12 April 1928. Col. Fitzmaurice was accompanied by two Germans, **Hermann Koehl** and **Baron Ehrenfield von Huenefeld**, and the flight took 36 hours. The first woman to fly solo across the Atlantic was the famous American aviator, **Amelia Earhart** (1897-1937), who made the journey on May 20-21 1932. Earhart was also the first person to complete the hazardous flight from Hawaii to California. The first round-the-world flight was completed by a US Army plane, one of four which started the journey, in 175 days in 1924. The flight did not include any long over-water sections, but only short flights from island to island. **Charles Lindbergh's** famous flight in the 'Spirit of St Louis', non-stop from New York to Paris, took him 33½ hours on 20 & 21 May 1927, and won him a prize of $25 000.

The flight of the world's first jet plane, a German Heinkel He 175, took place on 27 August 1939, almost two years before the first British jet. The German Messerschmitt Me 262, the first jet fighter, saw service with the German Air Force (Luftwaffe) in 1945. The first jet airliner, the de Havilland Comet, started service in 1952, while in 1959 Boeing introduced the 707, the first of Boeing's many successful jet passenger planes. The first supersonic airliner, the Anglo-French Concorde, began flying in 1969 and entered commercial service in 1976.

The first successful helicopter was made by the Russian-born US inventor, **Igor Sikorsky** (1889-1972), in 1939.

4.2 Do You Know

Compressed air can be used to drive machines.

Dentists' drills use compressed air to spin the vanes of a small turbine which turns the drill at high speed. The same principle was used by the American physicist, **A. A. Michelson** (1852-1931) to rotate the octagonal mirror in his apparatus to measure the speed of light.

Pneumatic drills operated by compressed air are used by road workers and builders to break up concrete, etc. The compressed air is applied alternately above and below a piston, moving it, and the blade of the drill which is attached to it, up and down.

The highest atmospheric pressure is found where the air is cold and dry.

Air is normally coldest and driest near the centre of large land masses, especially Central Asia in winter. The record high pressure of 1083.8 hPa was recorded in Agata, Siberia on 31 December 1968. The lowest atmospheric pressure is found in tropical storms. The record low pressure was 870 hPa in Typhoon Tip in the Pacific Ocean, west of Guam, on 12 October 1979.

A water barometer would have to be over 10 m high and a pump at the top of a well cannot pump water from a depth of greater than about 10 m.

Mercury is used in barometers because it is very dense, 13.6 times as dense as water. A barometer filled with water would need to be 13.6 times as high as a mercury barometer, or over 10 metres high (the barometer made by Pascal filled with red wine would have been even higher, which is why he needed glass tubes up to 14 m long). Otto von Guericke, the inventor of the first vacuum pump, made a water barometer with the top of the tube coming through the roof of his house. He had a wooden figure of a man floating on the water in the tube so that when the man was visible he could predict that the day would be fine, and when it was not visible, he could predict bad weather.

When water is being pumped out of a deep well or bore-hole using a vacuum pump it is the pressure of the atmosphere that pushes the water up the tube. Even if the pump at the top can create a perfect vacuum in the tube, the water will not rise more than 10 metres, the height of a water barometer. If the well is deeper than this, a pump at the top will not be able to lift water from it. For deeper wells a pressure pump may be placed at the bottom of the well to drive the water to the surface. Another possible alternative would be to seal the well and pump air down into it to increase the air pressure acting on the water, so that the greater air pressure could raise the water to the top.

Atmospheric pressure decreases with altitude, making it more difficult to breathe and also to cook food.

Pressure in a gas depends on the depth, but not in a simple way, because the density of the gas is not constant, but depends on the pressure. It is well known that the pressure of the air decreases as one goes up a mountain or in a plane. This is because one's depth from the top of the atmosphere decreases, and the weight of air pressing down on one is less. The pressure on the top of Mt Everest is only about $^1/_4$ of the pressure at sea level.

The decrease of air pressure causes difficulties with breathing when mountaineers try to climb the highest mountains. The first climbers to reach the summit of Mt Everest carried cylinders of oxygen on their backs to augment the oxygen in the air, but more recently many high mountains have been climbed without extra oxygen. Even at lower altitudes than Mt Everest the lack of pressure at height causes 'altitude sickness', experienced, for example, by travellers to the Andes in Peru, which reach heights of 4500 m. Those who live permanently in these regions are adapted to the lower atmospheric pressure by having a much larger number of red corpuscles in the blood, so that the reduced amount of oxygen in the air can be absorbed more efficiently by the blood. Anyone who lives for a time at high altitudes will find that the body adapts to the altitude by increasing the red corpuscle count. Athletes use this adaptation by training at high altitude, so that their blood can absorb and circulate more oxygen when they return to sea level. Doing this just before important events such as the Olympics gives them a performance advantage, especially in endurance events. Many of the successful athletes in events like the 5000 m, the 10 000 m and the marathon are natives of high altitude regions, mainly in Kenya and Ethiopia.

Another effect of reduced pressure at high altitude is that piston engines, which depend on the atmospheric pressure to push air into the engine, lose power at high altitudes. This is overcome in planes which use these types of engine by using supercharging, that is, using a type of compressor to force air into the engine.

The boiling point of water depends on the pressure, so that at high altitudes water boils at less than 100 °C. At a height of 1500 m water boils at 94.5 °C. This means that cooking foods by boiling, e.g. boiling an egg, takes much longer at high altitude.

4.3 Conceptual Approach

Density

Density is defined as mass per unit volume or mass divided by volume. The standard symbol for density is ρ. The unit of density used in school laboratory measurements is often the g cm^{-3} but the SI unit is, of course, the kg m^{-3}, and must be used in all calculations. Many students will probably need help to relate these two units.

1 kg = 1000 g

1 m = 100 cm $\Rightarrow$ 1 m^3 = 100^3 cm^3 = 1 000 000 cm^3 or 1 x 10^6 cm^3

Hence 1 g cm^{-3} = 1 000 000 g m^{-3} = 1000 kg m^{-3}

Thus, to convert g cm^{-3} to kg m^{-3} multiply by 1000 (10^3).

Students often have a poor concept of the size of 1 m^3. It may help to point out that the density of water is 1000 kg m^{-3}, i.e. the mass of 1 m^3 of water is one tonne.

To illustrate the usefulness of density, one can use the idea of a fair test. If you want to see which of iron or wood is heavier, and you compare the mass of a large block of wood with that of a small piece of iron, your students should say that that is not a fair test; that one must have the same volume of each. You can then go on to suggest that the volume be fixed at 1 cm^3 (or better, at 1 m^3) for all such comparisons, which leads to density being the unit used for comparison.

Pressure

Pressure at a point in a fluid is defined as force per unit area or force divided by area. The symbol for pressure is p and the unit is the pascal (Pa); 1 Pa = 1 N m^{-2}. The pascal is a relatively small unit - a force of 1 N exerted over an area of 1 m^2; for example, normal atmospheric pressure is 101 kPa.

Many people have difficulty distinguishing pressure from force as the two words are often used in everyday language as synonyms. It is said that 'the pressure which one exerts on a thumb-tack pushes it into the board', when it is simply the force exerted which pushes it in; or 'the pressure of the wind on the sails makes the yacht move'. It is necessary to make clear that pressure in science has a very definite meaning and does not mean force.

It is important to remember that pressure is a scalar quantity (see Chapter 5). It therefore cannot be said to act in any particular direction. Thus it is incorrect to say, for example, that the pressure at the bottom of a beaker of water acts downwards on the bottom of the beaker or that pressure in a fluid acts in all directions. It is the force which is associated with the pressure which is the vector quantity and acts in a particular direction. Thus the water exerts a downward force on the bottom of the beaker as a result of the force of gravity acting on the water; at a particular point in a fluid the forces exerted by the water is the same in all directions, i.e. the resultant force is zero. In this context care is needed in the use of language to avoid the suggestion that pressure and force are synonomous. For example, one should always refer to the pressure *at* a surface, etc., rather than the pressure *on* a surface.

Pressure increases with depth in a liquid. An experiment to show this uses a tall container with a series of holes in the side. When it is filled with water, the water comes out of the hole at the bottom with a much greater force than from the hole at the top, showing that the pressure is greater at the bottom.

Alternatively, one can use a thistle funnel, with its open end covered with a sheet of rubber, connected to a manometer or Bourdon gauge. The gauge or manometer will show that the pressure increases the deeper the funnel is placed in a tall container of liquid. If one uses liquids of different density, one can show that the pressure at a given depth (e.g. 10 cm) depends on the density of the liquid.

To calculate the pressure at a depth in a liquid

In a container of liquid, the pressure at any point results from the weight of the liquid above that point and the weight of the atmosphere, assuming that the container is open to the atmosphere. The pressure resulting from the weight of the liquid alone is known as the gauge pressure. So the actual pressure at a particular point in a liquid is equal to the gauge pressure plus the atmospheric pressure. In practice, even though it may not be explicitly stated, it is the gauge pressure which is normally in question.

To obtain an expression for the (gauge) pressure at a particular point in a liquid consider a container of liquid of density ρ, Fig. 4.1.

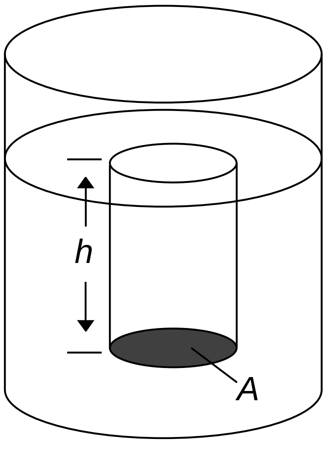


Fig. 4.1

Take the area A at a depth h as shown in the diagram.

Volume of liquid above A is $V = Ah$

Mass of liquid, $m = V\rho$

$= Ah\rho$

Weight of liquid, $W = mg$

$= Ah\rho g$

i.e. force acting on A is $F = Ah\rho g$

$$p = \frac{F}{A}$$

$$= \frac{Ah\rho g}{A}$$

$$= h\rho g$$

The pressure at a depth in a liquid is the product of the depth, the density of the liquid and the acceleration due to gravity. This equation may also be applied to gases if it may be assumed that the density is uniform.

The pressure in the sea or in a lake follows this formula. It is impossible to hide under water and breathe through a tube, unless one is very close to the surface (and so easily seen), as the pressure of the water around is too great to allow one to breathe, for the pressure of the air through the tube is not great enough to fill the lungs against the greater pressure of the water around. **Leonardo da Vinci** (1452-1519) designed a hose for breathing under water, but it was not successful for this reason.

Archimedes' Principle

Archimedes' principle states that when a body is immersed (submerged) in a fluid, the upthrust is equal to the weight of the displaced fluid. The upthrust, or buoyancy force, is a resultant upward force exerted by the fluid on the immersed body. This resultant force arises from the difference in pressure between the top of the body and the bottom (see below). One effect of the upthrust is that it is easier to lift a heavy object when it is immersed in, say, water than when it is out of the water. In other words, objects appear to weigh less when immersed in a fluid. For this reason the upthrust is sometimes referred to as the apparent loss in weight. If this term is used it is important to emphasise that it is an *apparent* loss in weight. The weight of a body is the force of gravity on it and this does not change when a body is immersed in a fluid.

In the case of a regularly shaped body immersed in a uniform fluid Archimedes' principle may be derived as follows. If a cylindrical body of cross-sectional area A and height h is immersed in a liquid of density ρ so that its top is at a depth d below the surface:

$$\text{Pressure at top of body} = \rho gd$$

$$\text{Force on top of body (down)} = pA$$

$$= \rho gAd$$

$$\text{Pressure at bottom of body} = \rho g(h + d)$$

$$\text{Force on bottom of body (up)} = pA$$

$$= \rho gA(h + d)$$

$$\text{Resultant force on body} = \rho g\,A(h + d) - \rho gAd$$

$$= \rho gAh$$

This is the upthrust on the body.

$$\text{Volume of body} = Ah$$

$$\text{Volume of displaced liquid} = Ah$$

$$\text{Mass of displaced liquid} = \rho Ah$$

$$\text{Weight of displaced liquid} = \rho gAh$$

Therefore upthrust = weight of displaced liquid.

Archimedes' principle and air

Archimedes' principle applies to all fluids, i.e. gases as well as liquids. Any body in air experiences an upthrust equal to the weight of the displaced air, but as this weight is so small, the effect is usually negligible. For example, if we consider the upthrust on 1 kg of water:

$$\text{Weight of water} = 9.8\text{ N}$$

$$\text{Volume of water} = 1 \times 10^{-3}\text{ m}^3$$

$$(\text{since density of water} = 1 \times 10^{3}\text{ kg m}^{-3})$$

$$\text{Density of air} = 1.29\text{ kg m}^{-3}$$

$$\text{Mass of displaced air} = 1.29 \times 10^{-3}\text{ kg}$$

$$\text{Upthrust} = \text{weight of displaced air}$$

$$= 1.29 \times 10^{-3} \times 9.8$$

$$= 0.0126\text{ N}$$

The only bodies for which the upthrust in air is generally significant are those of very low density, e.g. an inflated balloon, or which require very accurate weighing.

Law of Flotation

The law of flotation states that a floating body displaces its own weight of fluid. This law is a special case of Archimedes' principle. If a body is denser than the fluid in which it is immersed it sinks because the upthrust, which is equal to the weight of displaced fluid, is less than the weight of the body and so it experiences a resultant downwards force. If the body is less dense than the fluid then the upthrust when the body is completely immersed is larger than the weight of the body, so it experiences a resultant upwards force. As a result of this force the body rises to the surface and only remains immersed to the extent that the upthrust (or the weight of displaced fluid) is equal to the weight of the body. Since there is now no resultant force on the body it floats at this level.

The law of flotation also applies to bodies in air which are less dense than air. The main difference between bodies floating in air and those floating in water is that there is no clearly defined surface of air on which bodies float, as there is with water. Instead, the density of air gradually decreases as the height above the earth's surface increases, and a body less dense than air at sea level rises until its density is equal to the density of the surrounding air, where it floats because the weight of the displaced air is equal to the weight of the body. Another body of different density will float at a different height.

4.4 Experimental Approach

To Demonstrate Archimedes' Principle

A mass is hung from a spring balance, and its weight (in newtons) is found. If it is then lowered into water, its weight as shown on the spring balance will appear to decrease. The difference between the two readings is equal to the upthrust of the liquid on the mass. If it is lowered into different liquids, it will be seen that the upthrust varies from liquid to liquid. If liquids of known densities are used, it may be seen that the upthrust is proportional to the density of the liquid.

To measure the weight of the displaced liquid, the mass is put in an overflow can, and the liquid which overflows is collected in a weighed beaker. The mass of the liquid can then be found by weighing, and can be converted to weight by multiplying by g (9.8 m s^{-2}). The weight of the displaced liquid can be compared with the upthrust. It will not be exactly equal, due to experimental error, and especially due to surface tension effects in the overflow can.

An alternative, more direct, method of finding the weight of the liquid displaced using the overflow can is to tie a string around the top of the beaker with a loop across the middle so that it can be hung from a spring balance. The weight of the empty beaker and the weight of the beaker + liquid can be found directly by hanging from a spring balance, and the weight of the displaced liquid can be found by subtraction. This method has the advantage that, because the accuracy of a spring balance is less than that of a conventional balance, it is quite likely that the upthrust and the weight of displaced liquid will agree - at least they are both measured to the same level of accuracy.

Floating and Sinking

A body will float in a liquid if its density is equal to or less than the density of the liquid. This can be demonstrated using blocks of wood of different densities and trying to float them in liquids of different densities. Suitable liquids include water, methylated spirits, acetone (propanone), ether (ethoxyethane) and brine. The densities of the liquids may often be found on the label of the bottle.

Safety note

Acetone (propanone) and ether (ethoxyethane) are highly flammable. Do not allow any naked lights or sources of heat near while these are being used.

Law of Flotation

The demonstration is similar to that of Archimedes' principle. The weight of a body which will float is found, and the weight of the displaced liquid is found using an overflow can, as given above for the demonstration of Archimedes' principle. The two weights should be equal, within the limits of experimental error.

To Show that the Atmosphere Exerts Pressure

Vacuum pump method

Attach an empty container, of metal or plastic, to the vacuum pump, and turn on the pump. As it removes air from the container, the latter will collapse because the pressure outside is now greater than the pressure inside and so there is a resultant inward force on the sides of the container. (A plastic bottle which contained chemicals or soft drinks may be suitable for this experiment, but it may seem more impressive to students if a metal container is used.)

For this method, a filter pump is unlikely to remove enough air from the container to result in its collapse, unless the flow of water from the tap to which it is attached is very strong.

Steam method

Put some water in a metal container and, leaving the cap off, place the container over a Bunsen burner. Let the water boil for a few minutes so that the steam will drive out all the air from the container. Remove it from the heat and put on the top (or put a cork in the opening), then leave it to cool (or cool it under the tap, if preferred). The container will collapse as the steam in it condenses and leaves a partial vacuum behind.

Safety note

Ensure that the container does not contain any flammable liquid, or any substance which could produce noxious fumes when heated. Cans which contained paint stripper, wood preservative, etc., may be used, but only if they are well rinsed out using first a suitable solvent, followed by several rinses with water.

An alternative version of this method is to use an empty drinks can. When water has been boiled in it for a few minutes, it should be picked up with a crucible tongs and placed upside-down in cold water in a trough or sink, where it will collapse as the steam condenses.

Glass and card method

Take a jam-jar or gas jar (or any container with a level rim at the top - a beaker with a spout will not do) and fill it with water. Take a piece of thin card which is only a little larger than the mouth of the container and place it on the top of the container. Then, holding the card in place, turn the container upside-down. Let go the card; it will not fall off, because the pressure of the air under the card is greater than the pressure of the water above the card. (The experiment should be done over a sink, because the card will soften with the water and will eventually fall off.)

Measurement of Atmospheric Pressure

Atmospheric pressure is measured by barometers, of which there are two types, viz. the mercury barometer and the aneroid barometer.

The mercury barometer

A mercury barometer is made by taking a clean thick-walled glass tube, closed at one end, about 90 cm long, and filling it to the top with mercury. It is shaken and tapped to remove any air bubbles, then a finger is put on the open end and it is inverted into a dish of mercury. When the finger is removed, the level of mercury will fall in the tube, leaving a vacuum at the top. This is because the force arising from the atmospheric pressure at the bottom of the tube is not great enough to support the weight of mercury filling the tube. The vertical height from the level of mercury in the dish to the level of mercury in the tube gives a measure of the pressure of the air; it is normally about 750 mm, but it varies from day to day. If the tube is tilted, the mercury appears to move up the tube, but this is because the top of the tube gets lower; the vertical height as stated above remains unchanged.

Safety note

It is not recommended that a mercury barometer be made in the laboratory, because of the toxic nature of mercury. Mercury vapour is poisonous when mixed with air and inhaled. If mercury is spilled and collects in crevices, its vapour will collect in the room and, if the room is not well ventilated, the concentration of mercury vapour in the air at normal room temperature will come to exceed the recommended safety limit. Mercury which is spilled should be removed by combining it with powdered sulphur and then hoovering up the compound formed.

Atmospheric pressure is sometimes given in units of millimetres of mercury (mmHg), as measured from a barometer, giving values around 750. It is also commonly given, on home barometers, in inches of mercury (figures from 28-31 on such a barometer are inches of mercury).

The vacuum at the top of a barometer is sometimes called the Torricellian vacuum, after the Italian **Evangelista Torricelli** (1608-1647), who made the first barometer in 1643.

For accurate measurement, it is inconvenient that the level of the mercury in the dish varies as mercury flows in and out of the tube with the change of pressure, as it means that the scale to measure the height of mercury in the tube cannot be fixed, since its zero must be at the level of the mercury in the dish. To overcome this problem, a Fortin barometer is made with the mercury at the bottom in a flexible bag rather than a dish. Before taking a reading, the level of mercury in the bag is adjusted, using a screw at the base of the barometer, so that it just touches a fixed mark. This mark is the zero of the fixed scale, against which the level of mercury in the tube can be read to give the correct pressure, often using a vernier scale to get the reading to 0.1 mmHg (or 0.1 hPa if the scale is calibrated in hPa).

From the measurement of the height of mercury in a barometer, the atmospheric pressure can be calculated in pascal, the SI unit of pressure, using the formula for the pressure at a depth in a fluid.

In a mercury barometer, the pressure at the level of the mercury in the dish is the same inside and outside the tube - if they were not equal, the mercury would move in the tube until they were equal.

$$\text{Density of mercury} = 1.36 \times 10^4 \text{ kg m}^{-3}$$

$$\begin{aligned}\text{Atmospheric pressure} &= \text{pressure inside tube at the same level} \\ &= \text{pressure at a depth of 760 mm (0.76 m) in mercury} \\ &= \rho g h \\ &= 1.36 \times 10^4 \times 9.8 \times 0.76 \\ &= 1.01 \times 10^5 \text{ Pa}\end{aligned}$$

A pressure of 1×10^5 Pa was called 1 bar (from barometer), and atmospheric pressure was measured by meteorologists in millibars (mb). The bar and millibar are not SI units, so in the change over to SI units, the millibar was renamed the hectopascal (hPa); 1 hPa = 100 Pa = 1 mb. *Hecto* is not one of the recommended prefixes in the SI system of units (multiples and sub-multiples of standard units in the SI system are multiples of 10^3). However, meteorologists today measure pressure in hectopascals to preserve the equivalence with the millibar. Another non-SI unit which was used in connection with pressure is the torr (from Torricelli); 1 torr is a unit of pressure (used for measuring very low pressures, e.g. in an evacuated container) which is equal to the pressure of 1 mm of mercury (1 mmHg).

The aneroid barometer

The aneroid barometer uses the change in shape of an evacuated metal box as the pressure around it varies. The box is connected to a strong spring which keeps it from collapsing completely with the air pressure. When the pressure around it increases, the sides of the box collapse slightly, and the small movement is magnified by levers and made to move a pointer around a dial. When the pressure decreases, the spring pulls out the sides of the box and the pointer moves in the opposite direction. An aneroid barometer must be calibrated against a mercury barometer before it can be used to give readings of pressure, rather than just an indication that the pressure is increasing or decreasing.

An aneroid barometer has the major advantage that it does not contain mercury which is toxic and can spill (the word aneroid comes from the Greek meaning ‘without liquid’). It can also be made small in size, unlike the mercury barometer which must be at least 80 cm high. It does not have to be held upright, but will function in any orientation. However, it is not as accurate as a mercury barometer.

The aneroid barometer is attributed to **Lucien Vidie**, who patented it in 1844, but it is said that others had constructed similar devices earlier.

Aneroid barometers are used in barographs, in which the movement of the evacuated box is made to move a pen which traces out the changes of pressure on a piece of paper wrapped around a drum which is rotated by clockwork, so that the changes in pressure are related to the time at which they occur. A barograph thus produces a permanent record of the changes in pressure over a period of a day or a week.

4.5 Applications

Density - Separation of Mixtures

Differences in density are commonly used to separate the components of mixtures.

In creameries, the fact that cream is less dense than milk is used in separators to separate the cream from the milk, so that the cream can be used to make butter.

In fuel lines of engines and also of central heating systems, water traps are used to separate any water which may have become mixed with the fuel. A water trap consists of a small covered bowl through which the fuel passes. Any water falls to the bottom of the bowl and remains there until it is emptied by removing the bottom of the bowl.

Convection Currents

Convection currents in liquids and gases depend on the difference in density between hot and cold fluids. The hot fluid, being less dense, rises, while the colder fluid, being more dense, moves down to take its place. Thus heating elements (e.g. in electric kettles) are always placed low down in the appliance to allow the convection currents to move upwards. In contrast, the cooling elements in a refrigerator are usually at the top, so that the cold air flows downwards, and warm air rises to take its place and be cooled in its turn.

Diving

Watertight diving suits were developed by **Augustus Siebe**, who completed his first fully successful model in 1830. Divers must breathe compressed air when they go diving; a person depends on the pressure of the air to push air into his lungs, and if the pressure of the water around him is large, the pressure of the air he breathes must be equally large, or his lungs will not fill when he inhales. When a diver is working in a diving suit, the air is supplied by an air pump on the surface, and is carried down by a hose. Deep-sea divers in diving suits cannot go much beyond 100 - 150 m deep. For work at greater depths than this, divers work in rigid diving apparatus, in which they can breathe air at normal pressure, as the apparatus resists the pressure of the water.

Diving bells (large inverted cylinders, open at the bottom, full of air) were developed to allow divers to work under water. These were not new, for Aristotle records a type of diving bell being used in his time. The air in them would become compressed by the pressure of the water, so that the lower part of the bell would fill with water, but the diver in the bell could breathe the air in the upper part and then dive down to work below for a short time. Halley designed smaller bells, supplied with air from the main bell, which the diver could wear on his head, to allow him more freedom to move about.

Scuba divers carry compressed air in the air bottles on their backs. They cannot go deeper than 40 - 50 m when diving, because they have no other protection against the pressure.

When divers are working at great depths, breathing compressed air, they cannot come quickly to the surface, because the sudden decrease in pressure causes the ‘bends’, or caisson disease (so called because it was first observed in men working under pressure in caissons for underwater excavation). When a diver is breathing air at high pressure because he is in deep water, the gases in the air dissolve under pressure in the blood. The oxygen which dissolves is removed by the cells, but the nitrogen remains in the blood. When the pressure is quickly released the dissolved nitrogen will come out of solution and form bubbles, in the same way as carbon dioxide comes out of solution in aerated soft drinks when the can or bottle is opened and the pressure released. The bubbles of gas which form in the bloodstream cause intense pain in the joints and sometimes paralysis, and if a bubble gets into the heart it can cause the heart to stop. To avoid this two methods are used. Either the diver can come up slowly, stopping at intervals to allow the nitrogen gas to come out of the blood, or he can go into a decompression chamber at the surface, where the pressure is raised and then gradually lowered to atmospheric.

The pressure gets very great in deep water. Each 10 m depth adds an extra atmosphere to the pressure, so that a submarine at 500 m depth experiences a pressure of about 50 times atmospheric pressure. Submarines must be able to withstand this pressure; it has been known for a submarine to be crushed by the pressure when it has dived too deeply.

For deep-sea exploration, special submersibles have been developed to stand the enormous pressure. In the 1930s, the Americans **William Beebe** and **Otis Barton** built a thick-walled spherical submersible called the bathysphere (from the Greek bathos, meaning deep), which descended to a depth of 900 m. In 1954, **Auguste Piccard** built a bathyscaphe (Greek scaphos, meaning ship) for the French Navy. An improved model, built by the Italian Navy, was bought by the US Navy and, on 23 January 1960, Jacques Piccard and Lt Don Walsh descended in it to a depth of 10 916 m on the bottom of the Marianas Trench, the deepest trench on the earth's surface. Even at this great depth, they found life.

Hydraulics

Hydraulic systems depend on the fact that liquids are practically incompressible; when the pressure on a liquid is increased, its volume changes only very slightly (in contrast to a gas, which can be compressed). When a hydraulic pump applies pressure to a liquid at one end of a tube or cylinder, the liquid applies an equal pressure at the other end, and if there is a movable piston at the end of the tube, it can move and exert a force. Since a hydraulic system works with a constant pressure, the force exerted can be increased by increasing the area of the piston on which the pressure acts (force = pressure x area). For example, if a hydraulic system is working at 7×10^5 Pa, the force exerted by the fluid on a piston of area 1 cm^2 (1×10^{-4} m^2) would be 70 N, while that on a piston of area 10 cm^2 would be 700 N. This ability of a hydraulic system to exert a force of any size by a suitably sized cylinder and piston is of great use in many types of machinery used in industry, construction, transport, agriculture, etc.

Flotation

A ship made of steel (or even of concrete) can float in water because the average density of the ship is less than that of water, due to a large part of the space inside the hull being filled with air. While the density of the steel or concrete is greater than that of water, the density of the air is so much less that the average density is less than the density of water.

The depth to which a ship will float in the water depends on the density of the water, which varies with the salinity and with the temperature. To ensure that ships are not overloaded they have a mark, called a Plimsoll line, on the side. This indicates the maximum depth to which the ship may be allowed to sink. The Plimsoll line has different marks corresponding to different situations:

TF	=	Tropical freshwater
F	=	Fresh water
T	=	Tropical seawater
S	=	Summer
W	=	Winter
WNA	=	Winter North Atlantic

The marks are in this order down the side of the ship, as the density of the water increases from tropical fresh water to the water in winter in the North Atlantic, so that a ship loaded in tropical fresh water (e.g. on the Amazon in Brazil) to the TF mark should float at the W mark when it reaches northern waters in winter. The WNA mark gives extra freeboard (height above the waterline) for greater safety in storms in the North Atlantic in winter.

The Plimsoll line is named after **Samuel Plimsoll** (1824-1898), who was responsible for the passing of legislation which requires that all ships carry the mark named after him, which prevents them being so overloaded that they are in danger of sinking.

Hydrometers

Hydrometers are used to measure the density of liquids, using the law of flotation. A hydrometer consists of a long glass tube, the lower end of which is weighted. Above the weighted end, the tube is enlarged to provide a float.

When a hydrometer is placed in a liquid, the float keeps it from sinking, while the weight keeps it in a vertical position. In a liquid of low density, the hydrometer sinks almost to the top of the tube, because the volume of liquid it displaces must be large to make its weight equal to the weight of the hydrometer, while in a liquid of higher density, it displaces a smaller volume of liquid and floats with more of the tube out of the liquid. The tube has a scale marked on it which shows the density of the liquid, the lower density figures being at the top and higher figures lower down. The scale is not linear, and the figures give (usually) densities in kg m^{-3}; to get figures in g cm^{-3} they must be divided by 1000. Hydrometers provide a convenient and quick method of measuring the density of a liquid. They have been used to measure the density of milk in dairies, since the density decreases as the percentage of fat in the milk increases (special hydrometers for this are called lactometers), and are used to measure the density of acid in a car battery, which gives a measure of how fully charged the battery is.

Coffee Machines

In coffee machines the lower chamber, containing both air and water, is heated, increasing the pressure in it. This increased pressure causes the water to move up through the filter containing the coffee and into the upper chamber. Then the lower chamber is cooled, reducing the pressure there, and allowing the atmospheric pressure to cause the water to move down through the coffee again.

Milking Machines

Milking machines use a vacuum pump which reduces the air pressure in the machine. The reduced pressure in the liner of the cup holds the liner on the cow's teat, while at the same time a pulsator varies the pressure to alternately squeeze and relax the teat and so extract the milk.

Atmospheric Pressure and Weather

As the weather changes, the atmospheric pressure changes from day to day. This is because the density of air depends on whether it is wet or dry. Dry air has a greater density than damp air, surprising as it may seem. This can be understood if one thinks of damp air as air mixed with water vapour. Since water vapour is less dense than air, a mixture of the two gases must be less dense than air alone.

When the air over a certain point on the earth's surface is dry, the weight of air over it will be greater than if the air were damp, so that the air pressure is greater when the air is dry, and less when the air is wet. High atmospheric pressure means the air is dry, so that the weather is likely to be fine, while low pressure means the air is damp, so that one can expect cloud and rain. Students must be reminded that the link is between pressure and fine/wet weather, not hot/cold weather. While high pressure in summer brings fine, dry weather which is usually hot and sunny, in winter high pressure brings cold, frosty weather, because there is no blanket of cloud to insulate the earth during the long nights. Low pressure in summer may bring colder weather, but in winter it normally brings mild, wet weather.

Weather maps show regions of high pressure (called anticyclones) and regions of low pressure (called depressions or cyclones). The actual variation in pressure in Ireland is normally not very great; low pressure is usually around 950 - 980 hPa, while high pressure is 1010 - 1030 hPa. Often it is not the absolute pressure which is important in weather forecasting, but the change in pressure and also the rate of change of pressure. Falling pressure indicates that the weather is likely to become cloudier and wetter, while rising pressure indicates the arrival of finer and drier weather. A rapid change in pressure means that the wind will become strong, and very low pressure means that the weather will be stormy.

The dampness or dryness of the air is not the only factor which affects the pressure. The temperature of the air will also affect the pressure, as heating makes air expand and so reduces its density and pressure, while cooling makes it contract, resulting in higher density and pressure. During a storm, it may be noticed that the level of the water in a well rises, even though there may have been no rainfall. This is because the very low atmospheric pressure allows the water pressure to raise the level in the well.

Balloons

Balloons can be made to float in air by filling them with a gas which is less dense than the surrounding air, so that the total density of balloon and gas is less than that of air. Gases which are used include hydrogen, helium and hot air.

One can calculate the mass which a certain volume of a gas can lift into the air, knowing the density of the gas and of air. Hydrogen has a density of 0.09 kg m^{-3} and helium a density of twice that, compared to the density of air which is 1.29 kg m^{-3}. Thus 1 m^3 of hydrogen gas in a balloon will be able to lift 1.2 kg (1.29 - 0.09), while the same volume of helium will lift 1.11 kg. It is perhaps surprising that there is so little difference between hydrogen and helium; this is because both of them are very much less dense than air. In comparison, hot air at a temperature of 273 °C would have a density half that of normal air (0.64 kg m^{-3}), so that 1 m^3 of air at this temperature would lift 0.64 kg. Of course, the mass of the balloon must be included in the mass to be lifted; it is not taken into account in these simple calculations.

Heavier-than-air flight

Aeroplanes are able to fly because the movement of the wings through the air causes a difference in pressure between the top and the bottom of the wing. This difference in pressure is caused by the shape of the wing which is such that air moving along the top of the wing has farther to travel than air travelling along the bottom of the wing. The air on top of the wing must travel faster, and the increase in speed results in a drop in pressure above the wing (the change in pressure with velocity is called the Bernoulli effect). The difference in pressure produces a force which pushes the wing, and therefore the plane to which it is attached, upwards.

A simple demonstration of the Bernoulli effect is to suspend two bodies (e.g. two apples) from two strings so that they are close together. If one blows between the two bodies, they move closer together. The air moving between them increases in velocity, because it is travelling around the curve of the body, or because a large volume of air has to make its way in between the narrow gap between the bodies. The increase in air velocity causes a reduction in air pressure between the bodies, and the greater air pressure on the other side results in a force pushing the bodies together.

The difference in pressure above and below an aircraft wing is affected by the angle at which the wing is tilted to the airflow - the ‘angle of attack’ of the wing. The greater the angle of attack the greater the lift produced. However, if the angle of attack is increased too much, the airflow across the top of the wing changes from being smooth to being turbulent, and the wing loses its lift; the aircraft is said to stall.

Aeroplanes must be travelling at quite high speed before the wing produces sufficient lift to take it into the air; this is why planes need long runways (up to 4 km long) at airports. On helicopters, the wings are formed into a rotor which produces lift as it rotates, so that it can take off and land vertically. Helicopters have the disadvantage that they need a more powerful engine than a conventional plane of the same size, and so are much more expensive to run. They are also more expensive to buy, since the control system of a helicopter is much more complicated than that of an aeroplane.

Control of a helicopter is difficult. To alter the total lifting force and thus make the helicopter climb or descend, the angle of attack of all the rotor blades is changed at the same time. To make the helicopter fly forwards, the angle of attack of the rotor blades on the front of the rotor is decreased, and on the rear increased. This increases the lifting force at the rear, tilting the helicopter forward, and providing a component of the lifting force in the forward direction which makes the helicopter accelerate in this direction. When a helicopter is moving forward the velocity of the air around the advancing blade is larger than the velocity of the air around the retreating blade, so that the lift produced by the two sides differs. This can be counteracted by increasing the angle of attack of the blades on the retreating side and decreasing it on the advancing side, but this cannot be continued indefinitely, as too great an angle of attack on the retreating side will cause the blade to stall. This limits the maximum speed of a helicopter to about 300 km h^{-1}. The necessity to keep changing the angle of the blades as the rotor rotates makes the control system of the helicopter much more costly than that of an aeroplane and also makes a helicopter more difficult to fly.

By Newton’s third law, the forces exerted by the engine on the rotor produces equal and opposite reaction forces on the engine and therefore the helicopter, which would make the helicopter turn in the opposite direction to the rotor. These forces are counteracted by the tail rotor found on single-rotor helicopters. In twin-rotor helicopters the two rotors rotate in opposite directions.

4.6 Worked Examples

1. A U-tube contains oil of density 700 kg m^{-3}. If the difference between the levels on the two sides is 5 cm, calculate the difference in pressure.

$$\begin{aligned} p &= \rho g h \\ &= 700 \times 9.8 \times 0.05 \\ &= 343 \text{ Pa} \end{aligned}$$

2. A tank contains water to a depth of 3 m. Calculate the pressure at the bottom. (Density of water = 1000 kg m^{-3}.)

$$\begin{aligned} p &= \rho g h \\ &= 1000 \times 9.8 \times 3 \\ &= 2.9 \times 10^4 \text{ Pa} \end{aligned}$$

3. A fuel tank is used to store oil with a density of 700 kg m^{-3}. It is fitted with a circular trapdoor with a diameter of 1.5 m on one of its vertical sides. Find the force acting on the trapdoor when the level of the oil is 1.6 m above the top of the door.

$$\begin{aligned} \text{Depth of centre of door} &= 1.6 + 0.75 \\ &= 2.35 \text{ m} \end{aligned}$$

$$\begin{aligned} \text{Pressure at centre of door} &= 700 \times 9.8 \times 2.35 \\ &= 16121 \text{ Pa} \end{aligned}$$

$$\begin{aligned} \text{Area of door} &= \pi r^2 \\ &= 1.767 \text{ m}^2 \end{aligned}$$

$$\begin{aligned} \text{Force} &= pA \\ &= 16121 \times 1.767 \\ &= 2.8 \times 10^4 \text{ N} \end{aligned}$$

Boyle's law

The statement that pressure is inversely proportional to volume can be expressed mathematically as $pV =$ constant. This leads to the equation $p_1V_1 = p_2V_2$ for the situation where the pressure of a fixed mass of gas changes from p_1 to p_2, while the temperature remains constant.

4. The gas in a cylinder has a volume of 210 cm^3 when its pressure is 1.6 x 10^5 Pa. What would the volume be if the pressure changed to 1.2 x 10^5 Pa, while the temperature remained the same?

$$p_1V_1 = p_2V_2$$

$$1.6 \times 10^5 \times 210 = 1.2 \times 10^5 \times V$$

$$V = \frac{1.6 \times 10^5 \times 210}{1.2 \times 10^5}$$

$$= 280 \text{ cm}^3$$

5. A car tyre contains 2.5 litres of air at a pressure of 220 kPa. What volume would the air have at the same temperature when released from the tyre, if the air pressure outside were 100 kPa?

$$p_1V_1 = p_2V_2$$

$$220 \times 2.5 = 100 \times V$$

$$V = \frac{220 \times 2.5}{100}$$

$$= 5.5 \text{ litres}$$

6. In a Boyle's law apparatus, the volume of the air in the tube was 36 cm^3 when the pressure gauge read 1.5 x 10^5 Pa. Calculate the reading of the gauge when the volume of the air is 40 cm^3.

$$p_1V_1 = p_2V_2$$

$$1.5 \times 10^5 \times 36 = p \times 40$$

$$p = \frac{1.5 \times 10^5 \times 36}{40}$$

$$= 1.35 \times 10^5 \text{ Pa}$$

Note that in Boyle's law calculations, any units may be used for pressure and for volume. There is no need to convert pressure to pascals or volumes to metres^3; as long as the two pressures (or two volumes) given are in the same units, the unknown volume (or pressure), when calculated, will be in the same units as the known volume (or pressure).

4.7 Student Experiment

To Verify Boyle's Law

Boyle's law apparatus consists of a heavy-gauge glass tube partially filled with air, connected to a container of oil fitted with a pressure gauge. The pressure in the apparatus is changed by connecting a pump to it and pumping in air. This forces oil into the glass tube and increases the pressure of the air in the tube. The volume of air in the tube, and the pressure shown on the gauge, are read and recorded for several different pressures.

When carrying out the experiment using this apparatus, one should stress that the volume being measured in the tube is not the volume of oil, but the volume of the air enclosed by the oil - it is a common error for students to write in an account of the experiment 'the level of oil in the tube was read'. Attention should be drawn to the fact that the scale of volume behind the tube measures from the top down and not from the bottom up.

If one is following the usual experimental procedure and raising the pressure to the maximum value at the start, and then releasing the pressure gradually, stopping at intervals, it is necessary to take the following precautions. Firstly, time must be allowed for the temperature of the air in the tube to return to room temperature after each change in pressure (raising the pressure quickly increases the temperature, reducing the pressure lowers the temperature). Secondly, after the pressure has been reduced each time, the oil is allowed time to drain down from the tube before the reading of the volume is taken. If one wishes to obtain readings for pressures lower than atmospheric, it is possible to suck air out of the apparatus to get lower pressures.

CHAPTER 5

VECTORS AND SCALARS

5.1 Background

The origin of the topic was the formulation of the parallelogram law in 1586 by the Flemish mathematician, **Simon Stevin** (1548-1620). The development of vector mathematics by the American physicist, **Josiah Willard Gibbs** (1839-1903) and the English physicist, **Oliver Heaviside** (1850-1925), in the last quarter of the 19th century drew on the algebra of quaternions which had been formulated by the Irish mathematician **Sir William Rowan Hamilton** (1805-1865). Still further development of vectors led to the mathematics used by **Albert Einstein** (1879-1955) in his general theory of relativity.

5.2 Conceptual Approach

A vector quantity has direction and magnitude. For example, if a person travels a distance of 5 km from the school in a straight line, (s)he can end up anywhere on the circumference of a 5 km radius circle. To identify the actual finishing point, the displacement is specified, i.e. the direction in which the person travelled is given as well as the distance; displacement is an example of a vector quantity. Likewise, when a person applies a force to a body, either a push or a pull, it is impossible to do so without the push or pull being in a particular direction; force is also a vector quantity. Forces which have the same magnitude but are in different directions have different effects.

In contrast, a scalar quantity is completely defined by a magnitude without a direction - but we need to take care that students get the concept correct at the start. Using mass as the first example of a scalar quantity could lead to an incorrect idea of what a scalar is, for mass may still be confused with weight, which is a vector quantity, by some students. So, they could imagine that the direction of a scalar is always the same, rather than that it has no direction. Possibly the best scalar to use as a first example is temperature. When one has stated the magnitude of the temperature of a substance there is no direction which must be stated as well to completely define the quantity.

Representation of Vector Quantities

To represent a vector quantity diagrammatically, a line is drawn with length proportional to the magnitude of the vector quantity and with an arrow pointing in the same direction as the vector quantity. This allows geometry to be used to calculate the magnitude of unknown vector quantities, since the lengths of all lines in a vector diagram are proportional to the magnitudes of the vectors. In print, symbols for vector quantities may be given in ***bold italics*** to distinguish them from symbols for scalar quantities which are given in *italics*. In handwriting, a vector may be written with an arrow, e.g. $\vec{a}$, while its magnitude is written as $|\vec{a}|$.

Composition of Coplanar Forces

To add two vectors one can use the parallelogram law. To illustrate that the parallelogram law actually does give the sum of two vectors, we can imagine a helicopter hovering over a ship moving underneath, Fig. 5.1. A person walks across the deck of the ship from one side to the other. In a diagram the motion of the ship S is represented by the vector $\vec{ab}$, the motion of the person relative to the ship by the vector $\vec{ad}$, and the resultant motion of the person relative to the helicopter is the sum of these two vectors. The diagram shows that the resultant is actually the diagonal of the parallelogram, $\vec{ac}$.

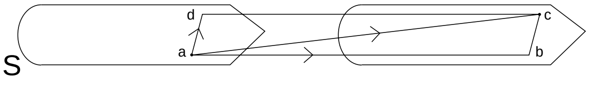


Fig. 5.1

Calculation of Resultants

When the two vectors are at right angles the magnitude of the resultant may be determined using Pythagoras' theorem. Having found the magnitude, the direction may be found by simple trigonometry. When the vectors are not at right angles the magnitude is found using the cosine rule. The direction is found by a re-application of the cosine rule or by application of the sine rule.

It may be pointed out to students that Pythagoras' theorem is a special case of the cosine rule. The cosine rule gives $a^2 = b^2 + c^2 - 2bc \cos A$, where A is the angle between b and c. If A is a right angle, then $\cos A = 0$ and the formula becomes $a^2 = b^2 + c^2$

It is also possible to find the resultant in a particular case by scale drawing instead of by calculation. This method is more tedious and generally less accurate.

It is also possible to use unit vectors in two perpendicular directions (***i*** and ***j*** vectors) to calculate the resultant. Students who do Applied Maths may be familiar with this method.

Resolution of Vectors

This topic may be introduced as the converse of addition of vectors.

Students can be reminded that it can be useful, sometimes, in a mathematical calculation, to replace the figure 8 by 5 + 3, and this does not change the value, since 8 = 5 + 3. In a similar way, a vector $\boldsymbol{a}$ can be replaced by two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ without altering the value if $\boldsymbol{a} = \boldsymbol{x} + \boldsymbol{y}$ by the parallelogram law (which is the method of adding vectors).

This means that resolution consists of constructing a rectangle, given the diagonal. A diagram will show that the two components must be $a \cos \theta$ and $a \sin \theta$, where $\boldsymbol{a}$ is the original vector and θ is the angle between the vector and some chosen direction. Perhaps the best example of this is to consider a force acting upwards at an angle on a body resting on a horizontal surface. When the force is resolved horizontally and vertically, the horizontal component is responsible for the acceleration of the body along the surface while the vertical component reduces the normal reaction.

Multiplication of Vectors

In mathematics, there are two different methods for multiplying vectors. One is the scalar or dot product of two vectors. The dot product is given by

$$\boldsymbol{a} \bullet \boldsymbol{b} = ab \cos \theta$$

where θ is the angle between the two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. The dot product of two vector quantities is a scalar quantity. For example, work is the dot product of force and displacement, i.e.

$$\begin{aligned} W &= \boldsymbol{F} \bullet \boldsymbol{s} \\ &= Fs \cos \theta \end{aligned}$$

where θ is the angle between the force and the displacement.

The second method for multiplying two vectors is the vector, or cross, product. The cross product is given by

$$\boldsymbol{a} \times \boldsymbol{b} = ab \sin \theta$$

where θ is the angle between the two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. The cross product of two vector quantities is a vector quantity.

For example, moment is the cross product of force and displacement, i.e.

$$\boldsymbol{M} = \boldsymbol{F} \times \boldsymbol{s}$$

$$= Fs \sin \theta$$

where θ is the angle between the force and the displacement. (Strictly speaking, moment is a pseudovector but the distinction need not concern us here.) Another example of the cross product is the equation for the force on a charge q moving with velocity $\boldsymbol{v}$ in a magnetic field of flux density $\boldsymbol{B}$. The force is given by

$$\boldsymbol{F} = q(\boldsymbol{v} \times \boldsymbol{B})$$

The magnitude of the force is given by

$$F = qvB \sin \theta$$

Where θ is the angle between the velocity and the magnetic flux density.

5.3 Worked Example

A body of mass 20 kg rests on a slope at an angle of 10° to the horizontal. What force is (a) tending to make the body slide down the plane, (b) pushing the body into the surface of the plane?

Weight of mass $= mg$

$= 20 \times 9.8$

$= 196$ N

Force down plane $= W \sin \theta$

$= 196 \times \sin 10^\circ$

$= 34$ N

Force into plane $= W \cos \theta$

$= 196 \times \cos 10^\circ$

$= 193$ N

5.4 Student Experiment

To Verify the Parallelogram Law

This experiment is carried out by applying three forces to a point and adjusting them until they are in equilibrium. When they are in equilibrium, their sum (as vectors) must be zero. The sum of any two of the vectors must then be equal and opposite to the third, to make the total sum of all three zero.

We use the parallelogram law to find the resultant of two of the forces in the experiment, and show that this resultant is equal in magnitude and opposite in direction to the third force. This proves that the resultant found by the parallelogram law is the sum of the two vectors - especially when the experiment is repeated several times and a similar result is found each time.

There are two ways of setting up the three forces required:

1. using three masses, strings and two pulleys - this is found in most text-books

2. using one mass, strings and two spring balances, as shown in the diagram, Fig. 5.2.

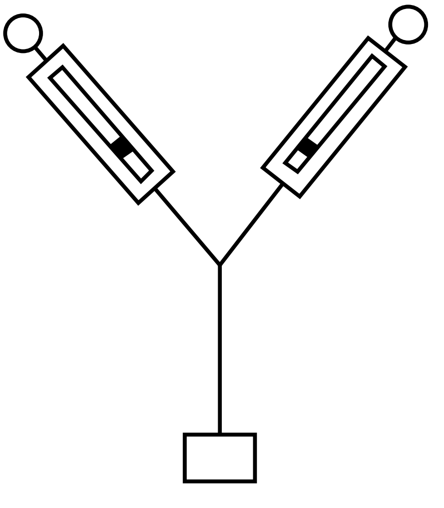


Fig 5.2

Using this method, it is essential that the spring balances are clamped so that they are in line with the strings, otherwise the pull at an angle will cause friction in the balances and give an incorrect reading. It is preferable to clamp on the bodies of the balances instead of using the handles at the top to ensure that the directions of the balances are correct.

In both methods, the apparatus should be disturbed by lifting up the central mass slightly and then releasing it. If the apparatus is free from friction, it should return to the same position as before (within the limits of experimental error). If it does not, but tends to stick in a different position, the apparatus must be adjusted to reduce friction where it occurs.

The simplest way of recording the direction of the strings is by using a force-board but, if this is not available, it can be done if one person holds a sheet of paper backed by a board (a hardbacked science notebook will do) behind the strings, while a second person marks the directions on the sheet of paper.

In the calculations, it is essential that the masses are converted to weights in newtons, for the law applies to forces (weights) which are vector quantities, but not to masses which are scalar quantities.

There are two parts to the calculation. First, the magnitude of the resultant of the two upward vectors is found, either by scale drawing or by calculation; as stated above, this should be equal to the magnitude of the third (downward) force, within the limits of experimental error (5% error is a reasonable accuracy for this experiment). Secondly, the angle which the resultant makes with one of the these vectors is found and, from this, the angle between the resultant and the third vector; this should be 180° (it should be possible to get it to ±10°).

MODULE 5

Heat and Temperature

CONTENTS

CHAPTER 1 TEMPERATURE

1.1 Introduction

Temperature is an indication of the hotness of something. We can tell by touching an object whether it is hot or cold. However, this is a subjective judgement, and is not very reliable. A simple experiment will demonstrate this. Take three basins of water, each holding the same quantity of liquid, but one containing hot water, a second containing tepid water, and the third one cold water. Place one hand in the hot water, and the other in the cold water. Leave them for a few seconds. Now, put both hands together into the basin of tepid water, and notice that both hands give different messages. The hand that came from the hot water now feels cold, whereas the hand taken out of the cold water feels warm. This experiment was first carried out over 300 years ago by an Englishman, **John Locke**.

We cannot rely on our skin as an accurate way of measuring temperature. The physiological sensations registered by receptors in our skin depend largely on their immediate past experience, as John Locke's experiment illustrates. So, we need thermometers with suitably defined scales for the reliable measurement of temperature.

1.2 Do You Know

Where was the coldest place on earth?

The lowest temperature recorded on earth was in Antarctica; the temperature was -89.2 °C. The highest temperature, 58 °C, was recorded in Libya.

The temperature limits for the existence of life.

Life on earth, in the form of bacteria and moulds, can exist between -260 °C and 170 °C, but mammals can only exist between -65 °C and 70 °C.

Normal body temperature is 37 °C.

Without oxygen, at this temperature, the brain dies after a few minutes, but if it is cooled down to 15 °C it can survive for 45 minutes; it can survive even longer at lower temperatures. During brain surgery, when blood is passed through a heart/lung machine, the body can be cooled sufficiently and then warmed up again.

Does the body temperature of animals which hibernate in winter drop?

Animals such as the squirrel which seek to escape the rigours of winter by hibernating have their body temperature lowered from 37 °C to c. 15 °C. They spend the coldest months of the year in this inactive state, curled up in a small hole underneath the ground, or in the hollow of a tree. Their breathing rate drops from 180 breaths or so per minute to c. 4 and the pulse falls from 200 per minute to 4 or 5. Hibernation is a state which may be described as being only slightly removed from death, as hibernating animals live on stores of body fat until spring comes.

Some interesting temperature values.

	Temp/°C
Surface of sun	6000
Oxyacetylene flame	4000
Electric arc	3500
Tungsten melts	3380
Filament of electric lamp	2500
Iron and steel melt	1500
White heat (metals)	1300
Gold melts	1063
Cherry red heat (metals)	900
Dull red heat (metals)	700
Bright coal fire	600
Lead melts	327
Sugar melts	160
Butter melts	31
Mercury freezes	-39
CO_2 snow changes to gas	-78
Oxygen boils	-183
Nitrogen boils	-196
Hydrogen boils	-253
Helium boils	-269
Absolute zero	-273

Table 1.1

Temperatures on different scales can have the same value.

The Celsius and Fahrenheit scales have exactly the same temperature at -40 °, i.e. -40 °F = -40 °C. The Celsius temperature that has the same numerical value as the Kelvin temperature is -136.5, though of course the signs are opposite, i.e. -136.5 °C = 136.5 K.

What is the difference between evaporating and boiling?

Evaporation is the escape of molecules in the form of vapour from the exposed surfaces of solids or liquids at all temperatures. When a liquid boils, bubbles of vapour are formed throughout the liquid. At any given pressure, boiling takes place only at a particular temperature.

Can a solid evaporate[1]?

The metal tungsten evaporates and the effect may be noticed in an electric lamp in which the filament is made of tungsten. Blackening takes place on the glass and is due to evaporation of the metal. The rate at which this takes place is negligible unless temperatures are close to the melting point of the metal. A lamp from a bicycle dynamo set sometimes shows this effect well.

What is the distinction between a gas and a vapour?

Both are matter in the gaseous state. However, we regard oxygen and nitrogen as gases, but steam as a vapour. This classification depends on what is called the ‘critical temperature’ of the substance in question. Every gas has a characteristic temperature above which it cannot be liquefied, no matter how great the pressure which is applied. This is called the critical temperature. So, a substance in the gaseous state below its critical temperature is referred to as a vapour, whereas above it, it is called a gas.

Examples of critical temperatures are given in Table 1.2.

	Temp/°C
Water	374
Ammonia	132
Carbon dioxide	31
Oxygen	-118
Nitrogen	-147

Table 1.2 *Selected critical temperatures*

Around room temperatures, then, steam is a vapour; oxygen and nitrogen are gases, whereas carbon dioxide can be either. To liquefy a gas, e.g. oxygen, it must first be cooled to -118 °C, or below, and then compressed.

[1]This is evaporation not sublimation. Some substances, e.g. iodine, change directly from a solid state to a vapour on heating. This process of converting a solid directly into a vapour on heating without any intermediate liquid stage is called sublimation.

What is the Joule-Kelvin effect?

If a gas is forced to move through a narrow opening, it expands, and its temperature changes. For most real gases at normal temperatures and pressures there is a cooling effect. However, some gases experience a heating effect as they flow through a narrow opening at normal temperatures and pressures. Examples of gases in this latter category are hydrogen and helium.

What happens if an object, with a hole in it, is heated?

If the object expands with increasing temperature the diameter of the hole also increases. In fact the hole expands at the same rate as if it were made of the same material as the object.

What is the temperature of an average Bunsen burner flame?

The gas used is generally propane/butane mixed with air which can give a flame temperature of up to about 900 °C. The inside, or primary, flame which is a pale turquoise in colour is surrounded by a colourless/pale lilac secondary cone (assuming sufficient air enters to give complete oxidation of the gas). The hottest part is at the tip of the primary cone, and the optimum gas/air mixture is 1:3. **Robert Wilhelm Bunsen** (1811-1899) was professor of chemistry at Heidelberg University, and is said to have designed the burner named after him in 1855.

What is the temperature of the filament of an electric light bulb, an oxyacetylene flame and an electric arc (carbon)?

The element tungsten (in olden times called wolfram, hence its symbol W) is heated electrically in a bulb which is usually filled with the inert gas argon. The temperature is 2600 °C (melting point of tungsten is 3380 °C).

When oxygen is combined with acetylene (ethyne) gas and ignited, a temperature of c. 3000 °C may be reached, depending on the composition of the mixture. It is used in welding and to cut through metal. A nitrous oxide/acetylene mixture gives a flame temperature of c. 2900 °C and an air/acetylene mixture gives a temperature of c. 2300 °C.

If a current of c. 10 A (at a voltage of c. 50 V) is passed through two carbon rods, in contact, an arc is struck between the rods as they are pulled apart. The current flow is not interrupted and an intense white light appears. Some ultraviolet radiation is also emitted. The temperature of the positive rod can reach between 3500 °C and 4000 °C. At one time lamps of this type were used for street lighting. **Sir William Siemens** (1823-1883) introduced this idea in Paris in 1879.

What is thermal paper?

Thermal paper is temperature-sensitive paper which changes colour at particular temperatures. Some temperature-sensitive paper is impregnated with cobalt chloride and is a pale pink/white colour which turns blue at 40 °C. Temperature-sensitive paper is also used as a moisture indicator, in which case it turns from blue to pink.

Very high temperatures are produced in furnaces.

In blast furnaces for the smelting of iron ore (in a reduction process using iron in the form of iron oxide, Fe_2O_3) preheated air, with minute oil drops, methane gas or pulverised carbon compounds (e.g. coke) are combined. This process can give temperatures up to 1500 °C. Carbon monoxide is initially formed and later is oxidised to the dioxide when hot air is blown through it. Limestone is usually added to help purify the metal.

In an induction furnace, even higher temperatures may be obtained. This type is used for smaller volumes of metal. A crucible is placed inside a coil of water-cooled copper tubing carrying a.c. at between 200 and 10 000 hertz. The intense alternating magnetic field set up inside the copper coil induces eddy currents in the metal which melt it. Most electrical furnaces are thermostatically set and controlled.

Although there is a lower limit to temperature values (absolute zero) there does not appear to be an upper limit.

In recent years higher and higher temperatures have been obtained with the use of lasers. As energy is concentrated in an extremely small volume, the temperature soars - values of up to 10^4 K may be reached. Nuclear reactions can give rise to much higher temperatures. The temperature at the centre of a hydrogen bomb explosion is c. 1 x 10^8 K. The interior of the sun is estimated to be at a temperature of c. 1.5 x 10^7 K.

1.3 Conceptual Approach

Temperature may be defined as the level or degree of hotness of something. It measures the intensity of hotness, but not the quantity of energy. It is a number on a chosen scale which expresses quantitatively whether an object is hot or cold. **Lord Kelvin** once said *'When you can measure what you are speaking about and express it in numbers, you know something about it.'*

Temperature is a measure of the average kinetic energy of the atoms or molecules in a system. Also, it indicates the direction in which energy will travel, when two objects are placed in contact. Energy flows from an object of higher temperature to an object of lower temperature, regardless of the relative amounts of energy contained in the objects involved.

If students sometimes find it difficult to distinguish between the words heat and temperature, it may help to give a few concrete examples, along the following lines.

a) If a red-hot poker is plunged into a bath of tepid water, energy passes from the poker to the water because the former is at a higher temperature, but the latter will contain more internal energy owing to its larger mass. Temperature then, not internal energy, is the decisive factor in determining heat flow.

b) Sparkler fireworks (fragments of white-hot metal) are at a much higher temperature, but have less internal energy than an iceberg. It is possible to have a lot of internal energy in a cool body and not so much in a very hot one.

c) If 1000 J of energy is supplied to a bath of water, the energy is shared between so many molecules that the average kinetic energy of the molecules will hardly increase at all, but if the same 1000 J is given to an iron nail it may become red-hot.

d) A very simple practical example could involve pouring hot water (not quite boiling) into two hot-water bottles; one until it is three-quarters full, and the other until it contains only a cupful or so. Both bottles will have water at approximately the same temperature, but the former will be much more effective in warming a bed.

e) When a small test-tube and a larger boiling tube, each full of water at room temperature and containing a thermometer, are placed in a water bath placed over a Bunsen burner and heated for the same length of time, they will show different temperature increases. The smaller one will be hotter and show a higher temperature on the thermometer.

f) A good analogy to use when teaching the heat and temperature section is that of water quantity and height. If a litre of water is placed in an inverted bell jar, Fig. 1.1, and connected by means of a rubber/glass tube to a burette containing 50 cm^3 of water, water will flow from the burette (higher level, but smaller quantity) to the bell jar (lower level, but larger quantity). When both levels are equal, the flow ceases. This idea is quite similar to that which happens with temperature (represented by water level), internal energy (quantity of water) and heating (movement of water).

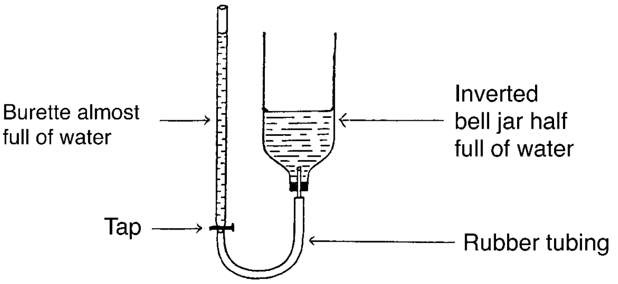


Fig. 1.1 *Temperature being compared with water level*

1.4 Experimental Approach

Note

Electric hot plates and electric kettles can be very helpful in this section.

Experiment 1.1
To Show the Distinction Between Internal Energy and Temperature

(a) Apparatus

Two beakers (1 small, the other twice or three times the size of the first); two thermometers; kettle full of boiling water.

Method

1. Warm both beakers initially, then half-fill each with the boiling water from the kettle. Record the temperature of each. Leave to stand for 15-20 minutes and then take temperature readings again.

2. The small beaker cools much more quickly than the larger one does. They both were at approximately the same temperature initially, but clearly contained different amounts of internal energy.

(b) Apparatus

Two beakers; hot water at 60 °C; two identical Bunsen burners; wire gauzes; two tripods; thermometers.

Method

Pour 200 cm^3 of water at 60 °C into one beaker, and 400 cm^3 into the other one. It is required to heat both of them to boiling point. Place them over the two Bunsen burners and record the time taken for this to happen. Though the temperature change is the same in both cases, different quantities of energy are required, as indicated by the different times taken.

Experiment 1.2
To Show Expansion of Metal on Heating

(a) Apparatus

Two retort stands; length of wire, about 1 m long; metre stick; a 200 g mass.

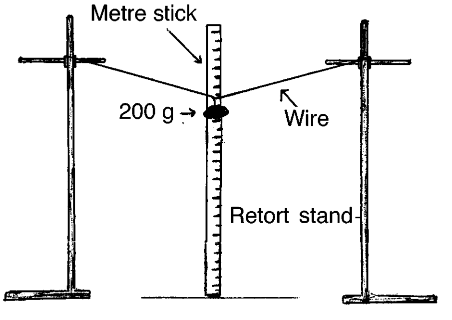


Fig. 1.2 *Expansion of metal on heating*

Method

Attach each end of the wire firmly to the two retort stands so that it is taut, Fig. 1.2. Tie the 200 g mass from the centre and measure its distance from the bench using a ruler.

Heat the wire using a Bunsen burner all along its length. Remove the Bunsen burner and immediately measure the distance of the mass from the bench. It can be seen that this distance is shorter, showing that expansion of the wire has taken place.

Note

This experiment illustrates why electricity and telephone wires sag more in summer.

Experiment 1.3
Simple Experiment to Show Water Expands on Freezing

Apparatus
Small glass bottle with screw cap; water; plastic box; freezer. The box must be sufficiently strong to contain any shards of glass which may be produced when the bottle breaks.

Method
Fill the bottle completely with water. Screw on the cap, seal in the plastic box and place in the freezer overnight.

Observation
The water, as it freezes and expands, will shatter the bottle.

Conclusion
Water expands on freezing.

Experiment 1.4
Simple Experiment to Show Expansion of Gases on Heating

Apparatus
Bottle; balloon with rim cut off; basin of cold water.

Method
1. Fill the bottle with hot water. Allow a few minutes for the glass to warm through and pour out.

2. Stretch the balloon over the neck of the heated bottle. The air in the balloon should expand from the heating effect of the hot bottle.

3. Now place the bottle with balloon in the basin, half-filled with cold water, and hold it in position.

4. As the air in the balloon now contracts, the balloon itself is pushed into the bottle by the pressure of the air outside.

5. If the bottle is now placed near a lighted Bunsen burner, the balloon will be seen to expand again.

Note
An alternative method involves inverting an empty bottle or round bottomed flask and immersing the neck of it under water. When heat is applied, e.g. using a hair dryer, the air expands and bubbles up through the water.

Experiment 1.5
To Show How a Geyser Works

Caution. Teacher demonstration only. Use protective screen.

Apparatus
The apparatus shown in Fig. 1.3 can be used to demonstrate how superheated water, accompanied by steam, shoots periodically into the air, when it is heated. The container used is a metal vessel, iron or tin, and might possibly be made by students as a project in a metalwork class. The experimental geyser, with its conical shape, could be anything from 0.5 m up to 1 or 2 m high, standing on four legs and having a space underneath to fit a Bunsen burner. The top should open out into a basin effect, as shown.

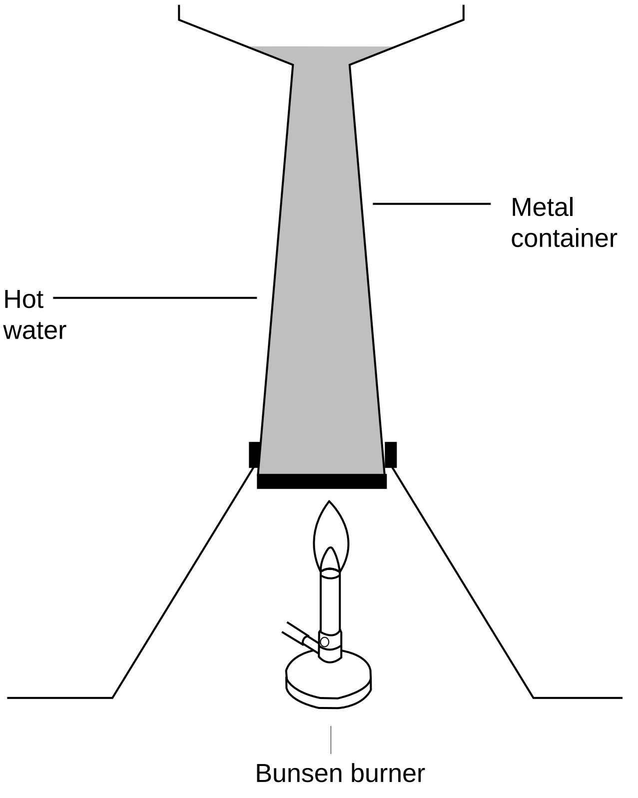


Fig. 1.3 *Apparatus to demonstrate geyser action*

Method

The container is filled completely with water, and a lighted Bunsen is placed underneath. Water immediately above the burner will boil at a higher temperature due to increased pressure caused by the water column above it, giving an eruption from time to time.

Experiment 1.6

Simple Experiment to Show that Thermal Stresses are Set up When Glass is Heated and Suddenly Cooled

Caution. Teacher demonstration only. Wear safety glasses and use protective screen.

Heat a piece of glass tubing with a Bunsen until it begins to bend under its own weight and then quickly plunge it into a beaker of cold water. The sudden contraction that follows the cooling produces stresses, and the glass shatters.

Experiment 1.7

To Show the Effect of Changing Pressure on the Boiling Point of Water

(a) Increased Pressure

Caution. Teacher demonstration only. Wear safety glasses and use protective screen.

Apparatus

A round-bottomed flask (this is stronger than a flat-bottomed one); stopper with two holes - one for a thermometer, the other for glass piping with attached rubber tube and clip; wire gauze; tripod; Bunsen burner; retort stand.

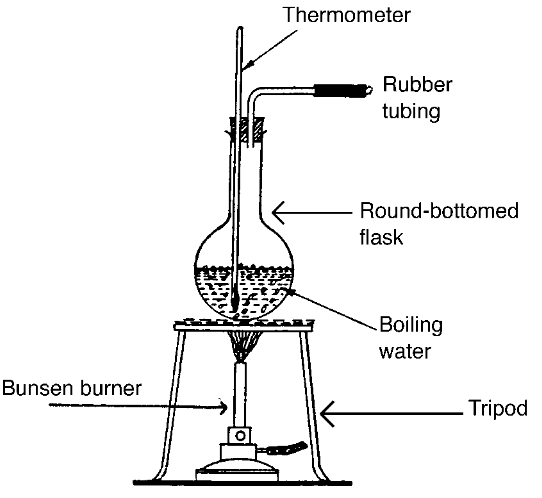


Fig. 1.4 *Demonstrating the effect of increased pressure on the boiling point of water*

Method

1. Half-fill the flask with pure water and bring to boiling point, having the clip on the rubber tube open. The temperature should be 100 °C (at atmospheric pressure), so any change in the pressure would change the boiling point.

2. Close the clip, and continue heating for a short period. (**NB**. This should be no more than 10-20 seconds, just long enough for the thermometer to register a higher temperature owing to increased pressure.) It can be seen that increasing the pressure raises the boiling point, maybe up to 105 °C, depending on the pressure generated.

Note

Adding anti-bumping granules to the water helps the water to boil more smoothly.

(b) Reduced Pressure

Apparatus

As for (a), but a vacuum pump may also be used.

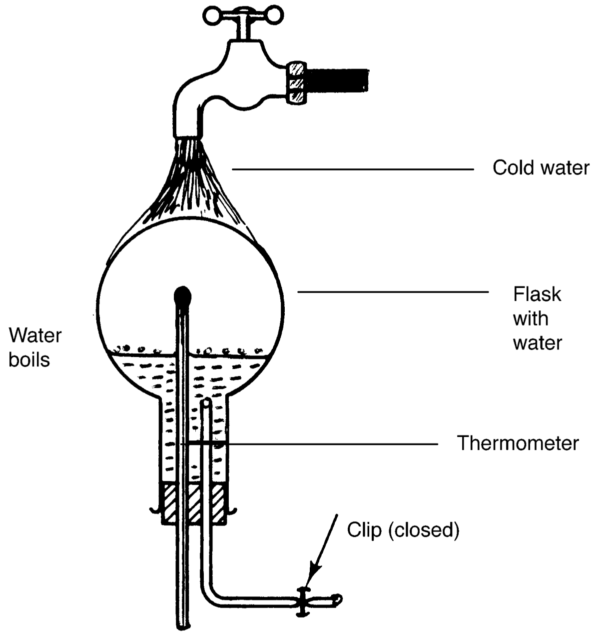


Fig. 1.5 *Demonstrating the effect of reduced pressure on the boiling point of water*

Method

1. Boil the water as before, with the clip open, allowing steam to drive out most of the air. Now close the clip, and remove the Bunsen burner.

2. Invert the flask, remove it from the retort stand and pour cold water over it. This condenses the remaining steam, giving a partial vacuum and thereby reducing the pressure.

3. Observe the reading on the thermometer. It will be less than 100 °C even though the liquid will still be vigorously boiling.

Note

At the top of Mt Everest, where the atmospheric pressure is only 26 cm Hg, water boils at 70 °C.

A vacuum pump may be used to produce a partial vacuum. If a mechanical pump is used the water will cause rusting, resulting in damage to the pump, unless some sort of vapour absorber is fitted.

(c) To Show That Dissolved Substances May Increase the Boiling Point of Water

Boil water in the flask having added salt (NaCl). The boiling point will be greater than 100 °C.

Note

A sensitive thermometer (reading to 0.2 °C or better) should be used.

1.5 Applications

Expansion in Everyday Life

The bimetallic strip

This is a strip of two dissimilar metals welded together along the entire length, which bends on heating, because one metal expands more than the other for a given rise in temperature. Common examples are copper/iron and brass/iron. A bimetal strip is the basis of some thermostats, and functions as a 'temperature switch' to make and break electrical contact. Because the strip bends or curves on heating and becomes straight again as it cools, an approximately steady temperature can be maintained. Some examples where it is used are a central heating system, electric cooker, blanket, toaster, etc. In the case of an electric iron, for instance, the circular knob on the top controls the gap width between contacts, and this regulates the heating level which is desirable for different fabrics. Bimetal strips are used also in flashing car direction indicators (a coil of wire connected to the car battery, and wound around the bimetal strip provides the heating), and in sets of flashing fairy lights used to decorate Christmas trees. Some thermometers also are based on the bimetal strip.

Gaps in railway tracks

Railway tracks some years ago were laid in separate sections, with gaps left between them. These spaces allowed for expansion effects due to hot weather, and prevented the buckling which could result in their absence. It was because of these gaps that trains made the 'clickety clack' sound often associated with train travel.

More recently, continuously welded track, fixed to concrete sleepers, is used where possible. As it is laid the track is stretched, so that in cold weather it is in a state of tension and in warm weather it is in a state of compression.

Metal bridges

Metal bridges have one end fixed to supports, and the other is set on rollers so that a certain amount of free movement is possible in hot weather. It is estimated that the Golden Gate Bridge in San Francisco is 20 cm longer in summer than in winter.

Oil pipelines

Oil pipelines in hot climates and heat pipes in power stations have Ω-shaped loops built into them at intervals, to allow for expansion.

Metal cables

Electricity wires and metal cables can be seen to sag more during hot weather.

Concrete roads

Concrete roads and footpaths are laid in sections with pitch placed between. In hot weather the pitch melts into a viscous tarry liquid which allows the concrete to expand without cracking.

Shipbuilding

In shipbuilding, red hot rivets are used to join metal plates together. As these cool, and contract, the segments are pulled together very tightly. This method may also be used to fit metal tyres on to railway carriage wheels. Alternatively, instead of heating the tyre the wheel may be cooled (by dipping in liquid nitrogen). This method has the advantage that it is less likely to de-temper the steel tyre.

Metal caps on glass bottles

A tight metal cap on a glass bottle may be loosened by pouring hot water on the neck of the bottle. Likewise a glass stopper stuck in the neck of a bottle may be released by rubbing the neck with a hot cloth. Both the heat of the cloth and the friction resulting from rubbing may free the stopper.

Hot liquid in cold glass containers

In jam-making for example, if hot jam is poured into cold jam jars, they may crack. This occurs because glass is a poor conductor so the inside of the jar expands faster than the outside. The jars should be well warmed beforehand in an oven.

Formation of scree

The tops of mountain ranges often have sharp, craggy rock pieces, called scree. These are often formed by rainwater freezing. Water running into cracks and crevices expands as it freezes, and acts like a wedge, splitting pieces off the rocks. When gardeners ‘spike’ or ‘rough-dig’ their gardens in late autumn, they are leaving the way open for winter frost to break up the soil. Ploughing also has the same effect.

Other examples

When filling an ice-cube tray, one should not fill it completely to the top, to allow for expansion.

A dent in a table tennis ball may be pushed out by placing it in a basin of hot water.

Because of expansion, it is very dangerous to heat any sealed vessel e.g. a tin of soup. It should always be opened before heating.

It can be noticed that furniture and roofs often creak at night after a warm day, owing to contraction as they cool.

At supersonic speeds, Concorde is almost 0.5 m longer owing to heating effects caused by friction between the air and the outer metal surface.

Liquid-in-glass thermometers, e.g. mercury and alcohol thermometers, are based on the expansion of liquids.

Low Temperature Effects

Superfluidity

The word was coined by the Russian physicist **Pyotr Kapitsa** (1894-1984) in 1938, but the phenomenon was first observed in 1930. When helium is cooled further below its boiling point (4.27 K), the liquid suddenly stops boiling at a sharply defined temperature (2.18 K), called the ‘lambda point’, and turns into a most unusual liquid. The behaviour of this liquid is best explained by considering it to be a mixture of two fluids - a normal liquid and a ‘superfluid’. As the temperature falls, the percentage of superfluid in the mixture increases from zero at the lambda point to 100% at 0 K. The superfluid has zero viscosity and can move into tiny cracks and openings so small that not even gases can penetrate them. It passes through the pores of an unglazed container into which it is put, and can spread out into a microscopically thin film over any solid matter it touches. This is referred to as ‘film flow’. This abnormal liquid also has an extremely high thermal conductivity. Strangest of all, this ‘unruly’ liquid cannot be confined in an open vessel, because it flows up the inside of its container, ‘creeps’ over the edge, and ‘crawls’ down the outside, where it can be seen dripping from the bottom of the vessel.

Superconductivity

Ordinary conductors of electricity become even better conductors as their temperature is reduced. This was noticed by the Estonian physicist **Heinrich Lenz** (1804-1865) in 1835, but much more work was done in this area by the Dutch physicist **Heike Kamerlingh Onnes** (1853-1926) in 1911. (Onnes was also the first to liquefy helium, which he did in 1908.) He used pure mercury in solid form, at very low temperatures, and measured its resistance (using $R = V/I$). There was a dramatic decrease in resistance, which happened quite suddenly at 4.2 K (called the transition temperature). The resistance fell to zero – the mercury had become a superconductor. This class of substance may continue to conduct an electric current almost indefinitely without the continued application of any source of emf. Because resistivity values are so low (less than $4 \times 10^{-25}\ \Omega$ m) there is virtually no heating effect, even when large currents are carried.

After working with mercury Onnes investigated other metals, and showed that indium, tin and lead also become superconductors at temperatures of 3.4 K, 3.7 K and 7.2 K, respectively. In one experiment, a current of a few hundred amperes was induced in a closed ring of superconducting lead, maintained at the temperature of liquid helium; it was found to be still flowing after 2 years with hardly any diminution. Since then, other superconductors have been discovered, some at higher temperatures. In 1973, an alloy of niobium and tin, Nb_3Sn, was found to be superconducting at 17.9 K; in 1987 other compounds were added to the list, with transition temperatures of 30 K and 77 K (the boiling point of nitrogen). But some substances only become superconducting at extremely low temperatures, e.g. the transition temperature for hafnium is 0.3 K, for zinc it is 0.88 K.

In 1957 **John Bardeen, Leon N. Cooper** and **John Robert Schrieffer** developed a theory known as the BCS theory which sought to explain the phenomenon of superconductivity. They suggested that the current in superconductors is carried by electrons which are linked together into pairs by attractive forces and so cannot dissipate energy through scattering, which is the usual cause of electrical resistance in conductors. They received the Nobel prize for their work some years later, in 1972.

In 1986 and 1987 compounds of barium, copper and yttrium came to notice, because they became superconducting at 90 K; so also did some ceramics, with superconducting temperatures of 120 K. All of these substances, with transition temperatures above the boiling point of liquid nitrogen, are much more useful in practice, as they are easier to use and much less expensive to keep. Liquid nitrogen keeps them sufficiently cold and superconducting, so avoiding the necessity of using liquid helium which is more expensive and inconvenient. It is thought that thermal vibration taking place at higher temperatures (above the transition temperature) breaks up the electron pairing and then normal resistance develops again.

The first experiments done on superconductivity raised hopes that very powerful electromagnets, motors, transformers and generators could be built, giving 100% efficiency. However, it was discovered that if a superconductor is exposed to a very high magnetic field its electrical resistance returns and it loses its superconducting properties. Because a current flowing in a conductor always generates a magnetic field, superconductors can carry only a limited current. However, modern superconductors are much less sensitive to the effects of magnetic fields and superconducting magnets are widely used, e.g. in magnetic resonance imaging (MRI) systems. In Japan, at the moment, experimental trains are being built containing superconducting magnets which are used to 'levitate' them over the tracks so that friction is greatly reduced and speeds of up to 480 km h^{-1} are achieved. (See also Electromagnetism module.)

Other effects

When some objects, which normally are flexible, are dipped into liquid air (below 93 K), they become hard and brittle. Flowers, pieces of fruit, rubber tyres, etc., shatter into tiny fragments when struck with a hammer.

Mercury turns into a strong, solid metal at low temperatures. A lead bell, cooled to the temperature of liquid air and struck, rings with a sweet sound.

When a body, e.g. a metal bar, is heated with its ends firmly fixed to prevent expansion, stresses are set up in the bar which may exceed the elastic limit and damage it. Conversely, if a hot solid cools down, and if it is not allowed to shrink, it may be weakened as tensile stresses beyond the elastic limit are set up. This is what happens when boiling water is poured into a thick glass tumbler. The inside layer heats and expands, and because glass is a poor conductor of heat, the outside is still cold so the vessel may crack (see experiment 1.6 section 1.4). One way of preventing this is to place a metal spoon in the glass which helps the conduction process; another is to use a thin-walled glass container. Alternatively, a container made of a 'heat resistant' material such as Pyrex may be used. Such containers have low expansion values and are widely used in laboratory work and in cooking.

Appendix 1.1
The Anomalous Expansion of Water

Most liquids, when cooled, contract and thereby become more dense. Water, however is an exception to this rule. It follows the usual pattern on cooling from 100 °C, until the temperature reaches 4 °C. At this temperature it has its maximum density and minimum volume, because below this value water starts to expand again on further cooling, so it becomes less dense.

In a pond, during very cold weather, water at the surface is in contact with cold air, so it cools down. When 4 °C is reached, because of the greater density, this water sinks to the bottom. As this happens, slightly warmer water from the bottom rises to the surface. This in turn is cooled and moves downwards, so convection currents are set up and proceed until the whole pond reaches a temperature of 4 °C. Further cooling at the surface then lowers the temperature of the water there to 0 °C. This remains at the surface because it is less dense than the water below at 4 °C. The water at 0 °C then freezes (at the surface) to ice, which floats because its relative density is 0.917. Because ice is a poor conductor of heat, it acts as a type of insulator between the cold air above and the slightly warmer water below it. One could say in these conditions that the temperature of a pond is graduated downwards, from 0 °C at the surface to 4 °C at the bottom. This abnormal behaviour of water is of immense significance for aquatic life - it means that fish can survive during a very cold spell, by moving down to the bottom of a lake or pond. It is only very shallow ponds that freeze through in cold weather - lakes hardly ever do. If water behaved like other liquids, ponds would freeze from the bottom upwards, with far-reaching biological consequences.

Not only does water expand on cooling from 4 °C to 0 °C, but the expansion that takes place when water freezes is even greater. (Every 100 cm^3 of water produces 109 cm^3 of ice.) This is what causes the bursting of pipes during very cold weather conditions, but any damage done does not become evident until a thaw sets in. It also explains why an ice-cube floats on water. An iceberg is really a giant ice-cube; only 10% of an iceberg is visible above water level. The following question might be an interesting one to put to students. If an ice-cube is floating in a glass of water which is full right up to the brim, does the water overflow as the ice melts?

The maximum density of water at 4 °C was first observed by **Hope** in 1804. He devised a piece of apparatus which is now known as Hope's apparatus, Fig. 1.6(a). It comprises a metal cylinder with a gallery around it half way up, into which a freezing mixture of ice and salt is put. Two thermometers above and below the gallery register temperature, and the values obtained are plotted on a graph, Fig. 1.6(b). The phenomenon may also be demonstrated very conveniently using a thermos flask as follows. Place some cold water and crushed ice in the flask, stir and leave to stand for a few minutes. It will then be found that the temperature of the water at the bottom of the flask is 4 °C while the temperature at the top, where the ice is, is 0 °C.

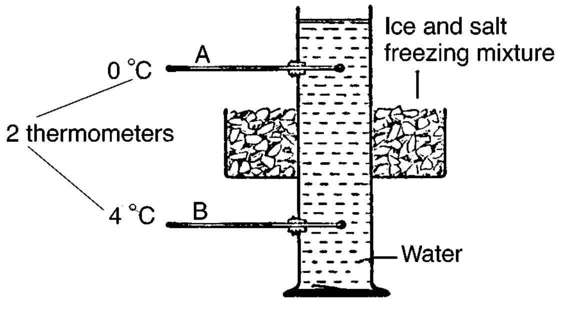


Fig. 1.6(a) *Hope's apparatus*

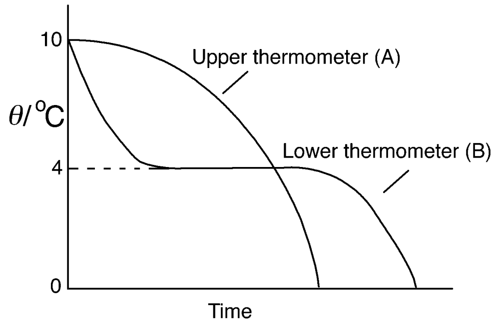


Fig. 1.6(b) *Variation of temperature with time at the top and bottom of Hope's apparatus*

The anomalous expansion of water is sometimes explained by the amalgamation of water molecules into groupings or clusters, e.g. H_4O_2, H_6O_3, H_8O_4. This takes place as the water warms slightly from 0 °C to 4 °C and causes contraction, which counteracts any expansion that might result from the small increase in temperature. Once the temperature reaches 4 °C, the expansion effect predominates.

Water is not alone in its unusual behaviour. The metals antimony and bismuth behave in a similar way. They too expand on freezing, which makes them ideal casting materials.

Appendix 1.2 Expansion Coefficients

Not all substances expand by the same amount as the temperature rises. Different materials have different expansion coefficients: linear (α_l) when dealing with the length of an object, (α_A) when referring to area or superficial expansion ($\alpha_A \approx 2\alpha_l$) and α_V when referring to cubic or volume expansion ($\alpha_V \approx 3\alpha_l$).

In considering the expansion of liquids the situation is more complex because, apart from the expansion of the liquid itself, the expansion of the enclosing container must be allowed for. The apparent expansion of a liquid or gas is the amount by which it expands when the expansion of the containing vessel is not taken into account. The real or absolute expansion is the amount by which a liquid or gas expands when the expansion α_l of the enclosing vessel is considered.

The linear expansion coefficient of a solid is defined as the increase in length, per unit length, per kelvin rise in temperature.

If l_0 is the length of a rod at 0 °C and l_θ is its length at a temperature θ, then the linear expansion coefficient, α_l, is given by

$$\alpha_l = \frac{(l_\theta - l_0)}{l_0\theta}$$

Thus, $l_\theta = l_0(1 + \alpha_l\theta)$.
Similarly, $A_\theta = A_0(1 + \alpha_A\theta)$ and $V_\theta = V_0(1 + \alpha_V\theta)$.

Table 1.3, p. 13, shows that mercury expands more than seven times as much as glass, and that alcohol expands more than five times as much as water, for a given rise in temperature.

The fact that two substances have the same expansion coefficients can sometimes be very useful. Steel and concrete have almost the same value, thus we can use steel rods with concrete to give reinforced concrete. Also platinum and glass have the same expansion coefficients. This means that in chemistry, when flame tests are carried out, the platinum wire used to carry a sample of salt does not separate from its glass handle when heated. All dentists know that the materials supplied to them for use in dental fillings have expansion coefficients that match natural tooth enamel.

One substance that expands very little indeed, i.e. is nearly temperature invariant, is a ferronickel alloy called invar. It expands by only a millionth of its length per kelvin rise in temperature. Invar was originally made by the Swiss-born physicist **Charles Edouard Guillaume** (1861-1938) and he received the Nobel prize in 1920 for his work. Invar contains 36% nickel, combined with steel, and is used whenever minimal expansion effects are required, e.g. in the balance wheels of watches and clocks and the pendulum of a pendulum clock.

Material	**Linear expansion coefficient, α_l / K^{-1}**	**Cubic expansion coefficient, α_V / K^{-1}**
Aluminium	2.3×10^{-5}	6.9×10^{-5}
Brass*	1.9×10^{-5}	5.7×10^{-5}
Copper	1.7×10^{-5}	5.0×10^{-5}
Iron/Steel*	1.2×10^{-5}	3.5×10^{-5}
Lead	2.9×10^{-5}	8.7×10^{-5}
Glass*	8.5×10^{-6}	2.5×10^{-5}
Platinum	8.9×10^{-6}	2.7×10^{-5}
Pyrex	3×10^{-6}	9×10^{-6}
Quartz	7.5×10^{-6}**	3.5×10^{-5}
Mercury		1.8×10^{-4}
Water		2.1×10^{-4}
Ethanol		1.1×10^{-3}
Air and most gases at atmospheric pressure		3.7×10^{-3}

Table 1.3 *Expansion coefficients at 20 °C (*Depends on composition. **Parallel to axis.)*

CHAPTER 2 THERMOMETRY AND THERMOMETERS

2.1 Background

History of Thermometers

As stated in section 1.1 our senses can give us an idea of whether something is hot or cold, but they do not provide us with an accurate measurement[1]. The temperature assessments of receptors in our skin are usually sufficient for our day-to-day needs - we know for example if we need to take off, or put on a jumper - but for scientific work and accurate measurement, thermometers are needed.

Instruments designed to measure temperature are called thermometers. The earliest ones were based on the fact that liquids and gases usually expand as they are heated, the forerunners of our present day liquid-in-glass instruments. **Galileo Galilei** (1564-1642) is credited with being the first person to devise a thermometer, in 1592. He called it a *thermoscope*. It consisted of a glass bulb (about the size of a hen's egg) at the end of a long glass tube which was about the thickness of a straw. To use it, Galileo warmed air in the bulb with his hands, having placed the tube vertically in a vessel of water with the bulb on top. On removing his hands the air in the bulb cooled and contracted and water rose up into the stem, depending on the temperature. With it, Galileo wrote that he could measure temperatures ranging 'from that of snow to the blood temperature of an ox'. (A form of this apparatus is currently widely used to demonstrate the expansion of gases.) There were two drawbacks to Galileo's thermoscope. Firstly, it had no graduations or markings. Secondly, it was affected by variations in atmospheric pressure as the vessel was open to the air; it acted as a combined thermometer/barometer. In 1611, another Italian, **Sanctorius of Padua** (1561-1636) made another version of the thermoscope, in which he used water as the expanding and contracting medium, and he designed it so that the bulb was underneath, with the tube pointing upwards and open at the end. He introduced a scale divided into 110 parts and claimed to be able to measure temperatures from melting snow to the heat of a candle.

One of Galileo's pupils, **Leopold de Medici**, constructed a thermometer which contained alcohol and was sealed after some air had been driven out. A French physician, **Jean Rey**, made a few thermometers using different liquids, e.g. alcohol, linseed oil, etc., around the year 1632. So also did **Ferdinand II**, Grand Duke of Tuscany, in 1644, and in Florence from 1657 to 1667 members of the Accademia del Cimento produced different types of sealed and graduated alcohol thermometers. Their high and low marks corresponded to the hottest and coldest days of the Tuscan year.

Another name linked with thermometry was a French physicist **Hubin** in the early 1670s. He used two liquids in his U-shaped thermometers. He chose high and low marks to correspond to the temperature of melting butter and the temperature of a Parisian wine cellar. **Newton** (1642-1727), in 1701, chose the freezing point of water and body temperature as his two fixed points - the interval between was divided into 12 parts. The thermometer he used contained linseed oil. It is said that it was the well-known Astronomer

[1] The skin can sometimes mislead us. If a person is blindfolded, and asked to distinguish between being touched on the arm by a hot iron, or an ice-cube, they can sometimes be confused as to which was used. The sensations can be mixed up.

Royal, **Edmund Halley** (1656-1742), a mathematician /astronomer, who first suggested using mercury in a thermometer, and that it was the French physicist, **Guillaume Amontons** (1663-1705), who actually used it initially. The Danish astronomer, **Ole Rømer** (1644-1710), also did some work on thermometers. He used diluted alcohol and set fixed points. To the melting point of ice he initially gave the value 7.5 and then later the value 8, and he assigned the value 60 to the boiling point of water. He assigned the figure 22.5 to blood temperature. Rømer's work is largely forgotten, but it was he who influenced **Daniel Gabriel Fahrenheit** (1686-1736), a German/Dutch instrument maker, living in Danzig. From 1709 to 1714 Fahrenheit used both alcohol and mercury as thermometric liquids and called the lowest temperature he could reach (using an unspecified mixture of ice and sea salt) 0 °F. The temperature of the human body he took to be 96 °F (we know now it is 98.4 °F on this scale). It was not until after his death that the values of the freezing and boiling points of water were set at 32 °F and 212 °F, respectively. The space between them was divided into 180 equal parts. Another scale was developed by a Frenchman, **René Antoine Ferchault de Réaumur** (1683-1757). He used spirit of wine diluted with water as his thermometric fluid and called the freezing point of water 0 °R, and the boiling point 80 °R. His scale was used for a time in continental Europe, mainly in Germany.

A much more famous scale, the centigrade scale (the Celsius scale since 1954), was introduced in 1742 by the Swedish astronomer **Anders Celsius** (1701-1744).

Celsius was professor of astronomy at Uppsala, and studied the aurora borealis (northern lights) and linked it with changes in the earth's magnetic field. He also did work on the size and brightness of stars. He realised that atmospheric pressure played a part in determining the fixed points, particularly the boiling point of water. In the beginning, he actually called the freezing point of water 100° and the boiling point of water 0°, and he was the first to use the symbol ° to represent a degree. It is said that one of his colleagues - **Stromer** - turned these figures upside down the following year to give us the values we use today. There were some scientists who claimed that Celsius did not develop this centigrade thermometer himself, and that credit for it should go to his fellow-countryman, the biologist **Carolus Linnaeus** (1707-1778). However, it is the name of Celsius that remains associated with the 0°-100° scale of temperature that is so widely used today. The Fahrenheit scale is still used occasionally, though increasingly rarely. To convert °F to °C the following formula is used: $°C = 5(°F - 32)/9$.

The degree Celsius is not an SI unit and temperatures for states colder than freezing water have negative values.

The well-known Manchester brewer, **James P. Joule** (1818-1889) (associated with the mechanical equivalent of heat) also had an interest in thermometry. He constructed a thermometer which was capable of reading to within 0.005 °F. On his honeymoon, in Switzerland, he used this to measure the temperature difference between water at the top and bottom of waterfalls. He showed that water at the bottom was indeed slightly hotter than that at the top. Potential energy present in the water at the top became converted to kinetic energy, later to internal energy, giving a rise in temperature, and some sound energy, as it hit the bottom.

With regard to other types of thermometer, it was **Sir Humphry Davy** (1778-1829) who, in 1821, discovered that the electrical resistance of metals is related to temperature. In 1871, the German scientist **Sir William Siemens** (1823-1883) first used platinum as a resistance material in a thermometer.

The Seebeck effect, which is the basis of the thermocouple thermometer, was discovered in 1821 by **Thomas Johann Seebeck** (1770-1831). He observed that an electric current is generated in a simple closed circuit consisting of two wires made from dissimilar metals when the two junctions of the wires are at different temperatures.

The constant volume gas thermometer was originally designed by a German professor of physics **Philippe Jolly** (1809-1884). It is now used as a standard to calibrate other thermometers.

The optical pyrometer, which is a radiation thermometer, used to measure high temperatures, had its beginnings in the work of the German physicist **Max Planck** (1858-1947). In 1900 he established a link between the temperature of an object and the spectral distribution of the radiation it emitted.

2.2 Temperature Scales

Thermometry is the science of the measurement of temperature. Temperature is a quantitative description of the state of hotness of matter. The temperature of an object - a figure ascribed to it - is not a fixed number, but depends on the type of thermometer used, and the scale adopted.

To establish a scale of temperature three steps are necessary.

1. Choose some physical property or characteristic of a body that varies continuously with increasing hotness, and stays constant at the same level of hotness. This is called a thermometric property. The thermometric property should change measurably and continuously as heat flows into or out of the body.

   Examples are:
   - the length of a liquid in a glass column;
   - the emf of a thermocouple;
   - the electrical resistance of a wire;
   - the product pV of a fixed mass of gas;
   - the pressure of a fixed mass of gas at constant volume, or the volume of a fixed mass of gas at constant pressure;
   - the colour of liquid crystals or incandescent substances, etc.

2. Select two reference or fixed points (but see p. 19). These are defined values of temperature identified by a particular occurrence which happens only at that specific temperature, e.g. the melting point of pure ice and the boiling point of distilled water at standard pressure. These standard degrees of hotness should be capable of being accurately reproduced and be accessible.

3. Divide the interval between the two fixed points (called the fundamental interval) into equal parts called degrees. The value of a particular thermometric property Y is measured at the fixed temperature points Y_0 at the ice-point, and Y_{100} at the steam point. If Y_θ is allotted to its value at some other temperature θ, then temperature is defined (on a centigrade scale) by the equation

$$\theta = \frac{Y_\theta - Y_0}{Y_{100} - Y_0} \times 100$$

   The denominator corresponds to the total change in the thermometric property between the upper and lower fixed points, while the numerator represents a change in the same property between the lower fixed point and the other temperature θ. The unspecified property Y in the equation above changes to $L, R, E, p,$ etc. according as length of liquid columns, resistance of metals, emf or pressure of a gas, respectively, are selected as the thermometric properties.

Note

It is possible to extrapolate the measured values of Y_θ outside the two fixed points. However, it must be remembered that the scale is meaningful only for those values within the range in which that thermometric property is measurable.

The boiling point and melting point of a liquid depend on the pressure. The effect of pressure change on the boiling point of water is shown in the following graph, Fig. 2.1.

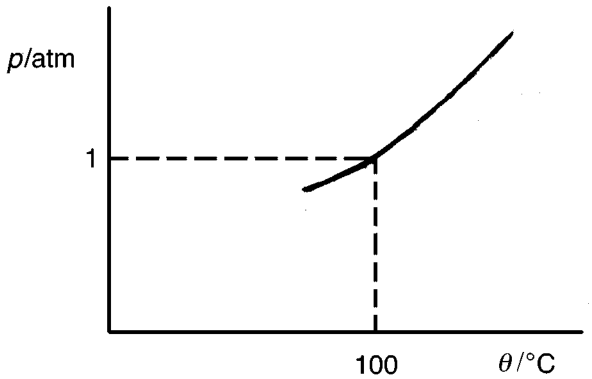


Fig. 2.1 Relationship between boiling point of water and pressure

The boiling point increases as pressure increases. Water and vapour can co-exist in equilibrium at all points on the graph.

The effect of pressure change on the melting point of ice is shown by the graph in Fig. 2.2. The melting point of ice decreases as pressure increases.

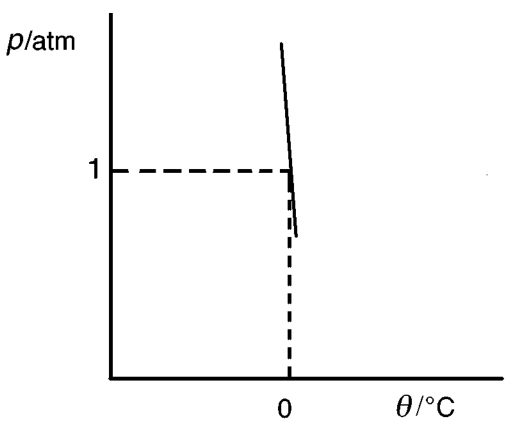


Fig. 2.2 Relationship between melting point of ice and pressure

When these two graphs are combined and extended, they meet at a point called the triple point of water. This graph is referred to as a phase diagram, Fig. 2.3.

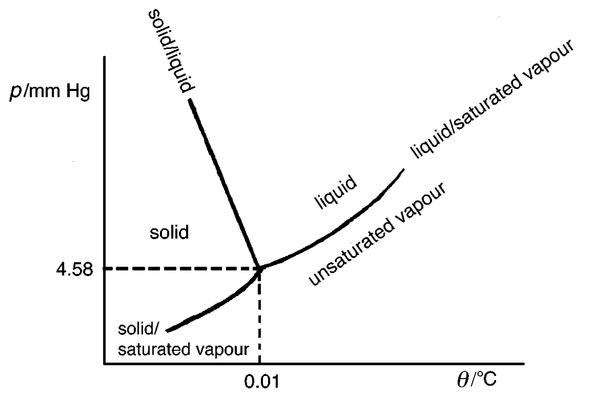


Fig. 2.3 Phase diagram for pure water

The **triple point** is the point of intersection. It is that unique temperature (273.16 K or 0.01 °C) at which the three phases of pure water - ice, liquid and water vapour - can co-exist in equilibrium. (Pressure value is 611 Pa which is equivalent to 4.58 mm of mercury. This is the pressure at which the melting point and the boiling point are the same.)

The Kelvin Scale of Temperature

It is not very convenient to have many temperature scales based on different thermometric properties since thermometers based on different thermometric properties may indicate different temperatures for a particular degree of hotness. There is need, therefore, for a fundamental or standard temperature scale.

In 1848 **William Thomson (Lord Kelvin)** (1824-1907) proposed an absolute thermodynamic temperature scale, which was based on the operation and efficiency of a reversible Carnot engine (see pp. 76-78) using the ideal gas. Kelvin defined temperature in terms of the ratio of energy flow into and out of a system (an ideal heat engine), by

$$\frac{Q_h}{Q_c} = \frac{T_h}{T_c}$$

where

Q_h = heat flow from a high temperature reservoir
Q_c = heat flow to a low temperature reservoir

and T_h and T_c are temperatures assigned to the hot and cold reservoirs.

The scale of temperature based on Kelvin's work in thermodynamics was called the absolute thermodynamic temperature scale, or the Kelvin scale, after its founder. The ideal gas scale (see below) is made to coincide with the thermodynamic scale by choosing appropriate fixed points.

In 1661 **Robert Boyle** (1627-1691) had stated that pV is constant for a fixed mass of gas at constant temperature. Kelvin showed that it was possible to define an absolute scale of temperature using the product pV of the ideal gas[2] as the thermometric property. An absolute scale is one in which temperature is proportional to the thermometric property, i.e. $T \propto Y$. Thus on the ideal gas scale,

$$T \propto pV$$

or

$$\frac{T}{T_1} = \frac{pV}{(pV)_1}$$

The ideal gas scale is an absolute scale and therefore only one fixed point is required since, from the definition, when pV is zero T is zero. The fixed point chosen is the triple point of water to which the value 273.16 is assigned. Thus,

$$\frac{T}{273.16} = \frac{(pV)_T}{(pV)_{\text{t.p.}}}$$

or

$$T = \frac{(pV)_T}{(pV)_{\text{t.p.}}} \times 273.16$$

The unit of temperature on this scale is the kelvin (K). At zero on this scale, i.e. absolute zero, molecular movement is at a minimum. (Kelvin had proposed the concept of absolute zero in 1851.) The third law of thermodynamics tells us that 0 K can never be reached but values down as low as 10^{-8} K have been achieved. The fixed point, the triple point of water, is extremely reproducible. A temperature change of 0.01 K (0.01 °C) would cause one of the three states of water to disappear. Either the water would freeze, or the ice would melt. For reasons of accuracy, precision and reproducibility, this fixed point has been accepted internationally. Water of the highest purity is distilled into a glass vessel that has its air removed before being sealed off. The cell is primed by using cooled brine to form a layer of ice on the inner wall.

For real gases, if the density and pressure of the gas used are extremely low, then the behaviour of real gases approaches that of the ideal gas. Real gas temperatures are obtained from the equation

$$T = \lim_{p \to 0} \frac{(pV)_T}{(pV)_{\text{t.p.}}} \times 273.16$$

In practice, the constant volume gas thermometer is used because a change in pressure with temperature at constant volume is more convenient to measure accurately than a change in volume at constant pressure. Corrections are made for the non-ideal gas used. The defining equation is

$$T = \frac{p_T}{p_{\text{t.p.}}} \times 273.16$$

The thermodynamic temperature scale is the fundamental temperature scale in science. Its unit, the kelvin, is the SI unit of temperature, and is defined as 1/273.16 of the temperature of the triple point of water. In 1967 the unit was changed from the degree kelvin (°K) to simply the kelvin (K). Since Kelvin proposed that pV of the ideal gas be taken as a thermometric property and defined a linear relationship between pV and T, the thermodynamic and ideal gas scales are identical and both are often referred to as the Kelvin scale. Both have the same unit, the kelvin, and both now use the same fixed point, the triple point of water.

[2] The ideal gas is one that obeys Boyle's law at all values of temperature and pressure. This imaginary gas would never liquefy. No such gas exists, however. Real gases approach the behaviour of the ideal gas at extremely low pressures.

The Celsius Scale of Temperature

The Celsius scale is defined in terms of the Kelvin scale by the equation

$$\theta/°C = T/K - 273.15$$

Thus, a change of temperature on the Kelvin scale is equal numerically to a change of temperature on the Celsius scale. The two scales are thus linked, with the degree Celsius being the same size as the kelvin. A change in temperature of 1 °C is equivalent to a change in temperature of 1 K, and there are 100 divisions between the ice and steam points of water on both scales.

Note

The Celsius scale is a centigrade scale. However, the equation on p. 17 gives only an approximation to the Celsius temperature. The closeness of the approximation depends on the thermometric property used, i.e. on the linearity of the change of the property with temperature on the Kelvin scale. For mercury-in-glass thermometers and for resistance thermometers in the range 0 °C to 100 °C the approximation is acceptable for most everyday work.

2.3 Thermometers

A thermometer is an instrument which uses a measurable property of a body that varies with temperature as a means of assigning a numerical value to the temperature. There are different types of thermometer; which one is chosen for use will depend on the range of temperature involved in any given situation, or the accuracy required.

Every thermometer records its own temperature, and absorbs thermal energy from the substance whose temperature it is measuring. Interaction between the thermometer and the substance in question leads to a state where both are at the same temperature - in thermal equilibrium. There is usually some time lag before this happens.

Characteristics of a good classroom thermometer

1. It must be accurate.
2. It must be quick to reach thermal equilibrium.
3. It must have a wide range to make it as useful as possible.
4. It must be able to reproduce the same reading as a second thermometer, assuming the same temperature conditions apply.
5. There should be an even change in thermometric property per degree change in temperature.
6. It must have a low heat capacity. It should take so little heat from the body whose temperature it is measuring that temperature is not appreciably changed.
7. It must be sensitive. This means that there should be a large change in the thermometric property for a small change in temperature. In the case of the liquid-in-glass thermometer, it ideally should have a large bulb with a narrow-bore capillary tube.
8. It should not be too cumbersome.

Advantages of using mercury in thermometers

1. It expands evenly as temperature increases.
2. It is a good conductor of heat.
3. It does not wet the glass.
4. It gives a direct reading.
5. It has a high boiling point (357 °C), therefore is good for high temperatures.

Disadvantages of using mercury

1. It is not very good for low temperatures - it freezes at -39 °C.
2. It cannot follow a rapidly changing temperature.
3. It cannot measure very small temperature changes.
4. It has a poisonous vapour which is emitted if the glass breaks.
5. It absorbs an appreciable amount of heat as it warms up.
6. The glass surrounding it does not expand evenly as temperature increases, and is slow to contract as it cools. To make the thermometer quick-acting, the glass should be quite thin so that heat can pass through it more easily. However this makes it more fragile.

Advantages of using alcohol

1. It is very good at low temperatures since its freezing point is -115 °C.

2. It is cheaper and its vapour is less toxic than that of mercury.

3. It is more sensitive than mercury - it is able to detect smaller temperature changes. Its expansion is six times that of the same volume of mercury for an equal rise in temperature.

Disadvantages of using alcohol

1. It is not as good a conductor of heat as mercury is, so it takes longer to reach the temperature of whatever substance it is in contact with.

2. It expands irregularly.

THERMOMETER	THERMOMETRIC PROPERTY	RANGE	ADVANTAGES	DISADVANTAGES	SPECIAL NOTES
Mercury-in-glass	Change in length of a thread of mercury	-39 °C to 357 °C (higher, up to 600°C if N_2 is added)	Direct reading Portable Very accurate in range 0 °C to 100 °C	Poisonous vapour if glass breaks Fragile Limited range	Most commonly used type of thermometer
Alcohol-in-glass	Change in length of a column of alcohol	-115 °C to 78 °C	Very good for lower temperatures	Wets the glass Not suitable for high temperatures Must be coloured Not very accurate	Commonly used in refrigerators and freezers
Constant volume gas	Change in pressure of a fixed mass of gas held at constant volume	-270 °C to 1500 °C	Very accurate Very sensitive Very wide range	Cumbersome Inconvenient and slow Corrections must be made	Used as a standard by which all thermometers are calibrated Used as standard on ITS-90 between 3 K and 24.6 K
Resistance	Change in electrical resistance	-200 °C to 960 °C	Accurate Wide range	Not very robust	Used to define temperatures on ITS-90 between 13.8 K and 962 °C
Thermistor	Change in resistance of a piece of semiconductor material	-160 °C to 600 °C	Small size Very sensitive Rapid response to fluctuating temperature	Not as accurate as the platinum resistance thermometer Range not as wide	Has a negative temperature coefficient of resistance
Thermocouple	Emf generated as a result of a temperature difference between two junctions	-200 °C to 1700 °C	Wide range Can be used in inaccessible places due to its small size Very good for rapidly changing temperatures Quick response time	Not very accurate	Different pairs of metals used for different temperature ranges
Radiation	Change of wavelength in radiation of hot bodies, colour difference	-10 °C to 2650 °C	The only thermometer capable of measuring temperatures above 1500 °C	Not direct reading Cumbersome Best acurracy I K	Readings calculated from Planck's radiation law Used as reference standard on ITS-90 for temperatures over 962 °C

Table 2.1 *Common types of thermometer*

3. It has a higher specific heat capacity than mercury.
4. It wets the glass, and must be coloured to make it more visible (usually red).

Water is unsuitable as a thermometric liquid

1. Its range is too short: 0 °C - 100 °C.
2. It is not very sensitive to heat - it has too high a specific heat capacity.
3. It behaves in a peculiar way at some temperatures - it contracts in volume on heating between 0 °C and 4 °C (by about 8%) and above 4 °C it expands on heating.
4. It wets the glass.
5. It must be coloured in order to see it clearly in glass.

Errors associated with liquid-in-glass thermometers

1. Usually calibration is correct only when the whole stem is in the material whose temperature is to be measured.
2. They must be calibrated when upright[3].
3. Outside pressure can sometimes force the meniscus up along the tube.
4. The glass stem also expands and may not be of uniform bore.

Notes

1. If the mercury column in a thermometer has separated, it can usually be re-united by cooling the bulb until all the mercury has retracted into it.

2. Thermal hysteresis refers to the fact that thermometers give different readings when cooled down to a particular temperature than when they are warmed up to that same temperature.
3. Nitrogen gas is sometimes put into the space above the mercury column in a mercury thermometer. This increases the pressure of the mercury and therefore its boiling point. In so doing the upper limit is extended from 357 °C to c. 500 °C.

4. When mercury thermometers are made (by warming and dipping in a bath of the liquid metal) they are often left aside for a period of time (about a year) to stabilise, before they are calibrated.

2.4 Other Types of Thermometer

The Clinical Thermometer

It is said that the physicist **Robert Boyle** was the first person to show that the human body in good health maintains a constant temperature. In 1858 the German physician **Karl August Wunderlich** introduced the procedure of frequent checks on body temperature, as an indicator of the course of disease.

The clinical thermometer is a specialised form of a mercury-in-glass thermometer, graduated from 34 °C to 42 °C (normal body temperature is 37 °C). It was invented by an English physician **Thomas Clifford Allbutt** in 1866. Its capillary tubing is of very small diameter but the wedge-shaped glass surrounding it magnifies the mercury thread into a broader band for ease of reading. It also helps to prevent the thermometer rolling when left on a table. The thicker glass also prevents children from biting through it. A constriction or ‘kink’ above the bulb is a very important feature - it breaks the mercury thread and prevents the mercury from going backwards into the bulb after it has been removed from the patient. It acts as a type of maximum thermometer; it has to be re-set with a flick of the wrist. Students sometimes ask why the mercury does not run back past the constriction, on contraction, when originally it was able to get through it. The reason for this is that the very narrow constriction only allows the expanding mercury through it under pressure. As the mercury cools it breaks at the gap because the intermolecular forces in it are not strong enough to drag it backwards into the bulb. The ‘flick of the wrist’ motion in an arc in the air helps to force the mercury back into the bulb. The clinical thermometer should not be sterilised in boiling water - an antiseptic solution is used instead. Two minutes should be allowed before reading the body temperature - it takes that amount of time for the thermometer to reach thermal equilibrium with the body.

[3] A mercury thermometer gives a lower reading in the vertical than in the horizontal position. This is due to pressure caused by the column of mercury increasing the size of the bulb, thus increasing the volume of mercury in it.

In more recent years, strips of plastic impregnated with liquid crystals have arrived on the market. They change colour depending on body temperature. In a particular type, the letter ‘N’ (normal) is formed when placed in contact with the forehead, while the letter ‘F’ (fever) appears at a higher body temperature. These strips, while convenient and safe to use with small children, are not as accurate as the mercury clinical thermometer.

One of the most recent thermometers which has become available is the 'E-Z Therm’ infrared radiation thermometer. A slender probe tip on this is inserted into the auditory canal of the ear by pulling the pinna upwards and backwards. Infrared emission is detected from the tympanic membrane and a reading is obtained in a time of just 3.3 seconds. The thermometer is battery operated; temperature readings are displayed digitally, the range being from 34.6 $^{\circ}$C to 42.4 $^{\circ}$C. The previous four temperature readings with the times at which they were measured during the previous 48 hours can also be obtained. Disposable sleeves are available for such thermometers. They prevent contact between patient and thermometer, thus avoiding cross infection and keeping the thermometer perfectly clean.

The Optical Fibre Thermometer

This is an extension of the optical pyrometer. An optical fibre of sapphire is blackened at one end and placed in or near the site where temperature is to be measured. At thermal equilibrium the black coating radiates black-body radiation, characteristic of its temperature, along the fibre. Temperature can be determined from an analysis of the spectral distribution. This method is sometimes used for measuring the temperature of a furnace.

Magnetic Thermometers

These are thermometers used to measure extremely low temperatures. The thermometric substance is a paramagnetic salt - either cerium magnesium nitrate or gadolinium sulphate. The thermometric property is their magnetic susceptibility (the degree to which a substance is susceptible to magnetisation). They become more sensitive the lower the temperature, and can measure temperatures below 0.002 K.

Thermal Noise

Electrons in a conductor move in a random manner, rather like molecules in a gas. Because the motion is random there may, at a particular instant, be more electrons at one end of the conductor than the other; a short time later this situation may reverse. There is therefore a random fluctuating voltage across the conductor. As with molecules in a gas, the kinetic energy of the electrons depends on the temperature and so the voltage fluctuations are also temperature-dependent. These voltage fluctuations show up as ‘noise’ or ‘static’ in electronic circuits, and the effect increases with increasing temperature. (It is for this reason that amplifiers at the focus of radio telescopes, etc., are cooled in liquid nitrogen or even helium.) The thermal noise thermometer is based on this effect, the randomly varying voltage being the thermometric property. The thermometer consists of two resistors of high-quality platinum linked to an electronic amplifier. One resistor is at a fixed reference temperature and the second is at the unknown temperature. Noise levels in the two resistors are compared and one is reduced until both are equal. From previous calibration work temperatures are determined. The possible range is wide (5 K-1050 K).

Gas Viscosity Thermometers

These thermometers work by measuring the resistance of a capillary tube to a flow of gas through it. Argon, at a particular reference temperature, passes through a variable length of narrow tubing. Its flow is compared with the flow through a ceramic capillary at a high temperature, as detected by a differential manometer. When the drops in pressure along the two capillaries are equal, temperature can be measured by an equation linking resistance with temperature. Previous calibration is of course always necessary.

Vapour Pressure Thermometer

This thermometer is used at very low temperatures and uses the vapour pressure of liquid helium. It is established that the saturation vapour pressure of liquid helium is directly related to the temperature of the substance. Temperatures as low as 1 K may be measured using the commoner isotope $^{4}_{2}He$. Even lower temperatures, down to 0.5 K, may be measured using the less common isotope $^{3}_{2}He$. However, they take a relatively long time to reach equilibrium, and the vapour pressure is affected by the presence of any impurities.

Acoustic Thermometers

The speed of sound as it passes through any medium is affected by temperature. Thermometers based on this fact are known as acoustic thermometers. If the frequency of the sound is above the normal range of human hearing, they are called ultrasonic interferometers. Waves are produced in a cylinder of variable length, containing helium gas at a known pressure, by a sound wave generator (quartz crystal ultrasonic oscillator). For specific lengths of the cylinder, standing waves are produced in it. By measuring the distance between nodes for different cylinder lengths the wavelength is determined. Knowing the frequency of the oscillator, f, the speed v may be calculated from $v = f\lambda$. The experiment is repeated at different pressures and then extrapolating to zero gives λ at zero pressure. Temperature can be calculated from an equation which relates wave speed to the temperature of the ideal gas.

Doppler Effect in Temperature Measurement

The broadening of spectral lines increases with increasing temperature. The magnitude of the frequency change, Δf, is proportional to the square root of the absolute temperature, T. This method can be used to measure the temperature of a very hot gas or plasma. The change in frequency is the thermometric property. (See module on Waves for more detail.)

Temperature Indicators

Some materials are useful as temperature indicators. They are usually available as wax-based crayons (thermal crayons) which give a single indication that a particular temperature has been reached. They are uni-directional (the colour does not change back again on cooling). They are quite accurate (c.1%), and the range available lies between 40 °C and 1800 °C. In the metal industry they play a significant part in estimating temperatures reached before welding is carried out.

2.5 Different Thermometers, Different Temperatures

If different thermometers using different thermometric properties are used to measure temperature, they may give somewhat different values. The discrepancies get bigger the higher the temperatures measured, e.g. 300 °C might be obtained with a platinum resistance thermometer, compared to 309 °C using a constant volume gas thermometer.

The separate scales must by definition coincide at the defined values of the fixed points, e.g. 0 °C and 100 °C, but they do not necessarily agree at the points in between. In getting different temperature values from different thermometers one must not assume that one value is more correct than another. All are correct according to their own particular scale. The discrepancy arises because thermometric properties do not vary in the same way as each other as the temperature changes. Different properties of matter have very little connection with one another, so we can hardly expect that different scales of temperature linked to these properties will keep in step with one another. Two scales of temperature will not agree unless the thermometric properties on which they are based vary with temperature in the same way. The change in length of a mercury thread for a given change in temperature is not necessarily proportional to the change in resistance of a wire, the change in pressure of a gas, or the change in emf of a thermocouple, produced by the same change in temperature. When this is examined from the point of view of a graph, the length of a mercury column plotted against the changing resistance of a wire, for example, would be non-linear. Taking a specific example, assigning a value of 50 °C, on any scale to the temperature of something suggests that the thermometric property used has a value which is mid-way between its 0 °C and its 100 °C values. But mercury-in-glass, resistance, gas pressure and emf do not all necessarily reach values which are halfway between their 0 °C and 100 °C values at exactly the same temperature.

It would be possible to obtain precisely the same temperature at all points from all thermometers, only if their thermometric properties were to vary uniformly with degree of hotness in exactly the same way. Every characteristic of matter used as a thermometric property introduces a scale of temperature that most likely will disagree with other scales between the fixed points hence the need for a standard thermometer against which other thermometers may be calibrated.

2.6 The International Temperature Scale

In practice, accurate measurement of temperature using a real gas thermometer is inconvenient, and requires painstaking laboratory work. So, to reduce the experimental difficulties involved and to standardise temperatures, the International Temperature Scale was adopted. It was first set up by the International Committee of Weights and Measures in 1927 and was revised or modified in 1948, 1960, 1968, 1975 and 1976. The most recent scale was decided on in 1990. This scale of temperature is based on the measured temperature values of a number of reproducible equilibrium states. In the beginning, six primary ones were chosen[4], then later some secondary points were added, bringing the number to eleven, and since ITS-90 there are now sixteen fixed points. These points are determined using substances of the highest purity, generally at their boiling points, freezing points or triple points. Except for the triple point of water, values are not defined but are based on experimental results, which change over the years as thermometry techniques improve. In most cases, pressures are taken to be one atmosphere, except for hydrogen which is measured at lower pressures.

The instruments used for interpolating between selected points and beyond the highest point (1357.77 K) are usually specified. Between 13.8033 K and 1234.93 K a platinum resistance thermometer is used. Optical pyrometers giving values based on Planck's law of radiation are used for temperatures above 1234.93 K, taking as reference the radiation emitted from a black body cavity. At very low temperatures (up to 5 K) ITS-90 is defined in terms of the vapour pressure of ${}^{3}_{2}He$ and ${}^{4}_{2}He$. Absolute zero cannot be attained in practice (third law of thermodynamics). Between 3.0 K and 24.6 K a constant volume gas thermometer is used.

The ITS recognises the thermodynamic temperature scale as the basic one, and is constructed so as to approximate thermodynamic temperatures. The unit used is the kelvin, which is defined to be 1/273.16 of the triple point of water.

[4]The six primary points originally selected were:

Boiling point of oxygen	90.18 K
Melting point of ice	273.15 K
Boiling point of water	373.15 K
Sulphur point (boiling point)	717.75 K
Silver point (melting point)	1233.65 K
Gold point (melting point)	1336.15 K

Substance	State	Temp/K
Hydrogen	Triple point	13.8033
Hydrogen	Boiling point[a]	17.035
Hydrogen	Boiling point	20.27
Neon	Triple point	24.5561
Oxygen	Triple point	54.3584
Argon	Triple point	83.8058
Mercury	Triple point	234.3156
Water	Triple point	273.16
Gallium	Melting point	302.9146
Indium	Freezing point	429.7485
Tin	Freezing point	505.078
Zinc	Freezing point	692.677
Aluminium	Freezing point	933.473
Silver	Freezing point	1234.93
Gold	Freezing point	1337.33
Copper	Freezing point	1357.77

Table 2.2 *Fixed points on the 1990 international temperature scale ([a] this boiling point is for a pressure of 33321.3 Pa. All other points, other than triple points, are for a pressure of 101292 Pa).*

2.7 Worked Examples

1. The resistances of a length of wire at the freezing point and boiling point of water are 4.6 Ω and 6.8 Ω, respectively. Its resistance at room temperature is 5.0 Ω. Calculate a value for the room temperature, based on the resistance of the wire.

$$\theta = \frac{R_\theta - R_0}{R_{100} - R_0} \times 100$$

$$= \frac{5.0 - 4.6}{6.8 - 4.6} \times 100$$

$$= 18\ ^\circ\text{C}$$

2. The length of a column of mercury in a thermometer is 3.2 cm at the freezing point of water and 26.5 cm at the boiling point. Calculate a value for the temperature when the length of the column is 15 cm.

$$\theta = \frac{l_\theta - l_0}{l_{100} - l_0} \times 100$$

$$= \frac{15 - 3.2}{26.5 - 3.2} \times 100$$

$$= 50.6\ ^\circ\text{C}$$

$$= 51\ ^\circ\text{C}$$

Note
The least accurate measurement is to two significant figures, hence the answer is restricted to two significant figures.

3. The pressure of the gas in a constant volume gas thermometer at the triple point of water is 680 mm of mercury and 705 mm of mercury at room temperature. Calculate the room temperature, as measured on the Kelvin temperature scale.

$$T = \frac{p_T}{p_{\text{t.p.}}} \times 273.16$$

$$= \frac{705}{680} \times 273.16$$

$$= 283\ \text{K}$$

4. Given that $\frac{p_{100}}{p_0} = 1.36$ calculate the temperature of absolute zero on the Celsius scale, where p_{100} is the pressure of a constant volume gas thermometer at the boiling point of pure water, and p_0 is the pressure measured at the ice point, both being measured at atmospheric pressure.

Temperature is defined on the Celsius scale using the constant volume gas thermometer as

$$\theta \quad = \quad \frac{p_\theta - p_0}{p_{100} - p_0} \quad \text{x} \;\; 100$$

The pressure of a gas at absolute zero is 0.
When θ is taken to be absolute zero, then p_θ is zero.

$$\theta \quad = \quad \frac{0 - p_0}{p_{100} - p_0} \quad \text{x} \;\; 100$$

$$p_{100} = \quad 1.36\, p_0$$

$$\theta \quad = \quad \frac{0 - p_0}{1.36\, p_0 - p_0} \quad \text{x} \;\; 100$$

$$\frac{-\,p_0}{0.36 p_0} \quad \text{x} \;\; 100$$

$$= \quad -278\ ^\circ\text{C}$$

Note
From the definition of the Celsius scale, the value for absolute zero on the Celsius scale is -273.15 °C.

2.8 Mandatory Student Experiment

Calibration of a Thermometer

The calibration of a thermometer involves the construction of a scale such that readings from the thermometer in a given case are consistent with those which would be obtained with a standard thermometer. This is normally achieved by comparing the readings from the thermometer to be calibrated with those from a thermometer which has previously been calibrated, directly or indirectly, against a standard thermometer (cf. p. 21). In the school laboratory the ordinary mercury thermometer may be used in this way to calibrate other thermometers.

Alternatively, a thermometer may be calibrated using suitable fixed points and an appropriate equation. For example, the equation on p. 17 may be used to calibrate a mercury-in-glass thermometer or a platinum resistance thermometer in the school laboratory since, over the range of temperature normally encountered, the thermometric properties on which these thermometers are based vary linearly with temperature as defined on the Kelvin scale. (More correctly, this process should be referred to as ‘graduating a thermometer’ since no allowance is made for the non-linearity of the variation of the thermometric properties with temperature.)

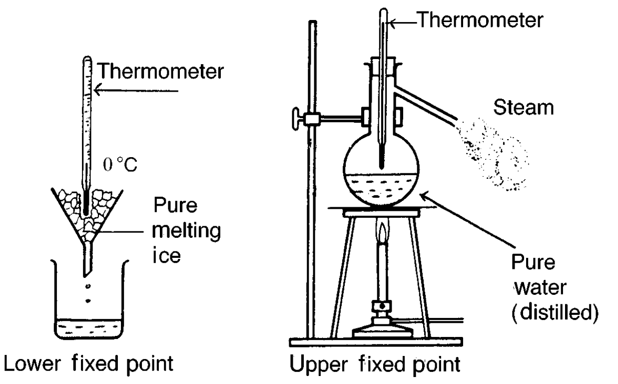


Fig. 2.4 *Calibrating a mercury thermometer*

Experiment 2.1(a) To Calibrate and Use a Liquid-in-Glass Thermometer, e.g. a Mercury Thermometer

Apparatus
As shown in the diagram. Ungraduated thermometers may be obtained from laboratory suppliers. Once the two fixed points are found and marked they can be used to measure temperature. The liquid mercury is chosen in preference to alcohol since the boiling point of alcohol is less than 100 °C.

Method
1. Place the thermometer bulb in pure crushed melting ice in a funnel. Allow some time for the level to stabilise and then measure the length of the mercury thread. This is the ice point, or 0 °C position.

2. Now place the thermometer bulb in the steam coming from boiling water. When the mercury level is constant again measure the length of the mercury column. This second point is the steam point or 100 °C position.

3. By measuring the different lengths of the mercury column when the bulb is in melting ice (l_0) , steam from boiling water (l_{100}) and a fluid which gives a length l_θ, the temperature of the fluid may be found from the equation:

$$\theta = \frac{l_\theta - l_0}{l_{100} - l_0} \times 100$$

Experiment 2.1(b) To Calibrate and Use a Thermocouple Thermometer

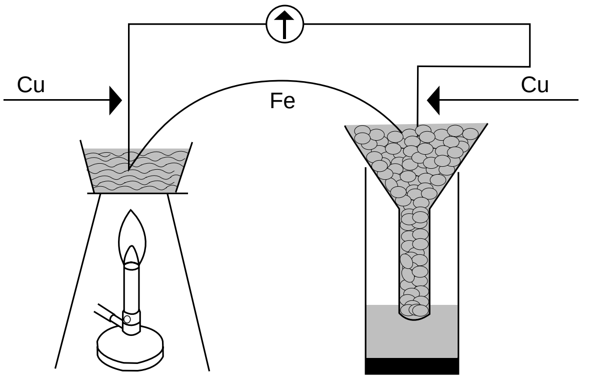


Fig. 2.5 Calibrating a thermocouple thermometer

Apparatus

As shown in the diagram.

Method

1. Check that the junctions are clean.

2. One of the junctions - the fixed or reference junction - is placed in crushed melting ice in a funnel and kept there for the duration of the experiment.

3. Place the other junction - the test one - also in melting ice. Note the reading on the meter. It should show 0 mV, because the two junctions are at the same temperature, so no emf is generated. This is E_0 (where E = emf).

4. Place a mercury thermometer with the test junction in the melting ice and heat gently. Record the thermometer reading and the emf at 1 °C intervals.

5. Plot a graph of emf against temperature. This graph may then be used as a calibration curve to convert emf readings to temperature values in °C.

Notes

1. The normal values that one could expect to find with the test junction in steam (100 °C) and, of course, the reference junction in ice might be in the region of 4-5 mV (depending on the metals used).

2. An interesting 'twist' to this experiment that a teacher may use, when finishing up a practical class, is to measure body temperature with a thermocouple. By getting some volunteer from the class to hold the test junction tightly in his/her closed fist, and observing the deflection produced, the body temperature may be obtained. The cold junction should be kept in melting ice.

3. Having calibrated the thermocouple thermometer it is instructive to graduate it using the equation on p. 17 and then to compare the values obtained for the temperature of, say, a beaker of warm water on the two scales.

Experiment 2.1(c)

To Calibrate and Use a Thermistor Thermometer

Apparatus

Thermistor (e.g. TH3); ohmmeter; beaker; round-bottomed flask.

Method

1. Place some pure crushed melting ice in the beaker, making sure the thermistor is well covered. Connect the terminals to the ohmmeter and record the resistance of the thermistor (R_0) when the reading stabilises.

2. Place a mercury thermometer with the thermistor in the melting ice and heat gently. Record the thermometer reading and the resistance at 1 °C intervals.

3. Plot a graph of resistance against temperature. This graph may then be used to convert resistance readings to temperature values in °C.

Notes

1. A thermistor is a device whose resistance changes rapidly with changing temperature. Most thermistors in school laboratories are semiconductor devices whose resistance decreases with increasing temperature.

2. The ohmmeter used to measure resistance is fast and very convenient but it is not as accurate as the metre bridge method.

3. Having calibrated the thermistor thermometer it is instructive to graduate it using the equation on p. 17 and then to compare the values obtained for the temperature of, say, a beaker of warm water on the two scales.

Experiment 2.1(d)

To Calibrate and Use a Platinum Resistance Thermometer

Apparatus

Some platinum wire is wound non-inductively[6] on an insulating frame of mica and placed inside a tube of quartz or hardened glass; Pyrex is suitable. Some glycerol or liquid paraffin is poured into the tube and this helps in heat transfer, but does not conduct electricity. Heating has to be done slowly so as to allow the platinum time to adjust to changing temperature. An ohmmeter or a metre bridge is used to measure the resistance.

Method

1. Place the thermometer in pure melting ice and connect its leads to an ohmmeter. Allow to stand for a few minutes before taking a reading of the resistance (R_0).

2. Place the thermometer in a large beaker of water and heat gently to boiling point. Allow sufficient time for the platinum to reach the temperature of the surroundings. Then take a reading from the ohmmeter (R_{100}).

3. Place the platinum in another vessel of warm/hot water, and again allowing it time to adjust, measure the new value of the resistance (R_θ).

4. Compare the result calculated with the direct reading from a mercury thermometer.

$$\theta = \frac{R_\theta - R_0}{R_{100} - R_0} \times 100$$

The metre bridge method can also be used here, as in 2.1(c). The platinum is inserted into one of the arms of the bridge and the bridge is balanced as usual.

Note

A less expensive wire such as iron is often used instead of platinum, and gives good results.

[6] Non-inductive winding is a method of double-winding the coils of resistors so that the magnetic fields caused by varying electric current cancel each other out and so produce no self-induced emfs.

Experiment 2.1(e)

To Calibrate and Use a Constant Volume Gas Thermometer

Apparatus

Constant volume gas thermometer; large beaker; barometer.

Method

1. Place the bulb of a constant volume gas thermometer in a large beaker of pure ice/water mixture. Allow sufficient time for the gas in the bulb to come to thermal equilibrium with its surroundings.

2. Adjust the flexible tube of mercury by raising or lowering until the mercury is at the reference point. Measure h_0 (difference in mercury levels). Read atmospheric pressure, H, from a barometer, expressing the pressure in mm of Hg.

$$p_0 = H \pm h_0$$

3. Place the bulb in steam from boiling water. Adjust until the mercury level is again at the reference point. Measure h_{100}.

$$p_{100} = H \pm h_{100}$$

4. Place the bulb in the substance whose temperature is required to be measured. Adjust until mercury is at the reference point. Measure h_θ.

$$p_\theta = H \pm h_\theta$$

Calculate temperature θ from the equation

$$\theta = \frac{p_\theta - p_0}{p_{100} - p_0} \times 100$$

Note

In experiments 2.1 (a), (d) and (e) a graphical method, rather than the formula, may also be used to find the unknown temperature. For example, in calibrating the mercury thermometer, Experiment 2.1(a), l_0 and l_{100} are plotted on a graph as shown in Fig. 2.6. These two points are then joined by a straight line since, *by definition*, there is a linear relationship between temperature and length of the mercury column. The unknown temperature, θ, is then found by finding the temperature on the x-axis corresponding to l_θ on the y-axis. A similar procedure is followed for the other types of thermometer described.

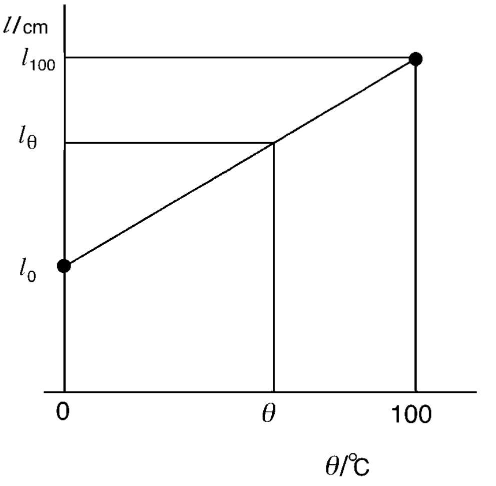


Fig. 2.6 *Using a graph to calibrate a mercury thermometer*

CHAPTER 3

TRANSFER OF HEAT

3.1 Background

States of Matter

The three common states of matter are solid, liquid and gas. A fourth state - the plasma[1] state - is now recognised by physicists. This, however, is only found at temperatures of the order of 10^{8} °C, where nuclear fusion might be possible. Plasma is matter in a highly ionised state, e.g. material in the sun. It can be looked on as 'fuel for fusion'. At such extremely high temperatures it cannot be contained easily and magnetic forces are needed to confine it in a so-called 'magnetic bottle' or tokamak (see module on Electromagnetism). Heated glass in molten form may be looked on as an intermediate state of matter - it is not really a solid or a true liquid. It, like tar or pitch, becomes soft gradually and melts over a range of temperature. This makes it easy for craft workers to work it into different shapes.

However, at more ordinary temperatures, we all tend to think of particular substances as being solids, liquids or gases. We classify copper as a solid, but at a temperature of 1080 °C it liquefies, and above 2580 °C, it exists as a gas. (The bright green colouring we see being emitted from copper sulphate in a flame test is due to the presence of copper in its gaseous form.) We think of mercury as always being a liquid metal, but at different temperatures it could exist in either solid or gaseous form. Similarly, nitrogen is usually found in gaseous form in the air but as a liquid at low temperatures it has a number of uses in medicine, e.g. the removal of warts and verrucae - its boiling point is –196 °C (77 K).

The question can be asked - what is the difference between the three fundamental states of matter? The answer must be that it depends on the arrangement of particles (atoms or molecules) in a substance and their degree of movement (at a microscopic level). In a solid, the particles are very tightly packed together, in a regular and orderly arrangement, with little freedom of movement. This movement can only be vibrational in nature, as the particles remain anchored in their basic positions, the particles oscillating in any of three dimensions. The energies of the particles, both kinetic and potential, collectively comprise the internal energy of the solid. Because atoms contain particles with electrical charge powerful electrostatic forces exist between them. One can visualise these forces as being like 'springs' connecting each unit to its neighbours. If, while vibrating, the particles come too close together, the forces are repulsive, while if the particles try to move farther apart, the forces are of attraction. It is these forces that are responsible for the properties of a solid:

- definite size and shape;
- inability to flow;
- very small compressibility.

The great tensile strength of some materials (e.g. steel, which can be used to reinforce concrete in construction projects) is due again to the large cohesive electrical forces holding the individual particles together.

When a solid is heated - its internal energy is increased - the amplitude of vibration of its particles

[1] Plasma is also a region of ionised gas in a discharge tube, containing approximately equal numbers of electrons and positive ions and in biology 'the liquid part of blood'.

is increased. When they move apart sufficiently, and their mutual attraction is reduced sufficiently, they can move around more easily, and change places with one another, i.e. the substance is now in a liquid form. Thermal energy has been used to partially break down or loosen bonds between neighbouring particles. It is because the particles can roll or slide over one another, moving freely amongst themselves, that the liquid can flow and has no shape of its own, taking the shape of any containing vessel. Liquids are almost incompressible, and we find some useful applications of this property, e.g. the hydraulic brake system in a vehicle where pressure is transmitted by brake fluid. In some liquids the forces between particles are so great that the liquids pour with difficulty.

When a liquid is heated the molecules become more energetic as more energy is added, and at a particular stage they can break away from their neighbours. At this stage the attraction between particles has little or no effect because of the large distances between them. They are independent of one another. The substance is now referred to as a gas or vapour (for the distinction between a gas and a vapour, see section 1.2). Gases, like liquids, are regarded as fluids, i.e. they are capable of flowing. The gaseous state is the most disorganised of the three states of matter. Gases take up much more space than liquids or solids, have no shape or volume of their own, and fill completely the vessel into which they are put. Because of the large distances between individual particles, gases can be compressed with ease. (Liquid air is approximately 750 times more dense than atmospheric air, and 1 cm^3 of water can yield a volume of 1700 cm^3 of steam at atmospheric pressure.) Most explosives work because certain chemical reactions produce a large quantity of gas very quickly and the force of this causes the violent explosion. The real difference then, between the states of matter is the degree of movement of their molecules. Solids and liquids are said to constitute the two condensed states of matter - gases involve larger distances between the individual particles.

3.2 Do You Know

What makes a kettle 'sing'?

Collapsing bubbles of steam cause the ‘singing’ noise that is sometimes heard before a kettle boils[2]. As the temperature of water in a kettle approaches its boiling point, water at the bottom of the kettle reaches 100 °C before water higher up does. Several tiny bubbles of steam form at the bottom and rise up through the water. They collapse in the slightly cooler liquid before they reach the surface. This happens because the pressure inside them becomes less than atmospheric pressure, and the collapse of these bubbles produces the ‘singing’ sound. Once the temperature of the whole liquid reaches the boiling point, the saturated vapour pressure inside the bubbles now equals atmospheric pressure. The bubbles now can reach the surface where they break and cause the familiar appearance of boiling.

Which is the better way to wrap food in foil?

One side of a sheet of aluminium foil is dull, and the other side is shiny and bright. If food is wrapped in it, e.g. a potato that one wishes to bake, then it is better to have the dull side facing outwards, as this will absorb energy more readily. However, if one wishes to keep sandwiches cool by wrapping them in foil, the shiny side should then be at the outside.

When did people first freeze food?

Freezing is one of the oldest methods used to preserve food. In cold climates, food was buried in ice to keep it for future use. However, in temperate climates, food was first frozen in England in 1851, when a salt and ice mixture was used. In 1908 an American, H. S. Baker, tried to freeze some fruits which were surplus to market sales, but some fruits lost their texture and shape in the process. The quick-freeze method, by which food was frozen in hours rather than days, was developed in Germany in 1916. It was in the following year that Clarence Birdseye used small containers to fast-freeze food. These were put on the market in 1919, marking the start of the frozen food industry.

[2] At about 50 °C, bubble formation first becomes noisy. The noise is caused by the collapse of bubbles of dissolved air as they are expelled and by steam bubbles which form on the hot heating coil collapsing on contact with the surrounding colder water. The noise changes near 100 °C.

What is frostbite?

Frostbite is an injury caused by exposure of body parts to low temperatures. Living tissue is damaged when tiny ice crystals form in the skin, or in the underlying tissues - nerves, muscles, etc. Areas of the body most likely to be affected are the extremities and exposed areas - the face, fingers and toes. The skin feels cold, painful and numb and is usually waxy-white in colour. If the condition is treated quickly, the outlook for complete recovery of the affected parts is good, but frostbite of long duration may cause tissue death (gangrene), resulting from impairment of the blood circulation. Apart from low temperatures, which may be aggravated by a cold wind, and reduced oxygen supplies at high altitudes, other factors may contribute to the onset of frostbite. These include insufficient food, dehydration, wet clothing, tightly fitting gloves or boots, injury with haemorrhage or fracture. The presence of a fever or intoxication with alcohol may further increase the heat loss from the body; in these cases the victim may be unaware of the seriousness of his/her condition. Treatment consists of warming the affected parts by immersion in warm (not hot) water (c. 40 °C) until circulation is restored and the tips of the extremities regain a pink colour. If warm water is not available the use of blankets is an alternative. Direct warming by a fire, a heater of any kind, or hot water bottle is not advisable as a sudden change in tissue temperature is as damaging to the tissue as the low temperature itself. Affected areas are bandaged to avoid infection and antibiotics are routinely prescribed.

Heatstroke - what happens?

If a person engages in strenuous exercise or does heavy physical work for long periods in hot weather, they may suffer from heatstroke. As the body overheats, more blood flows to the surface of the skin, so that increased evaporation of sweat, with the accompanying cooling effect, can take place. The greater blood flow to the skin results in a reduced blood flow to the brain. This causes a feeling of weakness, dizziness and nausea. Cramps may also be experienced due to dehydration and salt loss. In acute cases circulatory failure may take place, again due to dehydration. Water losses of about 7% from the body can be fatal while temperature values of 44 °C and over may cause brain damage. At such a body temperature, the skin feels hot, the pulse rate increases and the breathing becomes irregular. Victims should not be immersed suddenly in ice-cold water as this causes constriction of the blood vessels to the skin. This in turn causes increased blood flow to the interior of the body, which temporarily raises the body's core temperature still higher. A gradual lowering of the body temperature is the means of treatment of heatstroke. Cool mist dressings applied to the head and neck and fan-assisted ventilation will bring this about. Victims are encouraged to drink sufficient water, to take salt tablets and to move to a cooler location. Heat exhaustion is less severe than heatstroke. In both cases high humidity accompanying high temperatures makes a person feel even more uncomfortable due to lack of evaporation of perspiration.

Spontaneous combustion - what is it?

The scientific term for burning is combustion. When this takes place oxygen chemically combines with another element or elements. As oxygen occupies 21% of the air it is always available to react with various substances. In some reactions flames are visible, in others an explosion takes place; these are both rapid forms of combustion. Sometimes, however, combustion takes place very slowly - over a period of years in the case of iron rusting - yet this is still the combination of oxygen with iron. If, however, combustion takes place in such a way that any heat energy produced is retained then eventually a situation could arise where active burning starts. This is known as spontaneous combustion - the production of flames without addition of heat energy from an outside source. Situations where this might take place could be a hay barn storing very dry hay, a large coal depot or grain silo; even a number of oily cloths left in an enclosed dry space might start to burn. Oil slowly oxidises, causing a rise in temperature and eventually, in well-insulated conditions, a fire could start. Combustion, of course, takes place in our bodies all the time, when food is oxidised in the process of respiration, releasing energy and producing carbon dioxide and water vapour as waste products (metabolism).

Why does hot water freeze more quickly than cold water, if left out of doors, in very cold climates?

When equal quantities of hot and cold water, in open-topped vessels, are left outside, in places like Iceland or northern Canada, the hot water freezes before the cold sample. This is due to greater evaporation from the hot water, reducing the quantity of liquid remaining.

The Peltier effect

When a current flows across the junction of two dissimilar metals heat is either evolved or absorbed, depending on the direction of current flow. The effect is reversible - a heating effect becomes a cooling effect as the current changes direction. The phenomenon is called after the French physicist **Jean Peltier** (1785-1845) who discovered it in 1834. (The effect differs from the normal Joule heating that occurs whenever current flows in a circuit; this is unaffected by current direction.)

3.3 Conceptual Approach

Heat Transfer

Heat transfer is the passage of energy from a body or a place at a higher temperature to another at a lower temperature, and it occurs by three processes - conduction, convection and radiation. In some situations two of the processes, or all three, may be involved at the same time. However, it is usual to discuss them separately. When teaching a class, a teacher finds that this section usually presents little difficulty and most students know and understand the basic facts. The following is an analogy that students can identify with.

If a teacher at the top of a class wishes to return a copybook to a student sitting in the back row, he/she has three choices.

1. The teacher can hand it to a student in the front row, and ask the other students to pass it back, one by one, to the student in question (conduction).
2. The teacher can walk to the back of the class carrying the copy (convection)
3. The teacher could throw the copy over the heads of the other students (radiation) - but no teacher would ever do this!

This analogy helps convey the differences between the three methods, and can be a useful tool for the students in remembering these differences.

Conduction

Conduction is the method whereby solids are heated and, unlike what happens during convection, no net movement of matter takes place. It may be compared to a chain of persons passing buckets of water from a source to extinguish a fire. The individuals remain stationary, more or less, representing the molecules in a substance. The movement of the buckets represents the movement of heat.

The mechanism of conduction differs between metals and non-metals. In non-metals, heat energy is transferred by a series of atomic collisions or vibrations, whereas in metals, conduction takes place by both atomic vibration and random electron movement.

Conduction in non-metals

Atoms at higher temperatures vibrate or oscillate more energetically, i.e. with greater amplitude, than do their colder neighbours. But because they are locked in their basic equilibrium positions, and are bound to one another by electrostatic forces, they have little chance of moving apart. However, as atoms in the heated part of a substance are vibrating energetically they ‘jostle’ their immediate neighbours and spur them into action too. These in turn activate others and so energy is passed down the line. It is a slow process though, as the large size of the atoms (compared to electrons) prevents them from any great movement, so energy transfer is limited in most non-metals. There is also another view on this process. It is sometimes thought of in terms of waves passing through a material. ‘Packets’ of elastic ripples of very high frequency colliding with atoms of material are said to transfer energy to them. These wave packets are known as phonons (as distinct from photons in the photoelectric effect, but the basic idea is the same). In poor conductors of heat, electrons are tightly bound to their nuclei so no electron transfer of heat takes place.

Conduction in metals

The process detailed above takes place in metals, but another more effective method is involved also. Metals contain a large number of unbound or 'free' electrons, loosely connected to their parent atoms, which wander about through the atoms in a lattice. When a metal is heated these mobile valence electrons gain extra kinetic energy which increases their velocities. Because they are small and light compared with atoms of the material they travel quickly and transfer energy evenly to all parts of the metal, sometimes colliding with atoms as they move along. Metals are therefore good conductors of heat. Some are better than others and they are also good conductors of electricity. We can speak of a 'heat current' as we would of an electrical current.

The thermal conductivity, λ, of a substance is defined as the rate of heat flow through unit area of a sample of the substance, at right angles to the direction of flow, per unit temperature gradient.

Substance	λ/W m^{-1} K^{-1}
Silver	406.0
Copper	385.0
Aluminium	205.0
Brass	109.0
Steel	50.2
Lead	34.7
Mercury	8.3
Glass	1.1
Concrete	0.9
Water	0.56
Cardboard	0.21
Rubber	0.19
Wood	0.12-0.04
Silk	0.09
Fat	0.046
Fibreglass	0.042
Felt	0.04
Cork	0.04
Duck down	0.025
Air	0.024
Polyurethane	0.024

Table 3.1 *Thermal conductivities*

There are many applications of thermal conduction in everyday life. Some objects may feel cold to the touch if they are good conductors because they carry away heat from the body *rapidly*, so a concrete or tiled floor feels much colder to stand on than a carpeted one. A polystyrene cup feels warm to the touch because it conducts away barely any heat from the body. On the other hand, in a very hot room (e.g. Turkish bath), metal objects can feel very hot to the touch and may actually burn the skin. In a block of hot metal the atoms/molecules may vibrate rapidly, perhaps thousands of times each second. If one touches it with one's finger, the rapidly vibrating atoms cause the molecules of the skin to go into sudden and violent motion, resulting in the sensation of pain.

Convection

This method of heat transfer takes place in fluids (liquids and gases). It involves the actual movement of hot matter, which carries the internal energy with it from an object or place at a higher temperature to one at a lower temperature. The following are some of the distinguishing characteristics of convection.

1. Convection makes possible the transfer of heat over much larger distances than is the case with conduction. The atoms/molecules themselves move, carrying heat energy with them, whereas with conduction heat is transferred from particle to particle by collision, the particles remaining in their places.

2. Convection takes place only in fluids. It cannot take place in solids as the atoms/molecules are fixed in their positions relative to each other.

3. The effectiveness of convection can also depend on the viscosity of the fluid in question. Thin soup like chicken broth can be left over a gentle heat to simmer, but a thicker food like porridge tends to burn if it is not stirred while it cooks. This is because more viscous fluids are less able to transfer heat energy by convection.

4. No simple equation for heat transfer by convection can be formulated (unlike thermal conductivity in solids) as there are so many

variables involved, e.g. density, viscosity, type of surface, etc.

Two types of convection are recognised, viz. natural convection and forced convection.

Natural convection

Natural or free convection is the movement of fluids between areas of different densities, these differing densities arising from differences in temperature. As the fluid is heated, it becomes less dense, rises upwards and is replaced by cooler liquid or gas which flows in to take its place. This in turn is warmed and rises also, and so the process continues until all the fluid is at the same temperature. Heat transfer by natural convection usually takes place only in an upwards direction and cannot take place in the absence of gravity.

Forced convection

Forced convection arises when fluids are made to move by means of an external impetus, e.g. pump, fan, draught or wind, causing a more rapid movement of fluid. Newton recognised that hot objects cooled more quickly when they were placed in a draught. He based his law of cooling (the rate of heat loss from a body is proportional to the difference in temperature between it and its surroundings) on experiments carried out on hot liquids beside open windows.

Examples of forced convection include the following.

1. One could say that forced convection takes place in the human body, with the heart as pump and the warm blood as the circulating fluid.
2. Blowing over one's food to cool it is another example of forced convection.
3. In a domestic central heating system a pump is included to assist in the movement of hot water around a house, especially in a larger building.
4. Most car engines are cooled by convection and a pump is used to aid the circulation of water.
5. The term 'wind chill factor' describes the situation where a person's body loses heat more quickly when exposed to a cold wind.
6. The action of a hair-dryer is also an example of forced convection, as is an electric fan heater.

THE IRISH TIMES

WEATHER EYE

Warm feeling for wind chill

"THE only argument against an east wind," said the American poet James Russell Lowell, "is to put on your overcoat". Lowell would have known nothing of the concept of wind chill, since the term was not conceived until 1939, but he was obviously familiar with the general idea – that with low temperatures, the stronger the wind, the colder it seems to feel.

Even very low temperatures are quite bearable in calm conditions. You feel relatively comfortable because a layer of nearly motionless air encapsulates your body: its innermost layers adapt to the body temperature and the envelope insulates you like the "dead air" space between the panes of double-glazed window. When a breeze develops, however, it disturbs and ultimately blows away this insulating layer and thereafter each molecule of air carries away with it the heat it gathers as it touches you.

Used sensibly and appropriately, the concept of "wind chill equivalent temperature" as often used in the media, can be a useful one. Its popularity stems largely from the fact that at a superficial level it is very easy to understand. We have all felt chilled by a stiff breeze when standing in the cold and it is nice to be able to put a number on the extent of our discomfort.

It is appropriate enough to think in terms of wind chill in circumstances where humans or animals are exposed to the relevant conditions or when studying the problem of heat-loss from buildings in cold weather. Indeed the concept of wind chill can be applied in any situation where warm objects are exposed simultaneously to both wind and low temperatures. Quite simply, the cold wind carries the heat away.

But it is wrong to apply the concept to unheated, inanimate objects. The temperature of the radiator or a car, of an oil storage tank, or of growing plants – if they are dry – will not drop below the local air temperature no matter how strongly the wind blows. If they happen to be wet, their temperature may drop slightly below that of tbe surrounding air, because of heat lost through evaporation, but the concept of wind-chill in such cases is quite meaningless.

Wind chill also has ambiguities as far as humans are concerned. The extent to which we feel the cold depends on many other things besides heat and temperature. It varies depending on the amount of clothing we have on, on how old we happen to be, and on our physical, and indeed mental condition at the time.

Brendan McWilliams

Radiation

Radiation is the transfer of energy from one object to another by electromagnetic radiation. Energy transfer by electromagnetic waves was first described by the Scottish physicist **James Clerk Maxwell** (1831-1879) in 1865. Radiation travels to the earth from the sun (mostly through empty space) and consists of light and other wavelengths to which our eyes are not sensitive. The radiation mainly associated with heating is the infrared (IR). This was discovered in 1800 by the English astronomer **Sir William Herschel** (1738-1822) when he placed a blackened thermometer bulb just outside the red end of the visible spectrum. Because the waves are similar to light waves they exhibit its characteristics - reflection, refraction, interference, etc. The wavelength of course is longer (10^{-6} to 10^{-3} m), and the waves are more readily absorbed by matter.

When radiant energy falls on an object, some of it is reflected, some is transmitted and some is absorbed. It is the latter that causes the associated heating effect. Infrared rays give up their heat energy on collision with matter, causing greater vibrational movements in the molecules of the absorbing material and leading to an increase in internal energy levels and a rise in temperature.

Rate of emission

The rate of emission of radiant energy from hot objects was investigated by the Irish physicist **John Tyndall**[3] (1820-1893). Later an Austrian, **Joseph Stefan** (1835-1893), became interested and formulated a law which is now known as Stefan's law. It states that the total heat power, P, radiated by 1 m^2 of a blackbody[4] at temperature T is equal to σT^4, where σ is a constant known as the Stefan-Boltzmann constant, with a value of 5.67×10^{-8} W m^{-2} K^{-4}. Combining the results of Richie's experiment (1833), which compared the absorbing properties of substances, with Stefan's law, the total power P radiated by a body of surface area A and emissivity e (determined by the nature of the surface) at a temperature T is

$$P = \sigma e A T^4$$

Thus, the rate of emission of radiation from a hot body depends on the temperature of the body, the nature of its surface and the area of its surface (see p. 46).

With regard to the nature of the surface, it has been shown that some substances are also better absorbers than others. **Sir John Leslie** (1766-1833) devised an experiment which uses a metal box widely known as a 'Leslie cube'. The sides of the cube are all different, one is black, one white, one polished, one left untreated, and they radiate energy at different rates if the cube is filled with boiling water. It is accepted that the best absorber, and therefore the best radiator, is a dull, rough, matt surface, e.g. lampblack, and the worst is a bright, light coloured, highly polished one. Leslie used a differential air thermometer to measure the radiation emitted from each face of the cube (see section 3.7). **Benjamin Franklin** (1706-1790) also demonstrated the same fact in a very simple way. He placed some pieces of cloth, identical in every respect except in colour, on a surface of snow. As the pieces were warmed by the winter sunshine Franklin noticed that the dark-coloured ones had sunk deeper into the snow compared with the white and light-coloured cloths.

In general, a good radiator is also a good absorber and a poor radiator is a poor absorber (and therefore a good reflector).

Wavelengths or colours associated with radiation

A relatively cool emitter (e.g. a teapot of tea) or a hotter object (e.g. the base of a heated electric iron) radiate invisible waves of quite long wavelength. We can sense the warmth but cannot see it. However, in the case of an electric fire, red light is also emitted, so we both see the red element and feel the associated warmth.

The wavelengths emitted as a substance is heated depend on temperature. If a piece of platinum wire is heated in a dark room it at first simply feels warm, but with further heating it becomes red hot at approximately 500 °C; at first a dark red, then bright red. As heating continues the wire emits shades of orange, yellow, white, blue, and ultimately, at a temperature of around 1500 °C, it becomes so hot it may emit some ultraviolet radiation before finally

[3] Tyndall was born in Leighlinbridge, Co. Carlow. See section 4.1 for more detail.

[4] A blackbody is considered to be the 'perfect absorber'. It is described as an object that absorbs ALL the radiation, of every wavelength, that falls upon it. More specifically, it may be respresented as a black box with a small hole in one side. The idea behind this is that once radiant energy enters, it is confined or trapped inside.

melting. Stars which are hotter than the sun (e.g. Sirius) look blue not white. Glass is opaque to infrared radiation but transparent to light. A solution of iodine in carbon disulphide is opaque to light but transparent to infrared radiation.

The relationship between the maximum wavelength, λ_{max}, of the radiation emitted by a hot body and its temperature, *T*, is given by

$$\lambda_{max}\, T \;=\; \textit{a constant}$$

This relationship is known as **Wien's** law and the value of the constant for a blackbody is found to be 2.9×10^{-3} m K.

The colours associated with thermal radiation are employed in a number of ways. In thermograms, or 'heat photographs', a type of colour coding exists. Heat energy losses from a house are represented in different colours according to the amount of radiation emitted. In medicine, the presence of a tumour in the body may be detected because of the increased radiation caused by metabolic activity.

Police and rescue services too find thermal imaging helpful in some situations. A special camera can be used to detect the infrared radiation from a person trapped beneath the rubble of a collapsed building, or buried under snow after an avalanche. Military personnel also use infrared equipment to locate aircraft, detect or destroy missiles in the dark, etc.

In weather forecasting use is made of the fact that infrared radiation penetrates mist and fog.

A further use of infrared radiation is in the forensic units of police departments. 'Heat prints' are left by a person on a chair or bench on which they sat some hours previously.
Infrared lamps are widely used in the treatment of stiff joints associated with arthritis, and also in the healing of damaged muscles. The same type of radiation finds a place in the preparation and cooking of food. Sandwich makers use infrared waves to crisp the outside of bread and heat the filling; grilling takes place by radiation, as does the toasting of bread.

It is not only the sun which radiates energy. Every object that has a temperature above absolute zero continuously absorbs and emits radiation. This applies even to an iceberg. The earth receives heat energy from the sun; but conversely, it too must radiate in order to prevent overheating. By doing this a balance is achieved, whereby the earth radiates sufficient heat energy to maintain a uniform average temperature (see Appendix 4.1).

The same 'balancing act' takes place between a human being, or indeed any creature, and its surroundings. If a person with a body temperature of 37 °C is in a room at a temperature of 18 °C, then that individual radiates more heat energy to the objects in that room than he/she absorbs from them. However, to compensate for this, the body converts food (chemical energy) or stored body fat into thermal energy to maintain the normal temperature. If one stands beside a freezer containing frozen food in a supermarket, one feels cold because the body is losing more heat energy than it receives. Conversely, when standing beside a warm fire one feels hot, as more radiant energy is received than is emitted. There is always a net transfer of energy between objects unless they are at the same temperature.

Detectors of radiant energy

1. The skin detects the warmth associated with infrared radiation.

2. A liquid-in-glass thermometer is made more sensitive by blackening the bulb.

3. A thermopile is comprised of a number of thermocouples in series (see Chapter 2).

4. Photoelectric devices, e.g. phototransistor.

5. A bolometer. When first developed it consisted of a very narrow strip of blackened platinum. This was connected into one arm of a Wheatstone bridge. Incident radiation caused a rise in the temperature of the platinum strip, which in turn caused an increase in its resistance and unbalanced the bridge.

Thermistors now replace the platinum strip. Two are used, an 'active' one and a 'compensating' one. Radiation is allowed to fall on the 'active' one which unbalances the bridge.

Insulation

As stated previously, metals are generally good conductors of heat (and electricity) because of their free electrons. Non-metals are poor conductors of heat and so too are liquids (except liquid metals) and gases. When electrons are bound to individual atoms or molecules, only vibrational movement is possible and energy transfer is slow in these substances. They are called insulators. The term comes from the Latin word for island because, by surrounding a hot object with a poor conductor of heat, you thermally isolate it. The insulator acts as a barrier to the passage of heat. In other words it has a low thermal conductivity.

No insulator is 100% effective but a vacuum is regarded as being the best. This, however, only limits conduction and convection; transfer by radiation is still possible. This was considered by **Sir James Dewar**[5] (1842-1923), a Scotsman, when he designed his vacuum flask. So he silvered the inside walls to get around this problem. The flask was originally used to keep liquid oxygen cold (boiling point of oxygen is -183 °C). Examples of materials used as insulators include the following: air (non-moving, still or dead air), water, glass, cork, wood, wool, felt, PVC and all plastics, straw, glass wool or fibreglass, expanded polystyrene (contains a lot of air), asbestos, rubber and silk.

U-values

A list of the thermal conductivities of different substances is given on p. 35. Clearly, when selecting materials for construction, builders will choose those of low thermal conductivity in order to reduce heat energy losses. In practice, it is usual to use a figure for a particular structure, e.g. a brick wall, rather than take the values of the conductivities of its constituent materials. The figure used is called the *U*-value. It refers to Unit Heat Loss Rate, and its unit is W m^{-2} K^{-1}. The *U*-value may be defined as the rate at which thermal energy is conducted through unit area, per kelvin temperature difference between its two sides.

$$U = \frac{\text{thermal conductivity}}{\text{thickness}}$$

$$= \frac{\lambda}{d}$$

Example

The thermal conductivity of a type of brick is 0.84 W m^{-1} K^{-1}. What is the *U*-value of a layer of brick 0.4 m thick?

$$U = \frac{\text{thermal conductivity}}{\text{thickness}}$$

$$= \frac{\lambda}{d}$$

$$= \frac{0.84 \text{ W m}^{-1}\text{ K}^{-1}}{0.4 \text{ m}}$$

$$= 2.1 \text{ W m}^{-2}\text{ K}^{-1}$$

If one considers a wall composed of 3 layers of different material in contact with one another, Fig. 3.1, each having a different *U*-value and each of cross-sectional area *A*, heat energy will flow through in the direction shown.

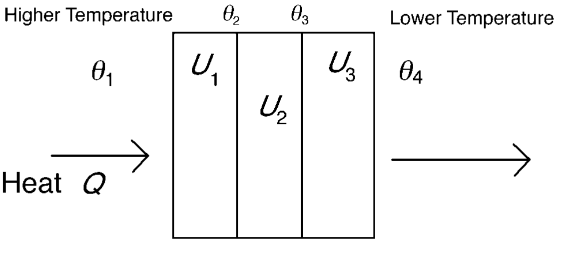


Fig. 3.1 *Heat flow through three different materials in contact*

[5] He also found that whenever ice is subjected to pressure, it melts. See 3.3 experiment 3.5.

At equilibrium, the rate of heat flow, Φ, across all 3 parts must be equal. Let the temperature at the boundaries be θ_2 and θ_3. From the definition of U-value (p. 39) the rate of heat flow is given by

$$\Phi = UA\,\Delta\theta$$

For material 1

$$\Phi_1 = U_1 A\,(\theta_1 - \theta_2)$$

$$\frac{\Phi_1}{U_1} = A\,(\theta_1 - \theta_2)$$

For material 2

$$\Phi_2 = U_2 A\,(\theta_2 - \theta_3)$$

$$\frac{\Phi_2}{U_2} = A\,(\theta_2 - \theta_3)$$

For material 3

$$\Phi_3 = U_3 A\,(\theta_3 - \theta_4)$$

$$\frac{\Phi_3}{U_3} = A\,(\theta_3 - \theta_4)$$

Since rate of heat flow across all three materials is the same

$$\Phi_1 = \Phi_2 = \Phi_3 = \Phi$$

Then

$$\Phi\left(\frac{1}{U_1} + \frac{1}{U_2} + \frac{1}{U_3}\right) = A\,(\theta_1 - \theta_4)$$

If the three materials could be replaced by one single material having a U-value of U_T which gives the same rate of heat flow Φ, then

$$\Phi = U_T A(\theta_1 - \theta_4)$$

or

$$\frac{\Phi}{U_T} = A(\theta_1 - \theta_4)$$

Thus

$$\frac{1}{U_T} = \frac{1}{U_1} + \frac{1}{U_2} + \frac{1}{U_3}$$

U-values combine in a similar way to resistances in parallel or capacitances in series.

Example

If the roof of a house has a U-value of 1.91 W m^{-2} K^{-1} what thickness of insulation, using material with λ = 0.033 W m^{-1} K^{-1}, would be required to bring the U-value down to 0.4 W m^{-2} K^{-1}?

U_{total} must be 0.4

$$\frac{1}{U} = \frac{1}{U_1} + \frac{1}{U_2}$$

$$\frac{1}{0.4} = \frac{1}{1.91} + \frac{1}{U_2}$$

$$\frac{1}{U_2} = \frac{1.91 - 0.4}{1.91 \times 0.4}$$

$$U_2 = 0.506$$

$$U_2 = \frac{\lambda}{d}$$

$$0.506 = \frac{0.033}{d}$$

$$d = 0.065 \text{ m}$$

i.e. 65 mm of insulation is required

The effect of adding insulation to a house is to reduce the *U*-values of the different areas.

		U-Value/W m^{-2} K^{-1}
Roof	Without insulation	2.3
	With insulation	0.4
Cavity Wall	Without insulation	1.6
	With insulation	0.6
Floor	Without insulation	0.9
	With insulation	0.6

Table 3.2 *Typical U-values of some structure*s

Payback time

This is the time in which the cost of insulating materials is recouped from savings made on energy consumption.

Method	**Payback Time**
Hot water lagging jacket	6 months
Draught proofing windows and doors	1 year
Roof insulation	1 to 2 years
Wall insulation	3 years
Adding an outside porch	30 years
Double glazing	40 years

Table 3.3 *Payback times for various types of insulation*

Adding double or triple glazing to one's house can be considered to be economic only when windows need to be replaced, or it is desired to reduce noise levels. By putting heavy curtains on windows, and drawing them, almost the same thermal effect as double glazing can be achieved.

Building regulations for Government grants require that *U*-values for houses must not exceed the following: roof - 0.4 W m^{-2} K^{-1}; walls - 1.1 W m^{-2} K^{-1}; floor - 0.6 W m^{-2} K^{-1}.

	U-value/ W m^{-2} K^{-1}
Solid wall (block and plaster)	2.3
Cavity wall (block/cavity/block)	1.6
Cavity wall (block/polystyrene board/block)	0.58
Cavity wall (timber frame with above)	0.46
Window (single glazed, metal frame)	5.7
Window (double glazed - 20 mm airspace, metal frame)	2.8

Table 3.4 *Typical U-values for various types of wall*

It can be seen from the figures in this table that the *U*-value of timber-framed cavity walls is lower than that of traditional concrete walls, and that more heat is lost through windows than through the same area of wall. The table also shows that double glazing halves the rate at which heat is lost through glass.

Building materials are often rated in terms of thermal resistance, the *R* value, which is defined as

$$R = \frac{d}{\lambda}$$

d = thickness of material involved

λ = thermal conductivity

Good insulators have low λ values and high *R* values. An advantage of using thermal resistance is that the *R* value for a number of layers sandwiched together is simply the sum of the *R* values of the separate layers.

$$R = R_1 + R_2 + R_3 \ldots$$

Some other definitions in this area are:

1. Thermal resistance $= \dfrac{1}{\text{thermal conductance}}$

2. Thermal conductance $= \dfrac{\text{thermal conductivity}}{\text{thickness}}$

3. U-value for a structure $= \dfrac{1}{\Sigma R}$

3.4 Experimental Approach

Experiment 3.1
To Show the High Conductivity of Metals

Apparatus
Coin - £1, 2p, or other; small piece of cloth, e.g. linen handkerchief.

Method
1. Wrap the piece of cloth tightly around the coin, stretching the fabric tightly.

2. Light a candle, and arrange for someone to touch the flame against the wrapped coin for a few seconds.

It can be seen that the cloth does not burn, as heat is conducted away by the metal so quickly that the handkerchief cannot get hot enough for it to burn.

Note
The Ingenhousz[6] apparatus, comprising metal rods protruding from a metal container into which is poured boiling water to heat the different materials, is also very useful.

Experiment 3.2
To Show Convection in Liquids

Apparatus
Flask fitted with stopper with 2 holes; bell jar; 2 lengths of glass tubing (one bent as shown in Fig. 3.2); colouring material, e.g. crystals of potassium manganate(VII) (potassium permanganate); retort stands; tripod; wire gauze; Bunsen burner.

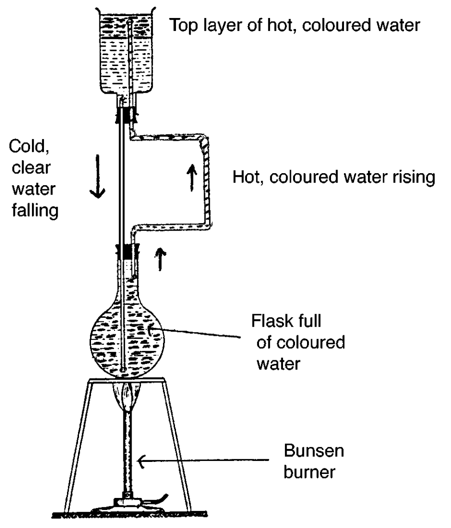


Fig. 3.2 *Convection currents*

Method

1. Place some of the crystals in the lower flask. Fill the apparatus with water and support with retort stands.

2. Heat the water in the lower vessel. Observe that the warm coloured water rises upwards, forming a layer of coloured water at the top of the upper container. After a time cooler clear water will be seen to move downwards. Eventually all liquid will be uniform in colour.

The principle of the domestic hot water system may be demonstrated using the apparatus shown. The lower flask represents the copper cylinder with an immersion heater. The bent tube represents hot water on its way upwards, which might be drawn off in bathrooms, etc.

Note
If a teacher is using a simple and familiar method of demonstrating convection currents (flask of water coloured with $KMnO_4$), it is often useful to add the permanganate by means of glass tubing. Place a length of glass tubing in the flask towards the centre, so that it touches the bottom. Drop one small crystal down through the tube. Any coloured water in the tube can be removed by putting one's finger over the top of the tube and lifting it out. If the flame of the Bunsen burner is kept at a low level it makes it easier to observe the convection currents formed.

[6] Jan Ingenhousz (1730-1799) was a Dutchman who did experiments on the conductivity of different materials.

Experiment 3.3
To Show Convection Currents in Gases

Method (a)

Apparatus
Gas jar; T-shaped piece of strong cardboard (as shown in Fig. 3.3); small piece of candle; taper.

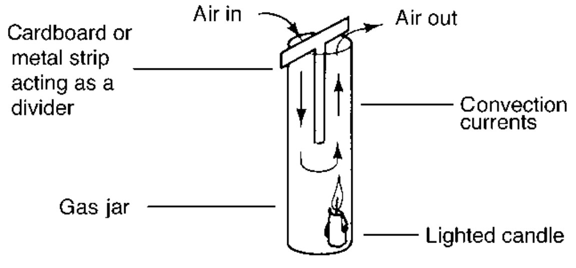


Fig. 3.3 *Convection currents in air*

Method

1. Place the candle in the gas jar and light it using a taper. The candle soon goes out.

2. Re-light it, but now place the cardboard across the gas jar making a central partition. The candle this time will keep burning because a circulation of air is now set up which keeps a steady supply of oxygen entering the gas jar.

Explanation

This resembles what happens when a fire burns in a fireplace. Hot air rising up the chimney creates a draught which draws colder air in from outside, thereby ventilating the room. The same idea was used to introduce a steady stream of fresh air into mines during the eighteenth century. Two shafts were sunk (downcast and upcast) and a fire was lit at the bottom of one of them (provided that no dangerous gas was present). The circulation of air set up ventilated the mine.

Method (b)

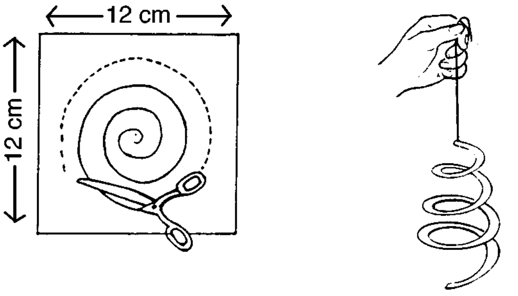


Fig. 3.4 *Convection currents in air*

A simple way to show the presence of convection currents in air is to draw a spiral on a square of paper as shown in Fig. 3.4 and then cut along the spiral as shown. By attaching a piece of cotton thread to the ‘snake’ the presence of convection currents may be investigated. The paper spiral can be held over a hot radiator, near an open fire, above a lighted Bunsen or candle, beside a window, etc., and a spinning motion indicates convection currents. The same effect may be achieved by using a ‘windmill’ of foil, pivoted on a thin piece of wood, held over hot objects. The rotating foil resembles a simple turbine and shows the movements of warm air. This principle is used to give a flickering effect in electric heaters in order to resemble glowing coal fires.

Method (c)

If an upturned paper bag is held with its open end over a heated toaster and then released it will be seen to move upwards with the convection currents set up.

Experiment 3.4
To Show the Cooling Effect of Evaporation

Method (a)
Caution. This experiment should be performed in a fume cupboard.

Apparatus
A small metal can (e.g. calorimeter); wooden block; some volatile liquid - ether, acetone or methylated spirits; long piece of glass or plastic tubing with some rubber tubing attached; a little water in a large beaker.

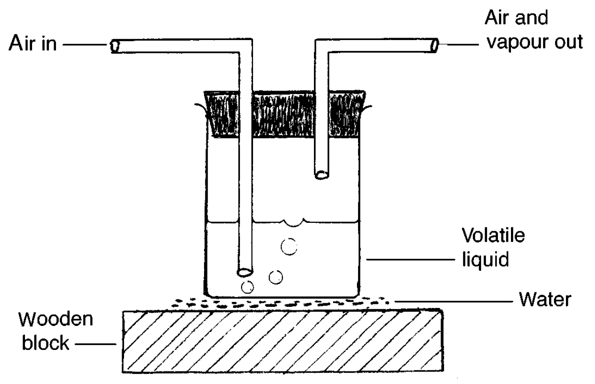


Fig. 3.5

Method

1. Dip the metal can into the beaker of water and, while some water drops are still adhering to it, stand it on a flat wooden block.

2. Pour a little of the volatile liquid into the can.

3. Using the long tube, blow air into the volatile liquid (do not have your face directly over this liquid while blowing; use a bent tube, or maybe a pump, if available).

4. As the volatile liquid evaporates, latent heat is taken from the liquid itself, then from the container and then from the water below it. This water soon freezes, causing the can and block to stick together, giving a firm bond. (Try lifting up the can!)

Method (b)

Caution. This experiment should be performed in a fume cupboard.

Apparatus

As shown in diagram, Fig. 3.6.

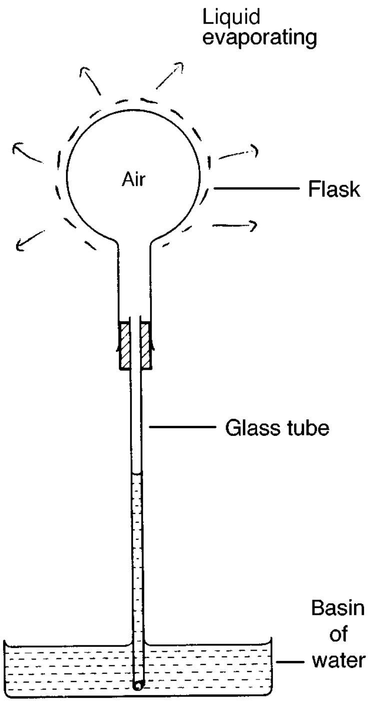


Fig. 3.6

Method

Invert an empty flask attached to a long glass tube (an air thermometer) over a beaker of water, so that the end of the tube dips well into the water. Support by using a retort stand. Pour a small quantity of a volatile liquid over the flask. It evaporates quickly, and in doing so takes the necessary latent heat from the flask and air inside. As the air contracts water is drawn up into the stem, demonstrating the fact that evaporation causes cooling and therefore contraction.

Method (c)

If a small quantity of a volatile liquid is poured on to a piece of cotton wool and dipped quickly on to the back of one's hand, it disappears immediately, leaving the hand feeling cold. The necessary latent heat needed for evaporation is taken from the skin. The faster a liquid evaporates the colder the hand feels.

Note

It is a good idea to rub a bit of cold cream or hand cream on afterwards to prevent skin irritations.

Method (d)

Place some wet tissue around the bulb of a thermometer. Read the temperature. Direct a hand-held hairdryer using the cold air setting (forced convection) on to the bulb. As the water evaporates the temperature registered on the thermometer is seen to drop, showing that latent heat is needed for evaporation.

Alternatively, dip the bulb of a thermometer directly into a volatile liquid, then remove it and watch the mercury fall in the thermometer.

Experiment 3.5

To Demonstrate the Process of Regelation

Apparatus

Length of thin copper wire; a few weights (total 30 - 40 N); newspapers; plastic basins; block of ice made a day or two beforehand (by pouring water into a baking tin, of dimensions approximately 20 x 10 x 4 cm, and allowing it to freeze in a freezer).

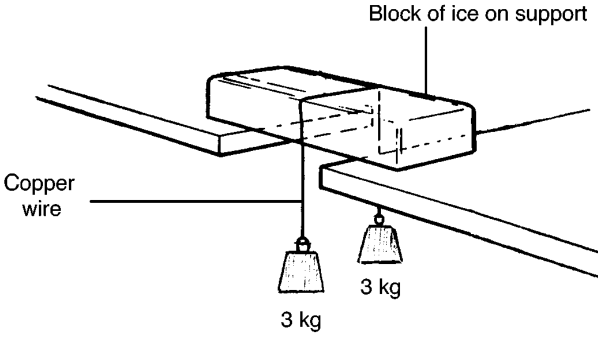


Fig. 3.7

Method

1. Move two school desks quite close together, but leaving a gap of 6-7 cm between them. Lay newspapers on them, and place a basin on the floor underneath the gap to collect any water that might drip during the experiment. Place the block of ice across the gap, resting on and supported by the desks, and insulated by the newspaper, Fig. 3.7.

2. Hang the bare copper wire, with weights hanging from it, on the ice as shown in the diagram. Leave for 30-40 minutes.

3. It will be seen that the wire has passed through the block of ice without, at the same time, dividing it in two. It cuts through, without separating the ice slab into two pieces.

Explanation

The thin copper wire exerts a large pressure on the ice. Due to its small area and large weight, the pressure is great, ($p = \frac{F}{A}$) so it lowers the freezing point (ice expands when it freezes, so increasing the pressure lowers the freezing point) to a temperature below 0 °C. The wire now sinks through the water formed. After the wire has passed through a section of the block, the water above it re-freezes (because it is no longer under pressure). This re-freezing is known as regelation.

Notes

1. As mentioned above, the wire must be narrow to give large pressure.
2. The wire must be made of material which is a good conductor of heat - nylon string will not work. When water above the copper wire freezes, it releases its latent heat which is conducted through the wire and helps to melt ice directly below it.

Experiment 3.6

To Show the Effect of Dissolved Solids on the Freezing Point

Apparatus

Ice-cube; beaker of water; length of cotton string; salt (NaCl).

Method

1. Place an ice-cube in a beaker of water. (Ask students if any of them might suggest a method whereby the ice-cube could be removed from the water without lifting it by one’s fingers, using cutlery of any kind, or spilling it out. They may or may not offer suggestions.)

2. Now wet the cotton string and sprinkle salt along each side of it. After a few seconds it will be possible to lift the ice out with the string.

Explanation

Salt lowers the freezing point of water, and so causes it to melt. This takes place in the vicinity of the string. Latent heat is drawn from the ice and the wet string. The string sinks into this newly formed water. When the water re-freezes the string is now embedded firmly enough in it to support the weight of the ice-cube as it is lifted out.

Experiment 3.7
To Show the Effect of Different Surfaces on Radiation

Apparatus
Leslie cube (available from laboratory suppliers); detector, e.g. thermopile (a number of thermocouples connected in series, with a cone to focus the radiation onto the thermocouples) with galvanometer, or differential air thermometer, or thermometer whose bulb is blackened.

The Leslie cube is a hollow metal cube of length of side c. 13 cm, with a lid, through which boiling water can be poured. One ought to keep the water boiling, i.e. keep a Bunsen or heater under it, otherwise as the thermopile is moved around the surfaces being compared are not at the same temperature. Leslie's original cube was made of copper but present day models are usually made of tin. The four faces have different finishes; one is rough matt black, one is dull white, one may be painted grey or left untreated, while the fourth is always highly polished. Sometimes two opposite faces may be identical, e.g. A and C dull black, with B and D highly polished. When the cube is filled with boiling water it will radiate energy, and this can be detected by some absorbing object placed in front of it.

Method
1. Place the detector about 3 cm away from each face of the cube in turn, noting the reading each time while keeping the same distance. It is found that the rough, matt black surface gives the highest reading, with the grey face next, then the white, then the highly polished face the lowest figure.
2. It is not just the black colour that is important; the finish or texture is also a factor, as a rough black surface radiates much more than a smooth black one.

Note
It must be remembered that the amount of radiation emitted depends on temperature, area, and the nature of the surface ($P = \sigma eAT^4$, cf. p. 37).

When Leslie performed this experiment in Edinburgh in 1804 he used a differential air thermometer to detect the emitted radiation. This is simply two similar glass bulbs connected by a narrow tube containing light oil or coloured water, Fig. 3.8. By placing the cube between the bulbs, more radiation from the black face produced greater expansion of air in one of the bulbs. This pushed the liquid downwards on that side and upwards on the other.

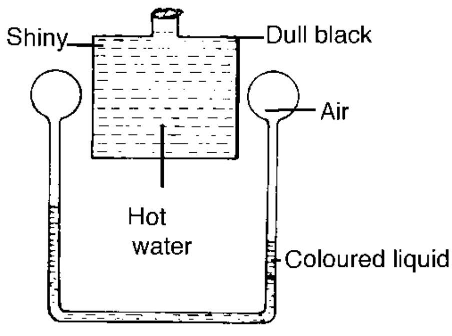


Fig. 3.8 *Differential air thermometer*

An infrared thermometer called a Digitron Infrared Detector is now available to measure the degree of radiation, or simply a mercury thermometer (graduated to read 0.2 °C and with a blackened bulb) will give reasonable results. The latter was used for the first time in 1800 by Herschel. He placed two thermometers in strong sunlight, one left as it was (shiny), the other coated with lampblack, and he noticed that the latter gave a higher reading.

Experiment 3.8
To Show that Dark Materials are Good Absorbers of Radiation

Method
Place two flasks directly underneath two 100 W bulbs or two infrared lamps. Each flask should contain a stopper with thermometer, as shown in Fig. 3.9. The bottom of one flask is left untreated, the bottom of the other one is painted black. Switch on the bulbs and leave the flasks in position for 20-30 minutes. Record the temperature as shown on each thermometer. It will be noticed that the blackened flask shows a higher reading, proving that black surfaces are better absorbers of radiant energy.

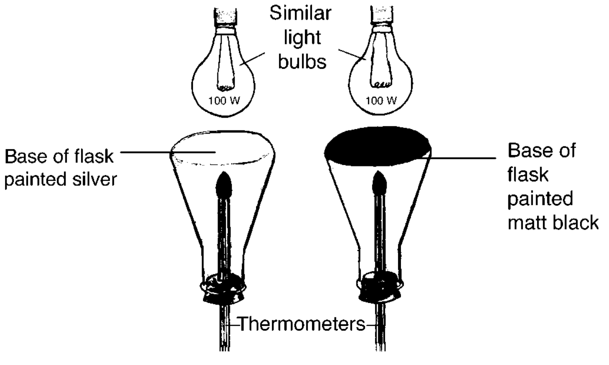


Fig. 3.9

Note
'St. Joan of Arc's' experiment using the backs of your hands as detectors of radiation from the underside of a heated gauze using blackened cooking foil over one hand and shiny foil over the other is also very effective.

Experiment 3.9
To Test the Insulating Properties of Different Materials

Method (a)
This is best done as a class experiment.

Apparatus
A number of small cans (e.g. calorimeters, all of the same size and made of the same material); the same number of large beakers; hot water (c. 90 °C); thermometers; different insulating materials, e.g. wool, cotton wool, small pieces of carpet, straw, felt, newspaper, foam, aeroboard, fibreglass, fur, feathers and air.

Method
Place each metal can inside a larger beaker, standing on a block of aeroboard. Use another block of aeroboard as a lid (punch a hole to let a thermometer pass through). Pack the space between each beaker and can with one of the above materials. Pour very hot water into each can and cover with a lid. Insert a thermometer and record the initial temperature in each case. Allow to stand for 15 minutes. Note the temperature readings in each case and compare values. It will be seen that some materials are more effective in preventing heat flow than others, although all are classified as insulators.

Method (b)

Apparatus
3 calorimeters with lids; pieces of aeroboard; strips of woollen material. Place the calorimeters on pieces of aeroboard. Wrap the outside of one tightly with strips of a woollen material. (Use rubber bands to keep the fabric layers close together.) For the second one, arrange to have the material loosely wrapped, thereby trapping plenty of air. Leave the third one unwrapped. Pour the same quantity of boiling water into each. Leave to stand for 15 minutes and again check the temperatures. It will be observed that the loosely wrapped calorimeter has the highest temperature, showing that trapped air is a good insulator.

Method (c)
Place two beakers on an insulating stand. Pour equal quantities of hot water at the same temperature into each. Record the temperature. Invert a small paper bag over one beaker. Cover the other with a sheet of paper. After 10 minutes, record the temperature in each by inserting a thermometer. It will be noticed that the beaker covered by the paper bag using air as an insulator is hotter than the other one.

3.5 Applications

Conduction

Metals are generally good conductors of heat, and non-metals are not. Liquids and gases are poor conductors. Students will remember the 'ice-cube experiment' which demonstrates the fact that water is a poor conductor of heat. Heat transfer in fluids is mainly by convection rather than conduction, except in the case of liquid metals.

Materials like copper, aluminium and stainless steel which are good conductors are used in cooking utensils; the handles of these utensils are made of poor conductors like wood or plastic.

A metal skewer in a kebab on a barbecue not only holds the food pieces together, but helps the heat pass through the food also.

We know from experience not to stir boiling soup with a metal spoon, as the heat energy would travel very quickly through it and burn our hand. Instead we use a wooden spoon.

In the electronics industry some heat-sensitive components are often mounted on a metal tray called a 'heat sink'. This removes surplus heat which would damage semiconductor devices. A heat sink is designed to have the maximum surface area and is usually painted matt black.

A blacksmith uses a long tongs to hold a red-hot horseshoe.

Convection

Sea breezes are caused by differential warming of the land and the sea during the day. The land, having a lower specific heat capacity than the water, becomes warmer, the warm air above the land rises and the cooler air over the sea moves in to take its place. At night the situation is reversed, with the land cooling more quickly than the sea, producing land breezes.

Winds are caused because equatorial regions are heated more than polar regions by the sun, setting up trade winds, etc., as colder air moves in to replace warmer, rising air.

The Gulf Stream is a large-scale example of a convection current, involving a transfer of warm water like a giant river flowing through the ocean. Because of it, Ireland, the UK and France enjoy relatively mild winters even though they lie as far north as Labrador, which has very cold winters.

Birds and gliders use convection currents over warm land (known as thermals) to help them soar into the air. The thermals give them a certain amount of lift to move them upwards without effort.

THE IRISH TIMES

WEATHER EYE

The Heat for Heights

ONE Friday in September 1991, a glider pilot, Brian Connolly from Bagenalstown, soared in his flimsy aircraft to a height of 24,000 feet above Dingle Bay in Co Kerry. His flight, which lasted four hours earned him the coveted international "Diamond" awarded for a gain in altitude of 5000 metres or more, and was a clear record for an Irish glider at the time: indeed the record may still stand. The question is, though – how did he do it?

There are two common ways in which soaring flight can be achieved. "Ridge soaring" is the most basic and the easiest, and uses the fact that air approaching a sizeable range of mountains does not go around the obstacle, but climbs over it, forced upwards in a gentle sliding motion. On the windward side of such an obstacle, and for some considerable altitude above it, there is a continuously rising body of air which can be used systematically by glider pilots to achieve a gain in height. It is an advantage when the mountain-range lies at right angles to the wind direction, so that the pilot has a long "band" of lift at his or her disposal.

On the day that Brian Connolly performed his feat in Kerry, a steady stream of wind was flowing northwards from Inch Strand, where the aircraft had been launched, to rise over the Dingle mountains. The lift thus provided carried the engineless aircraft far above the clouds, to heights normally associated only with commercial passenger-carrying jet aeroplanes.

The other common source of lift for gliders is the "thermal". Thermals are invisible currents of ascending warm air, which carry heat upwards through the atmosphere from patches of ground that have been heated by the sun. They are usually present underneath cumulus clouds, but they frequently occur well away from clouds of any sort: when a breeze is blowing, they do not remain static over the one spot, but are carried bodily downwind by the movement of the air.

A glider pilot may often seek to harness both sources of lift, perhaps availing first of the "ridge lift" from adjacent hills, while waiting for thermals to pass nearby as they are carried downwind through the valleys. The strongest lift occurs near the central core of a thermal, so to take maximum advantage of it, a pilot must fly round and round near the core, gaining height all the time in an exhilarating upward spiral – like Steve McQueen drawing circles in the sky in *The Thomas Crown Affair,* if you remember.

Brendan McWilliams

Sometimes they can stay aloft for several hours using these currents. Hot air balloons are also helped to stay aloft by convection currents. Hot air balloons first took to the air in 1784 piloted by the Montgolfier brothers in France.

A domestic water central heating system works by convection. The ‘radiators’ used in this system are incorrectly named, as heat energy is given off mainly by convection, rather than radiation. They are really convector heaters and act as heat exchangers.

Natives of countries where the average temperature is high wear loose flowing clothing, so that heat energy can be lost from the body by convection.

The flickering effect in an imitation coal or log fire is caused by convection. A bulb warms air which causes a disc to spin around, giving the illusion of flames.

In a bonfire, flames and sparks may be seen moving upwards, indicating the presence of convection currents.

Years ago, coal mines were ventilated by convection currents produced by a fire. Two shafts were sunk to assist in the movement of air. This principle may be demonstrated easily in the lab (see section 3.4).

An ice box is positioned at the top of a refrigerator, and the butter compartment is found high up on the door, to take advantage of the cooling effects of convection currents within the refrigerator.

Refrigerators have open shelves to allow circulation of air through them. For this reason, it is not wise to pack them too tightly with food. Similarly, a hot press should not have too many clothes filling it. The slotted shelves are there to allow warm air from the hot water cylinder move upwards to ‘air' the clothes.

Thicker insulation is needed in the attic of a house than in the walls of a house, because of the upward movement of warm air.

The air at the top of a hot oven is at a higher temperature than that at the bottom. If two apple tarts of the same size are placed in a hot oven, one near the top and the other near the bottom, the top one will be cooked before the lower one.

Convection currents cause ventilation of a room. As warm air ascends cold air is drawn in from outside to replace it, changing the air in the room continuously.

Black marks are sometimes noticed on a ceiling above a lamp. These are caused by dust particles being carried upwards in convection currents produced by the hot lamp.

Radiation

The ordinary domestic coal fire transmits heat by radiation. It does not do so by convection because the direction of smoke issuing from the fire indicates the presence of convection currents rising upwards. Neither can it be due to conduction, as air is a poor conductor of heat.

Pipes on solar panels are dark coloured so they will absorb the greatest possible amount of solar energy (because dark-coloured objects absorb more radiation than do bright-coloured objects).

People may wear white or light coloured clothes in summer to keep cool.

Houses situated in hot climates are usually painted white.

Aluminium paint is sprayed onto large oil storage tanks, or onto factory roofs.

Fire fighters wear suits of shiny reflective material when moving into a blazing building.

A sheet of foil attached to a window which faces the sun is effective in keeping a room cool in very hot weather.

Large tea or coffee urns in cafes and restaurants are highly polished so that they do not radiate much heat, thereby remaining hot much longer.

Marathon runners at the end of a race, accident victims suffering from shock, and premature babies are wrapped in blankets made from reflective material, so as to conserve their bodies' heat.

The Crookes radiometer works because one face of each vane is black and the other is shiny. As light energy falls on the vanes, the dark coloured faces absorb more radiation than the bright faces. As a result, air molecules rebound with greater velocity from the slightly hotter black faces and cause the spinning effect. The bulb is partly evacuated to assist in this.

In some cases it is desirable to have objects lose heat energy and so they are painted black. For example, the cooling fins at the back of a household refrigerator are blackened in order to radiate the maximum amount of energy.

Burglar alarms may be activated by the infrared radiation emanating from a burglar's body.

The sun's rays may be focused using a converging lens, on to paper, dry sticks or even the head of a match, resulting in ignition.

Frost is unlikely to form on cloudy nights due to cloud cover, which prevents temperatures becoming too low.

The curved shiny surface at the back of a radiant electric fire reflects radiated heat outwards. A person who is sitting in front of an electric fire on a cold day will notice how quickly they feel colder if somebody else stands directly between them and the fire.

Examples of insulation using air

People wear clothes to keep warm. The clothes themselves do *not* supply heat energy, they just help to retain that which the body produces. The wearing of clothes in thin layers is more effective, as more air is trapped. It is estimated that as much as 80% of the chemical energy of food which the body converts appears as thermal energy, and a large part of this would be lost to the surroundings were it not for the shield of insulation that clothes provide. Wool, because of its tiny trapped pockets of air, is a very good insulator. So too are string vests when worn in conjunction with windproof outer clothing. Cotton wool in sleeping bags, and the feathers or down in duvets, work on the same principle.

The insulation provided by a duvet is rated in terms of its tog value - the higher the figure given the better the insulation. The tog value is 'the temperature difference across a structure when the rate of loss of heat is 1 W m^{-2}. A new idea in duvets is the '3 in 1' or 'all year' quilt. It actually comprises 2 quilts, a summer one (4.5 tog), and a spring/autumn one (9 tog). These two may then be combined using Velcro strips to give a warm winter 13.5 tog quilt.

Tea cosies and indeed egg cosies are knitted in large open stitch to retain air.

Birds fluff out their feathers in cold weather to hold air, and therefore look fatter than they actually are.

Double glazing in windows reduces heat energy losses by having a layer of air between the two sheets of glass in every pane.

Snow might be regarded as an unlikely insulator. Gardeners know that small plants and seedlings are often protected from colder weather above the snow. And, of course, igloos are built of snow. There can be up to 20 K of a temperature difference between the inside of an igloo and the cold weather outside (due to air trapped between the layers of snow). Mountain climbers caught in a blizzard on snow-covered slopes often dig snow holes to shelter in.

Eskimos often wear fur coats with the fur next to the body, or they may use fur at the collar and cuffs of another coat to keep them warm.

Hot food, e.g. fish and chips, or cold food such as ice-cream is sometimes wrapped in sheets of paper to keep it hot or cold as the case may be. Both the sheets of paper and the air trapped between them are insulators.

In the making of the popular dessert 'baked alaska' a block of ice-cream, surrounded by meringue piled on and around it, is placed in a hot oven (230 °C) for about 3 minutes. The ice-cream still comes out of the oven quite cold after that time because the meringue has large quantities of air beaten in with the sugar and egg-white used, and it is this that insulates the ice-cream.

Other examples of insulation

Glass wool or fibreglass is used in attic insulation - a greater thickness is needed here than in the walls of the house because of the upward movement of the convection currents. Building regulations drawn up in 1992 recommended a minimum thickness of 150 mm of this material. It is never laid directly under the water tank in the attic, so that the water is less likely to freeze in cold weather. Lagging on pipes is widely used, and a lagging jacket on a hot water cylinder in the hot press is recommended.

Underlay underneath carpets helps improve insulation as well as helping to prolong the life of the carpet.

Draught excluders around badly fitting window frames and doors help to reduce energy losses. Because an appreciable amount of heat energy is lost from windows, heavy lined curtains drawn across in cold weather help to prevent heat loss from a room. Shutters on the outside have the same effect.

Cork or laminated cardboard table mats protect a polished wooden table from hot plates.

Oven gloves make it possible for roasting tins, etc., to be lifted out of a hot oven.

Gas/electric cookers and refrigerators have a thick layer of insulation around them, which is covered by a layer of aluminium foil.

Insulating grips on bicycle handlebars make them more comfortable to hold either in summer or in winter.

Wooden houses are warmer to live in in winter and cooler in hot summer weather and the same may be said of a house which has a thatched roof.

A cup of coffee keeps warm for a longer period of time if cream is added immediately after pouring, as the cream acts as an insulator.

Silk is a good insulator; members of the 1975 Mt Everest expedition team wore it in underwear.

Divers use water as an insulator in a wet suit. A small quantity of water forms an insulating layer next to the body. Body fat itself is also an insulator.

Polar bears, seals and whales can withstand extremely low temperatures because of their thick layer of fat or blubber.

During World War I, when energy was scarce, a 'hay box' (or Norwegian cooker) was used to cook food. It relied mainly on the insulating materials it contained - hay, straw, felt and still air. Meat and vegetables were quickly fried together and then quickly transferred to a casserole dish which was placed in the hay box. It was left to simmer for hours without any further heating being supplied.

Glass blowers in crystal factories use the fact that glass is a poor conductor of heat when they shape and manipulate molten glass after it is removed from the furnace.

Insulating materials are of special importance in space travel. The frictional heating that occurs when a space capsule returns to earth after a space mission may produce a rise in temperature of the order of thousands of kelvins. The nose of the spacecraft as it enters the earth's atmosphere experiences most of the heating effect so this is covered with a heat shield in the form of a honeycombed glass resin material. This layer melts and drips off, taking a large quantity of the heat energy developed with it (due to its large latent heat), thus protecting the crew.

From the list of applications above, one could be forgiven for thinking that insulators are always of benefit to us. However, there are many occasions when an insulating layer is undesirable. Two such cases are the following.

A layer of scale or ‘fur’ on the inside of some kettles and, more especially boilers, is an insulator. It tends to reduce the rate of transfer of heat and therefore the efficiency of the kettle or boiler.

The thin layer of air that remains between the bottom of a cooking utensil, e.g. a saucepan, on a hot plate also hinders heat transfer. This may be compensated for by using heavy pots and pans which make better contact with the stove and reduce the layer of air present.

CHAPTER 4

QUANTITY OF HEAT

4.1 Background

Since earliest times human beings have experienced the sensations received from touching hot and cold objects. They have made use of the effects of heat over many generations, but it was only about a hundred and fifty years ago that the true nature of heat was understood.

We know today that what is commonly called heat (meaning energy), but what should really be called internal energy, is simply one of many forms of energy. It is the commonest form available to us and most other forms of energy eventually revert to it. However, in the eighteenth century, people had strange ideas of what happened when things burned. They believed that burning involved the escape of a substance called phlogiston from the burning matter and that only matter which contained phlogiston could burn. This idea was known as the phlogiston theory (from the Greek word meaning flame) and it was proposed by a German chemist, **Georg Stahl** (1660-1734), in 1697. It was the well-known French scientist **Antoine Laurent Lavoisier** (1743-1794) who disproved it. Lavoisier showed that substances actually combine with oxygen from the air as they burn and thereby increase in mass, contradicting the phlogiston theory. Incidentally, it was Lavoisier who first proved that water is made up of hydrogen and oxygen, and that air is a mixture of nitrogen and oxygen. However, these very important discoveries did nothing to prevent his being guillotined, in Paris, on the orders of the Revolutionary Convention in the confusion following the French Revolution, probably because he had been a servant of the previous government.

At the beginning of the eighteenth century a Dutch physician, **Hermann Boerhaave** (1668-1738), described heat as a fluid that passed from one substance to another, and this idea was taken up by **Joseph Black** (1728-1799) in 1760, working in Glasgow University. He called the fluid caloric, and imagined it to be an invisible, weightless fluid, and suggested that it flowed into objects that were being heated. (It is from this that the unit calorie[1] comes; 1 cal = 4.2 joule or 1 kilocal = 4200 joule.)

One of the first people to suspect there was a link between heat and energy was the well-known English chemist **Sir Humphry Davy** (1778-1829) (of the miner's safety lamp fame). In 1797 he rubbed pieces of ice together (fastened by wire), showing that friction was responsible for melting the ice. But it was really in the following year, 1798, that an American, **Benjamin Thompson** (1753-1814), known as Count Rumford, began to suspect that there was a link between heat and energy. Rumford was War Minister in Bavaria and, while based in Munich, became interested in the boring of cannon in the army's arsenal, using horse-driven drills. He noticed that the metallic chips produced as the cannon were made became very hot and felt there was indeed a link between the friction produced in drilling and the hot metal. He further observed that a blunt borer produced fewer chips and more heat than a sharp one, so convincing him that the 'caloric' produced was not proportional to the quantity of material from which it was supposed to come. Rumford succeeded in getting water surrounding a borer in a gun barrel to boil. Eighteen pounds (8.2 kg) of water could be brought to boiling

[1] The calorie used in nutrition is the kcal.

point in less than three minutes. His findings were published in a paper at the time.

Over forty years later, in 1842, a German physician, **Julius Robert Mayer** (1814-1878), conducted experiments which linked heat produced to mechanical work and he wrote an account of his findings. (The same man also discovered that people in cold climates needed more food than those living in warm countries to keep the body temperature at a normal level.) But it was in the following years, 1843 to 1847, that a rich Manchester brewer, **James Prescott Joule** (1818-1889), started a series of experiments that proved beyond all doubt that mechanical work and heat energy were related. Somewhat more attention was paid to his work than to Mayer's since he was a physicist and had done more precise experimental work, but it was only after his meeting with Lord Kelvin that his experiments were taken seriously. As a young man, Joule suffered from ill health and his father, determined to encourage his son's interest in science, built a laboratory for him. The young man was educated at home for some years, and at one stage had the eminent chemist **John Dalton** (1766-1844) as his tutor. Some of Joule's most famous experiments, conducted over 5 years, used falling weights to churn up liquids (water, oil and mercury), using a paddle wheel system. The potential energy of the weights was converted to kinetic energy in the brass paddle, becoming internal energy of the water and the surrounding calorimeter and producing a rise in temperature. He found that a certain amount of mechanical work always produced the same amount of heat energy. The ratio of work done to heat generated was constant. There was a fixed rate of exchange. This ratio he called the mechanical equivalent of heat. He submitted his findings in a paper in Oxford in 1843. Overall, Joule spent thirty five years converting various kinds of work into heat. The apparatus from one of his experiments is on display in the Deutsches Museum in Munich, as indeed are some of Mayer's. It is not surprising, in the light of Joule's work, that the unit of work/energy is named after him. But it must also be said that in 1847 another well-known scientist, **Hermann von Helmholtz** (1821-1894), wrote a paper dealing with the transformation of energy. So by 1850 the principle of conservation of energy, one of the most basic ideas in physics, had been accepted.

Another well-known figure in the study of heat was a Frenchman, **Sadi Carnot** (1796-1832). His father was Minister of War under Napoleon, and encouraged the young Sadi to study physics and mathematics, and become an engineer. Carnot served for a time in the French army but his main interest was in thermodynamics, and in particular the study of heat engines. His work was largely theoretical and his ideas were presented in a paper entitled 'Reflections on the Motive Power of Heat'. Carnot accepted the caloric theory of heat. He was a quiet, retiring man and, although he died at the early age of 36 from cholera, is looked on as the father of thermodynamics. His name lives on in the Carnot cycle (see Appendix 4.2). Carnot's ideas were developed by a German theoretical physicist **Rudolf Gottlieb Clausius** (1822-1888), and also by **William Thomson** (1824-1907). In 1857 Thomson said of Carnot, 'He was the most profound thinker in thermodynamic philosophy in the first twenty years of the nineteenth century'.

William Thomson, later to become Lord Kelvin, was a mathematician, a physicist and an engineer. He was born in Belfast in 1824, the son of a professor of mathematics. He attended his father's lectures at the age of eight and attended the University of Glasgow at age ten. He published his first mathematics paper at age 16, and was appointed Professor of Natural Philosophy there at age 22 (a post he held for 53 years). He met Joule at the 1847 meeting of the British Association for the Advancement of Science in Oxford, and pointed out to his fellow scientists the importance of Joule's work. He was knighted in 1866, made a peer in 1892, choosing as his title Baron Kelvin of Largs, and retired in 1899. Altogether Kelvin published some 600 scientific papers and the SI unit of temperature, the kelvin (K), is named after him. The absolute scale of temperature also is called the Kelvin scale, and his name appears again in the phenomenon that he and Joule discovered, the Joule-Kelvin effect (see section 1.2, p 3). In 1848

Kelvin suggested the lowest temperature that could be reached was -273.15 $^{\circ}$C, 0 K or absolute zero. He accepted the principle of conservation of energy, concluded that heat only flows from a hot object to a colder one, suggested the idea of a heat pump, designed transatlantic cables and telegraphic instruments, wrote in detail on electromagnetism, suggested that the Niagara Falls could be used as a source of power, etc., etc. He died in 1907 and for his contributions to science was buried, next to Newton, in Westminster Abbey, London.

In the study of heat, one name which is of particular interest to Irish people is that of the physicist **John Tyndall** (1820-1893). He was born in Leighlinbridge, Co. Carlow, and gives his name to the 'Tyndall effect'. This deals with the scattering of light by dust particles in the air (the reason that the sky appears blue). His earliest work dealt with the magnetic properties of crystals and later he investigated the ability of various gases to absorb and radiate heat. In 1863 he published a paper on 'Heat as a Mode of Motion', and supported James Clerk Maxwell's theory that heat is the 'motion of molecules'. He also investigated the relationship between heat and humidity, which is of use to weather forecasters. In 1881 he showed that germ-free air does not cause food decay and he published in total 145 papers and 16 books. He was a friend and colleague of Michael Faraday, and died in Surrey in 1893.

The idea we have of heat today is that it is not a liquid, or a substance, but the form of energy transferred from a hot object to a cooler one as the result of the temperature difference between them. Heat flow has changed from being considered as 'caloric transfer' to energy transfer.

4.2 Do You Know

The high specific heat capacity of water ensures that the sea remains cooler than the land in warm weather.

This helps account for the popularity of seaside resorts and swimming pools, where people enjoy a cooling dip. Similarly, water in rock pools is rarely as hot as the surrounding rocks themselves, or as sand on a beach, despite all being subjected to the same amount of heating by the sun. This fact also accounts for sea breezes (see p. 48).

Water, because of its large specific heat capacity, is used in many types of heat exchanger.

Such heat exchangers range from industrial processes and central heating systems to the simple hot water bottle. One familiar example is the use of water in the cooling system of a car. It removes much of the heat energy that would otherwise cause damage to the engine.

Because so much of the human body is composed of water our muscles do not overheat.

If we engage in exercise or strenuous physical work, the body temperature does not rise excessively because of the body's high heat capacity. This, of course, is helped by the process of perspiration (see latent heat). The specific heat capacity of the human body (average) is taken to be 3470 J kg^{-1} K^{-1}.

The jam in a 'roly poly' pudding, the apple sauce in an apple pie or the mince in a mince pie feel much hotter than the surrounding pastry.

This is again due to the high specific heat capacity of water. The water content in the filling is higher than in the pastry and the thermal conductivity of the pastry is low.

What is a 'bomb calorimeter'?

It is a strong metal container, usually made of steel, which is used to measure the calorific value of fuels (the heat of combustion per kg). The fuel under test is placed in a crucible in a chamber which contains oxygen under pressure (to allow complete combustion), and is heated electrically. Surrounding the ignition chamber is a vessel containing water, into which a thermometer is inserted. A thick outer layer of lagging ensures that very little heat energy is lost to the surroundings. Calorific values are given in J kg^{-1}. Some sample values are given in Table 4.1.

Substance	Calorific value/J kg^{-1}
Natural gas	5.5×10^7
Coal	3×10^7
Petrol and fuel oil	4.6×10^7
Wood	1.4×10^7

Table 4.1 *Calorific values*

4.3 Conceptual Approach

Specific Heat Capacity

In the late 1750s Professor **Joseph Black** (1728-1799) of Glasgow University heated equal volumes of water and mercury in an oven for a certain period of time, and then checked their temperatures. He noted that the mercury was much hotter than the water, although they had both been heated in the same oven for the same length of time. It became clear to Black that different materials had different capacities to absorb and retain what he called the 'caloric fluid'.

To explain what had happened, Black used the analogy of different people's ability to consume and 'hold' liquor. Someone with a large alcohol capacity could take many drinks without much behavioural change, whereas the behaviour of another person with a lower alcoholic tolerance would change much more dramatically after imbibing just one or two drinks.

From Table 4.2, it is clear that materials differ from one another in the quantity of heat they need to warm them. The Swedish scientist **Wilcke** (1732-1796) coined the term 'heat capacity'. He dropped different hot objects into cold water and examined the temperature increases which resulted. Substances like copper show a greater change in temperature as they are heated for the same amount of time than do others. Water, for instance, seems to have a 'thirst' or 'appetite' for heat energy in that it needs more than ten times as much heat as copper in order to achieve a similar temperature rise. This is what we mean by specific heat capacity. The symbol '*c*' is used and the units are joules per kilogram per kelvin (J kg^{-1} K^{-1}). Liquids generally have higher specific heat capacities than solids, and gases have higher values still. Because gases tend to expand so much when heated, and because an expanding gas does work (which is equivalent to a certain quantity of heat) against the external pressure as it expands, the specific heat capacity will be greater if the pressure is kept constant. Hence a gas has an infinite number of specific heat capacities depending upon the conditions. Two are of special importance, viz. the specific heat capacity at constant volume (c_V) and the specific heat capacity at constant pressure (c_p). Approximate specific heat capacities are given in Table 4.2 below.

Substance	*c*/J kg^{-1} K^{-1}	Substance	*c*/J kg^{-1} K^{-1}
Aluminium	900	Ice	2100
Iron/steel	450	Wood	1700
Copper	390	Nylon	1700
Brass	380	Rubber	1700
Zinc	380	Marble	880
Silver	230	Concrete	850
Mercury	140	Granite	840
Tungsten	135	Sand	800
Platinum	130	Glass	670
Lead	130	Carbon	500
Hydrogen	14000	Ethanol	2400
Air	718	Paraffin	2100
Nitrogen	1040	Water	4186
Steam	2000	Sea water	3900

Table 4.2 *Specific heat capacities*

Heat Capacity

While specific heat capacity refers to the material from which an object is made, e.g. aluminium, heat capacity refers to a particular object (the whole bulk of matter in an object, e.g. a calorimeter). In other words, specific heat capacity is a characteristic of a substance, whereas heat capacity relates to a particular object. Therefore, the heat capacity of an object depends on both its *mass* and the *nature* of the substance or substances from which it is made. The symbol used is '*C*' and the units are joules per kelvin (J K^{-1}). The heat capacity of an object is defined as the heat energy needed to increase the temperature of the whole object by 1 K.

Latent Heat

When a substance changes phase from solid to liquid or from liquid to gas, energy is required. The heat energy which changes the state of a substance, without changing its temperature, is called latent heat (L). The word latent was first used by Joseph Black but Wilcke also independently observed the same phenomenon around the same time. Both worked with ice and snow during their investigations. It is sometimes called the heat of transformation. Latent heat means 'hidden' or concealed energy, in the sense that it does not show up on a thermometer. Usually, when a substance is heated, its temperature rises. However, when matter changes phase from solid to liquid, or liquid to gas, there is no actual temperature rise. The temperature of water, at normal atmospheric pressure, never exceeds 100 °C and a given mass of water at 0 °C has more heat energy than the same mass of ice at 0 °C.

Latent heat is energy used for loosening or breaking bonds between molecules and not for raising temperature. There are two forms, viz. latent heat of fusion, L_F , and latent heat of vaporisation, L_V. The specific latent heat, l, refers to unit mass of a substance, i.e. it is the quantity of energy required to change the state of 1 kg of that substance. The specific latent heat of fusion of ice is 0.336 MJ kg^{-1}. The specific latent heat of vaporisation of water is 2.26 MJ kg^{-1}. A simple calculation shows that it requires about 5 times more energy for 1 kg of water at 100 °C to boil dry than is required to raise the temperature of that water from 0 °C to 100 °C. At this point in our discussion, we may reflect on why indeed such a large quantity of heat energy is required to change the state of a substance. Let us refer back to section 3.1, States of Matter, and we will recall that in the case of changing from a solid to a liquid, energy is needed to:

1. increase the distances between atoms or molecules, pulling them apart;

2. reduce the number of bonds between neighbouring atoms or molecules so that they move farther apart and become a liquid.

When liquid changes state to become a vapour, energy is required for two reasons.

1. The term 'internal work' describes that work required to separate atoms and molecules.

2. 'External work' is required in pushing back the surrounding atmosphere to allow space for the atoms/molecules to escape as vapour.

The greater part of the energy is used in separating the molecules and only 10% is used in pushing back the atmosphere.

Vaporisation requires more energy than fusion, and the figures bear this out; for example, for water, l_F is only about 15% of l_V. Much less energy is necessary to separate the molecules present in ice to form water than is required to liberate the molecules present in water as steam is formed.

These figures do not refer simply to the energy which is absorbed in the transformation of a solid to a liquid and of a liquid to a gas. In the reversal of the sequence these figures also apply. For example, when 1 kg of water at 0 °C freezes to become ice at 0 °C it yields 0.336 MJ of energy. Similarly, with regard to condensation, we can say that when 1 kg of steam at 100 °C turns into water at 100 °C 2.26 MJ of energy is released. (The mandatory experiments to measure the specific latent heat of fusion of ice and vaporisation of water are covered in section 4.5.)

4.4 Applications

Specific Heat Capacity

Some electrical heaters contain oil which has a high boiling point and a relatively high specific heat capacity. Once the oil has been heated it retains the thermal energy for quite a while, cooling down only slowly.

In storage heaters, a core of high-density thermal blocks (like very hard building blocks), made from a material known as Feolite, is heated by electric elements during the night. These blocks are preferred to water, despite having a lower specific heat capacity, because they are more dense (and thus more compact) and do not leak. They are heated during the night using low-cost (off-peak) electricity and release the stored thermal energy during the day. The core is wrapped in an efficient insulating material within a slim metal casing. Most heaters have two controls - one to set the overnight input of electricity, another to control the output from the heater. Often storage heaters have a fan included which passes air over the heated bricks before expelling it into the room, thus assisting the heating process. Storage heaters maintain a steady room air temperature over long periods, warming the entire building fabric and reducing the risk of condensation. In some models, input control is efficient and automatically controlled by means of a room temperature sensor. It needs little or no adjustment once set. Output control, using a damper, is by an arrangement that allows some of the room air to pass through the storage core in order to give a boost to room temperature when required during the evening. Storage heater loadings range from 1.7 kW to 3.5 kW according to model. Storage heaters are a very popular form of home heating, providing steady warmth for living rooms, halls and landings. They are also ideal in homes where people are at home for long periods of the day.

The method of floor warming is suitable for new homes. This storage heating system, which makes use of off-peak electricity, can be used in new dwellings or those undergoing major conversion. The solid concrete floor in which the heating elements are laid provides continuous overall gentle heating which is safe and draught free. As well as providing radiant heat, air warmed by the floor rises and circulates by convection. Under-floor heating is invisible, silent and takes up no space in the room. There are no moving parts, so it is extremely reliable. Before the heating is installed the floor must be insulated above the damp course with 50 mm thick slabs of polyurethane or similar material. Control of the system is by a simple room thermostat which can also be linked to an outside weather sensing thermostat to provide more sophisticated control.

Historically, the same concept was used in warming pans. These consisted of pans containing live coals, ashes etc., which were placed in people's beds where they would release their heat, warming the blankets.

Latent Heat

The latent heat needed to change a liquid to a gas is needed not just at the boiling point but at all other temperatures also since water can evaporate at any temperature. The specific latent heat of vaporisation increases a little as the temperature falls. For example, it is approximately 2.46 MJ kg^{-1} at 25 $^{\circ}$C as opposed to the above-quoted 2.26 MJ kg^{-1} at 100 $^{\circ}$C. There are many practical applications of latent heat, and its effects in our day to day lives.

It feels colder during a thaw since energy is extracted from the atmosphere to melt the snow or frost remaining on the ground.

An ice-cube is more effective in cooling a drink on a summer's day than would be an equal mass of cold water. One could say that a little ice cools a lot of water.

If one wets one's finger in one's mouth and holds it up in the wind, the wind direction can be determined by the evaporation and subsequent cooling that takes place on the skin's surface.

A scald from steam is always more painful than that from boiling water. This is because steam condenses on the skin, and in so doing releases the latent heat of vaporisation. This heat energy causes vibrations which traumatise the skin tissue, creating the sensation of pain. (It is always advisable to cool the area with cold running water.) (See section 4.5, example 9.)

Large containers of water are sometimes left in a cellar in which apples, vegetables or tinned foods are stored. If the cellar temperature were to fall to, or below, 0 $^{\circ}$C, the water in the containers would freeze before the fluids contained within the food. As the water in the container freezes, the latent heat emitted may be enough to prevent the foods freezing. Vegetable sap contains salts and sugars and has a freezing point below 0 $^{\circ}$C.

Aftershave or skin tonic applied to the face makes the wearer feel fresher. Because these preparations contain alcohol, which is a volatile liquid, evaporation takes place quickly, and the skin thus feels cooler and fresher.

Campers often realise that it is more effective to wrap a damp cloth around bottles or cartons of milk to keep them cool, rather than placing them in a jar of water. The water evaporating from the cloth around the milk takes heat energy from the milk, thereby cooling it. The same idea was used by the Egyptians thousands of years ago. They realised that water could be kept cool by placing it in porous earthenware vessels. These containers allow sufficient liquid to escape and evaporate, and in so doing cool the water remaining. The same idea has been used with wine coolers, and also with pieces of muslin draped over jars.

Very often a nurse or doctor, before giving a patient an injection, will rub the area first with a volatile liquid. This acts as an antiseptic but also, as the liquid evaporates quickly, the skin is left cooler and slightly anaesthetised.

The cold feeling one has after emerging from a bath or shower is due to evaporation which extracts latent heat from the body. If the body is overcooled, e.g. in a draught, it is more susceptible to catching a cold or influenza.

A person can become badly chilled by being outside in damp clothes or shoes for a time. People remaining outdoors in wet clothes risk developing hypothermia.

Perspiration acts as a temperature control mechanism in the human body. One starts to sweat if the blood temperature rises by more than 0.5 K or if the temperature outside the body rises above about 31 °C. It is the natural process by which the body cools down. This mechanism is also based on latent heat. The centre in the brain involved in temperature control is known as the hypothalamus; when activated by a rise in blood temperature this stimulates sweat glands, with the resultant production of sweat. Evaporation of the sweat results in heat loss from the body surface. In addition, dilation of blood vessels in the skin takes place, giving the person a flushed appearance. This also has a cooling effect. These cooling mechanisms are designed to maintain the body temperature at the normal 37 °C. The temperature of the surface of the skin is 33 °C.

A device called the ‘Sizzle Stick’, which has been patented in the USA, claims to reduce significantly cooking times for roasting meat. The device is a hollow tube of metal which contains a wick along its length. Some water is added and the stick is inserted into the joint of meat. Heat energy is absorbed from the hot oven, which raises the temperature of the water in the lower, wider end of the tube. Eventually the water boils and in turning into steam absorbs latent heat. This hot steam rises through the tube, and as it passes through the cooler interior of the piece of meat, it condenses. As this happens, the latent heat which earlier was absorbed is released, and helps to cook the meat. The liquid water runs back along the tube, where the process starts once more. The action is that of a mini heat pump. The effectiveness of this tube over that of using a solid metal skewer (which heats by conduction) is due to the latent heat of vaporisation.

Which is more effective, to sponge the forehead with cold water or with tepid water? When teaching a class about body temperature, perspiration, etc., this question may arise in relation to a patient with a very high temperature. A number of students will always suggest using the cold water but, in fact, the tepid water is better. Because of its higher temperature, the tepid water will evaporate at a faster rate. It is then the latent heat absorbed from the body which cools the patient down.

A film of dew or frost may at times be noticed on the outer surface of a gas cylinder, as gas is being drawn from it. The hydrocarbons used (butane, propane or some such compound) are stored as liquids under pressure in the heavy cylinder. As they are released for use they vaporise, and the latent heat required is taken from the remaining liquid, cooling it down. Moisture from the air can then condense upon the surface of the cylinder.

Why does it take so long for a domestic refrigerator to form ice-cubes from a tray of water? The answer is that the refrigerator has first to cool down the water from room temperature to 0 °C. Then it must remove 336 kJ from each kg of water at 0 °C to change its state to ice at 0 °C. While water may cool quickly, it will freeze more slowly.

Because a dog does not perspire through its skin, its only way of cooling itself is through opening its mouth and letting its tongue hang out - again, latent heat is involved in the cooling process.

Sometimes, even though the temperature of the air is above freezing, ice can form on puddles on the road. The cooling effect which causes this is due to evaporation from the water surface and this can only be noticed when the air is very dry or when there is a good breeze.

Refrigerator/Heat Pump

A refrigerator is, in fact, a heat pump. It does exactly what the name implies - it 'pumps' heat from one place to another. A heat pump is a device that extracts internal energy from a body at a lower temperature and transfers it to a warmer one, i.e. it reverses the natural direction of heat flow. According to the second law of thermodynamics (see p. 75), heat will not flow of its own accord from a source at a lower temperature to one at a higher temperature. Kelvin was the first person to suggest, in 1851, that a heat pump was essentially a heat engine in reverse.

In the case of the refrigerator, energy is removed from both the air inside the cabinet and from any food present, i.e. the low temperature area, and transferred to the air outside the refrigerator, i.e. the high temperature area. As a result, the kitchen is slightly warmed, but this effect is not noticeable because of the larger volume of air and the heat capacity of the walls. The idea of a refrigerator was first developed by a Scottish printer, **James Harrison** (1816-1893), in 1837; while working in Australia he had noted the cooling effect of ether. The first domestic refrigerator was made in 1879 by **Von Linde** (German). In 1923 **Von Platen** and **Munters** first used 'freon' as the cooling liquid.

The refrigerator works by using latent heat. A volatile liquid is pumped through a system of closed pipes through a valve. A compressor maintains a higher pressure on one side of the valve than on the other. Because the pressure of the liquid is reduced when it passes through the valve it vaporises, and in so doing extracts the necessary latent heat from inside the cabinet. On the high pressure side of the valve the vapour liquifies once again, and in so doing releases the latent heat of vaporisation. We may notice that pipes at the back of a refrigerator are warm, and it is important to allow sufficient space for the circulation of air. As the refrigerator is thermostatically controlled, a slight increase in temperature in the cabinet will start up the electric motor, and the cooling process is repeated to maintain the pre-set temperature. In a gas refrigerator cooling is brought about by heating; a gas flame takes the place of the pump and creates circulation. This type of refrigerator is often called an absorption refrigerator. It is commonly used in caravans since it is silent.

The liquid used in a refrigerator is called a refrigerant. It has a high specific latent heat, and a low boiling point. In the past ammonia (NH_3), which has a boiling point of -37 °C, has been used. The most common liquids now in use are derivatives of methane (CH_4) and are generally referred to as CFCs. This is because these compounds contain fluorine and chlorine. Freon, the trade name for dichlorodifluoromethane (CCl_2F_2) is a non-flammable and non-toxic liquid which has a boiling point of -30 °C. Others include Carrene (CH_2Cl_2) and hydrochlorofluorocarbons ($HCFCl_2$). In recent years it has been accepted that some of these compounds damage the ozone layer surrounding the earth. The damaging effects of these compounds arise if the coolant system develops a leak or if they escape into the atmosphere when the refrigerators are discarded. Thus, more and more manufacturers are now choosing environmentally friendly or 'green' liquids, such as hydrofluorocarbons, which contain no chlorine and therefore cause less ozone damage.

A new type of refrigerator is based on an entirely different principle. It has been developed in France, and uses the Peltier effect (see p. 34) to achieve cooling. This effect was discovered in 1835. The Peltier effect refrigerator has become feasible with the use of semiconductors. If current is driven in reverse through a p-n junction, the junction absorbs energy and the surroundings are cooled. By incorporating this 'cold junction' into the cabinet of a refrigerator, low temperatures are obtained.

Heat pumps are used in other applications also. Buildings can be warmed in winter by taking heat energy from a low temperature source (the outside air, or water) and pumping it inside. Even though the air outside might be quite cold, the air still contains a considerable quantity of thermal energy (all matter with a temperature above absolute zero possesses some thermal energy). By using a heat pump we make use of this low grade energy and transform it into a more useful form. Trinity College, Dublin, is heated partially by heat collected from a nearby river and, similarly, the Royal Festival Hall in London obtains a lot of its heat energy from the Thames. In each case, electrical energy must be supplied to move the heat energy against its natural tendency. (A teacher could use the analogy of having to pump water uphill when teaching the principle of the heat pump.)

The term 'thermal comfort' has become important in recent years. This is used to describe the optimum conditions for human comfort regarding ventilation, temperature control and humidity levels, both in the home and in the place of work. Most heat pumps can be 'turned around' or reversed in summer, and used as air conditioners. In this situation, heat energy is removed from the lower temperature source (which this time is inside the building) and transferred outside to the warmer area (the hot summer air).

4.5 Worked Examples

1. Given that the specific heat capacity of water is eleven times that of copper, calculate the mass of copper at a temperature of 100 °C required to raise the temperature of 200 g of water from 20.0 °C to 24.0 °C, assuming no energy is lost to the surroundings.

$$\begin{aligned}
\text{Heat lost by copper} &= \text{heat gained by water} \\
(mc\Delta\theta)_{cu} &= (mc\Delta\theta)_{w} \\
m_{cu}c_{cu}(100 - 24) &= 0.200 \times 11c_{cu}\,(24 - 20) \\
76m_{cu} &= 8.8 \\
m_{cu} &= 0.11579 \\
&= 0.116 \text{ kg}
\end{aligned}$$

Note

The quantities given in the question have only three significant figures; this restricts the answer to an accuracy of about 1%.

2. Three litres of water at 100 °C are added to 15 litres of water at 40 °C. Calculate the temperature of the mixture. Take the mass of 1 litre of water to be 1 kg and the specific heat capacity of water to be 4.2 x 10^3 J kg^{-1} K^{-1}.

   Let the temperature of the mixture be θ.

$$\begin{aligned}
\text{Heat lost} &= \text{heat gained} \\
m_1c_1\Delta\theta_1 &= m_2c_2\Delta\theta_2 \\
3 \times 4.2 \times 10^3 \times (100 - \theta) &= 15 \times 4.2 \times 10^3 \times (\theta - 40) \\
100 - \theta &= 5\,(\theta - 40) \\
100 - \theta &= 5\theta - 200 \\
5\theta + \theta &= 200 + 100 \\
\theta &= 50\ ^\circ\text{C}
\end{aligned}$$

3. 1 kg of water at a temperature of 45 °C is mixed with 1.5 kg of alcohol at 20 °C. Find the final temperature of the mixture. Take the specific heat capacity of water to be 4200 J kg^{-1} $K^{-1}$and the specific heat capacity of alcohol to be 2400 J kg^{-1} K^{-1}. Assume no other exchange of heat occurs.

Let the final temperature of the mixture be θ.

1 kg of water cools from 45 °C to θ, giving $mc\Delta\theta$ = 1 x 4200 x (45 – θ)

1.5 kg of alcohol warms up from 20 °C to θ, absorbing 1.5 x 2400 x (θ – 20)

$$\begin{aligned} \text{Heat lost} &= \text{heat gained} \\ 1 \times 4200 \times (45 - \theta) &= 1.5 \times 2400 \times (\theta - 20) \\ 42(45 - \theta) &= 36(\theta - 20) \\ 1890 + 720 &= 42\theta + 36\theta \\ \theta &= 33\ ^\circ\text{C} \end{aligned}$$

4. If 0.70 kg of water at 18 °C is placed in a 1.0 kW electric kettle, how long will it take to boil the water? Take the specific heat capacity of water to be 4200 J kg^{-1} K^{-1}.

$$\begin{aligned} \text{Heat gained by the water} &= mc\Delta\theta \\ &= 0.7 \times 4200 \times 82 \\ &= 2.4108 \times 10^5\ \text{J} \end{aligned}$$

$$\begin{aligned} \text{Power} &= 1\ \text{kW} \\ &= 10^3\ \text{J s}^{-1} \end{aligned}$$

$$\text{Power} = \frac{\text{electrical energy supplied}}{\text{time taken}}$$

$$\begin{aligned} \text{Time taken} &= \frac{2.4108 \times 10^5}{10^3} \\ &= 241 \text{ seconds or 4 minutes} \end{aligned}$$

Note

In practice, the actual time would exceed the 4 minutes, as some energy goes to heat the material of the kettle, and some is lost to the surroundings.

5. A 500 watt electric drill is used to drill a hole in an aluminium block of mass 1 kg. If 80% of the energy used appears as heat in the metal find the increase in temperature of the aluminium in 20 seconds. Take the specific heat capacity of aluminium to be 910 J kg^{-1} K^{-1}.

$$\begin{aligned} 500 \text{ watt} &= 500 \text{ joule per second} \\ \text{Energy used in 20 s} &= 500 \times 20 \\ &= 10\ 000\ \text{J} \end{aligned}$$

80% of this is converted to heat in the aluminium block

$$\begin{aligned} 0.80 \times 10\ 000 &= 8000\ \text{J} \\ 8000 &= mc\Delta\theta \\ &= 1 \times 910 \times \Delta\theta \\ \Delta\theta &= 8.8\ \text{K} \end{aligned}$$

6. A car of mass 1800 kg and moving at 25 m s^{-1} is brought to rest by the application of disc brakes. Find the average increase in temperature of the brakes if each of the four brakes has a mass of 4.5 kg. Take the specific heat capacity of the brake material to be 680 J kg^{-1} K^{-1}, and assume that all the kinetic energy is changed into heat energy in the brakes.

$$\begin{aligned} \text{Kinetic energy} &= \tfrac{1}{2} mv^2 \\ &= 0.5 \times 1800 \times 625 \\ &= 5.625 \times 10^5\ \text{J} \end{aligned}$$

Four brakes, each of mass 4.5 kg gives a total mass of 18 kg.

$$\begin{aligned} \text{Heat gained by brakes} &= mc\Delta\theta \\ &= 18 \times 680 \times \Delta\theta \end{aligned}$$

Assuming all the kinetic energy is converted to heat energy

$$\begin{aligned} 18 \times 680 \times \Delta\theta &= 5.625 \times 10^5 \\ \Delta\theta &= 46\ \text{K} \end{aligned}$$

7. The temperature of 100 g of water was raised from 20 °C to 40 °C in 10 minutes by a heating coil of resistance 5.0 Ω and carrying a current of 2.0 A. Given that the specific heat capacity of water is 4.2 x 10^3 J kg^{-1} K^{-1}, calculate the amount of energy lost to the surroundings.

Energy supplied = I^2Rt

$= 2^2 \times 5 \times 600$

$= 12 \times 10^3$ J

Energy used = $mc\Delta\theta$

$= 0.1 \times 4.2 \times 10^3 \times 20$

$= 8.4 \times 10^3$ J

Energy lost to surroundings = $12 \times 10^3 - 8.4 \times 10^3$

$= 3.6 \times 10^3$ J

8. Water is being pumped through a central heating system at a rate of 1.15 m^3 per hour. The temperature of the water leaving the boiler is 58.0 °C and the temperature of the water returning is 48.0 °C. Calculate the power output of the boiler. Assume heat losses to the surroundings to be negligible. Take the density of water to be 1.0×10^3 kg m^{-3} and the specific heat capacity of water to be 4.2×10^3 J kg^{-1} K^{-1}.

Change in temperature of water = 58 – 48
= 10 K

Heat lost = $mc\Delta\theta$

$$\text{Density} = \frac{\text{mass}}{\text{volume}}$$

Mass = density x volume

Heat lost per hour = $1.15 \times 1 \times 10^3 \times 4.2 \times 10^3 \times 10$

$= 48.3 \times 10^6$ joule per hour

$$\text{Power} = \frac{\text{energy supplied}}{\text{time}}$$

$$= \frac{4.83 \times 10^7}{3.6 \times 10^3}$$

= 13.4 kW

9. Calculate the energy released when (a) 10 g water at 100 °C and (b) 10 g of steam at 100 °C, is spilt on the hand. Take the specific heat capacity of water to be 4200 J kg^{-1} K^{-1} and the specific latent heat of vaporisation of water to be 2.2 MJ kg^{-1}. Assume that the temperature of the skin is 33 °C.

(a) $Q = mc\Delta\theta$

$= 0.01 \times 4200 \times (100 - 33)$

= 2814 J

= 2.8 kJ

(b) The latent heat given out in changing from steam at 100 °C to water at the same temperature is

$Q = ml$

$= 0.01 \times 2.2 \times 10^6$

= 22 000 J

The heat given out when this condensed water drops in temperature from 100 °C to 33 °C is

$Q = mc\Delta\theta$

$= 0.01 \times 4200 \times (100 - 33)$

= 2814 J

So the total heat given out is = 25 kJ

10. When a falling hailstone is at a height of 2.00 km its mass is 2.50 g. What is its potential energy? Assuming that all of this potential energy is converted to latent heat during the fall calculate the mass of the hailstone on reaching the ground. Take the specific latent heat of fusion of ice to be 3.36×10^5 J kg^{-1} and the acceleration due to gravity to be 9.81 m s^{-2}.

Potential energy = mgh

$= 2.5 \times 10^{-3} \times 9.81 \times 2 \times 10^3$

= 49.05 J

The falling hailstone loses potential energy, and this is used to partly melt the hailstone.

$$\begin{aligned} ml &= 49.05 \\ m \times 3.36 \times 10^5 &= 49.05 \\ m &= 1.4598 \times 10^{-4}\text{ kg} \\ \text{Total mass of hailstone} &= 2.50\text{ g} \end{aligned}$$

$$\begin{aligned} \text{Remaining mass that reaches the ground} &= 2.50 - 0.1458\text{ g} \\ &= 2.354\text{ g} \\ &= 2.35 \times 10^{-3}\text{ kg} \end{aligned}$$

11. 0.30 kg of ice at 0 °C is added to 1.0 kg of water at 45 °C. What is the final temperature, assuming no heat exchange with the surroundings? Take the specific heat capacity of water to be 4200 J kg^{-1} K^{-1}, and the specific latent heat of fusion of ice to be 3.4 x 10^5 J kg^{-1} K^{-1}.

    Let θ be the final temperature.

    Heat lost by water = heat gained in melting the ice + heat gained in warming the ice water to a temperature θ

$$\begin{aligned} m_1c\Delta\theta_1 &= m_2l + m_2c\Delta\theta_2 \\ m_1c(45 - \theta) &= m_2l + m_2c\Delta\theta \\ 1 \times 4200 \times (45 - \theta) &= (0.3 \times 3.4 \times 10^5) \\ &\quad + (0.3 \times 4200 \times \theta) \\ 4200\,(45 - \theta) &= 1.02 \times 10^5 + 1260\theta \\ 1.89 \times 10^5 - 1.02 \times 10^5 &= 1260\theta + 4200\theta \\ 5460\theta &= 0.87 \times 10^5 \\ \theta &= 16\ ^\circ\text{C} \end{aligned}$$

12. A copper calorimeter of mass 180 g contains 450 g of water and 50 g of ice, all at 0 °C. Dry steam is passed into the calorimeter until a certain temperature, θ, is reached. The mass of the calorimeter and its contents at the end of the experiment increased by 25 g. If no heat was lost to the surroundings find the final temperature, θ. Take the specific heat capacities of water and copper to be 4200 J kg^{-1} K^{-1} and 390 J kg^{-1} K^{-1}, respectively. Take the specific latent heat of fusion of ice to be 3.36 x 10^5 J kg^{-1} and the specific latent heat of vaporisation of water to be 2.26 x 10^6 J kg^{-1}.

    Energy released when steam condenses at 100 °C is

$$\begin{aligned} Q &= ml \\ &= 0.025 \times 2.26 \times 10^6 \\ &= 5.65 \times 10^4\text{ J} \qquad \text{(a)} \end{aligned}$$

    Energy released when condensed steam cools from 100 °C to temperature θ is

$$\begin{aligned} Q &= mc\Delta\theta \\ &= 0.025 \times 4200 \times (100 - \theta) \\ &= 1.05 \times 10^4 - 105\theta \qquad \text{(b)} \end{aligned}$$

    Energy absorbed by 50 g of ice on melting at 0 °C is

$$\begin{aligned} Q &= ml \\ &= 0.05 \times 3.36 \times 10^5 \\ &= 1.68 \times 10^4\text{ J} \qquad \text{(c)} \end{aligned}$$

    Energy absorbed by 500 g of water (the original 450 g of water and 50 g of melted ice) in warming from 0 °C to θ is

$$\begin{aligned} Q &= mc\Delta\theta \\ &= 0.5 \times 4200 \times \theta \\ &= 2100\theta \qquad \text{(d)} \end{aligned}$$

    Energy absorbed by 180 g of copper (the calorimeter) in warming from 0 °C to θ is

$$\begin{aligned} Q &= mc\Delta\theta \\ &= 0.18 \times 390 \times \theta \\ &= 70.2\theta \qquad \text{(e)} \end{aligned}$$

    Heat lost = heat gained

$$\begin{aligned}
(a) + (b) &= (c) + (d) + (e) \\
5.65 \times 10^4 + 1.05 \times 10^4 - 105\theta &= 1.68 \times 10^4 + 2100\theta + 70.2\theta \\
5.02 \times 10^4 &= 2.275 \times 10^3\theta \\
\theta &= 22\ ^\circ C
\end{aligned}$$

13. A refrigerator converts 1.3 kg of water at 20 °C into ice at -15 °C in 1 hour. Calculate the total heat removed and the effective power of the refrigerator. Take the specific latent heat of ice to be 3.4×10^5 J kg^{-1}, the specific heat capacity of water to be 4.2×10^3 J kg^{-1} K^{-1} and the specific heat capacity of ice to be 2.1×10^3 J kg^{-1} K^{-1}.

Energy removed in cooling 1.3 kg water from 20 °C to 0 °C is

$$\begin{aligned}
Q &= mc\Delta\theta \\
&= 1.3 \times 4.2 \times 10^3 \times 20 \\
&= 1.092 \times 10^5 \text{J}
\end{aligned}$$

Energy removed in changing 1.3 kg of water at 0 °C to ice at 0 °C is

$$\begin{aligned}
Q &= ml \\
&= 1.3 \times 3.4 \times 10^5 \\
&= 4.42 \times 10^5 \text{ J}
\end{aligned}$$

Energy removed in cooling 1.3 kg ice from 0 °C to –15 °C is

$$\begin{aligned}
Q &= mc\Delta\theta \\
&= 1.3 \times 2.1 \times 10^3 \times 15 \\
&= 4.095 \times 10^4 \text{ J}
\end{aligned}$$

Total heat removed is

$$\begin{aligned}
Q &= 1.092 \times 10^5 + 4.42 \times 10^5 + 4.095 \times 10^4 \\
&= 5.922 \times 10^5 \text{ J}
\end{aligned}$$

$$\begin{aligned}
\text{Power} &= \frac{\text{energy in joules}}{\text{time in seconds}} \\
&= \frac{5.922 \times 10^5}{0 \times 60} \\
&= 164.5 \text{ W} \\
&= 0.16 \text{ kW}
\end{aligned}$$

14. Given that the specific heat capacity of air is 1.0×10^3 J kg^{-1} K^{-1} calculate how much energy must be removed from a refrigerator containing 0.15 m^3 of air to lower the temperature of the air from 20 °C to 4 °C (take density of air = 1.2 kg m^{-3}).

$$\begin{aligned}
\text{Mass} &= \text{density} \times \text{volume} \\
&= 1.2 \times 0.15 \\
&= 0.18 \text{ kg}
\end{aligned}$$

$$\begin{aligned}
Q &= mc\Delta\theta \\
&= 0.18 \times 10^3 \times (20 - 4) \\
&= 2880 \text{ J} \\
&= 2.9 \text{ kJ}
\end{aligned}$$

4.6 Mandatory Student Experiments

Helpful Hints

The following are some general guidelines to help improve the accuracy of one's results in carrying out the experiments on heat.

1. Most of the errors in these experiments arise from inaccurate temperature readings, so it is advisable to use a 0.2 °C thermometer. Arrange to have temperature changes as large as possible. However, this must be balanced against the need to minimise the temperature difference between the calorimeter and its surroundings.

2. Take temperature readings at eye level so as to reduce parallax error. A small magnifying glass is helpful.

3. Use a highly polished calorimeter; this reduces heat losses by radiation. (The word calorimeter, or ‘heat measurer’, was first used by Joseph Black - a flashback to the caloric theory.)

4. If possible, use a lid on the calorimeter; this reduces heat losses by convection and evaporation.

5. A stirrer, if used, should be of the same material as the calorimeter. These may be bought from suppliers, but it is possible to make one’s own by baring some copper wire and shaping it to fit the particular size of calorimeter used. Thermometers are widely used in practice as stirrers but mercury-in-glass ones often break, releasing the mercury which is toxic.

6. At the start of experimental work, it is useful to point out why theta, θ, is used to denote temperature rather than *‘t’*. Make use of the equation $VIt = mc\Delta\theta$ (where V is potential difference and I is current), which illustrates the point that ‘t’ is taken as the time, whereas theta is the symbol used for temperature.

7. Always stir the liquid in a calorimeter before taking a reading, and take the highest or lowest (whichever it may be) *steady* temperature.

8. In electrical methods, where a constant current is required, always include a rheostat.

9. When a heating coil is used it must always be completely covered with liquid. It is worth pointing out to students that they follow the same rule in their homes with the domestic electric kettle.

10. A polystyrene cup is a useful object in heat experiments. Because it has a negligible mass its heat capacity is negligible and almost all the heat energy goes into the liquid contained in the cup. This reduces the final calculations. It is interesting in a class situation, where different groups of students are working on the same experiment, to give calorimeters of different materials to various groups and then compare results at the end of the class.

11. Substances added to water in a calorimeter (metal pieces, ice, etc.) should be transferred quickly, but without splashing.

12. When time allows, experiments should always be repeated, and an average value of the quantity to be measured taken.

13. The experiment to compare the specific heat capacities of metal and a liquid (e.g. copper and water) is not a mandatory one for students. However, it can be a useful one to do if time permits as it conveys to a class the fact that water has a specific heat capacity more than ten times that of copper.

14. Insulation or lagging (draught proofing) is all-important in experimental work on heat, as quite a lot of thermal energy can be lost (or gained) to (from) the surroundings. Any good insulating material like polystyrene, cotton wool, felt or straw may be used. A piece of ‘aeroboard’ placed underneath the calorimeter as a stand will also help.

15. To minimise heat losses/gains to/from the surroundings, start with the calorimater a few degrees below room temperature and finish with it an equal amount above room temperature (or vice versa if the experiement involves cooling).

Experiment 4.1(a)

Measurement of the Specific Heat Capacity of Water

1. Joulemeter Method

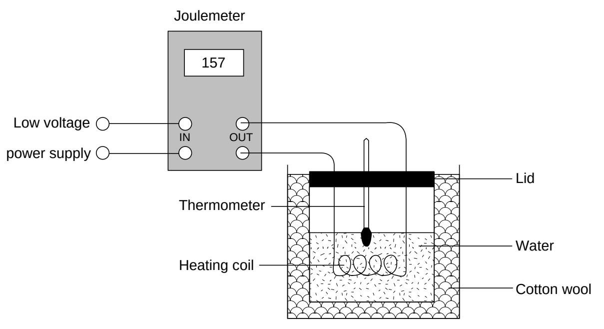


Fig. 4.1

Apparatus

As shown in the diagram, Fig. 4.1.

Method

1. Find the mass of the calorimeter and half fill with water. Find the mass of water. Make sure sufficient water is used to cover the heating coil. Place the calorimeter in a lagging jacket, stir the water, and record the initial temperature.

2. Zero the joulemeter and allow current to flow until a temperature increase of 10 K or so is achieved. Switch off the power, stir the water well and note the final temperature (the highest steady temperature reached).

A typical set of results might look like this:

Mass of calorimeter	0.080 kg
Mass of calorimeter + water	0.150 kg
Mass of water	0.070 kg
Initial temperature of water + calorimeter	15 °C
Final temperature of water + calorimeter	24 °C
Increase in temperature	9 K
Specific heat capacity of copper	390 J kg^{-1} K^{-1}
Energy supplied	2900 J

Energy supplied = energy gained by water + energy gained by calorimeter

$$Q = m_w c_w \Delta\theta + m_c c_c \Delta\theta$$

where m_w and m_c are masses of the water and the calorimeter and c_w and c_c are the specific heat capacities of water and copper (the material of the calorimeter), respectively.

Calculations

$$2900 = 0.07 \times c_w \times 9 + 0.08 \times 390 \times 9$$

$$2900 = 0.63 c_w + 280.8$$

$$c_w = 4157.46$$

$$= 4.2 \times 10^3 \text{ J kg}^{-1} \text{ K}^{-1}$$

Note

The joulemeter measures the electrical energy used directly in joules. It is basically the same as the domestic ESB meter, but the latter gives a reading in kilowatt hours. Some joulemeters have been given free of charge to schools by the ESB. These must be mounted vertically when in use. If a joulemeter is not available, the following method may be used instead. It is known as the ammeter-voltmeter method.

2. Ammeter-Voltmeter Method

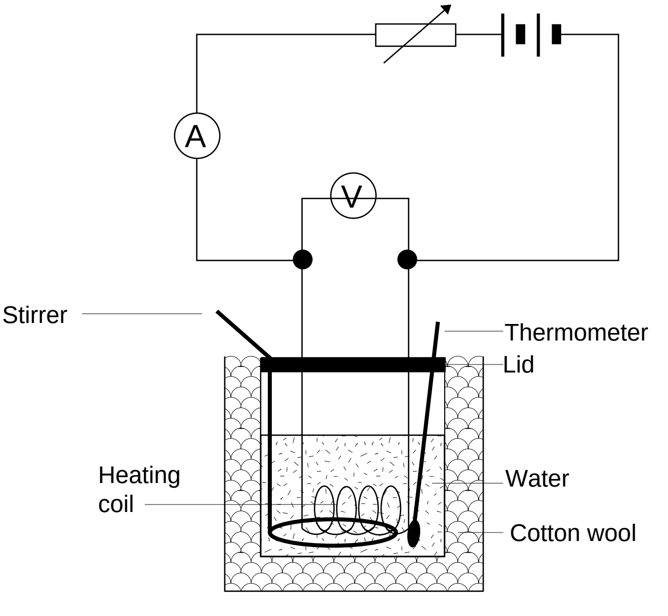


Fig. 4.2

Apparatus

As shown in diagram.

Method

1. Again, follow the procedure as detailed in the previous method, with regard to the mass of the calorimeter, covering the element with water, lagging jacket, etc. Stir and record the initial temperature of the water, θ_1.

2. Set up the electrical circuit, but do not switch on the current until everything else is ready. Note that the ammeter is placed in series with the heating coil, whereas the voltmeter is in parallel with it. Turn on the stopwatch and simultaneously switch on the current.

3. Allow a suitable current to flow through the heating coil (the optimum value depends on the resistance of the particular coil used). Keep the current constant by adjusting the rheostat. Allow the temperature to rise by about 10 K. Switch off the current and stop the clock at the same time. Measure the time, t, in seconds.

4. Stir the water in the calorimeter and record the highest steady temperature, θ_2. Allow the calorimeter and water to stand for $t/2$ seconds and now read the temperature again, θ_3. It will have cooled slightly. For greater accuracy, taking into account heat losses to the surroundings as per Newton's law, take as the final temperature $\theta_4 = \theta_2 + (\theta_2 - \theta_3)$. The corrected increase in temperature in the equation is now $\theta_4 - \theta_1$.

Calculations

Electrical energy supplied = energy gained

$$VIt = (m_w c_w \Delta\theta) + (m_c c_c \Delta\theta)$$

where
V = voltage across the coil
I = steady current
t = time
m_w, m_c = masses of water and calorimeter, respectively
c_w, c_c = specific heat capacities of water and material of calorimeter, respectively
$\Delta\theta$ = temperature increase

As all the other quantities are known, c_w may be calculated.

Note

A Joule's calorimeter is recommended for this experiment rather than ordinary calorimeters. It has a heating coil of constantan with an approximate resistance of 6 Ω. A close-fitting lid and stirrer are included and an outer lagged vessel is provided. It is used to investigate Joule's laws of heating.

Experiment 4.1(b)
To Measure the Specific Heat Capacity of a Metal

1. Joulemeter Method

Apparatus

As in previous experiment (Fig. 4.1) but with a block of the metal instead of the calorimeter of water.

Method

1. Two holes must be drilled in the metal - one for a thermometer, the other for a miniature heating element (12 V 50 W). The coil is often enclosed in a tube of stainless steel. A little oil, glycerine or Vaseline in these holes ensures good thermal contact.

2. Find the mass of the block and surround it with insulation as completely as possible.

3. Set up the circuit as above and zero the joulemeter. Record the initial temperature of the metal.

4. Switch on the current, allow the temperature of the block to rise by about 10 K and switch off the current.

5. Note the reading on the joulemeter.

6. Allow 1 minute for heat energy to spread uniformly throughout the block and record the highest temperature reached by the thermometer.

Sample readings might be as follows.

Mass of aluminium block	=	1 kg
Initial temperature of block	=	15 °C
Final temperature of block	=	22 °C
Joulemeter reading	=	6350 J

Calculations

Electrical energy supplied

= energy gained by aluminium block

$$W = mc\Delta\theta$$

$$6350 = 1 \times c \times 7$$

$$6350 = 7c$$

Specific heat capacity of aluminium,

$c = 9.1 \times 10^2 \text{ J kg}^{-1} \text{ K}^{-1}$

Note

These blocks of metal are available from suppliers in aluminium, copper, brass and steel.

2. Method of Mixtures

Apparatus

Calorimeter with lagging jacket; beaker; test-tube; tripod; Bunsen burner; thermometer.

Method

1. Follow the usual steps, as in previous experiments. Use a calorimeter one quarter filled with water, placed in a lagging jacket, with thermometer, etc.

2. Find the mass of a perfectly dry test-tube. Pour some copper pieces (turnings, flakes of copper or rivets - larger pieces) into the test-tube.

3. Find the mass of the test-tube with copper and, by subtraction, the mass of the copper pieces.

4. Place a plug of cotton wool loosely at the top of the test-tube. This helps to keep the copper dry. (Any moisture could make flakes of copper adhere to the glass.)

5. Place the test-tube containing the copper in a beaker of boiling water and allow to stand there for 10-15 minutes.

6. Having previously recorded the initial temperature of the water in the calorimeter, quickly, and without splashing, add the hot copper to it. Stir well and then read the highest steady temperature reached.

7. Again, assuming good insulation

   Energy lost by copper pieces = energy gained by water and by calorimeter

   $$mc_c\Delta\theta = mc_w\Delta\theta + mc_c\Delta\theta$$

   By using a copper calorimeter and taking c_w to be 4180 J kg^{-1} K^{-1}, the specific heat capacity of copper, c_c, may be found.

Sample results

Mass of calorimeter	0.082 kg
Mass of calorimeter + water	0.158 kg
Mass of water	0.076 kg
Mass of copper pieces	0.033 kg
Initial temperature of copper pieces	100 °C
Initial temperature of water in calorimeter	16 °C
Final temperature of water in calorimeter	19 °C
Specific heat capacity of water	4180 J kg^{-1} K^{-1}

$$mc_c\Delta\theta = mc_w\Delta\theta + mc_c\Delta\theta$$

$$0.033 \times c_c \times 81 = 0.076 \times 4180 \times 3 + 0.082 \times c_c \times 3$$

$$2.673\, c_c = 953.0 + 0.246\, c_c$$

$$c_c = 3.9 \times 10^2 \text{ J kg}^{-1} \text{ K}^{-1}$$

Note

The answer is only accurate at most to 2 significant figures as the least accurate measurement, temperature, is given to 2 significant figures.

Experiment 4.2

To Measure the Specific Latent Heat of Fusion of Ice

Apparatus

Calorimeter with lagging jacket; thermometer.

Method

1. Put some ice-cubes in a cloth or towel and crush them into fairly small pieces (less than 1 cm^3) using a hammer. Transfer to a small beaker, half-filled with water, and add a thermometer. Leave aside, and keep taking temperature readings until the mixture reaches 0 °C.

2. Meanwhile, find the mass of a calorimeter plus stirrer of the same material. Half-fill with warm water and find the mass of the water and calorimeter. Place in a lagging jacket and stand the calorimeter on a piece of aeroboard. Place a thermometer (0-50 °C preferably) in the calorimeter.

3. When the temperature of the warm water is about 10 K above room temperature stir and record the temperature. Add a piece of ice which has been carefully dried with filter paper. Be careful not to splash. Use a plastic tongs if available and, if not, two plastic 'takeaway' forks are a good substitute.

4. Only when one piece of ice has completely melted should another be added, stirring all the time. Keep adding ice until the temperature of the water in the calorimeter is about 10 K below room temperature. Do not let the temperature go more than 9-10 K below room temperature because water in the atmosphere may condense on the calorimeter, forming dew and giving up unwanted latent heat.

5. After the last piece of ice has been added and melted, stir again and record the final temperature. Now find the new mass of the calorimeter and contents. From this and earlier measurements, the mass of the newly melted ice, or ice water, may be found.

Notes

1. Again, a polystyrene cup may be used which would simplify calculations.

2. A few trays of ice are generally needed when this experiment is being performed in a class situation. Some pieces will be too small to use, and one finds that the ice melts rather quickly in a warm laboratory. Ice bags are particularly useful.

3. The fact of having water in the calorimeter warmed to 9-10 K above room temperature and then cooled to 9-10 K below room temperature acts as a form of cooling correction. Also, of course, the slightly warmed water assists in the rapid melting of the ice.

Sample results

Mass of empty calorimeter	0.0772 kg
Mass of calorimeter and water	0.1702 kg
Mass of water	0.0930 kg
Mass of calorimeter + water + melted ice	0.1894 kg
Mass of melted ice (same as mass of ice)	0.0192 kg
Specific heat capacity of water	4180 J kg^{-1} K^{-1}
Specific heat capacity of calorimeter material	390 J kg^{-1} K^{-1}
Original temperature of ice	0.0 °C
Initial temperature of warm water	29.4 °C
Final temperature of mixture	12.2 °C
Temperature rise of melted ice (ice water)	12.2 °C
Temperature drop of calorimeter and warm water	17.2 °C

The ice, in melting and warming up to the final temperature, absorbs energy twice.

1. It absorbs some to change its state from ice at 0 °C to water at 0 °C (latent heat).
2. It absorbs more to warm up from this newly formed water to the final temperature of the water in the calorimeter.

The energy for these comes from both the warm water and the calorimeter.

Energy gained by ice in melting and warming = energy lost by water and calorimeter

$$m_i l + m_i c_w \Delta\theta = m_w c_w \Delta\theta + m_c c_c \Delta\theta$$

$$0.0192\,l + 0.0192 \times 4180 \times 12.2 =$$

$$0.0930 \times 4180 \times 17.2 + 0.0772 \times 390 \times 17.2$$

$$0.0192\,l + 979.123 = 6686.33 + 517.86$$

$$0.0192\,l = 6225.06$$

$$l = 324\ 221.88\ \text{J kg}^{-1}$$

$$= 324\ \text{kJ kg}^{-1}$$

Note

The answer is only accurate to, at most, 3 significant figures as the least accurate measurement, temperature, is given to 3 significant figures.

Experiment 4.3

To Measure the Specific Latent Heat of Vaporisation of Steam

Apparatus

Round-bottomed flask fitted with a delivery tube; thermometer; steam trap; calorimeter with lagging jacket; Bunsen burner.

Method

1. Half-fill the round-bottomed flask with water, fit with the delivery tube and thermometer and heat the water until it boils. Reduce the Bunsen flame a little and allow steam to issue from the delivery tube. This tube should be slanted backwards towards the flask so that any condensing steam will run back to the flask. A steam trap should also be used; it really is a water trap as it holds any liquid formed. Some anti-bumping granules in the flask help to keep the water boiling smoothly and the cork in the flask should be tight-fitting. Steam pressure can build up during the experiment so, for safety reasons, sometimes an open tube is inserted in the flask instead of a thermometer. This helps to release some of the pressure. (The end of the tube in the flask should be under the surface of the water and it should be c. 1 m long to prevent boiling water escaping.) The flask of boiling water is a steam generator. The delivery tube should be well insulated. Cotton wool wrapped around and tied with string is effective, but can go on fire easily; fibreglass will not, but rubber gloves are needed to handle it safely. This must be done while the tube is cold. The last four or five centimetres of tube are left bare.

2. Meanwhile, find the mass of an empty calorimeter and stirrer. Half-fill it with cold water (cooled to 5 or 6 K below room temperature), find the mass of the water used and then place the calorimeter with thermometer into a well-lagged vessel. Stir the water and note the temperature just before adding the steam.

3. When steam has been issuing freely for a few minutes, and the initial condensation has warmed up the delivery tube, wipe the end of the nozzle with filter paper. Hold the nozzle supplying ‘dry’ steam just under the surface of the water in the calorimeter for a number of minutes while stirring to give equal distribution of heat. Allow the steam to enter the calorimeter until the temperature rises by 5 to 6 K above room temperature. (This cooling correction should minimise heat exchanges with the surroundings - heat gained in one half of the experiment approximately equals heat lost during the other half.) Stir the water in the calorimeter before taking the final temperature.

As the steam condenses, it loses heat energy in two ways:

1. by steam at 100 °C condensing to water at 100 °C (latent heat);

2. by the condensed steam cooling to the temperature of the mixture.

Note

Most problems arise in trying to get ‘dry’ steam, and in taking temperatures. Also, as steam enters the calorimeter and hits the water surface, some evaporation may take place, with subsequent cooling. Passing more steam into the cold water will give a larger mass of steam, a larger temperature difference, and therefore a more accurate answer.

Calculations are performed using the formula:
Energy lost by steam condensing and water cooling = energy gained by water and calorimeter

$$m_s l + m_s c_w \Delta\theta = m_w c_w \Delta\theta + m_c c_c \Delta\theta$$

Sample calculation

$0.00228l + 0.00228 \times 4180 \times 73.5$
$= 0.0831 \times 4180 \times 16.5 + 0.05314 \times 390 \times 16.5$

$$\begin{aligned} 0.00228l + 700.484 &= 5731.41 + 341.96 \\ 0.00228l &= 5372.88 \\ &= 2\,356\,526 \text{ J kg}^{-1} \\ &= 2.36 \text{ MJ kg}^{-1} \end{aligned}$$

Note

An approximate value for l can be got using an electric kettle and an electric balance. Place the electric kettle containing water on the electric balance and remove the lid (this ensures that the kettle continues to boil). Allow the water to boil for about 2 minutes then note the mass and time. Allow the boiling to continue for about 5 minutes and again note the mass. The mass difference equals the mass of steam evaporated. Substituting into $m_s l$ = power x time gives an approximation for l. The power rating of the kettle is normally stamped on it.

Appendix 4.1 Energy Sources

The Sun

The sun is our chief source of energy. Without it there would be no life on earth. Radiant energy from the sun supplies approximately 99.98% of our energy needs, not just at the present time, but as 'trapped sunshine' in fossil fuels produced millions of years ago. The sun is our local star - a mere 150 million km away! It is actually not a very big star. It is classified as being of medium size but the fact that it is relatively near to us makes it seem larger. It appears to us as an extremely bright solar disc, a great globe of incandescent gas, and we can look on it as a giant nuclear power station in the sky. Scientists such as **Joseph von Fraunhofer** (1787-1826) discovered the presence of various elements around the sun by the examination of the dark lines in the sun's spectrum. It is now thought that the sun consists of approximately 71% hydrogen and 27% helium, the remaining 2% being accounted for by other elements in gaseous form.

The sun belongs to the Milky Way galaxy and the part of the sun we see is called the photosphere. Its mass is estimated to be about 2×10^{30} kg and its diameter is roughly 1.4×10^6 km (more than 100 times the diameter of the earth). At its surface, the photosphere, the acceleration due to gravity is almost 28 times that on the surface of the earth. The temperature of the sun's surface is about 6000 °C, and estimates of the interior temperature give values of approximately 15×10^6 °C. These high temperatures produce thermonuclear reactions in the core in which lightweight atoms of hydrogen fuse, producing helium. (The word helium comes from the Greek word meaning sun.) As a result of the radiation emitted from the sun its mass decreases at the rate of 4 million tonnes per second. Astronomers calculate that the sun has been shining for about 5 billion years and that it should continue to shine for another 5 billion.

As far back as the fifth century BC the Greek philosopher Socrates recommended that houses be built in positions so as to receive the maximum amount of solar energy. The earth obtains 10^{22} joules of energy from the sun daily. The solar constant (the amount of energy falling normally on each square metre of the earth's atmosphere per second when the earth is at its mean distance from the sun) is 1.35 kW m^{-2}. The energy actually reaching the surface of the earth is reduced by cloud cover. In Ireland, the average figure is taken as approximately 115 W m^{-2}. Radiant energy from the sun can be collected in two ways, viz. passively (solar panels) and actively (solar cells).

In solar panels, water flows in blackened copper pipes laid on a matt black surface and absorbs solar energy. The panels are placed on roofs of buildings that face the sun and are tilted at a particular angle towards the sun. The panels are matt black to absorb the maximum amount of radiation and are covered with glass plates to trap the heat (greenhouse effect). The water, heated by the sun's rays, flows into a storage tank or into a heat exchanger.

Solar cells are photovoltaic cells, usually made of silicon, which convert light energy directly into electrical energy. Because each cell develops an emf of only 0.6 volt very many of them are used. Satellites and spaceships use solar power, and of course students will themselves be familiar with solar powered calculators.

In Mont Louis in the French Pyrenees, solar energy is harnessed using mirrors. The sun's rays are directed onto a small area by 3500 small mirrors, and temperatures of up to 3000 °C have been reached. The whole system is set so it can rotate and follow the path of the sun. Similar systems operate in Israel and in India. The same concept is used in a solar cooker. A large concave mirror set facing the sun on a hot day, and focused on a small piece of meat, can concentrate the rays sufficiently to cook it.

We obtain more radiant energy from the sun in summer than in winter for a number of reasons. The rays of the sun are more slanted as they hit the northern hemisphere in winter and are thus spread over a larger area, thereby producing less energy per unit area of surface. About 30% of the solar energy reaching the earth is reflected directly back into space by clouds or scattered by gases and dirt in the atmosphere. This figure is greater at the ice caps.

Approximately 47% of radiant energy is absorbed by the earth as heat or as thermal energy during the day. However, at night, especially on clear nights with very little cloud cover, much of this energy is radiated back into space. Clouds act like an insulating blanket around the earth and keep temperatures higher on a cloudy night than they would be on a clear night. Oceans, seas, lakes and rivers absorb about 19% of radiant energy, resulting in evaporation of water from the surface of the earth. This, when it falls as rain, can be used to produce hydroelectric power for the generation of electricity. Convection currents in the ocean (e.g. the Gulf Stream) which have a warming effect on climate, and winds arising from unequal heating of land and sea are due to solar heating.

The 'greenhouse effect' is the way in which the earth is heated as a result of the atmosphere trapping long wavelength infrared radiation. Short wavelength infrared and light radiation from the sun are transmitted through the atmosphere onto the surface of the earth. The earth re-radiates this energy but, because the earth is at a much lower temperature, the radiation has a much longer wavelength. Water vapour and carbon dioxide in the atmosphere absorb this long wavelength infrared radiation, preventing it from escaping back out into space. Thus the earth is maintained at a higher temperature than would otherwise be the case.

A small amount of energy is absorbed by green plants in the process of photosynthesis. This is possibly the most important chemical reaction of all. It provides food and oxygen for humans and animals. Plants are energy converters and are the first link in food chains, in which chemical energy is used by different organisms. Plants may also be used to collect solar energy. Plants absorb energy from the sun and store it, and this is released either through burning (cutting down a tree for firewood) or allowing the processes of fermentation and distillation to produce alcohol. Fast-growing plants which are used for energy production are known as biomass. Sugar cane, corn, oil seed rape, hemp, etc., are used. Vegetable oils can be used to power cars and farm machinery. Rotting plants produce methane gas and are widely used as energy sources in countries such as China and India.

Geothermal Energy

This type of heat energy comes from the earth itself. Some of it has accumulated due to condensation of bodies from which the earth was formed but mostly it arises from the radioactivity of thorium, uranium and their daughter products in the earth's crust and upper mantle. Geothermal energy was first developed for the generation of electricity in Tuscany, Italy, in 1904. It is now an important source of energy in many countries, including Iceland, New Zealand, Japan, China and the United States; current (1998) world wide production is of the order of 10^{10} W.

Temperature increases with depth below the surface of the earth, with the temperature of the core estimated to be almost 3000 °C. We are reminded of this increase in temperature every time that volcanoes erupt, or hot springs or geysers shoot boiling water into the air. Where there are gaps in the earth's crust due to tectonic activity, heat energy, produced mainly by friction (and radioactivity) melts the rock and flows to the surface. Most of this comes by conduction through rocks, but some comes by convection in the form of hot water and steam from springs and geysers and, much less frequently, as lava from volcanoes.

A geyser is a spring which from time to time erupts, spouting steam mixed with hot water upwards to a height of several metres.

The first explanation of geyser activity was given in 1847 by **Robert Bunsen** (1811-1899). Surface water seeps into deep vertical shafts, some as deep as 70 m, through gaps in the earth's crust. This water becomes heated to boiling point by the very hot rocks below (the water boils at a temperature of 130 °C because of the increased pressure caused by the large column of water above it). Convection currents are prevented by the narrowness of the shaft and, as the water is heated from below, rising bubbles of steam push hot water upwards. As a result of the expansion of the steam as the steam bubbles rise and the pressure on them falls, an eruption occurs. Geysers and hot springs, whether of water or of mud, go hand in hand with volcanic activity. Iceland is known as the 'land of frost and fire' due to the large number of volcanic eruptions that have taken place there - about 250 in the last 1000 years. Mount Hekla last erupted in 1991. Reykjavik uses geothermal energy to a very large degree; 85% of homes are heated by water coming from hot springs. Up to 200 litres of boiling water can be obtained per second from some of the larger ones. The word 'geyser' actually comes from a particular spring in the western part of the island. The 'Great Geyser' started erupting in 1294 but now seldom erupts unless it is encouraged to do so by adding soap solution. This is done to mark special occasions in Iceland and an eruption brought about by this method can shoot water to a height of 60 metres. Its neighbour, 'Strokkir' (the churn), spits its column of water to a height of 20 metres every five minutes or so. Another famous geyser is 'Old Faithful' in Yellowstone national park, USA (which contains more than 100 geysers and up to 4000 hot springs). This has been shooting water and steam into the air to heights of 37-46 metres for at least the last eighty years. The eruptions take place hourly, and each lasts for 3-4 minutes. As water from geysers is unusually rich in minerals, these become deposited in peculiarly shaped mounds around the spout holes and are characteristic of geyser regions. In New Zealand's north island, the geyser district occupies an area of approximately 13 000 km^2, which is about the area of Northern Ireland. It contains many boiling springs which provide heating for houses. Mud volcanoes are also found in the same place - they are referred to as 'mud pots'. In China warm water at a temperature of c. 50 °C is directly pumped into houses and factories to give central heating.

Sometimes fissures occur naturally and boiling water, accompanied by steam, rises upwards with considerable force. Alternatively, 2 holes may be drilled into the ground, 8 km deep and approximately 100 m apart. Cold water poured down at one side is heated by the very hot rocks underneath and returned as boiling water at the other side. The first such project was carried out in 1974 in New Mexico. If heated water can rise freely to the surface, i.e. if the gap is more or less vertical and not too deep, it gives rise to a hot spring but if the passageway is bent and twisted the water will be accompanied by steam under pressure, giving rise to what is called a geyser. At the present day, artificial geysers are found in a number of places, e.g. California, Italy, etc. However, geothermal energy is more expensive to harness than other sources at the moment but it may be used more widely in the future.

Chemical Reactions

People realise when they sit in front of a warm fire that energy is given off as heat. Some chemical reactions produce heat and are called exothermic reactions. Simple examples of exothermic reactions are the neutralisation of an acid by a base, the addition of concentrated acid to water, etc. Students in the lab will be familiar with the burning of hydrocarbons as an energy source when they use a Bunsen burner. The energy stored in a battery is also a form of chemical energy. And of course the processes of digestion and metabolism in the human body involve a number of chemical reactions and these produce heat and keep the body at the normal temperature of 37 °C. The main heat-producing chemical reaction is the combustion of fossil fuels. When these combine with oxygen much heat energy is released. Turf and coal seams are now found where once plants and trees grew and absorbed energy in photosynthesis. As these died and layer upon layer was added they were compressed together under great pressure. Oil was formed in the same way but it is mainly animal in origin. Marine plants and animals lived and died over a very long period of time and their remains, mixed with mud at the bottom of the sea bed, produced the oil we use today. Earth movements have since caused some areas of the sea bed to rise and become land so that oil deposits are found today under both land and sea. The word petroleum means ‘rock oil’. It is estimated that the complete combustion of 1 gram of petrol yields c. 46 kJ. For the last 200 years more and more non-renewable fuels have been taken from the earth, and what took millions of years to store is being used in a few hundred years. It is estimated that the remaining oil can last only about 30 years or so at the present rate of consumption, and that coal deposits will last for approximately 300 years. The solution for the future must lie in the increased use of renewable sources of energy like the wind, tides, solar power and nuclear energy, but at the moment we obtain about 90% of our energy needs from fossil fuels.

Appendix 4.2
Laws of Thermodynamics, Carnot Cycle, Heat Engines

First Law of Thermodynamics

Thermodynamics is the branch of physics that deals with energy transfer in the forms of heat and mechanical work. The word itself comes from Greek (*therme* - heat and *dynamic* - force or power). The flow of heat into and out of a system is examined, as is work done on, or by a particular system. There are four laws in thermodynamics. The first law may be written in the form

$$\Delta Q = \Delta U + \Delta W$$

where ΔQ is the energy supplied as heat, ΔU is the increase in internal energy and ΔW is the work done. If work is done by the system $\Delta W > 0$, if work is done on the system $\Delta W < 0$.

Second Law of Thermodynamics

This is more a statement of what cannot happen rather than of what actually can happen and can be stated in a number of ways. Clausius stated that while heat flows naturally from a hot object to a cold one it will not flow spontaneously from a cold object to a hot one. In the case of a refrigerator, which transfers energy from an area of low temperature to an area which is warmer, energy has to be supplied from an outside source; we notice that the motor starts up automatically from time to time, supplying the necessary electrical energy.

The second law may also be stated in terms of entropy. By entropy we mean the measure of disorder in a system. The word was first used by Clausius in 1850; the symbol used to represent it is *S*. The second law states that

$$\Delta S_{\text{universe}} > 0$$

meaning that the net change in the total entropy of the Universe is always greater than 0. The total entropy of any system and its environment will always increase as a result of any natural process.

Examples of entropy increasing are: glass breaking into many pieces (more disorder); an ice-cube melting (a less ordered arrangement of particles); an iron nail rusting; organic matter decaying into simpler substances; rocks crumbling to dust; a deck of playing cards, sorted into the different suits, falling on the ground; water evaporating into the air; eggs scrambled in cooking; etc.

In all of these examples, there is a definite direction involved - order becomes disorder, leading to an increase in entropy. In none of the above examples can matter, by itself, go back to its original state.

Third Law of Thermodynamics

This states that it is impossible to cool a substance to absolute zero. It is accepted that 0 K is unattainable. Temperatures extremely close to it have been achieved, but it is found that the closer one gets to it the more difficult it is to actually reach it. At this temperature, molecules have the lowest possible amount of energy - zero point energy.

Zeroth Law of Thermodynamics

This was developed in the 1930s after the first three laws had already been formulated. Because it comes logically before the others it was called the zeroth law. It relates to temperature. The law states that if two objects, A and B, are each in thermal equilibrium with a third one, C (e.g. a thermometer), then A and B are in thermal equilibrium with each other. **James Clerk Maxwell** (1831-1879) first introduced this idea and later, **Josiah Willard Gibbs** (1839-1903) developed it. They both recognised that when a hot and cold object come in contact with each other energy transfer takes place between them for a while until both reach the same temperature.

The Carnot Cycle/Heat Engines

Nicholas Leonard Sadi Carnot was a French engineer-physicist who became interested in heat engines when it was generally accepted that French engines were not as efficient as English ones. He made a particular study of these machines and devised a theoretical heat engine. It is known as a Carnot engine and the cycle of events that takes place in it is called the Carnot cycle. Because Carnot accepted the caloric theory one might argue that his work was based on a misconception, yet he established principles which still hold good today. Carnot's ideas were presented by him in 1824 in a paper entitled 'Reflections on the motive power of heat'. What is known today as Carnot's theorem states, 'All reversible engines working between the same temperature levels have the same efficiency'. Before Carnot it was thought that the important factor governing the efficiency of heat engines was the temperature of the incoming steam. But Carnot showed that it was in fact the difference in temperature between the input and output levels that mattered.

A heat engine is a physical system that uses the motive power of internal energy to produce useful mechanical work. The first stage involves the conversion of chemical energy from burning fossil fuel to produce internal energy. This step is practically 100% efficient but the second one is much less so. Less than half of this energy gives useful work; the remainder goes to waste in the surroundings, warming up the atmosphere. A working substance (e.g. steam or a hot gas) takes in energy, Q_1, at a high temperature, T_1. It performs useful work (mechanical energy), W. When it exerts a force on a moveable object, like a piston in a cylinder, motion is produced. The remainder of the energy, Q_2, is rejected at a lower temperature, T_2, into a sink or reservoir (e.g. the atmosphere). The working substance returns to its original state, at the high temperature source, and the cycle starts again. The same vapour or gas is used over and over and is always restored to its original condition.

Carnot died from cholera at the early age of thirty six and, because of the infectious nature of the disease, the original drafts of his papers, together with his personal belongings, were all burned. However, his work was continued by Rudolf Clausius and Kelvin (William Thomson). Carnot was a quiet, shy, likeable person, and, apart from his work in physics, had many other wide-ranging interests, including economics, literature, music, dancing, gymnastics and fencing.

Stages of the Carnot Cycle

The Carnot cycle is a reversible cyclic process involving a working substance which is taken through a series of changes or sequence of events. This substance (e.g. a gas) is enclosed in an insulated cylinder which has a moveable piston. Two reservoirs are, in turn, in good thermal contact with the cylinder, one at a high temperature, and the other at a lower temperature. Four stages can be identified: two reversible isothermal[2] processes and two reversible adiabatic[3] processes, Fig. 4.3.

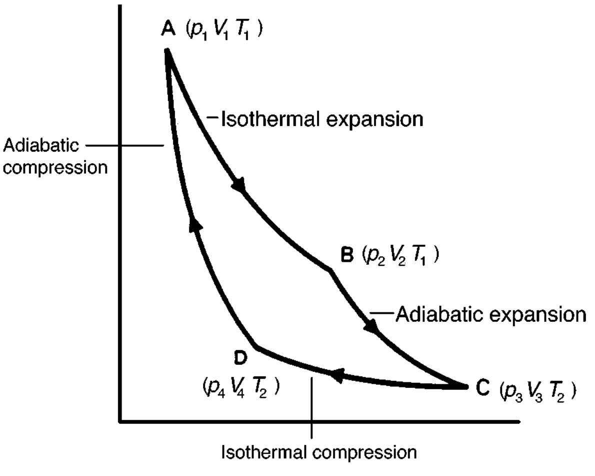


Fig. 4.3 *Carnot cycle for an ideal heat engine*

Stage 1 (A to B). The cylinder is placed in contact with the higher temperature reservoir and the mass of gas is heated and allowed to expand (an isothermal expansion) to a volume V_2. Since the temperature remains constant, the pressure decreases to p_2.

Stage 2 (B to C). The cylinder is insulated and the gas is expanded adiabatically. No heat enters or leaves the system at this stage and, as external work is done, the internal energy and therefore the temperature drops. The volume increases and the pressure decreases.

Stage 3 (C to D). The cylinder is now placed in contact with the reservoir at temperature T_2 and the gas is compressed isothermally. Its volume decreases while its pressure increases and heat flows out of the system.

Stage 4 (D to A). The cylinder is again insulated and the gas is taken back to its original temperature, volume and pressure by compressing it adiabatically.

So the gas is taken around a complete, closed, reversible cycle.

Types of Heat Engine

Steam engines

The basic idea of the steam engine is a very ancient one. It is said that an ancient Greek mathematician and physicist, Hero of Alexandria, showed that hot substances could be used to perform work. He invented a type of steam machine called the aeolipile. However, it was in the seventeenth century that heat engines were really invented. In 1681 **Denis Papin** (1647-1712), a Frenchman, noticing the pressure generated from steam, felt such power could be put to use. He designed a crude type of steam engine some years later in 1687 but it was really an English engineer, **Thomas Savery** (1650-1715), who invented the first real steam engine. This happened in 1698 and it was used to suck up water from mines. The principle behind this machine was that a partial vacuum is produced when steam in a chamber condenses. In 1705 another English engineer, **Thomas Newcomen** (1663-1729), added a piston, which moved up and down in a cylinder, to his engine and, when gears were introduced, the up and down motion was converted into a rotary one. The efficiency of these machines was low - as little

[2] This is a process which takes place at constant temperature. Supplying heat from outside results in volume and pressure changes within a system. As $\Delta U = 0$, $\Delta Q = \Delta W$.

[3] The word adiabatic comes from the Greek meaning impassable. An adiabatic process is one where no heat enters or leaves a system. Since $\Delta Q = 0$, $\Delta U = -\Delta W$. There is no exchange of heat with the surroundings. Two other processes in Thermodynamics are: Isochoric: Here, the volume is constant so no work is done. $\Delta W = 0$; $\Delta U = \Delta Q$. All the heat added to a system becomes the internal energy of the system.
Isobaric: In this case, the pressure is constant. $\Delta Q = \Delta U + p\Delta V$ where work done $\Delta W = p\Delta V$.
Note, in the above, U = Internal Energy Q = Energy as heat flow W = work done p = pressure V = volume

as 8% - but yet it can be said that they led directly to the Industrial Revolution. In 1764 a Scottish instrument maker, **James Watt** (1736-1819) was asked to mend a model of a steam engine and, as he did so, he came up with the idea of adding a condenser to it. This increased its efficiency to c. 20%. Another engineer **George Stephenson** (1781-1848) also played a large part in the development of steam engines. In 1825 he built, for the opening of the Stockton and Darlington railway, the first locomotive to achieve a speed of 15 m.p.h. Later, at the trials in connection with the opening of the Liverpool and Manchester Railway in 1829, he introduced a far superior and reliable locomotive, 'Rocket', which could reach a speed of about 36 m.p.h. while pulling a train. For this he received a prize of £500.

In 1838 the first ships crossed the Atlantic Ocean under steam power.

Steam turbines

In 1629 an Italian engineer**, Giovanni Branca**, made a basic steam turbine. So did a Swedish engineer, **Carl de Lavel** (1845-1913) and an Englishman **Charles Gordon Curtis** (1860-1953). However, the modern steam turbine was invented in 1884 by **Sir Charles Parsons**[4] (1854-1931) in Cambridge. Energy from the burning of fossil fuel boils water under increased pressure which produces steam at a high temperature. Jets of this steam are made to hit a set of blades attached to a shaft, causing the shaft to rotate. The energy of the steam at a high temperature, T_H, is partly converted to useful work and partly delivered as heat at a lower temperature, T_C. The efficiency is of the order of 40%. Most of our electricity is generated using steam turbines. Steam power has been used in shipping. The Cunard line was amongst the first shipping companies to use it. The Lusitania, which was sunk off Cobh on May 7 1915 by a torpedo from a German submarine, had steam turbines, as have present-day liners such as the QE2.

Internal combustion engines

Steam engines are external combustion engines. In the internal combustion engine heating takes place by the burning of a fuel in a cylinder (where energy conversion takes places) inside the engine, and not in a separate boiler outside as happens in a steam engine. An example is a petrol engine used in cars and other vehicles. In each cylinder of such an engine expanding hot gases resulting from combustion do work in moving a piston. The pistons are connected to a crankshaft, thereby producing rotary motion which is transferred to the wheels of a vehicle, via clutch, gears and shaft or chain. Again, as always happens, only part of the chemical energy converted does useful work; the remainder is carried away by exhaust gases at a lower temperature. Efficiency values are in the region of 25% - 30%. The cycle of movements inside a cylinder is repeated continuously, Fig. 4.4, p. 79. In order, they are:

1. intake - an inlet valve opens to take in an air/petrol vapour mixture;
2. compression - the piston compresses the mixture to c.1/8th of its original volume;
3. combustion - a spark (electrical discharge) from the spark plug ignites the mixture, and the resulting expanding gas moves the piston;
4. exhaust - the piston rises up to eject the waste gases through an exhaust pipe.

The first basic type of internal combustion engine was invented by **Jean Lenoir** (1822-1900) in Belgium in 1860. A German, **Nikolaus Otto** (1832-1891), built the first 'four stroke' engine in 1876 and then, in 1883, **Gottlieb Daimler** (1834-1900), also a German, produced a very successful model which again worked by the rapid combustion of petrol. The silencer which was added reduced the level of noise created by the engine. In 1886 **Karl Benz**, a German engineer, obtained a patent on a self-propelled vehicle powered by petrol and so began the automobile age. (The engine was named after his daughter, Mercedes.)

[4]The youngest son of the third Earl and Countess of Rosse from Birr Castle.

In 1878 a Scottish engineer, **Sir Dugald Clerk**, developed the two-stroke engine. In this type of engine the intake and compression strokes are combined, and the expansion and exhaust strokes also are combined. This version of the internal combustion engine is used in lawn-mowers, motor cycles, golf buggies, mopeds and outboard motors.

Diesel engines

One of the first people whose name was linked with a diesel engine was an Englishman called **Stuart** but the main person associated with this type of engine is the German who gave his name to it, **Rudolf Diesel** (1858-1913). He developed it in 1892, using a cycle in which fuel is added after air has been compressed. No spark plug is required, as combustion takes place spontaneously when the air becomes sufficiently hot after compression to ignite the fuel. A temperature is reached which is higher than the flash point of the heavy oil used. Air is compressed to about 1/16th of its original volume (rather than 1/8 in the petrol engine). The same four strokes take place in it as in a petrol engine but the efficiency is greater, c. 40%.

Jets

In this type of engine air is drawn in at the front, compressed, mixed with fuel (paraffin) and burned in a combustion chamber. Very hot gases are produced which are ejected with high velocity through an exhaust pipe at the rear. These fast moving escape gases cause the engine to move forwards (Newton's third law - the engine exerts a force on the gases and the gases exert an equal and opposite force on the engine).

Note
To show that the forward thrust is due to escaping gases from the rear, a balloon may be used. Blow up the balloon, hold it by the neck and then let it go.

Rockets
These work in the same way as jets (they also are internal combustion engines) but have to carry both fuel and oxygen. They can fly in areas where there is no air (outside the earth's atmosphere) but the oxygen required for combustion has to be taken along, as well as fuel (kerosene). Because of this, rocket engines are made in disposable sections. When a tank is empty it is jettisoned so that no unnecessary mass is carried. They are used in space exploration and for launching satellites in orbit around the earth.

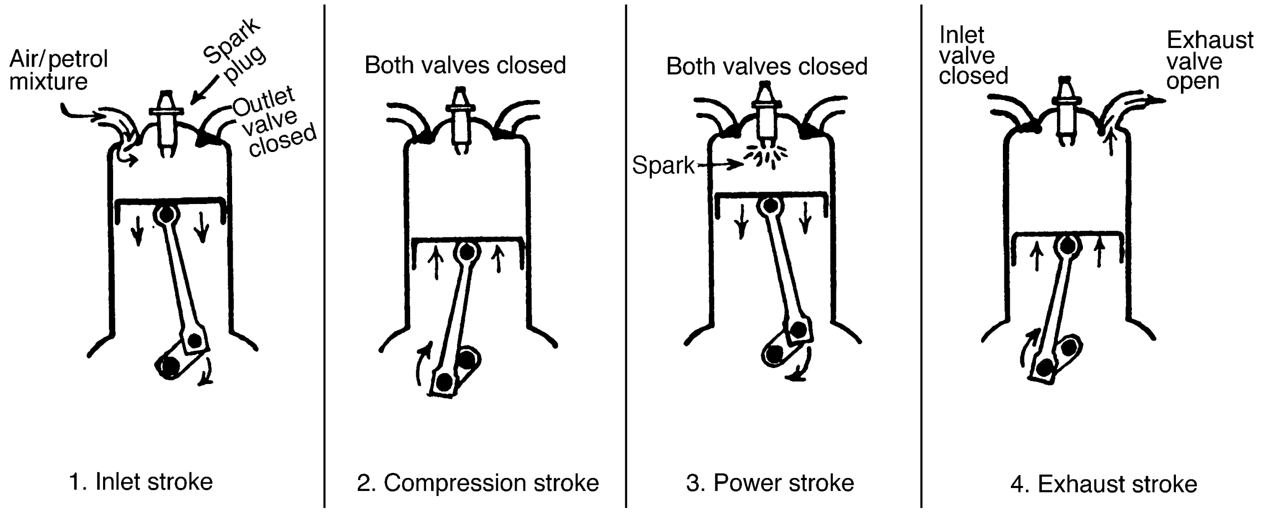


Fig. 4.4 *Four-stroke engine*

MODULE 6

Electromagnetism

CONTENTS

CHAPTER 1 MAGNETISM AND MAGNETIC FIELDS

1.1 Historical Background

Lodestone

Lodestone is a mineral that has aroused interest since ancient times because of its unusual properties. It has long been known that it attracts iron and that when suspended or floated it always turns to take up a north-south direction. Magnetic properties were known to the Greeks in classical times, and it is from Magnesia, in Thessaly (Greece), that the name ‘magnet’ originates. Lodestone is known to have been used as a navigational compass by sailors in the Mediterranean since at least the twelfth century, and is believed to have been used since the fourth century in China.

The modern development of magnetism can be said to have begun about 1600 with the publication by **William Gilbert** (1534-1603) of his book *De Magnete.* Gilbert, who was court physician to Queen Elizabeth I, observed that a magnetic needle hung freely about its centre of gravity pointed towards the North and dipped at about 70° to the horizontal in England, and that the angle of dip varied with latitude. He also performed many experiments and was the first to propose that the earth itself was a magnet and that this was what caused compasses to point in a north-south direction. To test his theory he constructed a permanent magnet in the form of a large lodestone sphere and showed that a small magnetised needle placed near the surface acts in the same way as a compass needle does at different points on the earth’s surface. For example, at the poles the needle pointed perpendicular to the surface of the lodestone, while halfway between, along the ‘equator’, the needle rested parallel to the surface. His book, *De Magnete*, is regarded as a classic in scientific literature because of its attempts to test speculation by detailed experiments.

Electricity and Magnetism

The first concrete evidence of a connection between electricity and magnetism was discovered in 1819 when Danish scientist, **Hans Christian Oersted** (1777-1851), found that a compass needle could be deflected when a wire carrying an electric current was placed over it. Almost sixty years later in 1878 the American physicist **H. A. Rowland** (1848-1901) discovered that a moving charged object also causes magnetic effects. The most curious aspect of Oersted’s discovery was the direction of the forces on the compass; the needle was deflected *perpendicular* to the wire, parallel to a tangent to the surface of the wire. Forces known previously had acted along the line joining the interacting bodies (e.g. electrostatic, gravitational forces). These forces between the current-carrying wire and the compass needle acted perpendicular to such a line. The magnetic needle was not attracted or repelled by the current in the wire, but was rotated by forces acting on its poles.

Prior to Oersted’s discovery, nobody suspected any link between electricity and magnetism. Consequently, his discovery opened up a whole new exciting subject for research, and scientists everywhere immediately began studying the magnetic effects of electric currents. Within a very short time, **André Marie Ampère** (1775-1836) in France showed that two parallel wires carrying electric currents exerted forces on each other, and in 1825 **William Sturgeon** (1783-1825) placed an iron core into a coil carrying a current and found that the magnetic effect of the current was greatly increased, thus discovering the electromagnet. The American, **Joseph Henry** (1789-1878) observed, in 1829, that when he switched off the current in a strong electromagnet a spark occurred. Thus was self induction discovered.

The Search for New Magnetic Materials

Throughout this century, the range of electrical appliances, and consequently the need for more and better magnets, has expanded rapidly. For example, today's family car may have up to 70 permanent magnets compared to one or two in the 1950s model.

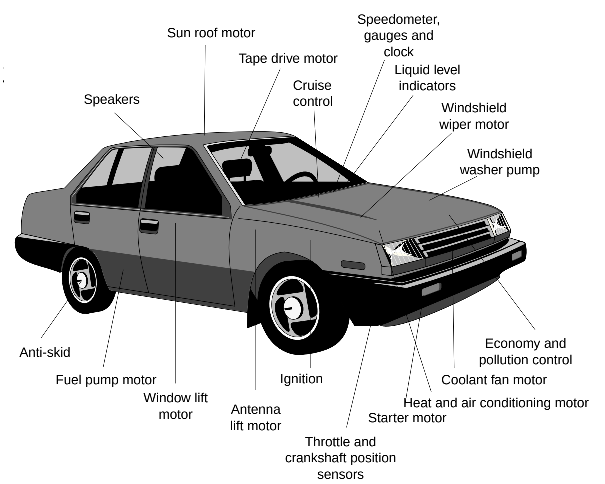


Fig. 1.1 *Some of the magnet-containing components in a modern car*

To meet this need, intensive research into new magnetic materials has continued and today's magnets are about 200 times stronger than the same size magnet in 1900. At that time, most magnets were made of steel. Then, in the 1930s, Alnico (an alloy of aluminium, nickel, iron and cobalt) magnets, which were much stronger, were developed (see p. 4). Ferrite magnets, manufactured from iron and barium oxides, were developed in the 1950s and are still used in large quantities in low-cost appliances such as magnetic signs and catches, domestic microwave ovens, loudspeakers and motors.

In recent decades, new families of rare-earth alloys which offer magnets that are much stronger and more permanent and which can be made into virtually any shape have been developed. Many of the discoveries in this field were made in Japan. However, in 1990 a nitrogen-containing compound ($Sm_2Fe_{17}N_3$) known as nitromag was discovered by a European research group working under Prof. J.M.D. Coey in Trinity College, Dublin and the race is on to develop useful permanent magnets based on this material.

The development of the new high-performance rare-earth magnets based on an alloy of iron, neodymium and boron has helped in the realisation of many new consumer products which could not otherwise have been possible, for example the walkman, the personal computer, the Laservision compact disc player, cordless power tools and the quartz analogue watch (which now accounts for almost half the world production of 700 million watches annually). Other major developments made possible by these new magnets include magnetic dentures, tiny 'brushless' motors, magnetically levitated (maglev) trains (see p. 7) and the replacement of hydraulic systems in aircraft and elsewhere by electrical ones. In addition, permanent magnets are now being used as a replacement for superconducting magnets in magnetic resonance imaging systems for medical diagnosis (see p. 16).

The Earth as a Magnet

For many centuries sailors and explorers have used the magnetic compass to help them navigate. Very recently we have discovered that some migrant birds and homing pigeons have magnetic sensors in their heads which help to guide them using the earth's magnetic field. We have also found out that the earth's magnetism forms a protective barrier against some of the charged particles that reach earth from space. In 1958 an American space satellite, Explorer 1, discovered the radiation belts, now known as the Van Allen belts, which surround the earth (see p. 7). These are composed of charged particles of cosmic and solar radiation which have been trapped by the earth's magnetic field. We are only just beginning to appreciate some of the ways in which the earth's magnetic field affects life on earth.

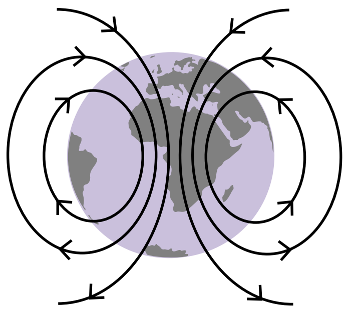


Fig. 1.2 *Earth's magnetic field*

It is believed that the earth's magnetic field is probably caused by electric currents circulating in the molten outer part of the iron-rich core of the planet, which is at a temperature of at least 2200 °C, Fig. 1.2. Note that it is the south pole of the earth's magnet that is located near the geographic north pole (which is why the north poles of other magnets are attracted in that direction).

The study of the magnetic fields of the geological past is called paleomagnetism (literally 'old magnetism'). Certain rocks, as they are formed, pick up permanent magnetisation in the direction of the earth's magnetic field at the time of their formation. Therefore, rocks carry a permanent record of the earth's magnetic field. The evidence shows that the magnetic poles have wandered over the years, and have undergone complete reversal at times, i.e. there were times when the pole located in the northern hemisphere was a north pole. This last occurred between 2.5 and 0.7 million years ago. (There is some speculation currently in scientific circles that another reversal is due in the next few hundred years. If this occurs, a number of species may become extinct, for example, sea creatures which depend on the earth's magnetic field to distinguish between up and down, or migrant birds who might then fly north rather than south for the winter.)

Angle of declination

The angle of declination, also known as the magnetic variation, varies from place to place as well as with time. At present the north magnetic pole is in northern Canada, approximately 1500 km from the true North Pole. So, for example, in the United States the angle of declination currently varies from about 25° E to 20° W, depending on location. In Ireland, the angle of declination varies from about 6° 30' W in the south east to about 8° 30' W in the west. (See also p. 5.)

Variations with time can be seen from the fact that in 1580 Gilbert measured the declination in London to be 11° E. It was 0° in 1657, reached 25° W in 1820 and decreased again to reach 8° W in 1990. The earth's magnetic north pole continues to drift eastward. In Ireland, this movement results in an annual variation of about 10' in the angle of declination. Magnetic compasses point to the magnetic north so navigators must know the magnetic declination at their particular location. Charts showing the magnetic declination are available for most parts of the world. On these charts points of equal magnetic declination are joined by lines known as isogonic lines; a line joining points where the magnetic declination is zero is called an agonic line.

Angle of dip

The angle of dip varies from place to place on the earth's surface. At the magnetic equator a magnetic needle, free to rotate in a vertical plane, will come to rest parallel to the earth's surface. As one moves north or south, the angle of dip increases until it reaches 90° at the magnetic poles, i.e. the needle rests perpendicularly to the earth's surface. It is for this reason that magnetic compasses cannot be used in the regions around the magnetic poles. The angle of dip in Ireland is approximately 70°. This means that, in Ireland, the vertical component of the earth's magnetic field is about three times its horizontal component.

One effect of the dip component of the earth's magnetic field is that iron posts, and indeed, retort stands, that have stood for a long time are usually magnetised. In the northern hemisphere, the upper ends of the posts have a south pole and the lower ends a north pole. In the southern hemisphere, the polarity is reversed. Another effect is that the north pole of an ordinary compass needle always dips towards the earth in the northern hemisphere and away from the earth in the southern hemisphere.

The earth's magnetic field is not symmetrical about the earth, but is distorted by the solar wind, which consists of streams of electrically charged particles emanating from the sun. The solar wind causes the magnetosphere, the region in space in which the magnetic field of the earth exerts a significant influence, to be flattened on the side facing the sun and elongated on the opposite side.

Elsewhere in the universe, the moon has no magnetic field, Venus has a weaker field than Earth, Jupiter's field is 20 000 times stronger than that of the earth, and many stars have strong magnetic fields. In the sun, the fact that the equator is rotating faster than the polar regions causes regular distortions of the solar magnetic field. These result in the creation of loops of magnetic energy called sunspots, and the release of intense streams of charged particles into space, some of which get trapped in the earth's magnetosphere (cf. Van Allen belts, p. 7) and which can cause short-wave radio interference. Maximum sunspot activity occurs in an 11 year cycle.

1.2 Do You Know

What causes magnetism in permanent magnets?

All magnetism is caused by the movement of electric charges. These charges are usually carried by electrons, whether through a wire or in an atom. In an atom the larger part of the magnetic moment is related to the spin of the electron. There is also a smaller magnetic moment related to the orbital motion of the electron.

In most materials the magnetic moments caused by the different electrons within each atom cancel each other out. In ferromagnetic materials, however, this does not happen, and many of these neighbouring magnetic moments align in a parallel direction forming a magnetic domain. These magnetic domains are visible under powerful microscopes. The domains are usually randomly oriented. If, however, they can remain aligned parallel to each other, a permanent magnet is produced.

What is ferromagnetism?

In 1825, William Sturgeon placed an iron core in a coil carrying a current, and found that the magnetic effect of the current was increased enormously. On switching off the current, the iron lost nearly all its magnetism. Iron, which can be magnetised strongly, is called a ferromagnetic material. Steel, made by adding a small percentage of carbon to iron, is also ferromagnetic. It is more difficult to magnetise than iron; however, it retains its magnetism after removal from a current-carrying coil.

A number of other elements are also ferromagnetic. The most common of these are nickel and cobalt. Alloys of ferromagnetic elements formed with each other and with other elements are also ferromagnetic. Such alloys are widely used in modern magnets. A modern alloy for permanent magnets, called Alnico, has the composition of 54% iron, 18% nickel, 12% cobalt, 6% copper and 10% aluminium. It retains its magnetism extremely well and, by analogy with steel, is therefore said to be magnetically very hard. Alloys which are easily magnetised, but do not retain their magnetism, are said to be magnetically soft. An example is mumetal, which contains 76% nickel, 17% iron, 5% copper and 2% chromium.

What is paramagnetism?

Paramagnetism is the quality of relatively weak attraction to magnetised bodies such as is shown by many substances, e.g. magnesium and tin. Paramagnetic behaviour occurs when an applied magnetic field causes the magnetic moments of the individual atoms to line up in such a way that the overall magnetic moment adds to that of the magnetic field.

What is diamagnetism?

Diamagnetism is characteristic of materials such as bismuth, antimony and their compounds which are repelled by magnetised objects. When diamagnetic materials are placed in a magnetic field, their induced magnetic moments line up in opposition to the applied magnetic field.

Why do we store permanent magnets with 'keepers'?

Magnets are stored with soft iron keepers across their poles to prevent them losing their magnetism. Magnetism is induced in the soft iron and so a closed loop of magnetic material is formed in which the magnetic dipole moments link up in closed chains. When the dipoles are linked together in this way they are much less likely to be knocked out of alignment. The strength of the links can be felt by trying to remove the keepers from a strong magnet.

How can magnets be demagnetised?

Anything which tends to shake magnetic dipoles out of parallel alignment in a magnet will reduce its magnetism.

a) Storing a magnet without keepers weakens its magnetism over a period of time.

b) Heating a magnet to a high temperature causes greatly increased vibration of its atoms which destroys any magnetisation of the material. Above a certain temperature, known as the Curie temperature (1043 K for iron), a magnet loses it magnetism.

c) Dropping and hammering magnets can knock the tiny magnetic dipoles out of parallel alignment.

d) If a magnet is placed in a solenoid carrying a slowly decreasing alternating current it is brought through ever decreasing hysteresis loops and can be completely demagnetised.

What is the origin of the term 'animal magnetism'?

Swiss-born **Franz Mesmer** (1734-1815), who was one of the first people to investigate hypnosis, believed that hypnotism was possible because of an invisible body fluid which he called 'animal magnetism' through which he could control his patients and cure their diseases. He even invented a special apparatus for this purpose!

It is from his name that the word 'mesmerised' is derived. His theory, of course, has long since been disregarded.

What is magnetic ink?

Magnetic ink is used on cheques so that machines in banks can read the cheque number and the account number and then automatically feed the information into the bank's computer. Cheques can be sorted in this way at a rate of nearly 2000 per minute. (cf. magnetic tape, p. 18.)

Why do airport runways have to be renumbered from time to time?

Airport runways worldwide are designated according to their direction relative to magnetic north. As the location of the north magnetic pole drifts, so, occasionally, must the runways be re-designated. For example, in December 1995 the 09/27 runway at Galway airport became 08/26. (The significance of this was that the direction of the runway, as indicated by a magnetic compass, changed from nearer 90°/270° to nearer to 80°/260° east of magnetic north, i.e. from >85°/265° E to <85°/265° E). Runways in Dublin and Shannon Airports have been re-designated also during the past fifteen years.

In addition to renaming runways, the Instrument Approach Procedure Charts used in air traffic control must be constantly updated, and new charts are prepared every few years. Every year, the magnetic variation assigned to each airport in the country is updated. The following are the values used at some airports in 1995.

Waterford	6° 49'	Cork	7° 46'
Dublin	7° 11'	Shannon	8° 10'
Connaught	8° 18'	Donegal	8° 26'
Kerry	8° 19'	Galway	8° 35'

Table 1.1 *Values of magnetic declination at selected airports in 1995*

There is an annual variation of 10' in the angle of declination at all airports in Ireland.

Why is soft iron generally used in the cores of electromagnets and transformers?

When the applied magnetic field is changing, the iron becomes magnetised and demagnetised. This involves aligning the magnetic domains and requires work to be done. This work appears as heat. Using soft iron, far less heat is produced in this way than with hard steel, as the magnetic domains are far more easily re-aligned in soft iron (see p. 4).

Why is a relay used with the starter motor of a car?

The starter motor in a car has the difficult job of turning a cold stiff engine when the lubricating oil is at its thickest. To provide the necessary power (up to 5 kW) from a 12 V battery requires a very large current. Since power = voltage x current, the current drawn can be as large as 400 A. Accordingly, the wires carrying this current must have as low a resistance as possible. For this reason, the starter motor is kept separate from the other circuits in the car and the wires in the motor itself and those connecting it to the battery are made of copper, are very thick and are kept as short as possible, i.e. the motor is as near the battery as possible.

When the ignition key is turned to start the car, a small current is passed through the coil of a relay which is fixed to the side of the motor. This closes the contacts and completes the starter motor circuit. In this way, the starter motor is operated remotely by turning the ignition key inside the car. Note that the relay on a car starter motor is usually referred to as ‘the solenoid’.

How do the doors on refrigerators and freezers close so well?

The plastic seal around the doors on refrigerators and freezers has a flexible strip inside it which is impregnated with powdered metal oxides. These are permanently magnetised and so the strip is attracted to the metal body of the refrigerator or freezer, ensuring a good seal.

What is a magnetic mine?

A magnetic mine is a form of non-contact mine usually laid on the sea-bed. It contains a delicate magnetic needle sufficiently sensitive to respond to the change in magnetic field when a ship passes overhead. This change causes an electrical contact to close and detonate an explosive charge. Magnetic mines took a great toll of merchant shipping in World War II.

What is a magnetic bottle?

A magnetic bottle is a configuration of magnetic fields used to contain the plasma in a nuclear fusion reactor. The magnetic fields are produced by very powerful electromagnets.

Extremely high temperatures, of the order of 10^8 K, are required in a fusion reactor. At such temperatures matter exists as a plasma, i.e. in the ionised state as electrons and positive ions. At a temperature of 10^8 K any solid container would vaporise and hence it is not possible to contain a high-temperature plasma in a material container.

A possible solution to this problem is magnetic confinement using magnetic fields shaped as shown, Fig. 1.3, to contain the hot plasma. The moving charged particles experience a force in the magnetic field which is at right angles to their velocity and to the magnetic field. The shape of the magnetic field is such that at all times this force directs the electrons inwards, thus confining the plasma. This arrangement is known as a magnetic bottle.

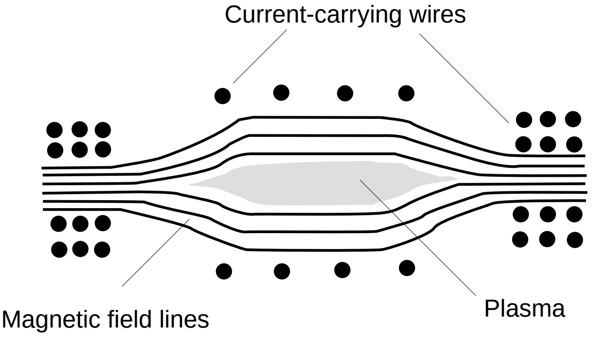


Fig. 1.3 *A magnetic bottle*

Unfortunately, some plasma tends to leak out of the ends of the magnetic bottles. More recently, a toroidal (doughnut-shaped) arrangement of magnetic field lines invented by the Russians and called a tokamak, Fig. 1.4, has been developed as a means of magnetic confinement of the hot plasma. In 1994, at the tokamak fusion reactor in Princeton University, scientists produced a controlled fusion reaction with a power output of 10.7 MW. In 1995 a plasma temperature of 510 MK was achieved in the Princeton reactor. In these experiments plasma currents of up to 3 MA were used. However, in all of the experiments the process still consumed more energy than it produced.

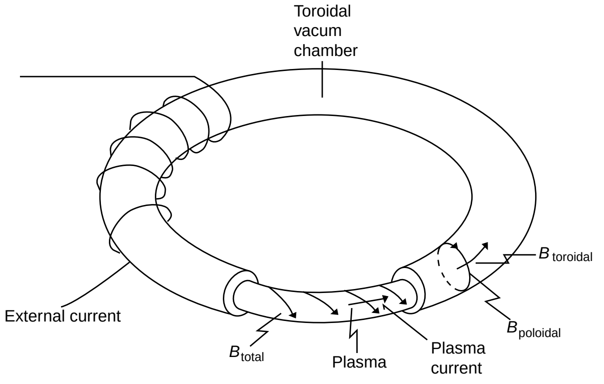


Fig. 1.4 *Tokamak*

What is a maglev train?

One of the more glamorous uses of magnets is in building mass transportation systems, based on magnetically levitated vehicles.

Maglev stands for magnetic levitation. A maglev train glides above a track called a guideway. The magnets used on the train may be either permanent or superconducting magnets. The guideway has a series of electromagnets which can be switched on and off as the train passes by. The interaction between the magnets on the train and the guideway provides all the accelerating and braking force, as well as the levitation force for the train as it glides above the guideway. Extra traction or braking is provided by varying the frequency in the electromagnet in the guideway. (Note: in some systems, the magnetic field due to the eddy currents induced when the train's magnets glide above the guideway is what provides the levitation.)

Because maglevs do not touch the guideway, and hence there is no friction, they offer greater acceleration, greater climbing ability, reduced noise and better performance in rain and snow. They are also more energy-efficient and cause less wear and tear on the tracks.

A maglev shuttle has been in use in Birmingham between the airport and the train station since 1984. However, greatest worldwide interest is in high speed maglev systems. Train speeds of over 500 km h^{-1} have been achieved in Japan. Plans for a high speed maglev route between Berlin and Hamburg (283 km) have recently been approved by the German parliament.

What are the Van Allen belts?

If a charged body is moving perpendicular to the lines of force in a uniform magnetic field, the force acting on it is always perpendicular to its direction of motion. Hence, the magnetic force does not change the speed of the charged body, but does change its direction. In fact, since the force is always perpendicular to the velocity, it acts as a centripetal force, causing the body to travel in a circle with uniform circular motion.

If the charged body has a velocity component in the direction of the magnetic field lines (but is not parallel to them), then the body will travel in a helical path as shown, Fig. 1.5.

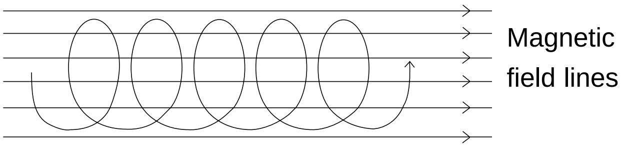


Fig. 1.5 *Charged particles follow a helical path in a magnetic field if their initial velocity is not at right angles to the field lines*

It is this spiral motion of charged particles in the earth's magnetic field that produces the Van Allen belts and the auroras. A stream of charged particles, from the sun and from outer space, continually sweeps past the earth. Many of these particles are deflected into spiral paths by the magnetic field of the earth, Fig. 1.6. Because the magnetic field is stronger near the earth's magnetic poles (i.e. the field lines are closer together), the particles are reflected and travel back and forth on spiralling paths between the poles. The extensive zones containing these rapidly moving, trapped particles are called Van Allen belts. At times of intense sunspot activity, the number of charged particles entering the Van Allen belts increases, causing them to overload. (These occurrences are known as magnetic storms.) When some of these particles approach the earth's magnetic poles and hit the atmosphere they excite the atoms in the gases, causing light to be emitted. This is what causes the aurora borealis (northern lights) and aurora australis (southern lights) which are often seen in the polar regions.

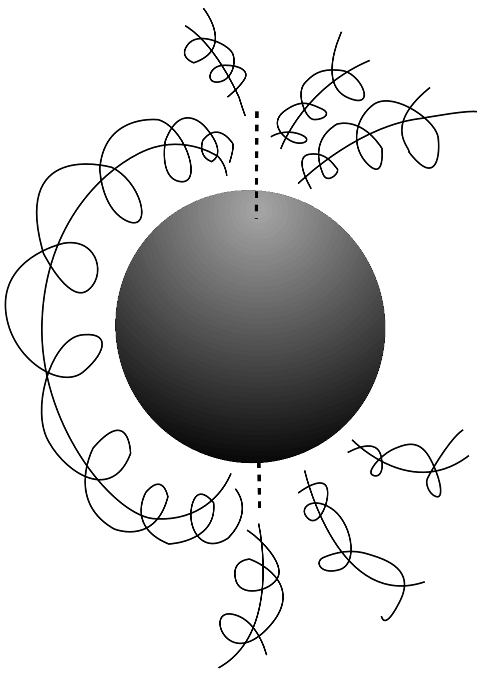


Fig. 1.6 *Charged particles travel in helical paths along the earth's magnetic field lines*

1.3 Conceptual Approach

The following concepts need to be understood.

1. All magnetic fields result from moving charges, and magnetic field lines are closed loops. The magnetic field lines are sometimes called 'lines of magnetic force'. The direction of a field line at a point is defined as the direction in which a free north pole would move if placed at that point.

2. In a magnetic field, the more tightly packed the lines are, i.e. the greater the number of lines per unit area, the stronger the magnetic field. This can be explained by looking at a diagram of magnetic field lines. Useful analogies can be drawn with contour lines on a map and isobars on a weather map.

3. The direction of the force on a current (or on a moving charge) in a magnetic field is perpendicular to the direction in which the current is flowing (or the charge is moving) and also to the direction of the lines of magnetic force. This concept may be introduced using Experiment 1.3.

4. The equation $F = IlB$, may be introduced using Experiment 1.3 to show that $F \propto I$. It can also be shown experimentally that $F \propto l$, the length of the conductor in the magnetic field (assuming the current is perpendicular to the lines of magnetic flux).

   So $F \propto I\,l$
   or $F = I\,lB$

   where B is a constant of proportionality whose value obviously depends on the strength of the magnetic field. B is called the magnetic induction, and is defined as the force per unit current per unit length of conductor in the direction at right angles to the magnetic field. B is a vector quantity and its unit is the tesla (T).

Note
The magnetic induction, B, is a measure of the *strength* of the magnetic field (just as the electric field intensity, E, is a measure of the *strength* of an electric field) because it determines the force on the conductor.

5. We noted above that the number of magnetic field lines per unit area is proportional to the strength of the magnetic field.

   So, number of magnetic field lines per unit area $\propto B$
   i.e. number of magnetic field lines per unit area $= kB$.

   By defining the value for the constant, k, to be 1, number of magnetic field lines per unit area is equal to B.

6. To introduce the idea of magnetic flux, Φ, consider an area, A, in a uniform magnetic field

of magnetic induction B. When the magnetic force lines are perpendicular to this area, Fig. 1.7, the total flux, Φ, through the area is defined as the product of B and A, i.e. $\Phi = BA$.

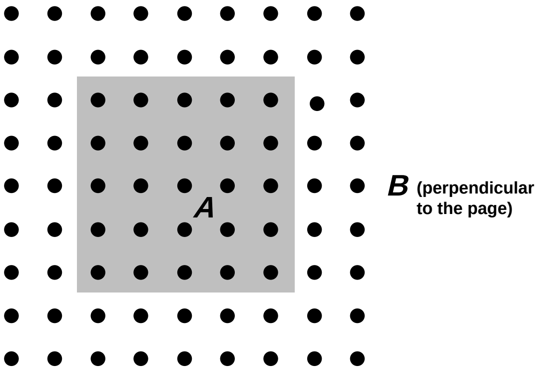


Fig. 1.7

The magnetic flux, Φ, can be visualised as the number of magnetic field lines passing through a given area. The number of magnetic field lines per unit area, i.e. B, is then usually referred to as the density of the magnetic flux or, more usually, the magnetic flux density.

7. Derivation of formula $F = qvB$.
Consider a section of conductor of length l through which a current I is flowing. If q is the total charge which carries the current in this section of the conductor, then $I = q/t$, where t is the time it takes the charge q to travel a distance l. The average velocity with which the charge flows is given by $v = l/t$, i.e. $l = vt$.

Substituting in $F = I\,lB$
gives $F = q/t \times vt \times B$
i.e. $F = qvB$

8. Definition of the ampere.
Electric currents exert forces on each other. The unit of electric current, the ampere, is defined in terms of these forces. These forces will obviously depend on the magnitude of the currents, the length of the wires, and the distance between them. The following is the definition of the ampere.

'The ampere is the current in a long straight thin wire that exerts a force of 2×10^{-7} N per metre on a neighbouring long parallel thin wire, 1 metre away in a vacuum, that carries an equal current.'

9. When introducing the concept of changing a galvanometer into an ammeter or voltmeter, a diagram such as that in Fig 1.8, is helpful in indicating that an ammeter consists of a galvanometer (which supplies the needle and scale) and a small parallel resistor (which allows a relatively large current to flow through the meter without burning out the coil). Some numerical problems should be introduced at this stage, (cf. p. 19).

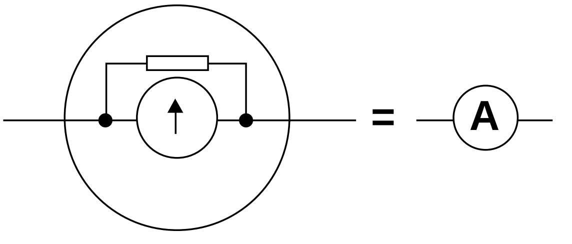


Fig. 1.8 *Galvanometer used as an ammeter*

Fig. 1.9 shows a simplified diagram of a multimeter (voltmeter and ammeter only). By turning the knob(s), resistors of different resistance can be added in series (voltmeter) or in parallel (ammeter) as required.

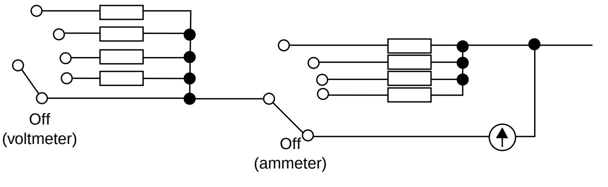


Fig. 1.9 *Galvanometer used as a multimeter*

Fig. 1.10 shows a diagram of an ohmmeter. The smaller the resistance connected across the terminals X and Y, the larger the current that will flow through the galvanometer. Thus, the ohmmeter gives maximum deflection for zero resistance. The variable resistor is connected to the 'set zero' knob.

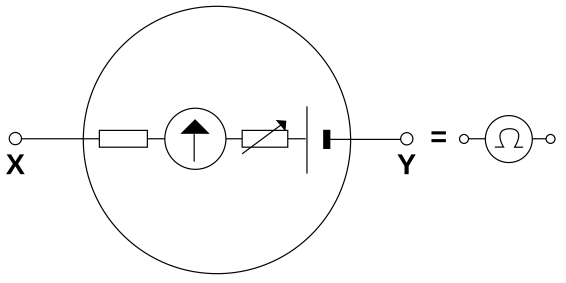


Fig. 1.10 *Galvanometer used as an ohmmeter*

1.4 Experimental Approach

Students will be familiar with magnets and with the idea of a force field creating action at a distance. The toppling of magnet A when magnet B approaches demonstrates this nicely (see Experiment 1.1).

Show that an electric current creates a magnetic field (see Experiment 1.2) and also that an electric current experiences a force in a magnetic field (see Experiment 1.3). These experiments show that electric currents have magnetic fields associated with them.

Look at the shapes of magnetic fields due to currents in straight wires, coils and solenoids. (The overhead projector is very useful for this purpose. Suitable coils and compasses for use with the projector are inexpensive and readily available.) Observe that the magnetic field lines are closed loops. Their direction is defined as the direction in which a free north pole would move, and can be predicted by the right hand rule.

Note the similarity between the shapes of the magnetic field due to a current in a solenoid and that of a bar magnet. Introduce the idea of all magnetic forces being due to currents or moving charges - magnetism in permanent magnets is due to the spin and the orbital motion of electrons (see p. 4). Ferromagnetic materials have many magnetic domains. These are groups of atoms (c. 10^{10} atoms) in which the magnetic moments of the individual atoms are aligned with each other. In an unmagnetised sample of ferromagnetic material the magnetic domains are arranged randomly. When the material is magnetised the domains are aligned with each other.

Experiment 1.2(a) introduces a new concept which students will not have met before, i.e. a force which does not act along the line between the two bodies, but perpendicular to that line. (Note: the compass is neither attracted to, nor repelled by, the wire - it is deflected perpendicular to it.) Use this concept, in conjunction with Experiment 1.3, to establish Fleming's left hand rule for the direction of the force on an electric current in a magnetic field.

Experiment 1.3 demonstrates very clearly that when the direction of the electric current is parallel to the magnetic field lines, it does not experience any force. This experiment is also a very effective method for introducing the equation $F = IlB$.

The force on moving charges in a magnetic field can be demonstrated by bringing a magnet near the screen of a cathode ray oscilloscope. (Take great care, as too strong a magnet can damage the CRO. In particular, **do not do this with a television screen.**)

Demonstrating Applications

A simple demonstration motor/generator, used as a motor, shows that a coil in a magnetic field experiences a torque. So, as well as demonstrating how a motor works, it also introduces the principle of the moving-coil meter.

A very good demonstration to introduce the loudspeaker is to get some speakers from old radios (easily obtained from your local TV dealer). Connect one to a signal generator at very low frequency (less than 1 Hz) and observe its vibrations. A table tennis ball or polystyrene beads placed on the cone of the speaker may be used to highlight the effect. Increase the frequency until the speaker produces an audible sound. Dismantle another and show the magnet and coil.

A simple circuit for demonstrating an electromagnetic relay is shown, Fig. 1.11. In a practical application of the relay only one battery or power supply would be used.

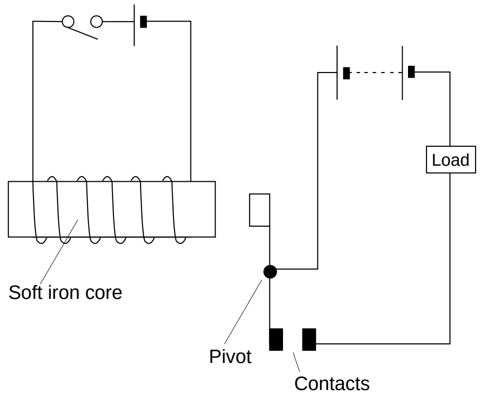


Fig. 1.11 *Simple electromagnetic relay circuit*

1.5 Experiments and Demonstrations

Experiment 1.1

Action at a Distance

Apparatus
Two magnets.

Method
Stand magnet A on its end and move magnet B towards it.

Observation
Magnet A falls without magnet B making contact - hence the idea of the magnetic field and force at a distance.

Experiments 1.2

To Show That Electric Currents Produce Magnetic Fields

Apparatus (a)
Straight wire, battery or power supply, plotting compass. If a power supply or accumulator is used a small resistor may be necessary to limit the current to c. 2 A.

Method
Place the wire in a North-South direction and place the compass above or below it. Connect the ends of the wire to the battery or power supply, through the resistor if necessary. Note that the compass needle is deflected and comes to rest perpendicular to the wire.

Apparatus (b)
Solenoid or coil of wire, battery or power supply, plotting compass. If a power supply or accumulator is used a small resistor may be necessary to limit the current to c. 2 A.

Method
Place the solenoid in an East-West direction and place the compass beside it, mid-way along its length. Connect the ends of the solenoid to the battery or power supply, through the resistor if necessary. Note that the compass needle is deflected and comes to rest parallel to the solenoid.

These are very simple, yet effective, demonstrations. Experiment 1.2(b) may be used to show that the magnetic field is strengthened by inserting an iron core into the solenoid.

Experiment 1.3

Force on a Current-Carrying Conductor in a Magnetic Field

Apparatus
U-shaped magnet, strip of aluminium foil, battery or power supply.

Method
Place the aluminium foil between the poles of the magnet, Fig. 1.12(a). Connect the ends of the foil briefly to the battery or power supply. Note that the foil experiences a force. If the force is downwards reverse connections to the battery or power supply. The force should now be upwards and the movement of the foil will be more obvious.

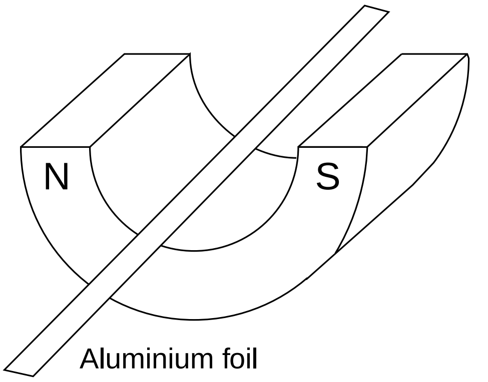


Fig. 1. 12(a) *Demonstrating the force on a current-carrying conductor*

With the arrangement shown in Fig. 1.12(a) the foil experiences a force which is perpendicular to the direction in which the current is flowing and also to the direction of the magnetic field lines. Reversing the current or turning the magnet around produces a force in the opposite direction.

If the foil is placed parallel to the field lines, i.e. from one pole to the other, no force is experienced. In practice a small force will usually be observed since the field is not uniform and it is therefore not possible to have the current parallel to all of the field lines.

To introduce the equation $F = I\,lB$, the force on the conductor may be determined as follows. For this part of the experiment it is better to use a rigid conductor, e.g. a copper rod. Fix the conductor firmly in a stand. Place the magnet on a top pan balance and set tare to zero. Increase the current from zero to a few amps, noting the reading on the balance for different values of the current. The reading on the balance is proportional to the force on the magnet and therefore to the force on the current. It will be seen that $F \propto I$.

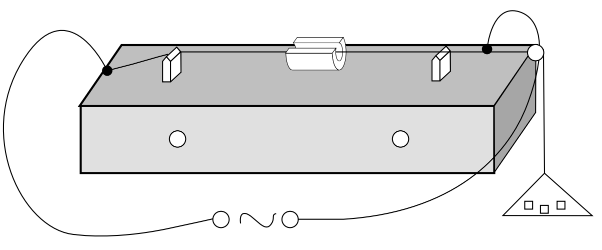


Fig. 1.12(b) *Using resonance to demonstrate the force on a current-carrying conductor*

A further variation on this experiment is illustrated in Fig. 1.12(b). The magnet is placed under the wire of a sonometer and the ends of the wire are connected to a signal generator. Again, the wire experiences a force, either up or down, when the current is switched on. However, since the current from the generator is alternating the force is also alternating. The frequency of the current from the signal generator is varied until the wire resonates. This occurs when the natural frequency of the wire is the same as that of the current from the signal generator.

(This experiment is also used to show how the natural frequency of a stretched string varies with the length and tension, cf. module on Waves.)

Experiment 1.4

To Remagnetise a Magnet

Place the magnet in the centre of a solenoid. Connect the solenoid to a battery or d.c. power supply.

Note

Many schools may have incomplete Norstedt Electricity and Magnetism Kits. Spare parts are available from laboratory suppliers.

1.6 Applications

Compasses

The magnetic compass has been used, in one form or another, for navigation for nearly 800 years. In its simplest form a compass consists of a magnetised needle mounted on a vertical axis at the centre of a graduated circular card. In the mariner's compass a magnetic needle is fixed to the underside of the compass card, so that the card rotates about a vertical axis. The assembly is housed in a glass-covered bronze bowl which is hung in gimbals so that the card remains horizontal despite the movement of the ship. A more common version of the mariner's compass has the compass card floating in a liquid. This increases friction and thus reduces the vibrations of the card resulting from the motion of the ship.

The traditional mariner’s compass soon proved unsatisfactory for use in aircraft. The use of spring-mounted bowls and cards of very small diameter reduced the problems but tilting still occurred in rapidly accelerating aircraft, resulting in false readings. The most important of these showed the aircraft to have turned through a greater or smaller angle than it actually did. This was known as the northerly turning error. The compass readings were also affected by the various metal components and electrical equipment in the cockpit. The magnetic compass has now been largely replaced by other types of compass, most recently the gyrocompass (see below).

The inductor compass

The traditional magnetic compass aligns itself with the horizontal component of the earth's magnetic field. The inductor compass, on the other hand, in effect measures the magnitude of the horizontal component of the earth's magnetic flux density and indicates the direction in which this quantity is a maximum.

One version of the inductor compass, the saturable-inductor compass, is based on metals which can be magnetically saturated very easily, i.e. metals in which it is easy to produce the maximum magnetic flux density. The magnetic flux density induced in a bar of such a metal in the earth's magnetic field is greatest when the bar is placed in a magnetic north-south direction. Thus, by measuring the magnetic flux density it is possible to determine when the bar is aligned with the earth's magnetic field. Compasses of this type have no moving parts and the readings are taken electronically and so may be fed directly to other instruments and navigational equipment.

The gyromagnetic compass

The erroneous readings given by magnetic compasses in a rapidly accelerating aircraft or speedboat can be avoided by keeping the compass mounting horizontal using a gyroscope. The compass may be a traditional needle type with the axis vertical and supported at both ends or it may be an inductor type (see above).

Another system combines the reading from a magnetic compass with observations of the orientation of the axis of a gyroscope. The magnetic compass gives reliable readings as long as an aircraft maintains a level course in a straight line in still air. However, the compass is useless during changes of direction or in rough air. The gyroscope, on the other hand, maintains its orientation in space under these conditions, even though its orientation may drift by several degrees over a period of one or two hours. A combination of observations from both instruments thus provides reliable directional information under all conditions.

Most modern navigation depends on the gyrocompass. This device is based purely on the properties of the gyroscope and is independent of the magnetic field of the earth. Therefore no adjustment need be made for magnetic declination or the proximity of magnetic materials.

The Fluorescent Light Tube

The fluorescent lamp is a gas discharge tube containing mercury vapour at low pressure. Electrons emitted from the hot electrodes collide with the mercury atoms causing them to emit radiation, mostly in the ultraviolet range. The inside walls of the tube are coated with a fluorescent material which absorbs the ultraviolet rays and emits visible light.

The fluorescent tube must be operated with an inductance coil (the choke) which, combined with a small auxiliary glow lamp (the starter), serves to ignite the lamp, Fig. 1.13. The glow lamp contains a bimetal thermal contact which closes when hot.

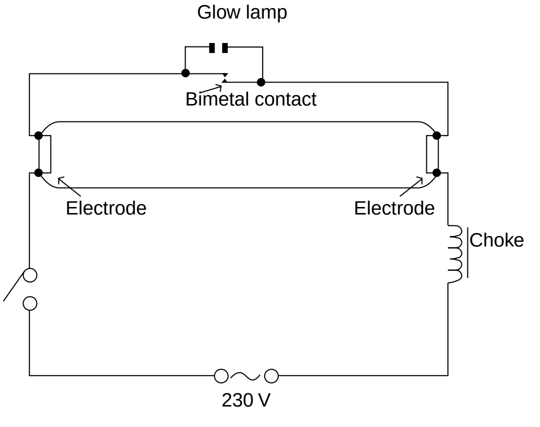


Fig. 1.13 *Fluorescent lamp*

When the current is switched on, the starter lights up (the bimetal contact is open). This causes the bimetal strip to heat up and close the contact which causes a short circuit through the starter, switching off the glow lamp. The bimetal strip now cools and breaks the contact. This causes a large back emf in the inductor which is sufficiently large (c. 10 kV) to begin the discharge in the fluorescent lamp itself.

The glow lamp is now bypassed and ceases to function (the bimetal contacts remain open). The inductor now serves a second function; it acts as a 'choke' to limit the current.

The fluorescent lamp is much more economical than the filament lamp (a 40 W fluorescent lamp gives approximately the same light output as a 100 W filament lamp) because the light is produced by excitation of the individual atoms of a vapour rather than by an incandescent body and so more of the energy is emitted as light rather than as infrared radiation. Also, because of its large light-emitting area it gives a pleasant light and produces only soft shadows.

Electric Motors

Motors are based on the principle that a current-carrying conductor in a magnetic field experiences a force. In small motors the magnetic field may be provided by a permanent magnet. In larger motors, however, the magnetic field is provided by a powerful electromagnet, whose coils are known as the stator coils or the field coils, Fig. 1.14. (Sometimes a 4-pole arrangement is used instead of just a pair of field coils.) Multiple coils are sunk into the surface of the soft iron of the rotor, so that the magnetic field passes from the iron core of the stator coils to the iron core of the rotor coils across the narrowest possible air gap. The commutator has a pair of segments for each coil in the rotor. The segments are arranged so that the current in each coil changes direction each time the coil is vertical. This ensures that the direction of the torque is always the same.

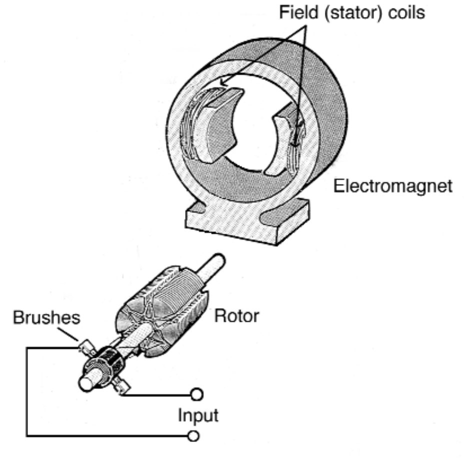


Fig. 1.14 *Electric motor*

Electric motors which run on alternating current

Alternating current (a.c.) changes its direction of flow very rapidly at a controlled rate. In the mains supply in Ireland, the current flows in each direction 50 times every second (the mains frequency is 50 Hz). This means that the current changes direction 100 times every second at a regular rate. It follows that if we use an alternating current for an electric motor, and provided the magnetic field is constant, a commutator is not required to reverse the direction of the current in the coils of the rotor to keep it turning because the current is already reversing itself.

So, a motor which runs on a.c. has two separate complete rings, called slip rings, instead of a commutator. Each end of the rotor coil is connected to a separate complete copper ring, and the same carbon brush makes continuous contact with the same ring. Each time the current reverses direction, the rotor must have turned half a revolution and be passing through the vertical position. In other words, if the current changes direction 100 times a second, the rotor must make exactly 50 revolutions every second to keep up with the rate of reversal of the current.
Such a motor is called a synchronous motor, because its speed (50 revolutions per second) is matched exactly, i.e. synchronised, to the frequency of the alternating current. A synchronous motor can only run at the one speed or not at all. This means that these motors will not self-start without special arrangements being built into their design. Once running, if they are slowed down by a load which is too heavy, they simply stop.

Motors with brushes and commutators

The simplest motors, both d.c. and a.c., need brushes and a commutator or slip rings to supply the current to the rotating coil or rotor. This type of motor can be improved in several ways.

In a series-wound motor the field winding (the coils of the electromagnet) of the motor is connected in series with the armature coils. Series-wound motors develop a greater starting torque because the armature current flows through the field coils. At the start, the back emf in the stationary armature is zero, so a large current flows, giving a very strong magnetic field. As the armature begins to rotate, the back emf increases, reducing the current. Series-wound motors are used in machines where a large starting torque is required, e.g. cranes.

A shunt-wound motor is one where the field winding is connected in parallel with the armature. In such a motor the current in the field winding is independent of the speed of the armature and so the strength of the magnetic field is constant. The speed of a shunt-wound motor is thus much more constant than that of a series-wound one. Shunt-wound motors are used for driving machine tools, and in other jobs where a steady speed is required.

It is interesting to note that a motor can burn out if the voltage *drops*. The explanation for this is the same as above - when the voltage drops, the motor slows down, the back emf decreases and so the current flowing in the armature increases.

Back emf in motors

When a motor is running there are two emfs in the rotor: an externally applied emf from a power supply and an internally induced emf due to the dynamo effect. The induced emf is less than the applied emf and, according to Lenz's law, opposes the applied emf. Because it opposes the applied emf it is usually referred to as a back emf (see Experiment 2.5).

The current which flows in the rotor coil can be calculated from the formula

$$\text{applied emf} - \text{induced emf} = I \times R$$

where R is the resistance of the rotor coil and the external circuit.

However, when the motor is not turning there is no back emf. (This occurs when the motor is just switched on or when a machine driven by the motor gets jammed.) The effect of this is to produce a much larger current in the rotor coil. This current can cause overheating and damage to the rotor coil.

Two methods are commonly used to protect motors from overheating.

a) A thermal cut-out switch is included in the circuit of the rotor coil, which causes the current to stop if the temperature rises. Many electric motors used in power tools are protected in this way.

b) A resistor is connected in series with the rotor coil to limit the current. When the motor reaches a certain speed this resistor is switched out.

Motors without brushes and commutators

Over 90% of all power motors used in the world are induction motors. In this type of motor, a rotating magnetic field is produced. This induces eddy currents in a metal disc (or cylinder) which cause it to rotate, and no brushes or commutator are required.

Induction motors have fewer moving parts and so are stronger and longer lasting. However, they only work with an a.c. supply (only changing currents produce induction effects), and their maximum speed is limited by the frequency of the supply. (A 50 Hz supply limits the motor to a maximum speed of less than 3000 r.p.m.)

Fig. 1.15 (p.16) shows the principle of a 3-phase induction motor. The diagrams show how a 3-phase current produces a rotating magnetic field in the stator which is wound with three sets of coils, one for each of the three phases. The graph shows how the three phases reach a maximum in turn. The magnetic field at six successive instances, (1) to (6), is shown. For simplicity, only the winding carrying maximum current is shown at each instant. The other windings carry smaller currents which add to the strength of the field.

The rotating magnetic field in the stator causes induced currents in the rotor which make it rotate with the field.

Tiny motors

The new high performance rare-earth permanent magnets have made possible the construction of tiny motors with sufficient power to drive such

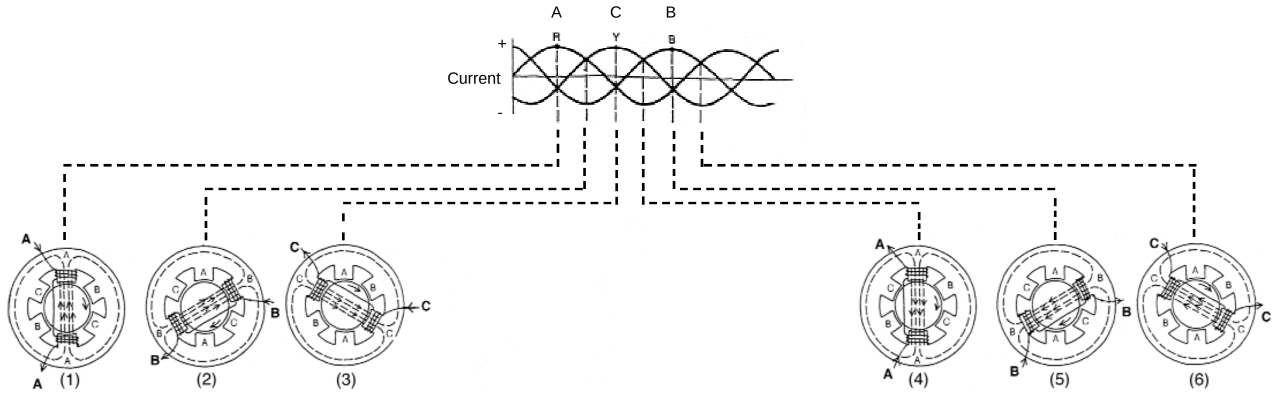


Fig. 1.15 *3 - Phase induction motor*

devices as the hands in quartz watches, and to position accurately the read/write heads in computer disc drives. Common designs are the stepper motor and the pancake motor (see below).

These tiny motors are 'inside out' motors in the sense that they have a permanent magnet rotor inside and a copper wound stator outside, the reverse of the traditional small motor design. Speed of rotation is adjusted very precisely by using electronic pulses from digital controllers.

Pancake motors

The flat 'pancake' motor is used where axial length is restricted and where nothing but a very thin motor will do. Pancake motor designs use an axial magnetic field rather than a radial field to produce useful power. In these designs the rotor takes the form of a disc with embedded magnets interacting with fixed planar windings.

The field strength achievable with the new magnetic materials means that the rotating magnetic disc need only be a millimetre thick, or even less, for small brushless motors.

Windings embedded in slots or etched windings have become a common approach for some flat motor designs. Pancake motors with an output rating of between 0.1 W and 15 kW are now being produced.

Research is ongoing into even smaller motors using nanotechnology. These motors could be as small as a grain of sand and have parts whose dimensions would be of the order of nanometres.

Magnetic Resonance Imaging (MRI)

The nuclei of many atoms possess spin (similar to electrons) and therefore have a magnetic moment. (Magnetic moments associated with nuclei are much smaller than those of electronic origin.) So, when the nuclei are placed in a strong magnetic field they will align themselves in the direction of the magnetic field.

If a pulse of electromagnetic radiation of the precisely correct frequency is applied, some of the nuclei absorb energy and flip through 180° so that their spin magnetic moment is exactly opposite in direction to the magnetic field. The required frequency is in the radiofrequency (RF) part of the electromagnetic spectrum. When the pulse ends the nuclei return to their equilibrium state with the emission of RF radiation. This radiation can be used to build up an image of the body containing the nuclei which emitted it. This technique is widely used in medicine and is known as magnetic resonance imaging (MRI). Because of its abundance in the body, hydrogen is the nucleus most used in imaging.

For parts of the body such as the brain, spinal cord and heart, MRI is superior to conventional scanning methods. It does not use X-rays, has no known dangers or complications and is non-invasive. It offers an early and easy diagnosis of many conditions, including cancer, and is a help in planning the most effective treatment.

Using the technique of magnetic resonance spectroscopy MRI instruments can be used to detect and measure chemicals within the body without having to take any samples - not even blood samples. The emitted RF spectrum is compared with the known spectra of various chemicals. This process is also used widely, other than in medicine, in chemical analysis and in research in nuclear physics.

Powerful magnetic fields which are uniform over a large area (big enough to accommodate a patient) are required. Very large electromagnets, whose coils are made of superconducting niobium-titanium alloy, are used. This is only superconducting at a temperature of -269 $^{\circ}$C (i.e. 4 K), so it must be immersed in liquid helium. The replenishment of the liquid helium is the major cost in the operation of MRI.

There are about half a dozen MRI units in hospitals in Ireland. The MRI unit in the Mater Private Hospital in Dublin is involved, as part of a European project, in research into schizophrenia and cancer.

The techniques used in MRI are known in physics as nuclear magnetic resonance (NMR). The term magnetic resonance imaging was coined by the medical profession to avoid anxiety among patients who feared they were being treated with nuclear radiation.

Miniature Circuit Breakers

A miniature circuit breaker (MCB) is a device which protects a circuit from current overload, i.e. it has a similar effect to a fuse. However, unlike a fuse, it can be reset rather than having to be replaced when a fault occurs. MCBs are used in modern house wiring circuits instead of fuses (see module on Electricity).

Large circuit breakers, e.g. those used to protect power stations, are sometimes immersed in oil to quench any arc which might form across the terminals when the current is interrupted.

Telephone Handsets

The telephone earpiece is simply a small version of the standard loudspeaker. It consists of a coil of wire placed over a permanent magnet. The coil is attached to a small diaphragm which is usually made of plastic. When an alternating current flows in the coil it experiences a varying force and moves in and out. The resulting movement of the diaphragm produces a series of rarefactions and compressions, i.e. a sound wave, in the surrounding air.

Two types of microphone are in use in modern telephone handsets. One consists of the same arrangement as the earpiece, i.e. diaphragm, coil and magnet, but worked in reverse. A sound wave causes the diaphragm to move in and out and this in turn causes a similar movement of the coil. Since the coil is moving in a magnetic field an emf is induced in it (see Chapter 2) and as a result an alternating current flows in the coil. This current corresponds to the original sound vibrations and is transmitted along the telephone wires to the local exchange. There it is redirected to the receiver of another telephone.

The more modern telephone handset microphone is based on a variable capacitor. Movements of the diaphragm result in changes in the capacitance of the capacitor which in turn result in an alternating voltage corresponding to the original sound vibrations, and ultimately an alternating current is transmitted to the local exchange.

Previously, telephones used a carbon-granule type microphone which was originally patented in 1878 and which changed only slightly in over 100 years of use. This consisted essentially of a small box with flexible sides and containing loose carbon granules. Sound waves entering the microphone caused the sides of the box to vibrate in and out. This movement alternately increased and decreased the density of the carbon granules and consequently increased and decreased their total resistance. This variation in the resistance caused a corresponding variation in the magnitude of the current flowing through the granules. This varying current carried the signal to the exchange as in the modern version of the device.

Telephone Transmission Lines

Telephone cables or transmission lines now carry communications of three main types; voice, data and visual telecommunications, e.g. TV pictures.

The rapidly expanding demand for telephone lines has led to many new developments which are aimed at carrying more communications with improved reliability. These include the following.

- Many conversations can now be carried along the same wire.
- Ocean cables and satellites are increasing the telephone links around the world.
- Integrated-circuit electronic systems have replaced the relatively unreliable electromagnetic relays in telephone exchanges.
- Glass fibres, carrying messages as pulses of light, are rapidly replacing copper wires. The glass fibre transmission lines are cheaper and can carry more communications than a much thicker and heavier multi-stranded copper cable. (See module on Light, p. 11.)
- Signals are increasingly transmitted along lines in digital form, rather than analogue form.

Recording Heads

Recording heads are small electromagnets with a very tiny gap (typically c. 0.3 μm in width) which are used to 'read' and 'write' on magnetic tapes and discs. Recording tape for use in audio and video recorders contains a thin layer of magnetic oxide on a thin plastic tape. During recording, the audio or video signal voltage is sent to the recording head, which acts as a tiny electromagnet, Fig. 1.16, that magnetises the tiny section of tape passing over the gap in the head at each instant. In playback, the changing magnetism of the moving tape at the gap causes corresponding changes in the magnetic field within the soft-iron head, which in turn induces an emf in the coil (Faraday's law). This induced emf is the output signal that can be amplified and sent to a loudspeaker (or in the case of a video signal, to a television set). In audio and video recorders, the signals are usually analogue - they vary continuously in amplitude over time. The variation in the degree of magnetisation of the tape from point to point reflects the variation in amplitude of the audio or video signal.

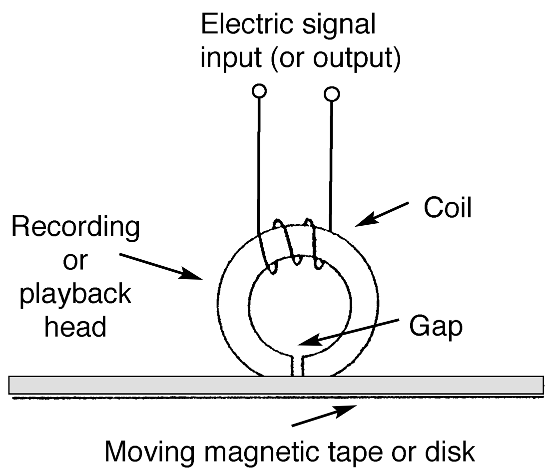


Fig. 1.16 *Recording/playback head*

Digital information, such as used in the process of reading and writing on computer discs (floppy discs or hard discs) is managed using heads that are basically the same as described above. The essential difference is that computer signals are not analogue, but are digital, meaning that only two values are possible. The two possible values are usually referred to as 1 and 0. These digital signals may then be stored, e.g. on compact disc or CD-ROM, for reproduction later. (In the CD player or CD-ROM drive these digital codes, which are embedded beneath the plastic surface of the disc, are read by a laser beam.) The magnetism on the storage disc or tape is also digital, with the magnetic moments of the tiny magnetic elements on the disc pointing either 'up' or 'down'. The writing and reading of a disc depends on changing magnetic fields, and emfs in the head, as described above, for analogue recording.

Research into new and better media for information storage is progressing rapidly. One research project at the University of Limerick concerns the development of magneto-optical discs. These make use of the fact that the direction of polarisation of light is rotated by magnetic fields. Such discs promise very high capacity (up to 1 gigabyte) and are about ten times faster than present floppy discs.

1.7 Worked Examples

1. (a) Why does an electric charge entering a uniform magnetic field perpendicularly, travel in a circular path?

   (b) If an electron, travelling at 10% of the speed of light, enters the earth's magnetic field at a location where the flux density is 10 μT, in a direction perpendicular to the lines of magnetic flux, what is the radius of its circular path? (Speed of light, $c = 3 \times 10^8$ m s^{-1}, charge on the electron, $e = 1.6 \times 10^{-19}$ C, take mass of electron = 9.1×10^{-31} kg.)

   (a) Because the force on the charge is always perpendicular to the direction in which the charge is travelling it acts as a centripetal force.

   (b)
$$F = evB$$
$$evB = \frac{mv^2}{r}$$
$$r = \frac{mv}{eB}$$
$$= \frac{9.1 \times 10^{-31} \times 3 \times 10^7}{1.6 \times 10^{-19} \times 10 \times 10^{-6}}$$
$$= 17 \text{ m}$$

2. A galvanometer has a coil of resistance 100 Ω and a full scale deflection of 2 mA. Calculate the resistance required to convert it to (a) an ammeter reading up to 1 A, (b) a voltmeter reading up to 10 V.

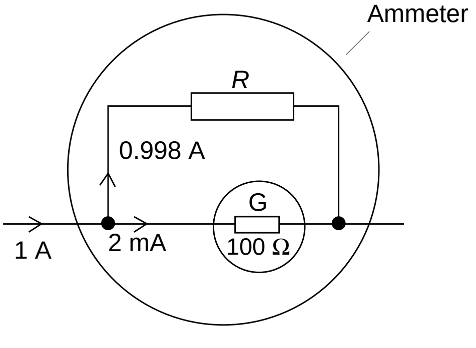


Fig. 1.17(a)

(a) Let R be the resistance of the parallel resistor required, Fig. 1.17(a). The current through the resistor is

$$I = 1 \text{ A} - 2 \text{ mA} = 0.998 \text{ A}$$

P.d. across the resistor = p.d. across the galvanometer

$$2 \times 10^{-3} \times 100 = 0.998R$$

$$R = 0.2\ \Omega$$

(b) Let R be the resistance of the resistor to be connected in series, Fig. 1.17(b). When the current is 2 mA (i.e. the galvanometer gives a full scale deflection), the p.d. across the voltmeter must be 10 V.

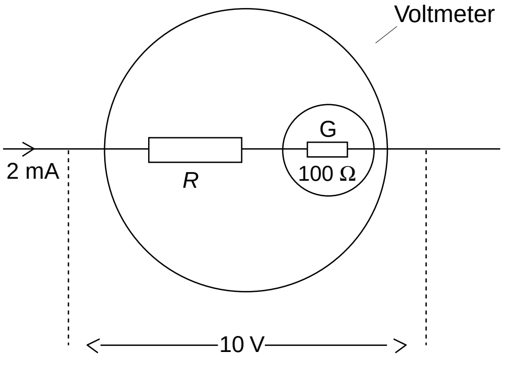


Fig. 1.17(b)

$$V = IR$$
$$10 = 0.002 \times (R + 100)$$
$$0.002R + 0.2 = 10$$
$$R = 4900\ \Omega$$

CHAPTER 2

ELECTROMAGNETIC INDUCTION

2.1 Historical Background

The work of **Michael Faraday** (1791-1867), generally regarded as one of the greatest of all experimental scientists, ushered in the electrical age. From his work on electric currents and magnetism, he developed the concept of 'lines of force' or 'field lines'. Using this concept, he went on to build an 'electromagnetic rotator', Fig. 2.1, which was the first device for producing a continuous motion using an electric current - the first motor.

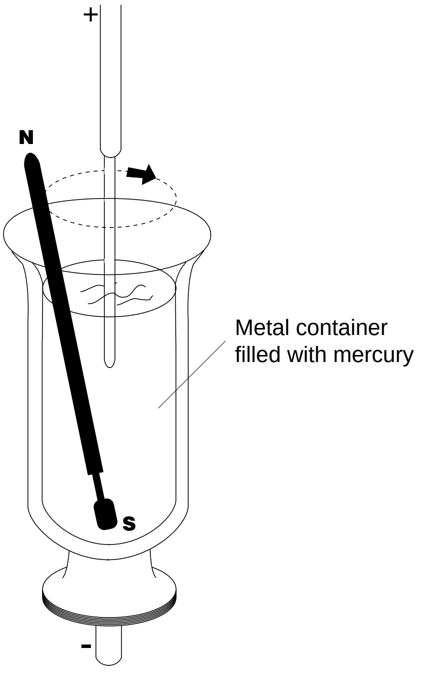


Fig. 2.1 *Faraday's 'electromagnetic rotator'*

Faraday's major discovery was electromagnetic induction. After Oersted and Ampère had shown that steady currents produce steady magnetic fields, it seemed reasonable to assume that steady magnetic fields should produce electric currents, and scientists all over the world involved themselves in research. It was not until 1831 that Faraday discovered, almost accidentally, that only *changing* magnetic fields produce electric currents. Using the apparatus shown, Fig. 2.2, he noticed that a current flowed in the secondary coil only when the current was switched on or off in the primary. Following further experiments with magnets and conductors moving relative to each other, Faraday enunciated his law of electro-magnetic induction. (Joseph Henry, in America, discovered electromagnetic induction at the same time as Faraday, and quite independently, but Faraday is usually credited with the discovery as he was the first to publish his results.)

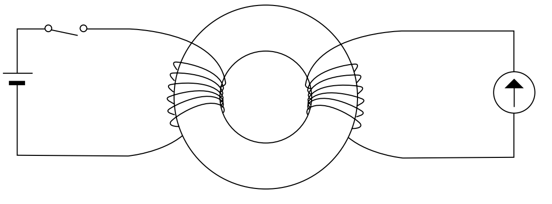


Fig. 2.2 *Faraday's demonstration of induced currents*

The discovery of electromagnetic induction is regarded as initiating the modern electrical age because it predicted that electricity can be produced by mechanical means, and there is a vast supply of mechanical energy available from many sources such as wind, falling water or continuous mechanical motion produced, for example, by a steam engine. His own 'Faraday disc dynamo', Fig. 2.3, was the first electric generator to produce a continuous current using mechanical energy, albeit very inefficiently. Despite many efforts, however, it was not until the 1870s that designs for practical large-scale generators were perfected.

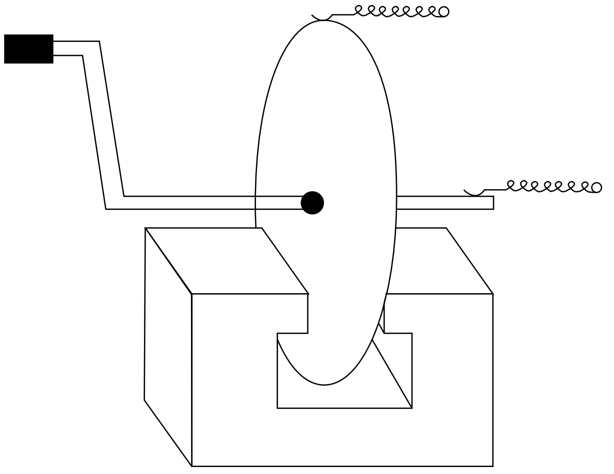


Fig. 2.3 *The 'Faraday disc dynamo'*

In Ireland, electromagnetic induction was investigated by **Nicholas Callan** (1799-1864). Fr Callan, who was professor of Natural Philosophy in St Patrick's College Maynooth, invented an induction coil in 1836 and went on to build a number of such coils, one of which he demonstrated at the Royal Society in London in 1837. The largest of these coils, built in 1837, could produce a spark 38 cm long. The secondary coil of this apparatus was wound manually with about 46 km of fine wire, all insulated by hand with beeswax and gutta-percha. Callan also invented a form of induction motor and suggested that it might be used in the electrification of the Dublin to Dun Laoghaire railway line. (**Heinrich Ruhmkorff** (1803-1877), a German instrument maker who resided in Paris, began producing induction coils in 1851 and is often erroneously credited with the invention of the induction coil.)

The unification of the theories of electricity and magnetism was achieved mathematically in 1873 by the Scottish physicist **James Clerk Maxwell** (1831-1879). Maxwell also predicted the existence of electromagnetic waves and identified light as an electromagnetic phenomenon. It was not until the development of atomic theory in the early part of this century that an explanation of magnetism in terms of electrons was possible.

The Electrification of Cities

Throughout the nineteenth century, efforts were made to provide street lighting for the growing cities of the industrialised world. In 1813, the first gas lights were installed on Westminster Bridge in London. The electric arc lamp had been discovered in 1801 by **Sir Humphry Davy** (1778-1829), and by the late 1800s dynamo-powered arc lamps were used in many European cities. One of the earliest public demonstrations of the electric arc lamp took place when the GPO in O'Connell Street (then Sackville Street), Dublin, was illuminated from Nelson's Pillar for the visit of Queen Victoria and Prince Albert in 1849. However, it was the invention in 1879 by **Thomas Edison** (1847-1931) of a practical light bulb that provided the impetus for the rapid development of more and better systems of electricity generation and transmission. The availability of cheap electricity led in turn to many other uses such as electric trams (from 1879), electric underground trains (from 1890 in London) and the electric elevator, or lift (from 1880) which made high rise buildings practical, as well as many new household and industrial uses.

A.C. Versus D.C.

The electrification of American cities began in the early 1880s following the invention of the carbon filament light bulb in 1879 by Thomas Edison. Initially, direct current was used, but in 1886 an alternating current system was used to provide street lighting in Buffalo, New York. About this time a major public controversy took place in the United States concerning the relative merits of d.c. and a.c. systems. The a.c. system has the advantage that a.c. is cheaper to produce because less sparking occurs using slip rings than in the commutator of a d.c. dynamo; a.c. is also cheaper and more efficient to transmit because its voltage can be easily stepped up or down. (The transformer had been invented by Joseph Henry in 1838.) However, there were fears that a.c. systems were unsafe because of the high voltages used for transmission. These fears were greatly exacerbated when, in 1890, New York State adopted electrocution as the means of capital punishment! However, the a.c. systems continued to expand without any spectacular accidents, and the dominance of the a.c. system was assured when the decision was made in 1893 to use a.c. for the new hydroelectric plant at Niagara Falls, using generators designed by **Nikola Tesla** (1856-1893).

Interestingly, in recent years d.c. has again gained popularity for transmission of electrical energy over long distances and also for transmission by underground cable. There are a number of reasons for this. Firstly, it can be transmitted at higher voltages than a.c. This is because with a.c., the effective, or r.m.s., voltage is less than the peak voltage, and the breakdown of insulation or of air is determined by the peak voltage. In addition, energy losses during transmission are lower for a given current with d.c. This is because a.c. produces alternating magnetic fields which cause induced currents in nearby wires resulting in loss of transmitted energy.

There is also a capacitance effect between the high voltage cable and the ground. With a.c. this results in energy being used to constantly re-align the molecules in the dielectric (air or insulation) between the cable and the ground. In the case of underground cables, where the distance between the cable and the ground is much smaller, this effect is far more significant, making d.c. transmission more economical, even over relatively short distances. For example, the underground link between France and England uses d.c. The use of d.c. also reduces the need for maintaining synchronicity between generators.

2.2 Do You Know

Is it easy to bug a telephone?

If the telephone wires are separated, and a wire leading to headphones laid alongside one wire, Fig. 2.4, then you can easily eavesdrop! This happens because the changing currents in the telephone wire cause induced currents in the headphone circuit. The bug would work much better if the wires were wrapped around a soft iron core! With optical fibre cables, bugging using this principle is not possible.

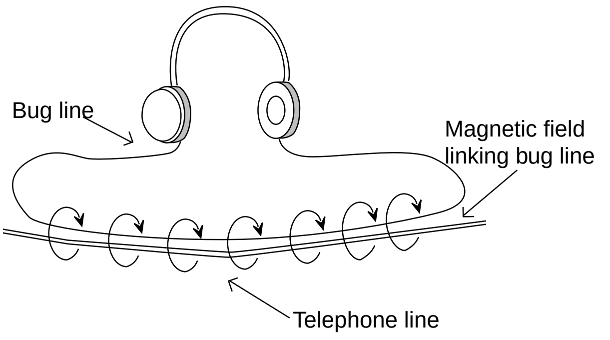


Fig. 2.4 *Bugging a telephone wire*

How does a seismograph work?

A seismograph, or geophone, uses the principle of electromagnetic induction to convert the motion of the earth, whether due to an earthquake or to an explosion (such as for mineral prospecting or for detecting a bomb test), into an electrical signal. A coil of wire, wound on a soft iron core, is fixed to the case of the seismograph while a magnet is suspended from the case by a spring, Fig. 2.5.

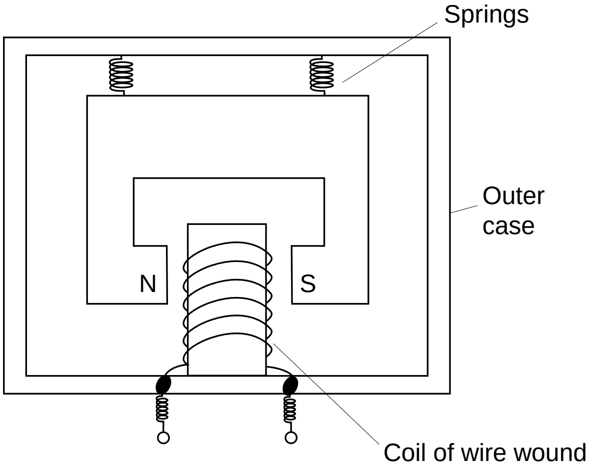


Fig. 2.5 *A seismograph*

When an earth tremor occurs the coil moves with the earth while the magnet tends to remain stationary, and the relative movement of the coil and magnet produces an induced emf in the coil, the nature of which provides information about the tremor.

How do large generators differ from small dynamos?

Small dynamos use permanent magnets to generate the magnetic field. Large ones have electromagnets whose coils are fed by a little of the dynamo's own output current.

The very large a.c. generators in power stations are called 'alternators'. In these, strong electromagnets rotate, driven by a turbine; these are called the 'rotor'. The coils in which the output voltage is generated remain stationary and are called the 'stator'.

This is a better design for large generators because it eliminates the need for brushes and slip rings (or commutator) for the large output current produced. Wires from the stator carry this current straight out to the consumers. This arrangement is particularly suited to the generation of high voltage a.c. as the sparking that occurs at high voltages would destroy the brushes.

The spinning rotor's electromagnets are supplied with the small direct current they need through brushes and slip-rings. That direct current usually comes from a small d.c. dynamo on the same spinning shaft as the big generator.
The alternators used in all modern cars are similar to those used in power stations.

Motors can also act as generators.

One of the difficulties for an electricity supply company such as the ESB is that electricity cannot be stored in large quantities. Yet the ESB must have sufficient capacity to meet the highest point of demand in any 24 hour period. As the daily peak period lasts for only a short time, it means that without a pumped storage station such as the one in Turlough Hill in Co. Wicklow, expensive generating plant would be out of operation for many hours each day.

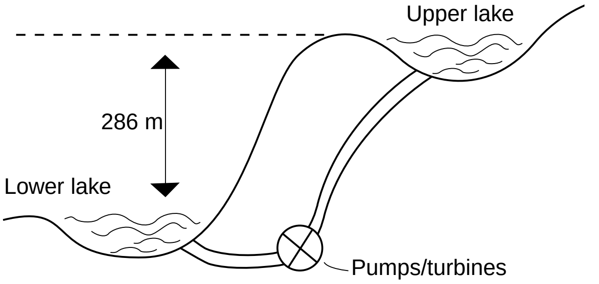


Fig. 2.6 *Pumped storage system*

Electric generators and electric motors have the same basic structure and the same device can, in principle, be used as either a motor or a generator. This fact is used in the operation of a pumped storage station. During the periods of low demand, e.g. at night, spare electric power from the system is used to drive large electric motors which drive pumps to pump water from a lower lake to an upper lake, Fig. 2.6. Then, during peak periods, water is allowed to flow back down again, rotating the 'motor' in the opposite direction so that it now acts as a generator and feeds extra energy into the system. This makes the whole system more economical and more flexible.

What is regenerative braking?

Some electric trains use regenerative braking on downhill slopes and when slowing down. This means that the motor functions as a generator, thus providing the braking effect, as well as making for greater economy, because the energy produced is fed into the supply line.

Regenerative braking is also used on *Le Shuttle*, the trains which travel the Channel Tunnel, achieving energy savings of up to 20%. There is also less wear on the brake blocks and less dust produced from their wear. Another important environmental consideration is that large transfers of energy to the environment are avoided.

What are the advantages of transmitting electrical energy at high voltages?

Since power = voltage x current ($P = VI$) the higher the voltage the lower the current for a given power output. The lower the current the less energy is lost as heat in the cables carrying the current (Joule's law: $P \propto I^2$). Thus the use of high voltages makes the distribution of electrical energy more efficient, and therefore more economical. A further advantage of using high voltages is that the lower currents allow thinner wires to be used for the cables. These are less expensive to produce and represent a smaller load on the pylons.

Due to insulation difficulties voltage cannot be increased without limit, but voltages as high as 400 kV are used for overhead transmission cables.

What is a dead-beat galvanometer or ammeter?

This is a moving-coil meter in which the coil is wound on a light-weight metal (usually aluminium) frame. As the coil turns between the poles of the magnet the frame also turns in the magnetic field and so emfs are induced in it. Large eddy currents are induced in the frame because of its very low resistance and, according to Lenz's law, the directions in which these currents flow are such that their effect opposes the movement of the coil. The coil and its pointer therefore turn smoothly and come to rest reasonably quickly, without overshooting the reading and oscillating about it. This effect is described as dead beat.

When an ammeter or galvanometer is being transported, a short length of wire of low resistance should be connected between its terminals. This joins the moving coil in a closed circuit. The damping effect of the eddy currents in the metal frame and the induced current in the closed circuit of the coil is sufficient to stop the coil almost completely from swinging about and being damaged.

What are non-inductive coils?

In some circuits containing coils, self-induction is a nuisance. A non-inductive coil is one designed to eliminate self-induction. This is done by doubling the wire back on itself before winding it into a coil, Fig. 2.7. As the current then flows in both directions through each part of the coil, the magnetic field produced is negligible. Such coils are used in rheostats and resistance boxes.

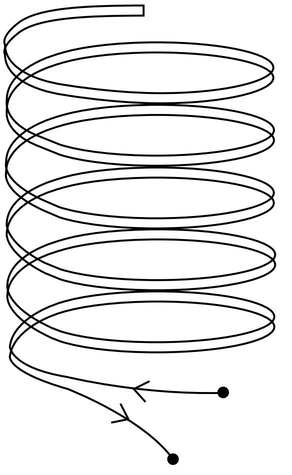


Fig. 2.7 *Non-inductive winding*

How is such a large emf induced when the current in a coil is interrupted?

When the current in a coil is interrupted by breaking the circuit, a large emf is induced. This can be seen in Experiment 2.4, or in an induction coil (p. 28) where a spark passes across the gap and energy is liberated in the form of heat and light. This energy has been stored in the magnetic field of the coil, just as the energy of a charged capacitor is stored in the electrostatic field between its plates. When d.c. current is first switched on, the back emf induced in the coil opposes the rise of current. The current flows against the back emf and therefore does work against it. The total work done is stored as energy in the magnetic field of the coil. It is released when the current collapses; for then the induced emf tends to maintain the current, and to do external work of some kind.

Ghost signals on the transatlantic telegraph cable.

When the transatlantic telegraph cable was first laid in about 1860, strange 'ghost signals' were often detected on the cable, even when no signals were being sent. It turned out that these were caused by the magnetic storms which occur in the earth's magnetic field. These caused a changing magnetic flux in the closed loop made up by the long cable and the earth and the induced currents were detected as the ghost signals.

Because the transatlantic cable was so long, small currents could flow back and forth from one end to the other. The great length of the cable enabled it to act like a capacitor where a.c. currents can flow back and forth from one plate to the other.

Safety with induction coils.

Induction coils can produce emfs of several hundred kV. Voltages of this magnitude can be very dangerous as the sparks produced by any voltage greater than 6 kV may generate X-rays in low pressure discharge tubes. So, induction coils should only be used with suitable screening, or when the voltage produced is less than 6 kV.

2.3 Conceptual Approach

1. Definition of electromagnetic induction: when the magnetic flux through a loop is changing, an emf is induced in the loop. Note that the loop does not need to be a conductor, although in practice it frequently is. However, the law applies equally to loops in insulators, including a vacuum. (See experimental approach outlined below.)

2. Lenz's law – the direction in which an induced current flows is such that its effects oppose the change which created it – is an example of the principle of conservation of energy. When a magnet is being pushed into (or being pulled out of) a coil, the person moving the magnet does work pushing the magnet in (or pulling it out). This work supplies the necessary energy to drive the current in the coil. A force is necessary to move the magnet in the coil because the effect of the induced current opposes the motion of the magnet – Lenz's law.

3. While Faraday's law applies to any loop, conducting or non-conducting, Lenz's law applies only when there is a closed circuit and current flows.

4. The equation $$E = -\frac{\mathrm{d}\Phi}{\mathrm{d}t}$$

   may be used as a statement of both Faraday's and Lenz's laws. The negative sign is a convention which indicates that the effect of the induced current opposes the change in the magnetic flux. It should be emphasised that the negative sign has no mathematical significance.

5. Generators and motors have the same structure, but generators change mechanical energy into electrical energy, while motors change electrical energy into mechanical energy.

6. The operation of generators, transformers and induction coils is based on the principle of electromagnetic induction, i.e. a changing magnetic flux produces an emf. Motors, meters and loudspeakers are examples of appliances based on the principle that a current in a magnetic field experiences a force. (Students often confuse these principles.)

7. The principle of the induction motor is explained in terms of Lenz's law. The changing magnetic field caused by moving the magnet past the disc, Fig. 2.8, causes currents to be induced in the disc. These currents oppose the changing magnetic field, i.e. they exert a force on the magnet which tries to stop it moving. Of course, there is an equal, but opposite, force exerted on the disc and, since the magnet does not stop, this force causes the disc to follow the magnet.

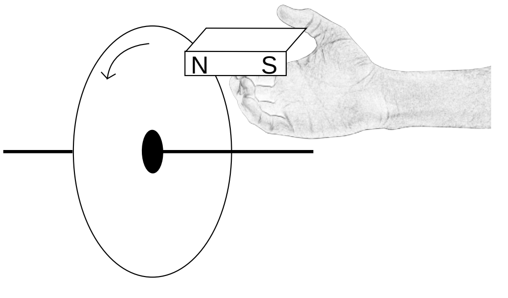


Fig. 2.8 *Demonstrating the principle of the induction motor*

8. The resistance of a piece of wire to the passage of electrical current depends on
   - length;
   - cross-sectional area;
   - resistivity of the material from which it is made;
   - temperature and other physical conditions (e.g. tension in a wire).

   If the length of wire is wound into a coil and a soft iron core inserted into it, a phenomenon called inductive reactance occurs. The coil and core assembly is often called a choke and it offers a special type of resistance to a.c. This has nothing to do with the factors affecting resistance mentioned above. This type of resistance is known as a reactance. The combination of a reactance and a resistance is called an impedance. Inductive reactance is produced as follows.

   When an a.c. current flows in the coil a magnetic field is generated. Because the current is changing continuously the magnetic field is also changing continuously. The coil is therefore in a continuously changing magnetic field and so an emf is induced in it. According to Lenz's law this induced emf opposes the applied emf and so the current in the coil is reduced. The induced emf in this case is often called the back emf and the process whereby a changing current in a coil causes a back emf in the coil is called self-induction.

9. In the transformer, an a.c current in the primary coil causes a changing magnetic flux. This changing flux is confined almost exclusively to the soft iron core because the permeability of soft iron is so much greater than that of air. The changing flux thus links the secondary coil which is also wound around the soft iron core and so an emf is induced in this coil. This process, whereby a changing current in one coil causes an induced emf in another coil, is called mutual induction.

10. The transformer increases or decreases a.c. voltages, while the induction coil primarily changes low voltage d.c. into high voltage pulses of d.c. (see p. 28).

11. The power output in a transformer is the same as the power input, assuming that there are no energy losses. Since power = voltage x current it follows that, if the voltage is stepped up, the current is stepped down and vice versa. (See Experiment 2.3.)

12. Most alternating currents vary sinusoidally, so what is the effective value of such an a.c.? The accepted definition is that one ampere of an a.c. is that current which produces the same power in a purely ohmic resistor as one ampere of d.c.

    Because the alternating current varies sinusoidally,

    $$I = I_o \sin 2\pi f t$$

    where I is the instantaneous value of the current at time t, I_o is the maximum, or peak, current and f is the frequency of the a.c. supply.

    $$\text{Power, } P = I^2 R$$

    Therefore the instantaneous value of the power is

    $$P = (I_o \sin 2\pi f t)^2 R$$

    The average value of the power is

    $$P = \frac{I_0^2 R}{2}$$

    since the average value of $(\sin 2\pi f t)^2$ over a full cycle is $1/2$.

    For the equivalent d.c.,

    $$P = I^2 R$$

    By comparison with the corresponding expression for a.c.,

    $$I^2 = \frac{I_0^2}{2}$$

$$I = \frac{I_0}{\sqrt{2}}$$

I is called the root-mean-square current, I_{rms}

It can be shown similarly that

$$V_{rms} = \frac{V_0}{\sqrt{2}}$$

In summary, the rms current provides the same power in a purely ohmic resistor as a direct current of the same magnitude.

13. The induction coil is perhaps best explained by examining it piece by piece. A primary coil, which consists of a small number of turns of fairly thick wire, is wrapped around a laminated soft iron core and is attached to a 'make-and-break' mechanism as shown in Fig. 2.9(a).

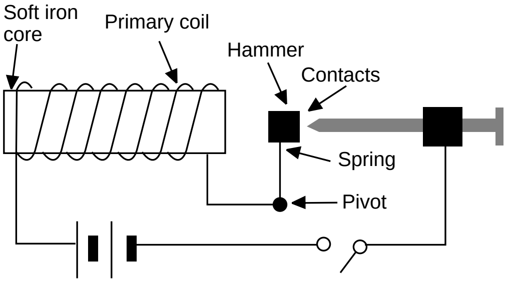


Fig. 2.9(a) *Primary circuit of induction coil*

When the coil is connected to a low voltage d.c. source, the coil becomes magnetised and attracts the hammer (called the armature) towards the core. This breaks the contact and stops the current flowing in the coil, causing the magnetic field to collapse. The armature is no longer attracted to the soft iron core and is pulled back from it, closing the contact again. The current again begins to flow in the coil, and the process is repeated. This occurs very quickly.

Because the current in the coil is fluctuating, the magnetic flux in the core is constantly changing. Thus self-induction produces a back emf in the coil. As the current is increasing in the coil this back emf opposes the increase. So the current increases relatively slowly, resulting in a back emf which is not too large (≤ supply voltage) since

$$\left\{ E = N\frac{d\Phi}{dt} \right\}$$

At the break, however, the current stops suddenly, causing the magnetic field to collapse quickly and so causing a large back emf.

This would cause a spark to occur at the contacts, resulting in the contacts getting hot and melting. By putting a capacitor across the make-and-break contacts as shown, Fig. 2.9(b), this spark is greatly reduced since the charge tends to flow into the capacitor rather than across the gap.

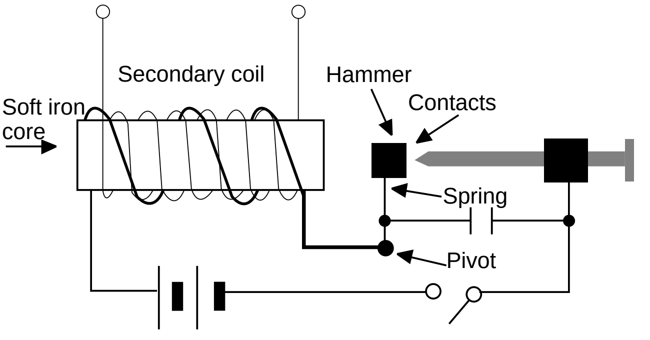


Fig. 2.9(b) *Complete induction coil with capacitor across contacts*

A secondary coil of very many turns of thin insulated wire is wrapped over the primary coil, Fig. 2.9(b). The changing magnetic flux in the primary coil now links the secondary coil. When the contacts open the change is very rapid and because the number of turns in the secondary coil is very many times the number of turns in the primary, the induced emf is very large since

$$\left\{ E = N\frac{d\Phi}{dt} \right\}$$

It may be big enough to cause a spark to occur between the ends of the secondary coil. Emfs of many thousands of volts can be readily obtained from such an induction coil, using a primary source of 6 V to 12 V.

The two coils and the make-and-break mechanism form the essential parts of the induction coil.

The emf at the 'make' will be opposite to and, as explained above, will be much smaller than that at the 'break'. So, in effect the emf induced in the secondary coil is in one direction. Fig. 2.9(c) shows the output voltage from an induction coil.

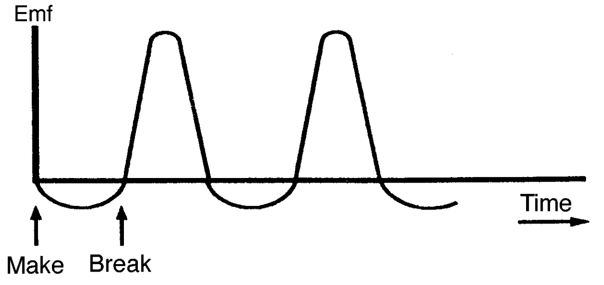


Fig. 2.9(c) *Output of induction coil*

2.4 Experimental Approach

When introducing electromagnetic induction, strongly emphasise that an emf is induced in a particular loop only when the magnetic field (or magnetic flux) through the loop is *changing*. A current is induced in the loop only if the loop is a conductor and the magnitude of the current will then depend on the resistance of the loop. In most cases in practice the loop is a coil of wire.

Note

The magnetic flux can be changing in three ways.

1. The magnet can be moving and the loop stationary (as in a bicycle dynamo) or moving more slowly (as in a simple induction motor, see p. 25).

2. The magnet can be stationary and the loop moving (as in a simple generator).

3. The loop can be near a varying current (as in a transformer).

It is well worthwhile demonstrating each of the above cases.

1. Move a magnet into a coil and note the direction in which the current flows as a result of the induced emf (see Experiment 2.1). Note its direction as the magnet is withdrawn. Turn the magnet around and repeat.

2. Stand the magnet on its end and move the coil up and down over it and again note that there is an induced emf.

3. Place the coil near another coil in which a smooth d.c. current is flowing. Note that there is no induced emf except when the current is switched on or off. Put a rheostat into the circuit and adjust the current in the primary coil. Note that an emf is induced in the secondary coil when the current is varying in the primary.

4. Place the coil near another coil in which an a.c. current is flowing. Note again that an emf is induced. Observe the effect of varying the distance between the coils. Note also the effect of linking the coils with (a) some iron rods e.g. retort stand rods, (b) a soft iron core. (This demonstrates that magnetic flux is carried much better in soft iron than in air, i.e. soft iron has a high magnetic permeability.)

When demonstrating the a.c. generator (a bicycle dynamo works very well), connect it to a d.c. milliammeter and rotate it slowly by hand. It can be seen clearly that the current changes direction twice for every complete rotation. Investigate the relationship between the speed of rotation and the emf (current) induced (use an a.c. meter in this case).

If the generator has a handle so that it can be rotated at constant frequency, the graph of output current against time can be shown using the oscilloscope. This is a good introduction to the concept of alternating current.

It is worth showing that if a d.c. motor is rotated manually it acts as a d.c. generator, since it is a coil rotating in a magnetic field. This shows that the motor and generator have the same basic structure (cf. p. 26).

The principle of the transformer is very easily demonstrated using a demountable transformer (see Experiment 2.3). Use this apparatus also to demonstrate the concept of mutual induction.

It is worthwhile, and interesting, to open up a variable low voltage power supply and look at the circuit which the current follows: through the fuse, the switch, the transformer, the a.c. output terminals and, through the bridge rectifier, to the d.c. terminals. The tapping of the secondary coil and the earthing of the metal case are also easy to see.

Experiment 2.4 demonstrates clearly the concept of self-induction.

To demonstrate peak and r.m.s. voltages, connect a low, smooth d.c. voltage to the Y-input of the oscilloscope. Measure the displacement. Then connect an a.c. voltage of the same nominal magnitude to the Y-input and note the peak voltage. Calculate the r.m.s. voltage.

2.5 Experiments and Demonstrations

Experiment 2.1

To Demonstrate the Principle of Electromagnetic Induction and Faraday's Law

Connect a coil to a centre-zero milliammeter, Fig. 2.10. Move the magnet into and out of the coil. Note that the needle is deflected when there is relative movement between the coil and magnet. Note the direction in which the needle is deflected. (A digital meter may be used but the concept of changing direction may not be conveyed as strongly to pupils as with an analogue meter.)

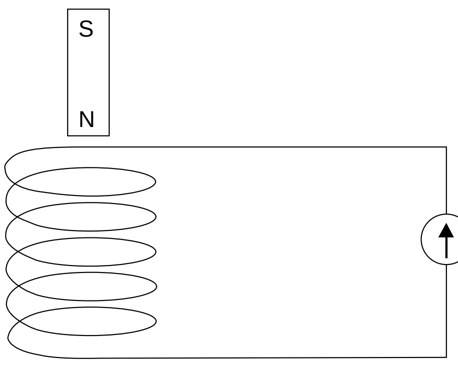


Fig. 2.10 *Demonstrating Faraday's law*

Move the magnet quickly into the coil. Note the deflection on the needle. Now move the magnet in slowly and note that the deflection is smaller, demonstrating Faraday's law.

Use coils with different numbers of turns. Move the magnet in and out at the same speed. Note that the deflection is proportional to the number of turns, i.e.

$$\left\{ E = N \frac{\mathrm{d}\Phi}{\mathrm{d}t} \right\}$$

Experiment 2.2

To Demonstrate Lenz's Law

Hang the aluminium plate from point X, Fig. 2.11(a), and allow it to swing freely with no magnet nearby. Note the time taken to come to rest. Now allow it to swing freely between the poles of a magnet as shown. Note that the plate now comes to rest much more quickly. This is because currents are induced in the plate as it moves in the magnetic field. These currents oppose the changes in the magnetic field (i.e. the movement of the plate) which produced them. This demonstrates Lenz's law. The currents produced in the plate are known as eddy currents. Note that aluminium is a non-magnetic material so the forces evident can be explained only in terms of induced currents.

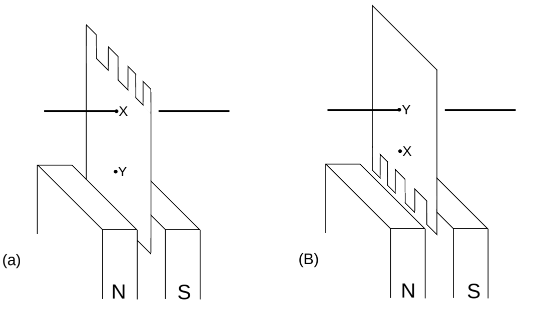


Fig. 2.11 *Demonstrating Lenz's law - method 1*

Now hang the plate from point Y and repeat, Fig. 2.11(b). Note that the damping effect is less in this case. This is because eddy currents are reduced. Note that this shows why laminating the core of a transformer reduces energy loss.

Lenz's law may also be demonstrated by hanging an aluminium cylinder between the poles of a magnet, Fig. 2.12. First, allow the aluminium cylinder to rotate freely away from the magnet and note the time taken to come to rest. Now place it between the poles of the magnet and repeat. Note that it comes to rest much more quickly. This is because the induced eddy currents oppose the movement of the cylinder. This, as well as demonstrating Lenz's law, illustrates the principle used in damping a moving-coil galvanometer.

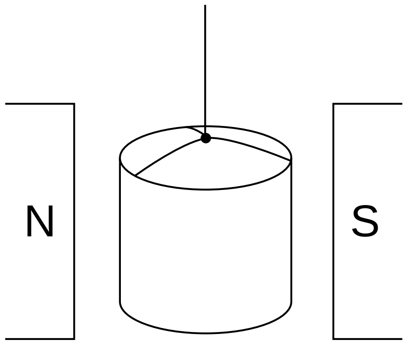


Fig. 2.12 *Demonstrating Lenz's law - method 2*

A further method of demonstrating Lenz's law involves suspending an aluminium ring from a long thread and moving a strong magnet in and out of it. As the magnet is pushed into the ring the ring is repelled and as it is pulled out again the ring is attracted, i.e. it tries to follow the magnet. If a ring with a break in it is used, it does not follow the magnet. Note that aluminium is a non-magnetic material so the forces evident can be explained only in terms of induced currents. (This experiment is not as easy to perform satisfactorily as the previous ones, unless a very strong magnet is used.)

The jumping ring experiment also demonstrates Lenz's law (see p. 32). Other experiments which demonstrate Lenz's law include the induction motor (p. 26 & 34) and back emf in a motor (p. 34).

Experiment 2.3
Experiments Using the Demountable Transformer

Safety Note

Connect only coils specially designated for that purpose to the mains supply. When using the mains supply with the primary coil use only designated secondary coils. Do not attempt to step up the voltage when using the mains supply as this would produce dangerously high voltages, and might also damage the insulation of the coils.

If using 4 mm plugs or other loose wires to connect the coil to the mains supply, use a safeblock, and make sure that everything is connected up before plugging into the mains. Always remove the mains plug before making any changes to the circuit. Safeblocks are quite inexpensive and are available from laboratory suppliers.

In general, care should always be taken with any circuit containing a coil. Even when the supply voltage is low, induced voltages can be very large since the induced voltage is proportional to the number of turns in the coil as well as the rate of change of flux through the coil.

Step-up and Step-down Transformer

Use a 12 V a.c. source for the primary coil. Use various secondary coils. First measure the primary voltage. Calculate the expected secondary voltage and then measure the actual secondary voltage.

The observed output voltage is a little less than calculated because of energy losses in the transformer. These are caused by (a) incomplete soft iron core, where bottom and top (sometimes called the 'U' core and the 'I' core) join, (b) eddy currents in the core, (c) resistance of the wires in the coils, (d) hysteresis, i.e. the energy lost in magnetising and demagnetising the iron in the core.

Inductive Heating

Set up the arrangement shown in Fig. 2.13. The circular copper dish acts as a secondary coil of one turn which has a very low resistance, so large currents are induced and a lot of heat generated. If some water is placed in the dish it will begin to boil very rapidly. If some solder is placed in the dish, it will melt. This principle is used in instant heat solder guns. (This experiment uses mains supply. See note on safety.)

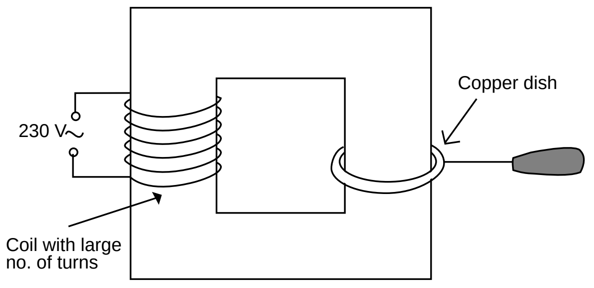


Fig. 2.13 *Inductive heating*

Spot Welding

The welder consists of a few turns of heavy gauge conductor as a secondary coil which is fitted with a pair of insulated handles, Fig. 2.14. By squeezing the handles, the two contact points are brought together. These contacts should be cleaned with sandpaper before use to remove any oxidation.

The materials to be welded are held between the contacts as the handles are squeezed together gently for a few seconds. The large current generated heats the strips at the point of contact until they become very hot and the metal softens. The handles are now squeezed more tightly pressing the two metal strips into each other. On releasing the pressure and cooling, a 'spot' weld results. (This experiment uses mains supply. See note on safety.)

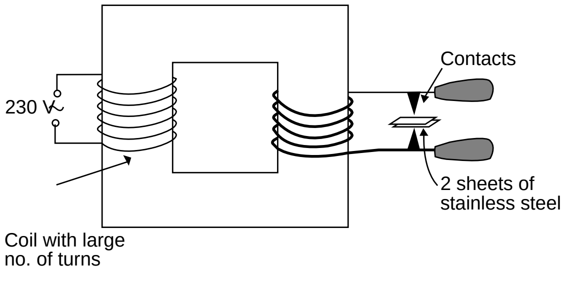


Fig. 2.14 *Spot welding*

Notes

1. Thin sheets of stainless steel or high carbon steel are most suitable for spot welding as they have high resistance and so more heat is produced for a given current.

2. This type of welding is called resistance welding (cf. arc welding, p. 37).

Jumping Ring

The straight iron core (the I-core) is placed on top of one limb of the U-core, Fig. 2.15, and the aluminium ring is placed over it. When the current is switched on in the coil, there is a sudden change in the magnetic flux linking the aluminium ring. This induces an emf in the ring and a very large current flows in the ring (perhaps several hundred amperes because the resistance of the ring is very low). The direction in which the current flows in the ring is such that it produces a magnetic field opposing the inducing magnetic field. The ring will jump into the air. If the ring is held to prevent it jumping, it gets very hot. If a ring of identical shape and size, but with a gap in it, is used, neither movement nor heating is observed. This illustrates the distinction between the induced emf and the induced current. The emf is the same in both cases (since rate of change of flux is the same) but the induced current in the second case is practically zero because of the very high resistance of the air gap in the ring. (This experiment uses mains supply. See note on safety.)

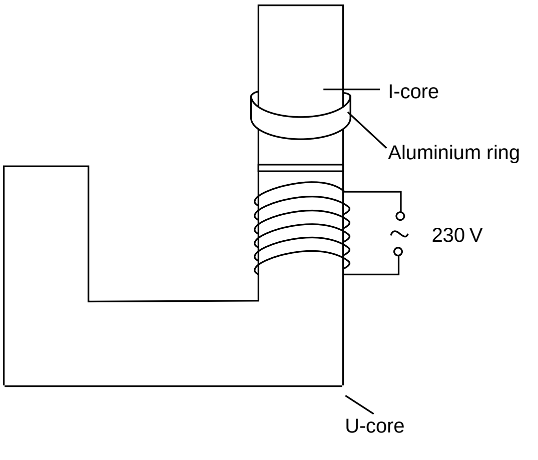


Fig. 2.15 *Jumping ring experiment*

Experiment 2.4

Self Induction

Back Emf When D.C. Is Switched On

Apparatus

Two small bulbs (e.g. 6 V, 0.3 A), coil (approx. 1000 turns) with soft iron core, rheostats, 6 V d.c. supply.

Procedure

Set up the circuit as shown in Fig. 2.16. Adjust R_2 until both bulbs are lighting. Adjust R_1 until both have equal brightness. (Do this when A is just lighting.) Switch off the current. Now close K and observe that B lights up more quickly than A.

Explanation

Bulb A is slower to light because of the back emf generated in L which opposes the increasing current.

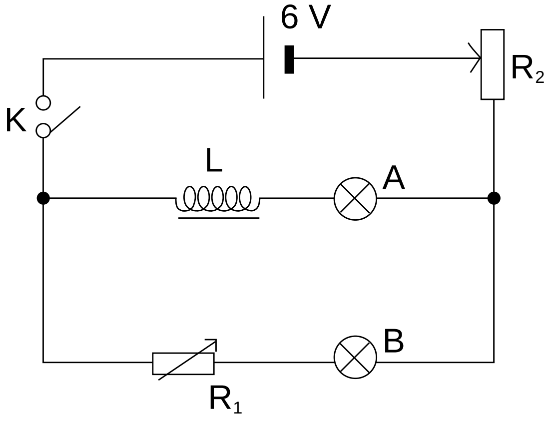


Fig. 2.16 *Demonstrating back emf in a coil*

Back Emf When D.C. Is Switched Off

Apparatus

Coil (approx. 1000 turns) with soft iron core, neon bulb (mains type), 6 V d.c. supply.

Procedure

Set up the circuit as shown in Fig. 2.17 with the switch K closed. Now open K, and observe that the bulb flashes.

Explanation

The sudden change in magnetic flux in L when the current is switched off induces a large back emf which causes the bulb to flash. (It takes a voltage of c. 90 V to make the bulb conduct.)

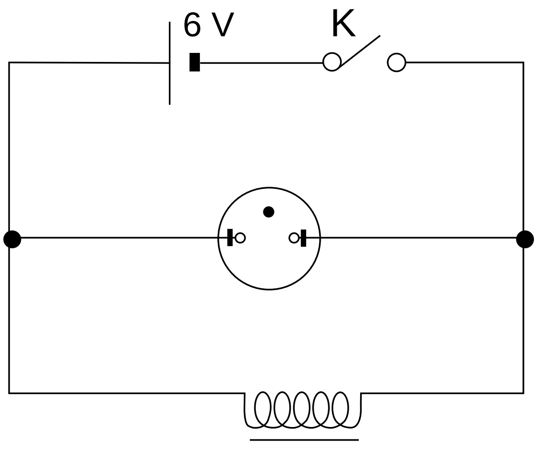


Fig. 2.17 *Demonstrating self-induction*

To Show Self-Induction With A.C.

Use the same apparatus and circuit as in Fig. 2.16, but replace the d.c. with an a.c. source. B is seen to light more brightly than A. Note also the difference in the brightness of A with and without the soft iron core.

Explanation

Though the coil has the same ohmic resistance as R_1, its impedance to a.c. is greater (cf. p. 27).

Alternative Method of Showing Self-Induction

Apparatus

Small bulb (e.g. 6 V, 0.3 A), coil (approx. 1000 turns) with soft iron core, 6 V a.c. supply.

Procedure

Set up the circuit as shown in Fig. 2.18. Note the brightness of the bulb when the iron core is not used. Now gradually insert the core into the coil. As the core is inserted, the bulb becomes dimmer.

Explanation

The soft iron core increases the strength of the magnetic field in the coil. Since the frequency of the supply is fixed (50 Hz) it follows that the rate of change of flux through the coil is increased. The induced emf is therefore increased and so the impedance of the coil is increased.

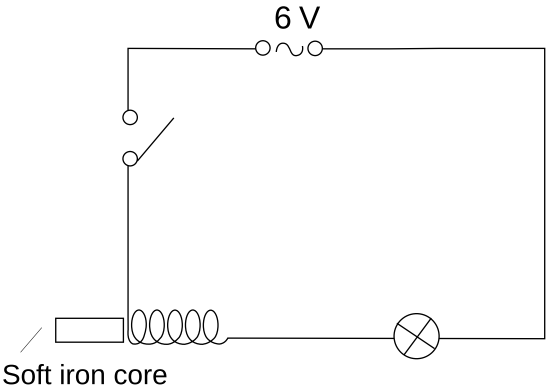


Fig. 2.18 *Effect of iron core on induced emf*

Note

This experiment demonstrates the principle of one type of dimmer switch, e.g. as used in stage lighting.

Inductance and Frequency

Apparatus

Signal generator, coil (approx. 400 turns) with soft iron core, bulb (e.g. 6 V, 0.3 A).

Procedure

Set up the circuit as shown in Fig. 2.19. Vary the frequency of the signal generator, keeping the applied voltage constant. Note that the bulb gets dimmer as the frequency increases.

Explanation

At higher frequencies, the self-inductance in the coil increases because the magnetic field is changing more rapidly.

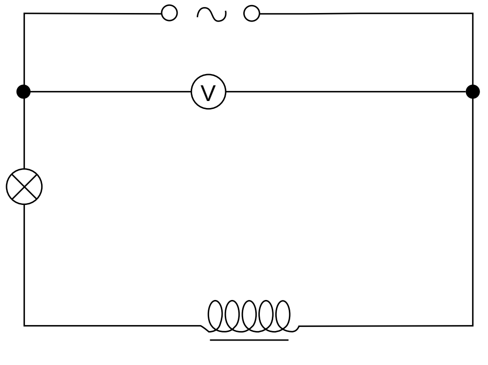


Fig. 2.19 *Self-inductance increases with frequency*

Experiment 2.5

Back Emf in Motors

Connect a small motor to a d.c. source and measure the current when the motor is turning. Now slow down the rotation of the motor by pressing against its shaft and note that the current increases, and reaches a maximum when the motor is stopped. (Allow current to flow through the motor for only a very short time when it is stationary so as not to burn out the coil.)

The back emf in a motor can be clearly demonstrated as follows. First connect the terminals of the motor to a centre-zero voltmeter and then turn the motor mechanically, e.g. by hand, so that it acts as a generator. Note the direction of rotation and the direction of the induced emf as indicated by the voltmeter. Now connect the motor, in series with an ammeter, to a d.c. power supply so that it turns in the same direction as before. Note the direction of the current. It is found to be in the opposite direction to the current which flowed when the motor was driven as a generator. But, since the coil in the motor is rotating, it acts as a generator and an emf continues to be generated. This generated emf is called a 'back' emf because it is in the opposite direction to that which produces rotation. (This is in accordance with Lenz's law.) The back emf is proportional to the speed of rotation (in keeping with Faraday's law).

Experiment 2.6

To Construct a Model Induction Motor

Apparatus

Two coils (approx. 600 turns) with soft iron cores, 5 μF capacitor (non-electrolytic), aluminium or copper calorimeter (or aluminium can) free to rotate on vertical support.

Procedure

Place the two coils at right angles to each other as shown in Fig. 2.20 and connect up the circuit as shown in the diagram. Balance the calorimeter or can carefully on a sharp-pointed support so that it can rotate easily. When the current is switched on the can will rotate steadily.

Explanation

The capacitor causes the currents in the coils, and therefore the resulting magnetic fields, to be out of phase with each other. Thus, the magnetic field due to the current in one coil reaches a maximum before the other, and so on. The direction of the resultant magnetic field of the two coils therefore rotates through 360° as the a.c. goes through one cycle. The rotating magnetic field induces currents in the can which, by Lenz's law, cause the can to follow the magnetic field.

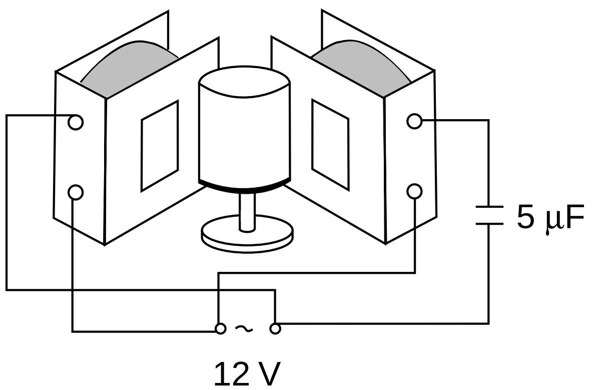


Fig. 2.20 Demonstrating the principle of the induction motor

Experiment 2.7

To Demonstrate the Principle of the Moving Iron Meter

Apparatus

Coil (600 turns, no iron core), 2 iron wires (or wire nails), 6 V power supply.

Procedure

Connect the coil to the power supply and place the iron wires inside the coil. When the power supply is switched on the wires roll apart. This is because they become magnetised and their north and south poles are side by side, so they repel. This experiment also works if an a.c. source is used, because even though the magnetic poles are constantly changing from end to end in each wire, like poles are always side by side.

Some ammeters are based on this principle. Two plates of iron are placed inside a coil, one fixed and the other connected to a pointer and scale. Such an ammeter can be used to measure a.c. as well as d.c. current, unlike a moving-coil instrument which can only be used for d.c. (A moving-coil meter can be used to measure a.c. indirectly by first converting the a.c. to d.c. using a rectifier. This method is normally used in analogue multimeters.)

Experiment 2.8

To Demonstrate the Principle of the Induction Coil

Apparatus

Two coils (approximately 30 turns and 12 000 turns), iron core, 1 μF capacitor, 1.5 V battery, switch, spark plug.

Procedure

Set up the circuit shown in Fig. 2.21. When switch K is opened, the plug sparks. The circuit in Fig. 2.22 is similar to that used in the ignition system of older petrol-engined cars.

Caution

Do not touch the plug when the circuit is being broken. Keep the spark gap at 2 mm or less. Otherwise, the large voltages induced could damage the secondary coil.

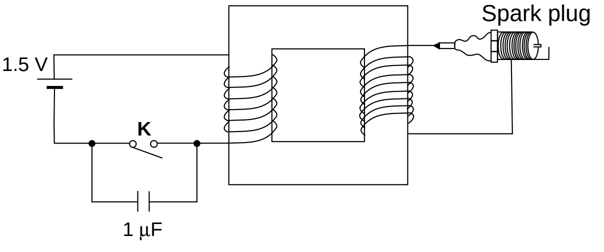


Fig. 2.22 Demonstrating the principle of the induction coil as formerly used in a petrol-engined car

2.6 Applications

Transformers

Commercial transformers are usually designed as shown in Fig. 2.22. The secondary coil is wound over the top of the primary, often with an earthed screen between them. Both coils are wound on the central limb of a double magnetic loop. This arrangement gives maximum flux linkage between the coils and therefore the most efficient transfer of energy from one coil to another.

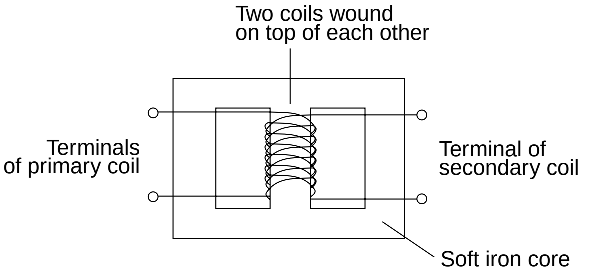


Fig. 2.22 *Simple commercial transformer*

Tapped secondary coils

Tapped secondary coils have a number of connections corresponding to different numbers of turns, Fig. 2.23. A variable low voltage (LT) power supply, such as those used in school laboratories, uses a tapped secondary coil.

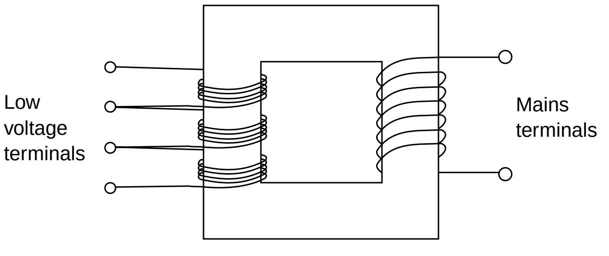


Fig. 2.23 *Transformer with tapped secondary coil*

Autotransformers

An autotransformer uses only one coil. The input voltage is fed across the full coil and a lower output voltage is taken between one end and an intermediate point. Alternatively, if the input is applied at an intermediate point and the output taken across the full coil an increased voltage is obtained.

The Speedometer

One type of speedometer uses a simple induction motor, (see p. 26). A cable driven from the gearbox of the vehicle (or the wheel of a bicycle) causes a permanent magnet to rotate close to an aluminium disc. The rotating magnet induces eddy currents in the disc which cause the disc to follow the rotation of the magnet. The induced current, and hence the torque on the disc, is proportional to the speed of rotation of the magnet and therefore to the speed of the car. The disc is connected to a spring which produces a torque to oppose that on the disc. The disc therefore comes to rest when the two torques acting on it are equal. According to Hooke's law the torque produced by the spring is proportional to the angle through which it has turned. Therefore the angle through which the disc moves before coming to rest is proportional to the speed of the car. A pointer is attached to the disc and rotates on a scale calibrated in kilometres per hour or miles per hour. Most modern cars use an electronic speedometer.

The Heart Pacemaker

One type of heart pacemaker works by having a secondary coil implanted inside the chest wall, with electrodes leading from it to the heart muscles. The primary coil is strapped to the chest just above the coil. Pulses of d.c. current are passed through the primary coil and these cause pulses of emf to be induced in the secondary coil, which in turn stimulate the heart muscles.

Heavy-duty Dimmer Switch

In a heavy-duty dimmer switch an inductor with a moveable iron core is used in series with the lamp to control the current and hence to control the light output of the lamp. In an a.c. circuit a back emf is induced in an inductor and this reduces the current flowing in the circuit. The larger the inductance of the inductor the larger the induced emf and the smaller the current. The inductance in turn depends on the position of the iron core; the further the core is pushed into the coil of the inductor the larger the inductance. The advantage of using an inductor rather than a resistor is that an inductor has a relatively small resistance and so little energy is dissipated as heat. Rather, the energy is stored temporarily in the magnetic field of the inductor before being converted back to electrical energy. The principle of the dimmer switch is demonstrated in Experiment 2.4, p. 32.

Note

Dimmer switches used in home wiring use a different principle. They are essentially electronic switches which switch the current off and on twice in each cycle of the a.c. mains. The larger the proportion of the time for which the current is switched off the lower the average light output. This type of dimmer switch is not suitable for use with fluorescent lamps because the average voltage across the tube is insufficient to maintain the discharge.

The Anode in an X-ray Tube

In an X-ray tube most of the kinetic energy of the electrons is converted to heat when they strike the anode. This causes a rapid rise in temperature at the point where the electrons strike the anode; it also results in pitting of the anode at this point. To minimise these effects the anode in most modern diagnostic X-ray tubes is shaped as shown in Fig. 2.24 and rotates very rapidly at about 50 revolutions per second. This helps to spread the heat and to prolong the life of the anode. When the machine is finished in use, the rotating anode is slowed down rapidly. This is necessary because if it were allowed to slow down gradually, various components of the X-ray machine would resonate when the frequency of rotation of the spinning anode was the same as the natural frequency of the component concerned. The braking of the anode is achieved by electromagnetic induction. To slow the anode, electromagnets (not shown in the diagram) near the anode are switched on. This results in large eddy currents being induced in the anode. In accordance with Lenz's law, the direction of these currents is such that the forces produced in the magnetic field oppose the motion of the anode.

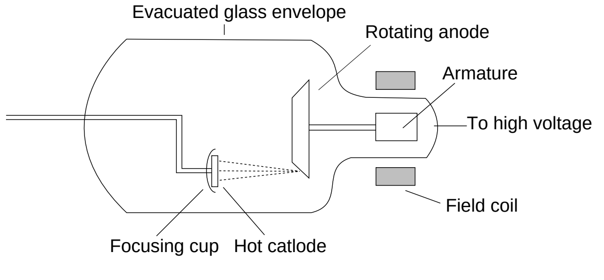


Fig. 2.24 *X-ray tube with rotating anode*

Emergency Brakes

Emergency braking in some kinds of vehicles is based on the principle of electromagnetic damping. A thick metal disc mounted on a wheel axle rotates between the poles of a powerful electromagnet which is not turned on. In an emergency, a large direct current is passed through the electromagnet. The rotating disc now experiences a strong braking effect, thus slowing down the vehicle.

The Electric Arc Welder

The electric arc was discovered by the British chemist, Sir Humphry Davy in 1800. An arc is a continuous electric discharge, between electrodes, in a gas at low pressure or in open air.

To start an arc, the two electrodes (usually carbon) are brought into contact, and a large current passes through them. This causes the electrodes to become very hot and, when they are separated, the air between them is ionised and current continues to flow between them. This current is mainly carried by electrons flowing from the negative to the positive electrode, but also in part by positive ions flowing in the opposite direction. Because of this, about twice as much heat is produced at the positive as at the negative electrode.

Because of the great heat produced, arcs are also used to produce very high temperatures in furnaces and for welding.

The arc welder consists of a transformer to step down the mains voltage to between 10 V and 50 V, and a tapped choke to control the output current, Fig. 2.25.

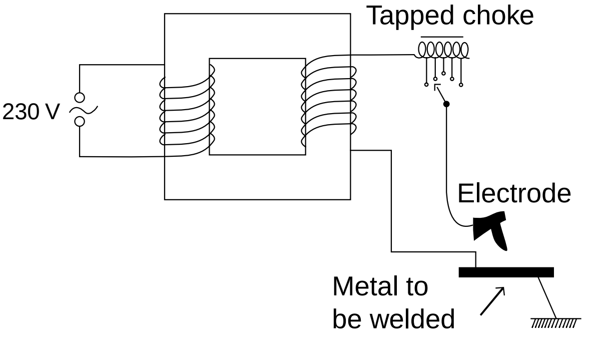


Fig. 2.25 *Arc welder*

When the arc is struck (i.e. when the welding rod, or electrode, touches the work), a short circuit causes a large current to flow and the tip of the electrode begins to melt.

The arc is produced by a large electric current (up to 500 A or more) jumping across the gap between the electrode and the joint to be welded. The heat causes the edges of the metal to melt and, at the same time, additional molten metal is transferred across the arc from the welding rod (electrode). This molten mass of metal cools and solidifies into one piece (cf. p. 31).

Note

Arc welding can be carried out using either a.c. or d.c. When using a.c., an equal amount of heat is produced at both electrodes. However, when using d.c., roughly twice as much heat is produced at the positive electrode as at the negative. This is because with d.c. the electrons which carry most of the current collide with the positive electrode, whereas with a.c. an equal number of electrons impinge on each electrode.

2.7 Worked Examples

1. A loop of wire is moving to the right in a magnetic field whose field lines are perpendicular to the page in the region shown in Fig. 2.26. The dimensions of the loop are 50 cm x 60 cm. It is moving at a constant speed of 4.0 m s^{-1} and it has a resistance of 20 Ω. The strength of the magnetic field is 0.80 T. Calculate the current flowing in the loop when it is in positions (a), (b) and (c). Calculate also the work required to move the coil from (a) to (c).

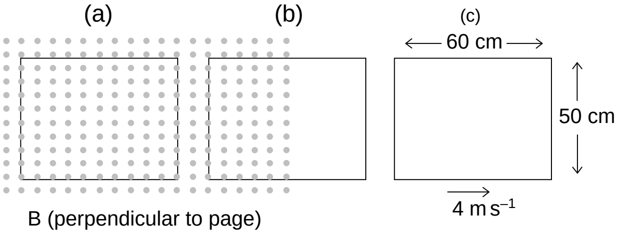


Fig. 2.26

At (a) there is no current in the loop because, even though the loop is moving, the total magnetic flux through it is not changing.

At (c) there is no change in the magnetic flux through the loop, and hence, again, no current is induced.

At (b), however, the magnetic flux through the loop is changing, so an emf is induced, and hence an electric current flows.

Induced emf, E, is given by

$$E = \frac{d\Phi}{dt}$$

$$= \frac{d(BA)}{dt}$$

where A is the area of the loop and t is the time taken for the loop to cross the boundary of the magnetic field.

$$t = \frac{0.6}{4}$$

$$= 0.15$$

$$E = \frac{0.8 \times (0.5 \times 0.6)}{0.15}$$

$$= 1.6 \text{ V}$$

$$I = \frac{E}{R}$$

$$= \frac{1.6}{20}$$

$$= 0.08 \text{ A}$$

The work done in moving the loop at constant speed from (a) to (c) is equal to the internal energy generated in the coil, i.e.

$$W = I^2Rt$$

$$= 0.08^2 \times 20 \times 0.15$$

$$= 0.019 \text{ J}$$

2. A transformer has a primary coil of 400 turns and a secondary coil of 10 000 turns. Calculate the output voltage from the secondary coil when the primary coil is connected to a 12 V supply.

$$\frac{V_{out}}{V_{in}} = \frac{N_{secondary}}{N_{primary}}$$

$$\frac{V_{out}}{12} = \frac{10000}{400}$$

$$V_{out} = 300 \text{ V}$$

3. What is the peak voltage of the mains supply (rms voltage is 230 V)?

$$V_{peak} = V_{rms} \times \sqrt{2}$$

$$= 230 \times \sqrt{2}$$

$$= 325 \text{ V}$$

MODULE 7

Mechanics II

CONTENTS

CHAPTER 1

GRAVITATION AND CIRCULAR MOTION

1.1 Background

Getting To Know Our Neighbours

For thousands of years the motions of the heavenly bodies, the sun, moon, stars, planets and comets, have been carefully observed and recorded. No other field of science has had such a long accumulation of data as astronomy. The most basic celestial cycle as seen by the ancients was, of course, day and night, caused by the sun travelling in an arc from east to west across the sky. The annual variations in the sun's altitude were also well known, Fig. 1.1. The stars were observed also to travel in an east to west direction, new ones rising in the east while others set in the west. In the Northern Hemisphere the stars appear to rotate counter-clockwise around Polaris, the North Star, but the star patterns appear to be unvarying; they are called 'fixed stars'.

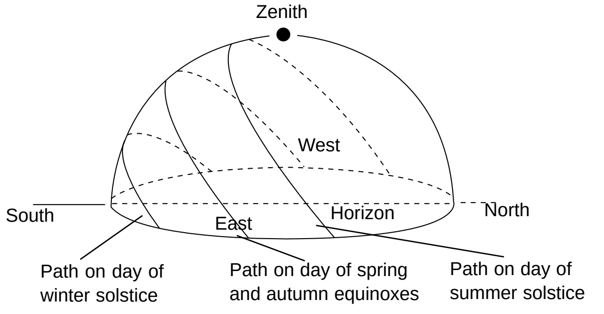


Fig. 1.1 *Apparent path of sun in Ireland at solstices and equinoxes*

The moon also was observed to travel in an east to west direction. A full moon occurs when the moon is opposite the sun, and rises in the east at sunset. Each night after that, it rises nearly an hour later, and appears less round. After about fourteen days, when the moon is passing near the line between the earth and the sun and rises at the same time as the sun, we cannot see the moon at all.

In addition, a number of 'wandering stars' were known. These were the planets Mercury, Venus, Mars, Jupiter and Saturn, all of which can be seen with the naked eye. (The word planet comes from the Greek word for 'wanderer'.) They were called wanderers because they were seen to move against the background of the 'fixed' stars. They normally rise in the east and set in the west, but at certain times their motion is in the opposite direction. (This is called 'retrograde motion', and occurs at times when the planets are at maximum brightness.)

These astronomical motions dominated the patterns of life for people long ago. Human activity was regulated by the presence or absence of daylight, and by the warmth which changed with the seasons. Once people began to settle in villages, and became dependent on agriculture (about 10 000 years ago), they needed a calendar to enable them to plan their ploughing, sowing and harvesting. This became a matter of survival, because seeds planted too early or too late will not produce good yields.

These observations on the motions of the stars and planets led people to question what the earth was, and its place in the universe. To the Greeks, it appeared that the earth was large, solid and permanent (and flat), while the heavens contained many small, distant objects and seemed to move around the earth. It seemed logical to deduce that the earth was at the centre of the universe. The early Greek philosophers from the time of **Plato** (427-347 BC) were among the first people to seek explanations for the universe on the basis of the observations that had been made up to that time. They developed a model for the universe in which the moon, the sun, the planets and the stars were contained in large, spherical shells which rotated about the earth at different and varying speeds. This geocentric (earth-centred) view of the universe was further developed by **Aristotle** (384-322 BC) who proposed a total of 55 such shells. Yet, while this model provided a general explanation for the motion of the celestial bodies, Aristotle observed that it did not predict their locations very accurately.

The fact that the sun, moon and stars apparently orbited the earth in circular paths led the Greeks to regard the circle as the most perfect geometrical figure, and circular motion as natural motion. The sun and moon were clearly spherical themselves, giving added support to this approach. With our space-age perspective, we may regard these ideas as hopelessly naïve, but a moment's reflection will indicate that many of our daily needs can be met by this simple model of the universe - a flat earth around which everything else revolves.

Not all Greek philosophers agreed with the geocentric model of the universe. One such dissenter was the astronomer, **Aristarchus of Samos** (310-230 BC). Aristarchus suggested that if the sun were regarded as being at the centre of the universe, with the earth, planets and stars revolving around it, a much simpler explanation for the observed universe was possible. With this theory, he explained very simply the observed retrograde motion of the planets, and explained the variations in their brightness. Many of the ancient Greeks favoured the philosophic concept of a moving earth, but Aristotle's opinion carried great weight.

During a period of about 400 years after Aristotle, many Greek astronomers made important contributions to further developing his model. By far the most influential of these was **Ptolemy** (c. 100-170 AD). Ptolemy was based in Alexandria where, for many centuries, a great library, perhaps the ancient equivalent of a modern research institute, flourished. He used a variety of new mathematical devices to explain the observed motions of the heavenly bodies, and his model was able to make predictions regarding their positions with considerable accuracy. It also successfully explained the variations in the brightness of the planets. In 150 AD, he published an extensive treatise in thirteen volumes, *The Great Syntax*, on his calculations of the motions of the sun, moon, and planets. He was also the author of the *Geography*, the leading map of the known world for many centuries.

To the modern mind, Ptolemy's system seems quite far-fetched. However, in his time and for 1400 years afterwards, it was the only system which was capable of producing astronomical tables of the required accuracy for calendar-making and navigation. His goal in his work was to develop a mathematical technique which would permit accurate calculations. He had no interest in philosophical considerations, and physics was considered a branch of philosophy. In short, his scheme was widely used for the very pragmatic and convincing reason that it worked!

With the decline of secular learning accompanying the collapse of the Roman Empire, the work of Ptolemy and others was temporarily lost to the Western world. In 640 AD, Alexandria was captured by the Muslims and the library was eventually burned down by order of Caliph Omar. Fortunately, some of the scientific and mathematical works there had been copied and were available to philosophers elsewhere in the East. Eventually Ptolemy's work was reintroduced into Europe. These works greatly influenced European thought from the 12th century. The Dominican monk, **Thomas Aquinas** (1225-1274) blended major elements of Greek thought and Christian theology into a single philosophy. His work was widely accepted in Europe for several centuries, a period when there was very little questioning of science because such questioning was also seen as questioning theology.

Copernicus and the Heliocentric System

The Renaissance was a time of extraordinary change and discovery. The absolute authority of the Roman Catholic church was being questioned by Martin Luther in Wittenberg and by John Calvin in Geneva; Henry VIII was arguing with the Pope, although Henry was, perhaps, motivated by considerations other than cosmology. A great flowering in art and architecture was taking place in Italy and throughout Europe. Ptolemy's ancient *Geography*, which had been based in part on the reports of sailors and merchants, was grossly inaccurate beyond the boundaries of the Roman Empire. The new geographical knowledge of the fifteenth and sixteenth centuries, gleaned in part from the voyages of Columbus and others, and the discovery of the New World, undermined confidence in Ptolemy as a reliable map-maker, and thus cast doubt on his astronomical system as well. It was in these times that the Polish astronomer, **Nicholas Copernicus** (1473-1543), lived and worked.

In his great masterwork, *On the Revolutions of the Heavenly Spheres (De Revolutionibus Orbium Coelestium)* (1543), Copernicus proposed a model for the universe that places the sun at its centre. His theory was clearer and simpler than the Ptolemaic model but, apart from this, had no clear scientific advantages over the geocentric theory. There was no known observation that was explained by one system and not by the other.

Copernicus had no proof that his heliocentric (sun-centred) system was correct. To him it was both clearer and more pleasing than the geocentric system. Others felt differently. Common sense seemed to suggest that the earth was flat and that everything revolves around it, not an easy point of view to abandon. Nor, in that period of history, were people accustomed to the idea of using experiments or observations of nature to distinguish between different philosophical systems. However, the main difficulty arose because Copernicus was unable to persuade most of his readers that his heliocentric system reflected the mind of God as closely as did the geocentric system. The removal of man from the centre of the universe was difficult for scientists as well as for theologians. It was widely believed that the universe was made for man's benefit and that human beings were special.

Tycho Brahe, the Man With the Metal Nose

Tycho Brahe (1546-1601) was a Danish astronomer who studied the motions of the heavenly bodies from an early age. He concluded that both Ptolemy and Copernicus had relied on tables of planetary positions that were inaccurate. He devoted his life's work to making more detailed and accurate observations and measurements of astronomical phenomena. He received generous sponsorship from the King of Denmark who gave him his own island where he built his 'Castle of the Heavens' containing four large observatories, a library and a laboratory where many scientists, technicians and students came to study astronomy. Though the telescope had not yet been invented, Tycho and his students designed large-scale instruments and improved the precision of astronomical measurements by a factor of 10 (the precision of such instruments had been frozen at 10 minutes of arc for 15 centuries). This enabled him to make far more detailed observations than had been made previously.

Tycho was a colourful character, rather vain, who wore a metal nose as a result of a duelling injury. He surveyed the heavens, measuring the positions of the stars and planets with a precision far greater than anyone before him, and in his lifetime gathered a vast collection of precise and accurate data, from which Kepler in particular extracted the hidden science.

Kepler and his Laws of Planetary Motion

In the last few years of his life, Tycho moved to Prague, where he hired a young German assistant called **Johannes Kepler** (1571-1630). Kepler was given the task of determining in precise detail the orbit of Mars. This was a particularly difficult problem which had not been solved by Tycho or his other assistants. For a year and a half Kepler analysed Tycho's observations of Mars, struggling to fit them to the system of orbits described by Copernicus, but without success. The problem lay in the circular orbits that the Copernican system still used. Eventually Kepler was forced to conclude that the orbit was not a circle. So he tried a variety of other geometrical shapes, until he finally found the ellipse. This worked, not just for Mars, but for all of the planets. He stated this as his law of elliptical orbits: *The planets move in elliptical orbits about the sun, with the sun at one focus of the ellipse.*

Another problem that had vexed astronomers was the fact that the planets moved faster in their orbits at some times than at others. This had never been satisfactorily explained until Kepler observed that, with the planets moving in elliptical orbits, the fastest motion occurs when they are nearest to the sun. This led to his law of areas: *Imaginary lines from the sun to the planets sweep out equal areas in equal times as the planets move about the sun* (Fig. 1.2).

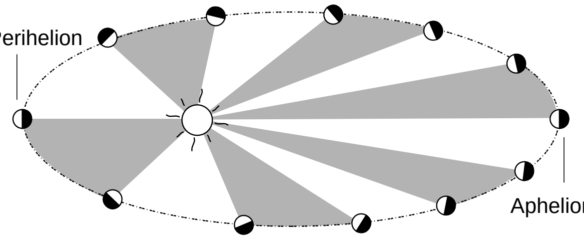


Fig. 1.2 *Path of planet around sun illustrating Kepler's law of areas*

Kepler published these two laws in his book *Astronomia Nova* in 1609. He now had an accurate description of how the planets moved, but he was certain there must be a harmony in those motions that would explain why the planets moved the way they did. Ten years after publishing his first two laws, he found the relationship that applied to the motion of all planets. This is Kepler's third law, the law of periods (also called the harmonic law): *the cube of the average distance from the sun divided by the square of the time required to traverse the orbit is a constant, the same for every planet.*

By means of Kepler's three laws, the elliptic law, the law of areas, and the harmonic law, a planet's motion can be predicted far into the future. Only three corrections are applied today: the masses of the planets affect the predictions slightly; one planet will disturb the motion of another; and a slight correction must be made in the case of Mercury's orbit because of an effect predicted by Einstein's theory of relativity.

Kepler's work is very significant in the history of science. Apart from the overthrow of the idea of circular motion as being the natural motion of heavenly bodies, the statement of his empirical laws in the form of mathematical equations helped to establish the use of the equation as the normal form of stating laws in physical science. (How many students have since regretted his insight!)

It is interesting to note that up to this time, the various theories of planetary motion were developed in order to describe the motion and to make accurate predictions of the positions of the planets. The causes of planetary motion had not been looked for in any detail. Heavenly bodies were assumed to be completely separate in composition from earthly bodies. Their motions were simply taken for granted and the idea that they might obey the same physical laws as earthly bodies had not been considered. Kepler was one of the first to seek to explain, rather than just describe their motions, though he met with little success in this regard. It was not until Newton enunciated his law of universal gravitation that a single theory explained the motion of heavenly bodies as well as earthly phenomena.

Galileo and the Foundation of Modern Science

Galileo Galilei (1564-1642) became aware of the telescope shortly after its invention in the Netherlands. In 1610 he made some telescopes of his own and became the first scientist to use this instrument to study the night sky. With his telescope, Galileo made a number of amazing observations. He discovered that the 'cloud' of the Milky Way was composed of many individual stars. He was the first to observe craters on the moon and was able to estimate the height of its mountains, and he discovered sunspots and four of the moons of Jupiter. Each of these discoveries dealt a major blow to the geocentric model of the universe. For example, the fact that the moon had craters and the sun had sunspots meant that heavenly bodies were not all perfect celestial spheres. The satellites of Jupiter contradicted the geocentric belief that all bodies circled the earth. Unlike most scientists of his day, Galileo was a master at publicising his own work. He wrote in Italian rather than Latin, hence appealing to a wider audience. His great works, *Dialogue Concerning the Two Chief World Systems* (1632) and his later book, *Discourses Concerning Two New Sciences*, were written in the form of a dialogue with three characters, one arguing for the Copernican view, one for the Ptolemaic view and a detached observer who, naturally enough, was usually swayed by the arguments of the former! His writings, however, brought him into conflict with the Church authorities of the day when he insisted that the Copernican theory was true, rather than just one of several possible methods to compute the planetary motions, for this was held to be contrary to the teachings of the Bible. His books were placed on the *Index of Forbidden Books* where they remained, along with those of Copernicus and Kepler, until 1835. Eventually, though he was an old man, Galileo was arrested and imprisoned for a few months. At his trial, he was threatened with torture, forced to confess to holding and teaching forbidden ideas and to denounce the Copernican theory. In return, he was sentenced to only perpetual house arrest.

As well as his work in astronomy, Galileo observed that the period of a pendulum was independent of the amplitude of the swing, and deduced the formula for the periodic time. He did this by using his pulse to measure the time of swing of a chandelier in Pisa Cathedral. (It was due to him that pendulums came to be used as clocks. Indeed, the best clocks in the seventeenth century were pendulum clocks.) He showed that the speed of a falling body was independent of its weight, disproving Aristotle's view, and he proved that an object follows a parabolic path when thrown into the air at an angle other than 90^{o} to the horizontal. As well as his great discoveries, Galileo's legacy resulted from his methods - unwilling just to make observations and then devise theories to explain them, he went further and devised experiments to test and refine his explanations. In this he was a pioneer of the *experimental method.* (See also Galileo and the Pendulum, p. 29.)

Newton and the Law of Gravitation

Isaac Newton was born on Christmas Day in 1642, the same year that Galileo died. He began his studies at Cambridge University in 1661 where he developed a new branch of mathematics called calculus. (Calculus was also developed independently by **Gottfried Leibnitz** in Germany.) Many of his greatest discoveries were made at home during the years 1665-7 when the Black Plague caused the university to be closed, but they were not published until later. Recognised as one of the greatest scientists of all time, Newton was one of the first scientists to use quantitative or mathematical arguments to test scientific ideas.

Newton's greatest achievement, the enunciation of his law of universal gravitation, can be seen as the synthesis of much of the work of scientists over the previous century. 'If I saw a little farther than others', he wrote in a letter, 'it was because I was standing on the shoulders of giants'. (Some scholars believe this remark was a snide reference to **Robert Hooke** who was small of stature and with whom Newton had many public arguments.) Kepler had described the motions of the planets, and Galileo had demonstrated that falling objects are accelerated towards the earth. Newton's great achievement was to show that both of these phenomena could be explained by the same principle, namely that: *All objects attract each other with a force which, for point masses, is proportional to the product of the masses and inversely proportional to the square of the distance between them.*

Aristotle didn't need to ask why objects, such as apples, fall. To him, downward motion was simply the natural motion of the apple. However, for Newton, straight-line motion at constant speed was natural motion; any other motion was unnatural. Thus, a falling apple, which is accelerated, must be caused or forced to fall. This idea made it necessary to invent the concept of *gravity*, a force that reaches up from earth and pulls objects downward. Newton's greatest achievement was to develop a quantitative theory of gravity which extended from earth to heavenly bodies which had been supposed to be essentially different from our own planet, and to show that the same force of gravity could explain the falling apple and the motion of the planets. Thus Newton unified the heavens with earth and altered forever our perception of humanity's place in the universe.

Newton combined the skills of the theoretical and the experimental scientist. As well as his work in mechanics, he also made ingenious pieces of equipment - he made the first reflecting telescope - and he showed that white light was made up of several different colours.

Limitations of the Newtonian Model

Newton's 'clockwork' model of the universe proved remarkably successful in providing explanations for the observed universe. However, we now recognise that in relation to the motion of the galaxies, there are discrepancies between predictions and observations. Although the forces within galaxies appear to be consistent with Newton's model, this may not be the case for forces between the galaxies. Even within the solar system, observations of the orbit of Mercury differed slightly from what was predicted by Newton's theory of gravitation. At the other end of the scale, Newtonian mechanics fails completely to explain the observed behaviour of atoms and sub-atomic particles.

Attempts to modify Newtonian mechanics in order to explain these discrepancies were unsuccessful. Just like Ptolemy's theory in an earlier time, the Newtonian theory would have to be modified. Further explanations were required. At the macroscopic level, Einstein's theory of general relativity (1915) provided a new explanation for gravity which accurately predicts the orbit of Mercury, as well as explaining other anomalies in the Newtonian model. At the sub-atomic level quantum mechanics was found to provide much more accurate predictions than Newtonian mechanics.

However, we still use Newton's theory for all practical purposes because the difference between its predictions and those of general relativity and quantum mechanics is very small in the situations that we normally deal with. (Newton's theory also has the great advantage that it is much simpler to work with than either relativity or quantum mechanics!)

Some Landmarks in Space Exploration

That an artificial satellite could be placed in orbit around the earth was envisaged long before it became possible to do so. The development of rockets for military purposes in the 1940s led to the construction of more and more powerful rocket engines, and these made it possible for the Russians to place earth's first artificial satellite, called Sputnik 1, in orbit on 4 October 1957, followed by a much larger one, Sputnik 2, on 3 November 1957. (The name Sputnik, in Russian, means companion or fellow-traveller.) Sputnik 1 had a periodic time of 96 minutes, while Sputnik 2 took 104 minutes to orbit the earth. It carried a dog to study the effects of orbital flight; the dog died after about 7 days. American space experts were astonished at the ability of the Russians to place heavy satellites in orbit - Sputnik 2, together with its final stage, had a mass of about 500 kg. The first American satellite was launched in January 1958 - it had a mass of only 15 kg. The Russians and the Americans launched many satellites in the following years. Some notable ones included Tiros 1, the first satellite to observe the earth and the weather; Sputnik 5, which was successfully brought back to earth along with its 'crew' of two dogs and six mice; and Telstar, the first communications satellite, which in July 1962 relayed TV signals across the Atlantic. In July 1963 Syncom 2 was successfully launched, the first of many communications satellites in geostationary orbit. In the 1970s the Russians placed the Salyut space station in orbit, and the US followed with Skylab. In the 1980s the Russians replaced Salyut with the Mir space station. The US space shuttle Columbia made its maiden flight in April 1981, while the European Ariane rocket was first launched in June 1981. Satellite television was seen in Europe for the first time in May 1987, when Sky started broadcasting from the Astra satellite. The most noteworthy of the many scientific satellites in orbit is the Hubble space telescope, launched in 1991. Other satellites are used for earth sensing (taking detailed pictures of the earth's surface), e.g. Landsat, or for determining the position of a receiver on the earth, e.g. GSP (Global Satellite Positioning) satellites. Many other satellites are used for military intelligence.

Manned space flight began with Vostok 1, launched by the Russians in April 1961, which carried Yuri Gagarin, a Russian army officer, into space. He completed slightly more than one orbit and landed safely. In August of the same year, Major Gherman Titov completed over 17 orbits in Vostok 2. The first American to orbit the earth was John Glenn (later a US Senator) in February 1962.

The first woman in space, Valentina Tereshkova, flew in Vostok 6, launched in June 1962. The Russian Alexei Leonov made the first spacewalk in March 1965. Many astronauts have spent long periods in orbit, up to several months, especially in the Salyut and Mir space stations.

The first spacecraft to leave the earth's gravitational field was the Russian Luna 1, launched in January 1959, which went close to the moon and continued into orbit around the sun. Luna 2, launched in September 1959, crashed on the moon, while Luna 3, launched the following month, sent back pictures showing the far side of the moon. The first successful soft landing on the moon, allowing instruments to send back data from the surface, was achieved by Luna 9 in February 1966, while Luna 10, in March 1966, was the first to go into orbit around the moon. The next major step was to send a space capsule around the moon and back to earth; this was achieved by the Russian Zond 5 in September 1968. The US Apollo program succeeded in sending a manned spacecraft around the moon and back (Apollo 8) in December 1968, followed by the first manned moon landing (Apollo 11) in July 1969, when Neil Armstrong and Edwin ('Buzz') Aldrin became the first men to walk on the moon. Altogether six Apollo craft reached the moon between July 1969 and December 1972, carrying a total of 12 men to its surface.

Beyond the moon, spacecraft were sent to the other planets. The US spacecraft Mariner 2, launched in August 1962, passed close to Venus and sent back measurements of the planet's surface temperature, etc. The Russian Venera 3 crashed into Venus in March 1966, while Venera 4 landed by parachute on the planet in October 1967. Venera 9 sent back the first pictures from the surface of Venus in October 1975. Mariner 10 passed by Venus on its way to Mercury, near which it passed in March 1974, sending back the first close-ups of this planet. In the opposite direction, Mariner 4 made the first successful fly-by of Mars in June 1965, and Mariner 9 went into orbit around the planet in November 1971, while the US Viking 1 sent back the first data from the Martian surface after soft-landing in July 1976. Beyond Mars, the US Pioneer 10, launched in March 1972, flew past Jupiter, while Pioneer 11, launched in April 1973, used Jupiter's gravity to send it to a fly-by of Saturn in September 1979. Pioneer 11 lost contact with earth in September 1995, while Pioneer 10 remained in contact, at twice the distance of Pluto, until April 1997, when it was decided that the data it was sending was not worth the expense of keeping the control centre running. Voyager 1, launched in October 1977, flew by Jupiter and Saturn, sending back the first pictures of these planets, while Voyager 2 passed close not only to these two planets but to Uranus and Neptune as well. It was launched in September 1977 and passed Neptune in August 1989, the signals coming back to earth taking over 4 hours to reach us. In the next few years, the two Voyagers will send back signals as they pass the edge of the solar system, where the effects of the sun finally end, and head out into interstellar space. They will be able to tell astronomers where the sun's influence effectively ends, and will also sample the interstellar material and send back details of this. Their power supply systems will probably last until about the year 2020.

1.2 Do You Know

How does gravity affect the atmospheres of different planets?

Knowing the masses and the radii of the planets, the acceleration due to gravity on each can be calculated (see table, p. 54). The different values of g are partly responsible for the different atmospheres on the planets, the other important factor being their surface temperature which is determined largely by their distance from the sun. When the value of g is large, it is difficult for gases to escape from the surface of the planet, but when it is smaller, gases can escape much more easily, particularly those whose molecules have lower mass, since they have higher velocities. The temperature at the surface of the planet, largely dependent on the distance from the sun, also affects the velocity of the molecules and hence their ease of escape. The result is that Mercury, with a high surface temperature and low gravity, has no atmosphere, as molecules of all gases on Mercury can easily reach escape velocity and leave the planet. Similarly the moon, with its very low value of g, has almost no atmosphere. Venus and Earth, which have reasonably large values of g, have retained the heavier gases, carbon dioxide in the case of Venus and nitrogen and oxygen in the case of Earth, but lighter gases like hydrogen have long ago escaped into space, if they ever existed in the atmospheres of these planets. Jupiter, being much colder and with a high value of g, has retained almost all its gases, so it has a very dense atmosphere of lighter gases, mainly hydrogen, methane and ammonia. The planets beyond Jupiter have similar atmospheres because of their very low surface temperatures and reasonably large values of g.

Why do the stars twinkle?

When you look at a star in the night sky, you are looking through the earth's entire atmosphere. This is made up of many layers of air, all of which are moving and are at various temperatures and therefore have various refractive indices. The effect of this is to make the stars appear to twinkle, when actually they are shining with a steady light. (This effect is similar to the shimmering that is seen near road surfaces on hot days.) This, of course, makes accurate observations more difficult, and explains why many observatories are located at high altitude. In recent years, the Hubble Space Telescope which orbits the earth in a satellite and hence is free of atmospheric distortion, is capable of making more accurate observations.

Which is more useful, the sun or the moon?

'Which is more useful, the sun or the moon?' inquired Kuzma Prutkov, a delightful character created by Alexei Tolstoy. 'The moon', he answered, 'because it shines at night, when it is dark; while the sun shines only in daytime'.

A scale model of the solar system.

If we made a model of the solar system on a scale of 1 to 2 x 10^8, then the earth would be the size of an orange, and the moon would be the size of a grape orbiting the orange at a distance of 2 m. The sun would be 750 m away, and would have a diameter of 7 m.

How old is the solar system?

Our galaxy is estimated to be about 12 x 10^9 years old, while the sun and planets, are much younger. They are estimated to be only about 4.5 x 10^9 years old.

Every second, 5 x 10^{13} tons of hydrogen are converted to helium in nuclear reactions in the core of the sun. However, the mass of helium produced is slightly less than that of the hydrogen, so that the mass of the sun is decreasing by 4 million tonnes every second. This mass is converted into radiant energy. The sun has been emitting radiant energy and losing mass at this rate for several billion years. Someday, the sun will run out of nuclear fuel and collapse. But there is no need to panic yet; this will not happen for another 5 x 10^9 years.

In some of the older stars, more complex reactions are causing elements heavier than helium to form. In converting one element into another the stars, young and old, are succeeding where the medieval alchemists failed.

Where is the edge of the solar system?

The solar system extends to where the gravitational field of the sun is challenged by the fields of nearby stars. This is far beyond the orbit of Pluto, out in the region where the great cloud of comets lies.

What is the shape of planetary orbits?

When a planet orbits a star in an elliptical orbit, the centre of mass of the two bodies lies close to one of the foci of the ellipse. The same principle applies in the case of moons orbiting a planet. The sun is so massive in relation to the planets that its centre of mass virtually coincides with the focus.

All the planets move in a counter-clockwise direction around the sun (as seen from the north). Most of the moons of the various planets in the solar system revolve in the same direction, but four of Jupiter's moons and one each of Saturn and Neptune revolve in the opposite sense. Such motion is called *retrograde*, a word also used to describe the apparent backward motion of a planet seen in the sky (see p. 1).

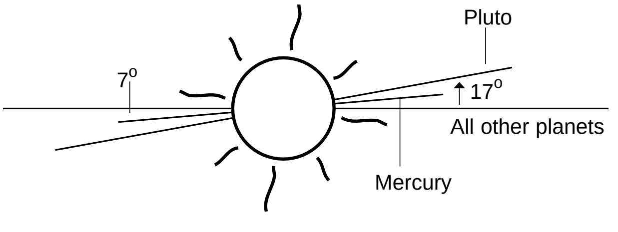


Fig. 1.3 *The orbits of all planets except Mercury and Pluto lie in almost the same plane*

The two planets that stray farthest above and below the plane defined by the earth's orbit are Mercury and Pluto. Their orbits have **inclinations** of 7° and 17°, respectively (Fig. 1.3). All of the other planets have orbital inclinations within 4° of the earth's orbit. The orbits of Mercury and Pluto also have the greatest **eccentricity**. That is, the ellipses along which these planets move are more elongated than the paths of the other planets.

What is a perturbation?

All satellites (planets, moons, comets, etc., as well as artificial satellites) are influenced not only by the gravitation of the parent body about which they orbit, but by all other bodies in the solar system as well. These other bodies cause small variations in their orbits, called perturbations. In the case of man-made satellites, small adjustments must continually be made to their orbits to take account of these perturbations.

How is gravity understood in Einstein's theory of relativity?

By the early years of this century, there were a number of observations which could not be explained on the basis of Newton's gravitational theory. Among these were the following.

- The long axis of the orbit of the planet Mercury was seen to rotate about the sun at a rate of about one degree in ten thousand years. This could not be explained according to Newtonian mechanics.

- According to Newton's theory of gravity, objects attract each other with a force that depends on the distance between them. This means that if one moves one of the objects, the force on the other one would change instantaneously. In other words, gravitational effects should travel with infinite velocity. However, Einstein's special theory of relativity (1906) stated that nothing can travel at a speed greater than the speed of light in a vacuum.

Einstein made a number of unsuccessful attempts between 1908 and 1914 to find a theory of gravity that was consistent with special relativity. Finally, in 1915, he proposed what we now call the theory of general relativity. This theory predicted previously unexplained observations such as those above. Equally importantly, however, the new theory entirely altered the basic concepts used to explain gravitation. No longer can we accept force or action at a distance as philosophically valid. In Einstein's concept any mass curves, or warps, the space-time continuum. The effects of gravitation can be then considered to be a result of this curvature of space-time produced by the presence of matter: the curvature results in measurable consequences on matter, which we call the acceleration of gravity and the force of gravity.

Among the predictions of Einstein's theory are the following.

1. Clocks run more slowly near massive bodies, i.e. clocks near the sun will run slowly compared with identical clocks farther away from it.

2. In the presence of gravitational fields Euclidean geometry no longer applies to physical objects. For example, in the arrangement of three rigid rulers to form a triangle, the sum of the subtended angles will not equal 180^o.

3. Light does not travel in straight lines, the rays being deflected by gravitational fields.

Is the universe expanding?

Another major problem not addressed by either Newtonian mechanics or general relativity was an explanation of why the galaxies did not collapse in upon each other due to the fact that the only forces acting upon them were the attractive gravitational forces. One possible explanation would be if the universe were constantly expanding, i.e. if the galaxies were constantly moving away from each other. In 1929, the American astronomer **Edwin Hubble** (1889-1953), using the 100-inch diameter telescope at the Mount Wilson Observatory in California, found conclusive evidence that the universe is indeed expanding and changing in time.

One conclusion from this is that at earlier times objects would have been closer together. In fact, it seemed that there was a time, called the 'big bang', about ten or twenty thousand million years ago, when the universe was infinitesimally small and infinitely dense. This discovery finally brought the question of the beginning of the universe into the realm of science.

What are the fundamental forces of nature?

Modern science recognises four fundamental forces in nature.

1. The **gravitational force** is the weakest of the four forces.

2. The **weak nuclear force** has a short range (of the order of magnitude of the nuclear diameter), and is not well understood because it is difficult to detect. It is the force involved in beta decay within the nucleus. It is stronger than gravity, but weaker than the remaining two forces.

3. The **electromagnetic force** is responsible for many easily observed phenomena and so is quite well understood. It is stronger than gravity and the weak nuclear force. All contact forces, (pushes, pulls, friction, etc.) come under the heading of electromagnetic forces. For example, the force your fingers exert on a pencil is the result of electrical repulsion between the outer electrons of the atoms of your fingers and those of the pencil.

4. The **strong nuclear force** holds the nucleus together.

Newton, with his concept of gravity, unified a wide range of forces that were thought of as different in his day. He saw a person's weight, the force holding earth in its orbit around the sun, the force holding the moon in its orbit, the force pulling a falling apple to earth and many more as aspects of a single force, viz. the force of gravity. In the nineteenth century the electric and magnetic forces were brought together as the electromagnetic force. In the 1960s it was proposed that the electromagnetic force and the weak nuclear force are in fact a single force - the electroweak force. Work is continuing on the search for a single theory which will unify all the fundamental forces of nature.

What is the zodiac?

The imaginary line which the sun follows in its annual journey through the constellations is called the ecliptic. The moon, in its monthly journey around the earth, deviates by only a few degrees from the ecliptic, thus appearing to travel through the same constellations. The region around the ecliptic through which the sun and moon appear to travel is called the zodiac. The sun moves through the zodiac once every 365.26 days, the moon once every 29.5 days.

What causes the phases of the moon?

A new moon occurs when the moon is between us and the sun. We cannot see the 'new moon' (position 1, Fig. 1.4) because the side facing earth receives no sunlight. It rises and sets with the sun and is only above the horizon in daytime. During the nights after the new moon, a thin 'crescent moon' is visible in the western sky at dusk, growing and moving further east each evening. The first quarter (position 3) occurs roughly a week after 'new moon' when the crescent has grown to a semicircle. At this point the moon rises in the middle hours of the day and becomes visible near nightfall, setting around midnight. The 'full moon', when the complete circle of the moon's face is lit (position 5), occurs about two weeks after 'new moon', and the moon rises near the time of sunset. The 'third quarter' moon, when the moon again appears as a semicircle, but with the opposite side lit compared with the 'first quarter' moon (position 7), occurs roughly a week after full moon. It is not so commonly seen, since it now rises near midnight and is most clearly visible before dawn.

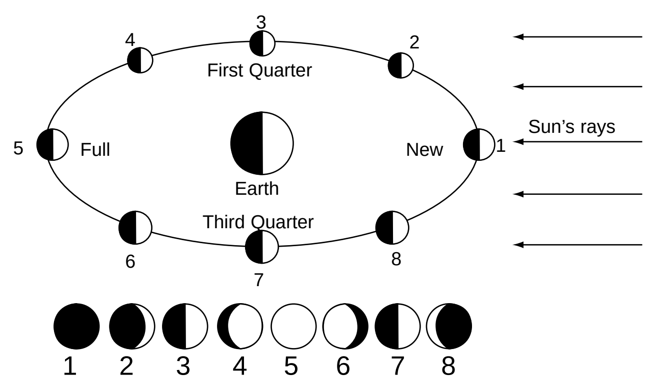


Fig. 1.4 *Phases of the moon*

The full moon occurs when the moon is on the opposite side of the earth to the sun; the full moon rises near sunset and sets near sunrise. Since the full moon is opposite the sun, it is seen high in the sky in winter and low in the summer (i.e., the opposite to the sun). This explains those bright winter nights which produce such spectacular visual effects, especially with snow and frost.

When the waning moon again appears as a semicircle, with the curved edge facing east, this is called last quarter. And finally, the night before the next new moon, we have a thin crescent in the morning sky, rising just before the sun.

What causes eclipses?

An eclipse of the sun occurs when the moon is exactly between the earth and the sun. This can only occur when a new moon is exactly lined up between the earth and the sun, Fig. 1.5(a). When this happens, the moon's shadow falls on the earth, and the moon blocks out the sun's light completely.

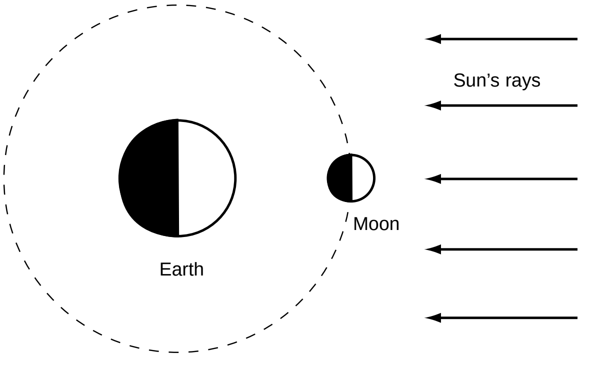


Fig. 1.5 *(a) An eclipse of the sun*

An eclipse of the moon, Fig. 1.5(b), occurs when a full moon passes through the earth's shadow. However, the earth's atmosphere refracts light into the shadow region so that, during an eclipse, the moon appears as a dull, copper disk in the night sky. An eclipse of the moon can be seen from the entire night side of the earth, whereas an eclipse of the sun is only visible along the narrow path of the moon's shadow.

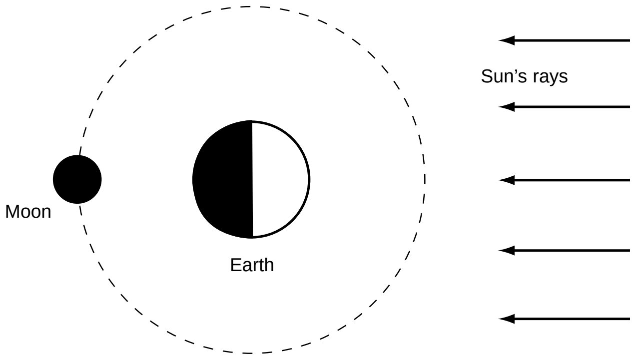


Fig. 1.5 *(b) An eclipse of the moon*

What are comets and meteors?

Comets are members of the solar system, and travel in orbits about the sun which are much more elliptical than those of the planets. Solar radiation causes particles to vaporise off the icy nucleus of the comet, and it is these particles which produce the comet's tail. This tail can be millions of kilometres long, and always points away from the sun.

Sometimes the orbits of comets are altered by close encounters with a planet (most often Jupiter, the largest planet). For example, Brook's comet (1887V) had a period of 29.2 years before its encounter with Jupiter on 20 July 1886. Afterwards, its orbit shrank in size, and its period changed to 7.1 years.

Meteors are closely connected with comets and are often the debris lost by the main body of the comet. When meteors enter the earth's atmosphere they are visible as shooting stars. This is because friction with the earth's atmosphere slows them down, converting their kinetic energy into heat.

A bright meteor is sometimes called a fireball. If it explodes it is called a bolide and if it lands on the earth's surface it is called a meteorite.

Asteroids are minor planets that orbit the sun, mainly in the region between Mars and Jupiter. They range in diameter from a few to several hundred kilometres. They may be the remains of a planet that disintegrated or failed to attain a stable existence. Some astronomers believe that the retrograde outermost satellites of Jupiter and Saturn may be captured asteroids.

Which has greater attraction for the moon, the earth or the sun?

It might seem, since the sun does not pull the moon out of its orbit around the earth, that the force of attraction between the earth and the moon is greater than that between the sun and the moon. But it is not so. Using Newton's law of gravitation, it may be shown that the sun attracts the moon with a force more than twice that between the earth and the moon. Why then does the moon not circle the sun rather than the earth?

The answer is that, of course, the moon does primarily orbit about the sun just as the earth does, and the earth's attraction causes a perturbation of that orbit.

How may artificial 'weight' be created in a space station?

'Weightlessness' in a space vehicle is highly inconvenient to an astronaut in many ways as indicated on p. 23. A suggested solution to this problem would be to build a large space station in circular form, and to rotate it so that the outer rim, acting as the floor, would have to apply a radial force to the occupants or any objects inside to keep them moving in a circle. To an astronaut, the effect of this force would be indistinguishable from the effect of the force of gravity, thus allowing eating, drinking, and working in comfort. The speed of rotation would determine the 'weight' of the astronaut in such a station. If an astronaut were to walk around the space station in the direction of rotation, his/her 'weight' would increase while, if he/she walked in the opposite direction, it would decrease. At the very centre of the station the astronaut would, of course, be 'weightless'.

Why may an orbiting satellite be considered to be continuously falling?

Consider a person firing rockets horizontally from the top of a very high mountain, Fig. 1.6. Paths A, B, C, D, and E represent possible trajectories for the rocket as the initial horizontal velocity is increased.

If there were no gravity, the rocket would travel in the direction towards F. So in every case, even E where the rocket is travelling in a circular orbit, it is continually ‘falling’ towards the centre of the earth. Hence, the apparent weightlessness experienced by an astronaut in orbit around the earth is similar to the apparent weightlessness experienced by a person in a freely falling lift.

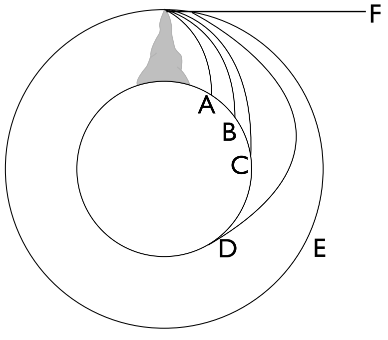


Fig. 1.6 *A rocket fired at different speeds from the top of a mountain*

If the rocket were fired at a velocity greater than that needed to put it into circular orbit E, it would travel in an elliptical orbit around the earth. The greater the velocity, the greater the eccentricity of the orbit. If the velocity were sufficiently great, the rocket would escape from the earth completely, travelling in a parabolic path.

What is meant by the escape velocity of a planet?

The escape velocity described above is exactly the speed required for a ball projected from the surface to escape entirely from the planet. At this speed, the energy is so large that gravity, pulling backward, can slow but never quite stop the outward motion. For the earth, the escape velocity is approximately 40 000 km h^{-1} or just over 11 km s^{-1}. On giant Jupiter the escape velocity is 60 km s^{-1}, which explains how Jupiter can retain an atmosphere of hydrogen while the earth cannot.

Escape from a planet becomes easier if the mass of the planet is small. For example, Phobos, a small satellite of Mars, has an escape velocity of only about 17 m s^{-1}. This means that you could throw a ball into space from Phobos, or you could ride into space on a motorcycle, if you could find a smooth highway to serve as a launch pad.

What is the exact period of a geostationary satellite?

A geostationary satellite is one which appears to remain stationary over a particular point on the earth, i.e. its period is the same as the period of rotation of the earth. In daily life, we measure time with reference to the sun. The solar day, 24 hours, is defined as the time between successive passages of the sun over a particular meridian. However, in this time, the earth has revolved more than 360^{o}, because the earth itself is travelling around the sun. The earth rotates 0.986^{o} per day (i.e. 360^{o}/365.25 days) around the sun, i.e. the earth rotates 360.986^{o} about its axis in 24 hours. Therefore, it takes 23 h 56 min and 4.1 s to rotate through 360^{o}. Hence, this is the period of a satellite in geostationary orbit.

How may the mass of the earth be found?

A century after Newton had proposed his law of gravitation, the English scientist, **Henry Cavendish** (1731-1810) demonstrated the gravitational force of attraction between small bodies a few centimetres apart. Two lead balls, each about 5 cm in diameter, were attached to the ends of a lightweight rod suspended by a long fine wire as shown, Fig. 1.7. Two larger lead balls, about 20 cm in diameter were placed as shown, almost touching the small balls. The gravitational attraction between the balls caused the system to rotate to a new equilibrium position determined by the stiffness of the wire. Cavendish was able to measure the gravitational force between the balls as being approximately 5 x 10^{-7} N and hence he was able to determine a value for G. Knowing G and the radius of the earth, r, it is possible to find a value for the mass of the earth. Consequently, Cavendish’s experiment was popularly known as ‘weighing the earth’.

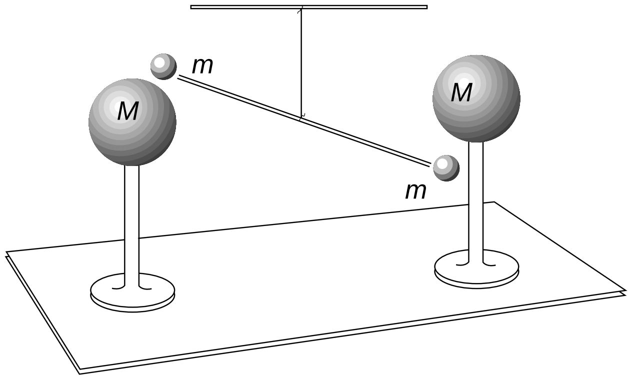


Fig. 1.7 *Cavendish's experiment*

Why do you not get wet swinging a bucket of water in a vertical circle?

If an object is swung in a vertical circle with speed, $v > \sqrt{rg}$ (see p.19) then the gravitational force is less than the centripetal force required to keep the object travelling in a circular path, and hence the object continues travelling in a circle. In the case of the bucket of water, the gravitational force is insufficient to pull the water out of the bucket. If the bucket is swung more slowly so that $v > \sqrt{rg}$ then you would get wet!

The same reasoning explains why an aircraft pilot, without being strapped into the cockpit, can 'loop-the-loop' without falling out.

Why do charged particles in magnetic fields move in circular paths?

When an electrically charged particle, with charge q, is moving in a magnetic field of magnetic flux density B, with velocity v perpendicular to the magnetic field, it experiences a force

$$F = qvB$$

acting always perpendicularly to its velocity. This provides the necessary centripetal force, hence causing the particle to travel with uniform circular motion. Note that when a particle is travelling with uniform circular motion, its instantaneous velocity is tangential to the circle, while the centripetal force is radial, towards the centre, i.e., the force is always perpendicular to the velocity. (See also module on Electromagnetism.)

Why do car wheels need to be balanced?

If motor car wheels are not properly balanced, the car will experience strong vibrations at particular speeds which can be unpleasant and can affect steering control. (Even when a wheel is balanced statically, it may not be balanced dynamically.) Wheel wobble is checked on a dynamic wheel balancing machine, with balancing masses placed on opposite sides of the wheel rim.

The doggie and the spin dryer.

When the spin drier rotates rapidly, the water is thrown out tangentially as shown, Fig. 1.8. The same happens when a dog shakes himself after coming out of the water, even though it may appear that the water is being thrown off radially.

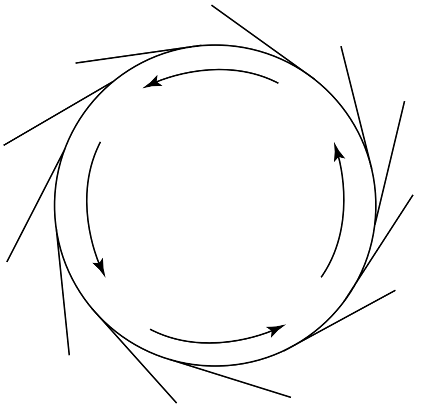


Fig. 1.8 *Water being expelled from a spin drier*

Why do you tend to topple over in a bus when it rounds a corner?

If you are standing in a bus when it begins to round a corner, you will topple over unless you hold on to something. What happens is that your body continues travelling in a straight line. However, your feet, which are on the floor of the bus, follow the circular path followed by the bus, hence causing you to fall over.

Likewise, in the case of a car travelling around a corner, you may experience a sensation of being pushed outward. What is happening, however, is that you continue to move in a straight line, while the car moves in a circular path. To make you go in the curved path, the back of the seat (friction) or the door of the car (direct contact) must exert a force on you. This force exerted *on* you *by* the seat or door provides the centripetal force necessary to make you travel in a circular path.

Why are roads sometimes banked at curves?

To make a car travel around a curve, an inward force must be exerted on it. This force is supplied by the friction between the tyres and the road. If the road is icy, so that the friction is reduced, or if the car is travelling too fast, then the force of friction is insufficient to supply the centripetal force necessary for the car to travel in a curved path and, instead, it skids off in almost a straight line.

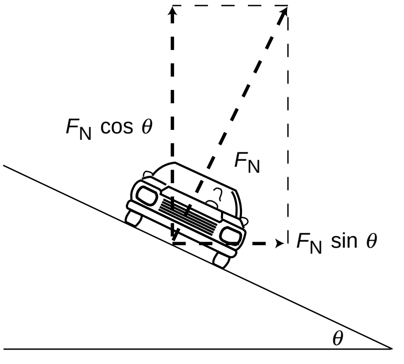


Fig. 1.9 *Car travelling in a circular path on a slope*

If the curve is banked, the chance of skidding is reduced because the normal force exerted by the road on the car has a component which acts toward the centre of the circle, Fig. 1.9. This reduces the need for friction. For a given angle of banking, there will be a particular speed at which the horizontal component of the normal force, $F_N \sin\theta$, is just equal to the force required to give a vehicle its centripetal acceleration. This occurs when

$$F_N \sin\theta = \frac{mv^2}{r}$$

The speed at which this occurs is called the 'design speed'. At this speed, there is no need for any friction between the tyres and the road to make the car travel around the curve.

1.3 Conceptual Approach

Gravitation

When introducing Newton's law of gravitation, its historical context ought to be introduced also (see section 1.1). In fact, the formulation of this law by Newton ended the belief that had existed for over 2000 years that the heavenly bodies were made of a substance different from, and obeyed laws different from, the earth. The fact that the mechanistic universe described by Newton seemed to explain the known facts so well led in turn to a new way of seeing everything, from philosophy to social organisation.

It is also useful to explain Newton's law by appealing to the common sense of the pupils. Once given the explanation that all mass particles attract each other, they will find it easier to understand that the force of attraction, F, between them depends on the masses of each of the particles, and depends inversely in some way on the distance between them. (An example of an attractive force which varies similarly with the square of the distance can be demonstrated by using two magnets.) By this reasoning, when two bodies of masses m_1 and m_2 are a distance d apart, then Newton's law states that:

$$F \propto m_1$$
$$F \propto m_2$$
$$F \propto \frac{1}{d^2}$$

It follows that $$F \propto \frac{m_1 m_2}{d^2}$$

or $$F = \frac{G m_1 m_2}{d^2}$$

where G is a constant of proportionality called the gravitational constant.

The question of why two pens on a desk, for example, do not appear to attract each other (or at least do not roll towards each other) is explained by the weakness of the gravitational force. The gravitational force only becomes significant when at least one of the bodies has a large mass. This also affords an opportunity to explain that gravity is one of the four fundamental forces of nature (the others being the electromagnetic force and the weak and strong nuclear forces, see p. 10) and that, while gravity is the weakest of all the forces, it is what maintains the motions of the solar system and the universe.

The idea of weight as a force is one that needs to be reinforced. (In the everyday experience of pupils, the word 'weight' is often used instead of the word 'mass': for example, a bag of sugar is often said to 'weigh' one kilogram.) The weight of a body on earth can be defined as the gravitational force exerted on it by the earth. Since force causes acceleration, weight also causes acceleration. On the earth's surface, the magnitude of this acceleration is approximately 9.8 m s^{-2}. (This varies slightly from place to place, according to the geological nature of the area and owing to the fact that the earth is not a perfect sphere. Of course, its value also decreases slowly with height above the earth's surface.) In general, the letter g is used to represent the acceleration caused by the force of gravity.

From the equation

$$F = \frac{Gm_1m_2}{d^2}$$

the weight of a body of mass m at the earth's surface is

$$W = \frac{GmM}{r^2}$$

where G is the gravitational constant, M is the mass of the earth and r is the radius of the earth. Since these are constants (assuming the earth to be spherical)

$$W = (\text{constant}) \times m \qquad \text{(i)}$$

From Newton's second law, this constant is equal to the acceleration of the body at the earth's surface, and is usually represented by the letter g.

i.e. $\qquad W = mg$

The use of the word 'gravity' is not very useful for students of physics, and can often lead to confusion about what exactly is meant. It will avoid many complications if pupils always refer to 'the force due to gravity' or 'the acceleration due to gravity' as appropriate.

From above, it can be seen that the value of g is given by

$$g = \frac{GM}{r^2} \qquad \text{(ii)}$$

Substituting the values for G, M and r, the value of g turns out to be 9.81 m s^{-2}. This value varies according to location on the earth's surface (see above).

Notes

1. Equation (ii) above can also be used to calculate the value of the acceleration due to gravity at distances above the earth's surface, or to calculate the acceleration due to gravity on the moon or other planets. See Worked Examples, p. 27.

2. From equation (i) the constant, g, is the force per unit mass acting on a body at a point in the gravitational field. In this sense, it can be thought of as the gravitational field intensity, analogous to the field intensity in an electric field which is defined as the force per unit charge at a point in the electric field.

Uniform Circular Motion

Uniform circular motion means motion at constant speed in a circle. This is a concept which pupils will recognise from everyday experience and which builds on concepts in physics which they have already met. They must understand that this type of motion involves acceleration, since the direction of the velocity is changing continuously. (Hence the need for a force to maintain the object in a circular path.)

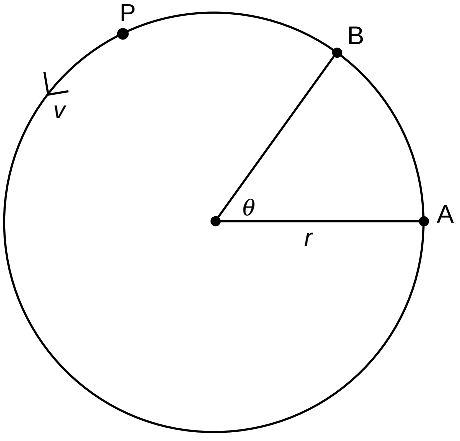


Fig. 1.10

To derive equations for centripetal acceleration, it is first necessary for pupils to understand the concepts of radian measure, angular displacement and angular speed. Introduce these simply from Fig. 1.10 which represents a particle P, travelling in a circular path of radius r with a constant speed v. (Indicate here to pupils that, though the particle is travelling with constant speed, it does not have constant velocity as the direction is continuously changing.) If the particle moves from A to B in time t, the angle θ is the angular displacement of the particle. If the angle θ is measured in radians, then its magnitude is defined as

$$\theta = \frac{|AB|}{r} \qquad \text{(i)}$$

where |AB| is the length of the arc AB.

Note

If θ (Fig. 1.10) measured 1 radian, |AB| would be equal to the radius; if θ measured 2 radians, then |AB| would be twice the radius, etc. For a complete circle, distance travelled would be $2\pi r$, where r is the radius of the circle. Hence the angular displacement is $2\pi r/r = 2\pi$, i.e. 2π radians $\equiv 360^{o}$.

The angular speed, ω, of the particle is defined as the rate of change of the angular displacement. Hence, for constant angular speed,

$$\omega = \frac{\theta}{t} \qquad \text{(ii)}$$

The periodic time, or period, T, of the particle is the time taken to complete one full revolution. Since this represents an angle of 2π radians, it follows that

$$\omega = \frac{2\pi}{T} \qquad \text{(iii)}$$

Since the linear speed, v, is given by

$$v = \frac{|AB|}{t}$$

it follows that

$$v = \frac{r\theta}{t}$$

$$= r\omega \qquad \text{(iv)}$$

This equation gives the relationship between the angular speed and the linear speed of the particle.

Centripetal Force and Centripetal Acceleration

The word ‘centripetal’ is derived from Latin *peto* (I seek), and means ‘centre seeking’. (The word ‘petition’ is derived from the same root.)

Deriving the equations for centripetal acceleration and centripetal force involves quite a high level of abstract mathematical reasoning. It is advisable, therefore, to introduce the topic by referring to the pupils’ existing knowledge and experience, and to link the mathematical processes into that.

Pupils will know from experience that if they twirl an object tied to a string in a circular motion they will feel the tension in the string exert an outward force on their hand. They will understand that this tension, in turn, exerts an inward force on the object, and that it is this inward force which keeps the object travelling in a circular path. Alternatively, pupils may understand the situation better if it is presented in terms of an object being twirled on the end of a spring. The increase in the length of the spring gives a visual indication of the forces acting on the object and the hand. This may be more readily understood than the concept of tension in a string.

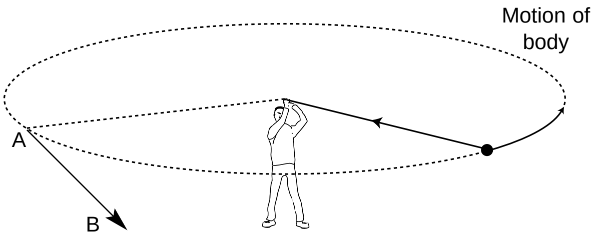


Fig. 1.11 *If string breaks when body is in position A body moves in direction AB*

It needs to be emphasised to pupils that if the force ceases to act (for example, if the string or spring snaps), the object will not continue on its circular path but will fly off at a tangent to it, Fig. 1.11. This is what happens when a hammer thrower releases the hammer in athletics. The inward force is called the centripetal force. Again, it needs to be emphasised to pupils that the centripetal force is not a separate force but the resultant force towards the centre of the circle when a particle is travelling in circular motion.

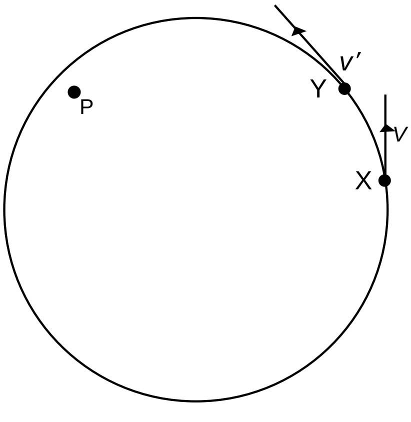


Fig. 1.12

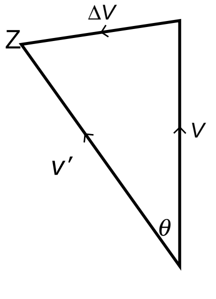


Fig. 1.13

Fig. 1.12 shows that the velocity of the particle, P, changes from v to v' as it moves from X to Y. This change in velocity is represented by the vector Δv, (Fig. 1.13). Let the time taken to travel from X to Y be t. The smaller the value of t, the smaller the magnitude of Δv, and the nearer the angle at Z becomes to 90^o. Hence, for instantaneous change, Δv is perpendicular to v, i.e. the change in velocity, and hence the acceleration, is directed towards the centre of the circle.

Since the angle at Z is almost 90^o,

$$\Delta v = v \sin\theta = v' \sin\theta \qquad \text{since } |v| = |v'|$$

$$\Delta v = v\theta \qquad \text{since, when } \theta \text{ is small, } \sin\theta \approx \theta$$

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

Thus, from eqn. (iv)

$$a = r\omega^2 \qquad \text{(v)}$$

or

$$a = \frac{v^2}{r} \qquad \text{(vi)}$$

These are the equations for centripetal acceleration.

Since

$$F = ma$$

the corresponding equations for centripetal force are

$$F = mr\omega^2$$

and

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

(See Appendix A, p. 52, for alternative derivation of formula for centripetal acceleration using calculus.)

Notes

1. It is important for pupils to understand that a 'centripetal force' is not some new kind of force. The term merely describes the resultant force towards the centre of the circle. For example, in the case of an object, tied to the end of a string and being swung in a horizontal circle, e.g. on a smooth surface, the centripetal force is equal to the force exerted *on* the object *by* the string. This force, in turn, is provided by the force which your hand exerts on the string. The force *on* your hand is equal in magnitude, but opposite in direction, to the centripetal force.

2. It is very important also for pupils to realise that there are not two forces acting inwards on a body travelling in a horizontal circle, *viz*. the force exerted by the string *and* the centripetal force. The only inward force is the force exerted by the string. This understanding becomes particularly important in studying the case of a body travelling in a vertical circle. To help to avoid misunderstandings, it is best to use diagrams which show only the actual forces acting on a body in circular motion.

Objects Travelling in a Vertical Circle

When a body is swung in a vertical circle on the end of a string two forces act on it. These are the tension, *T*, in the string and the weight of the body, *W (=mg)*. The forces acting on the body at the top and at the bottom of its path are as shown in Fig. 1.14.

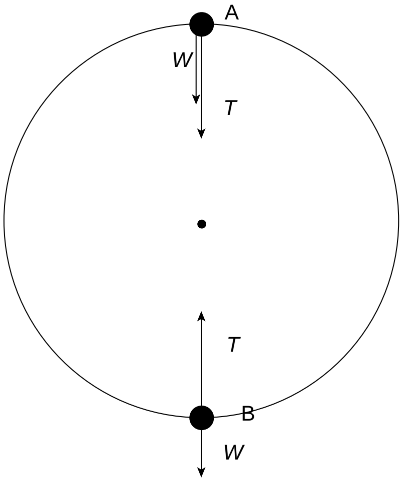


Fig. 1.14

At the top of the path, A, both forces act towards the centre, i.e. at A

$$T + mg = \frac{mv^2}{r}$$

Obviously, the bigger the value of *v*, the greater the tension, *T*, in the string; as *v* decreases, so *T* decreases. The tension in the string is zero when

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

$$\text{i.e.} \quad v = \sqrt{rg}$$

This is the minimum speed required for the object to travel in a vertical circular path. If the speed is less than this, then the weight acting towards the centre at A is greater than the centripetal force required to keep the body travelling in a circular path, and hence the body 'falls' out of the circular path (Fig. 1.15).

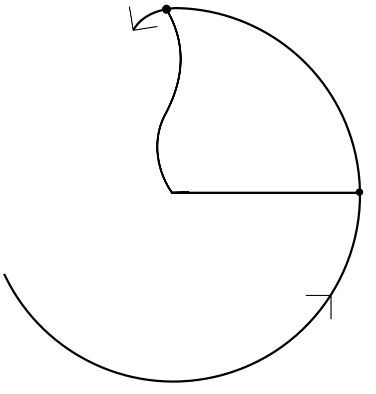


Fig. 1.15

At the bottom of the path, B, the tension and the weight are in opposite directions so, at B

$$T - mg = \frac{mv^2}{r}$$

Satellites, Moons, Planets and Circular Orbits

Since a satellite (or moon or planet) revolves around a central body (e.g. the earth or the sun) in an approximately circular orbit some force must act on it to maintain its circular motion. This force is the gravitational force exerted on it by the central body. From this starting point, the solutions to most questions concerning satellites are easily deduced. In this case the gravitational force is the resultant force on the orbiting body and therefore

$$\frac{GMm}{r^2} = mr\omega^2$$

where M is the mass of the central body, m is the mass of the satellite, and r is the radius of the satellite's orbit.

Substituting from eqn. (iii)

$$\frac{GMm}{r^2} = \frac{mr4\pi^2}{T^2}$$

where T is the periodic time of the satellite. Thus,

$$T^2 = \frac{4\pi^2 r^3}{GM}$$

This equation can be applied to any satellite in a circular orbit, e.g. the earth or other planets about the sun, or artificial satellites around the earth. The equation gives the relationship between the period, the mass of the central body and the radius of the orbit. If the mass of the central body is constant then the square of the period is proportional to the cube of the radius of the orbit - Kepler's third law.

A satellite in geostationary orbit is one which is always above the same point on the earth. Such a satellite must be directly above the equator, and must have a period of approximately 24 hours (see p. 13).

Interesting topics to introduce at this juncture are weightlessness (p. 23) and escape velocity (p. 13).

Note

The derivations performed above apply to circular orbits. The orbits of the planets and of many satellites are elliptical, but not far removed from being circular, and the formulae derived above can be applied with reasonable accuracy. They do not apply, however, for comets or satellites with very elliptical orbits.

1.4 Experimental Approach

Experiments to measure the acceleration due to gravity by free-fall methods have been described in the module *Mechanics 1*. For measurement of the acceleration due to gravity using a simple pendulum, see p. 35.

Circular motion is not a part of the Leaving Certificate Physics course that lends itself to laboratory demonstrations, nor are there any prescribed experiments from this section of the course. However, it is a section of the course where pupils' experience can be called upon in order to relate the theory to their understanding of the world around them.

Experiment 1.1

To Demonstrate the Relationship Between Centripetal Force and Rate of Rotation (radius of rotation constant)

Apparatus

Glass tube, string, fixed mass, selection of smaller masses (e.g. washers), paper clip.

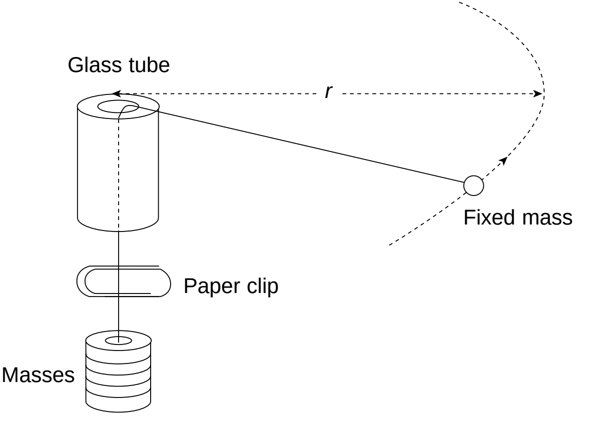


Fig. 1.16

The glass tube, Fig. 1.16, should have smooth ends. The fixed mass is whirled round at a constant rate so that the paper clip remains stationary just below (not touching) the end of the glass tube. The centripetal force is equal to the weight of the hanging masses. The rate of rotation can be found by timing a fixed number of revolutions using a stop watch. From this find the periodic time, T, for one revolution, and the rate of rotation, f, is then given by

$$f = 1/T$$

Repeat for different values of m, where m is the total mass of the hanging masses. A graph of f^2 against m should give a straight line through the origin.

Experiment 1.2

To Demonstrate the Relationship Between Centripetal Force and Radius of Rotation (rate of rotation constant)

Apparatus

As in previous experiment.

Because it would be difficult to find the exact weight to keep the rate of rotation the same for different radii, the following procedure is used.

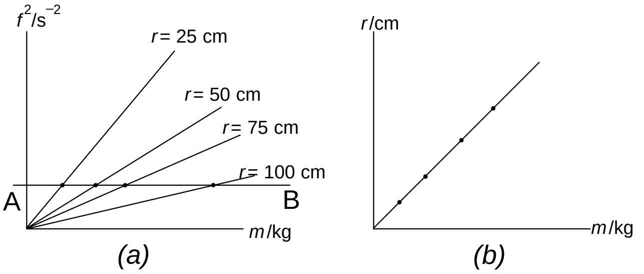


Fig. 1.17 *(a) Graph of f^2 against m for different values of r, (b) graph of r against m*

The experiment is carried out as in the previous experiment using different radii of rotation, and a series of straight lines is obtained as shown, Fig. 1.17(a). The values of m for each radius which correspond to a particular value for f^2 are now found. This is done by drawing a horizontal line AB as shown, and finding the values on the x-axis corresponding to where AB cuts each line. A second graph is then drawn of the radius, r, against m, Fig. 1.17(b). This should be a straight line through the origin.

1.5 Applications

The Gravimeter

A gravimeter is an instrument which can detect small variations in the gravitational acceleration (gravimeters today can detect variations in g to 1 part in 10^8). It is widely used for mineral and oil prospecting, as unusual underground deposits can be detected by the presence of local gravitational variations.

The value of g is not constant at all points on the earth's surface. It is affected by mountains and valleys, and by the fact that the earth is not a perfect sphere and bulges at the equator. The earth's rotation also has an effect on the apparent value of g. The value of g varies from about 9.78 m s^{-2} at the equator to approximately 9.83 m s^{-2} at the poles.

Additionally, however, very tiny local variations can be detected in the value of g. These are called 'gravity anomalies' and are caused by rocks of different densities and the presence of other irregularities, including mineral and oil deposits, in the earth's crust. Such measurements are used by geophysicists to investigate the structure of the earth's crust, and in mineral and oil exploration. For example, oil is often found under salt domes which have lower density than rocks in the earth's surface. Hence, slightly smaller values for g than in surrounding areas have sometimes led to the discovery of oil. Likewise, higher local values of g may mean that deposits of ores of heavy metals are present, as these have higher than average density.

Since Newton's time, differences in the value of the acceleration due to gravity at different locations have been recognised by timing the same pendulum in different places. The English physicist, **Henry Kater**, in 1817 was the first to make absolute measurements of g using a reversible pendulum. If a rigid pendulum has two points of support about which the period of swing is the same, then the distance between these two points is equal to the length of a simple pendulum having the same periodic time. Kater's pendulum was the most accurate way of measuring g until the 1950s when electronic timers made it possible to measure accurately the time taken by a body in free fall to fall through a fixed distance.

Other types of gravimeters that have been used include ones using vibrating strings where the frequency of vibration is determined by g, and the superconducting gravimeter where the height of a magnetically levitated superconducting disk depends on the value of g. (See module on electromagnetism.) Very accurate gravimeters using lasers as light sources for interferometers, where the falling body reflects the beam of light back on itself, have been developed and can measure g with an accuracy of one part in 10^8.

From the time of Newton, measurements of differences in the acceleration due to gravity were made by timing the same pendulum at different places. During the 1930s, however, static gravimeters based on free-fall measurements replaced pendulums for detection of local variations in the value of g. Spring gravimeters balance the force of gravity against the elastic force of the spring, using electronic means to achieve high sensitivity. Vibrating string gravimeters in which the string's vibration frequency is determined by g also have been developed. Apollo 17 astronauts used such a device to conduct a gravity survey of their moon landing site.

Governors of Steam Engines

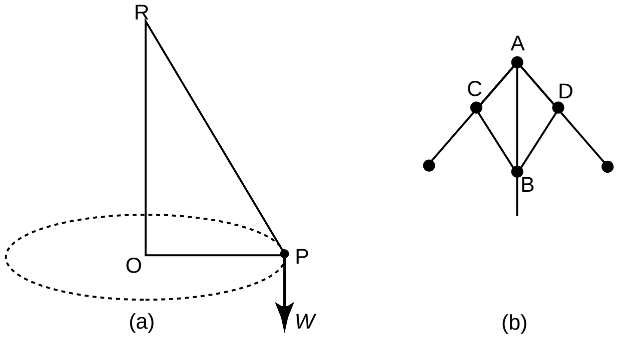


Fig. 1.18 *Conical pendulum*

When an object is swung round as a conical pendulum, Fig. 1.18(a), the greater the angular velocity, the greater the height it rises from its rest position. This principle is used to regulate the supply of steam to an engine which is required to rotate a shaft at constant speed.

Two light rods with weights attached are hinged at A to a vertical shaft which is rotated by the engine, Fig. 1.18(b). Two other rods, BC and BD, are hinged to a collar at B which can slide up or down the shaft as shown. A lever attached to B opens or closes a valve thus regulating the amount of steam to the engine. (It is arranged so that when B rises it closes the valve.)

When the speed of rotation of the shaft increases, the weights on the rods rise and pull B up, thus shutting off some of the steam and slowing down the engine. If the speed decreases, B is lowered, letting in more steam and speeding up the engine.

Balancing Rotating Masses

In engines of various types, objects are frequently attached to rotating shafts. When a mass m is attached at a radius r to a shaft rotating with an angular velocity ω a centripetal force of magnitude $mr\omega^2$ is required to maintain the mass in circular motion. This force is supplied by the shaft which, in turn, experiences an equal and opposite force. This force will produce considerable stresses in the shaft and can result in wear of bearings and cause vibrations which are unpleasant and even dangerous. These effects can be counteracted by placing a second mass, m_B, Fig. 1.19, opposite to the original mass so that the two forces on the shaft cancel each other out.

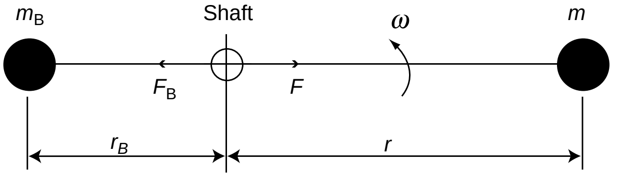


Fig. 1.19

The Wall of Death

Found in carnivals or funfairs, the 'wall of death' is a large upright drum inside which people stand with their backs against the wall. The drum is rotated at an increasing rate and at full speed, when the floor is pulled downwards, the occupants are 'stuck' to the wall.

The reason that the people do not slide down the wall is that the wall, which is travelling in a circular path, exerts a force inwards, F_i, (i.e. a centripetal force) on the occupants, which forces them to travel in a circle. (The occupants exert an equal and opposite force, F_o, on the wall.) The force of gravity, i.e. the weight, W, acting on the occupants acts downwards. The frictional force, F_f, acts in the opposite direction to W, and is proportional to the normal reaction, F_i, i.e.,

$$F_f \propto F_i$$

$$\text{But,} \qquad F_i = mr\omega^2$$

$$\Rightarrow \qquad F_f \propto \omega^2$$

Thus, if the angular speed, ω, of the drum is great enough, the frictional force is greater than the weight of the occupants, and so they do not slide down the wall.

The Centrifuge

A centrifuge is a device used to sediment materials quickly or to separate materials with different densities. In a typical centrifuge machine, test-tubes are held in the centrifuge rotor, with the top of the test-tubes nearest the axis of rotation. When the rotor is rotated at high speed the test-tubes are approximately perpendicular to the axis of rotation. In this position, the bottom of the test-tube exerts a force on the liquid at the bottom, providing the centripetal force required to keep the liquid moving in a circle. This force is transmitted through the liquid, decreasing in magnitude towards the top of the test-tube, i.e. nearer the centre of the circle. There is thus a pressure differential between the top and bottom of the tube analogous to that which exists in a test-tube placed vertically at rest. Less dense material therefore moves towards the top of the tube while more dense material moves towards the bottom.

A centrifuge is used to separate substances which do not separate quickly under gravity. The centrifuge speeds up the process.

Weight and Apparent Weightlessness

Let us, at this point, define the term weight precisely. The term is defined differently in different texts. We will use the more common definition, viz. weight is the force of gravity on a body (cf. p. 15).

If you were to stand on a spring weighing machine at the North or South poles, then the reading on the machine would be your weight. This would not be the case, however, at other points on the earth's surface because of the rotation of the earth. A body at the equator is rotating in a circle. This requires a centripetal force acting towards the centre of the earth. In this case there are two forces acting on the body - its weight, W, and the force, R, exerted on it by the weighing machine (this is the value registered by the weighing machine).

Thus

$$W - R = \frac{mv^2}{r} \quad \text{where } r \text{ is the radius of the earth}$$

or

$$R = W - \frac{mv^2}{r}$$

Hence, if you stood on the weighing machine at the equator, the reading it would give would be reduced.

To understand this more clearly, consider the following thought experiment. Suppose a person was standing on a weighing machine at the equator when, suddenly, the earth began to rotate more rapidly on its axis (ignore any atmospheric effects this might cause). As the rate of rotation increases, the centripetal force needed to keep him/her travelling in the circular path increases, and so the reading on the weighing machine is less, i.e. the person *appears* to have less weight.

If the earth's speed of rotation continues to increase, then a certain critical speed will be reached where the necessary centripetal force will be just equal to the total gravitational force. At this point the person will be *apparently* weightless, even though the full gravitational force still continues to act.

For an astronaut in a satellite circling the earth, the situation is similar. The gravitational force acting on the astronaut is equal to the centripetal force required to keep him/her travelling in a circular orbit. Hence, the astronaut exerts no force on the 'floor' of the satellite, and experiences apparent weightlessness.

If you were standing on a weighing machine in a lift which was in free fall, then you would not exert any force on it, i.e. you would appear to be weightless. Indeed, any time our bodies are unsupported we experience apparent weightlessness even though our weight is still *mg*.

In a space vehicle, because the astronaut is circling the earth at the same rate as the vehicle itself, he/she exerts no force on the floor of the spacecraft. The same applies to other objects in an orbiting space vehicle so that they do not 'fall' when released. Hence, anything not in use must be firmly fixed. One operation made difficult by weightlessness is drinking, since liquids will not pour. In addition, controlled movement is possible only by the use of handrails.

It is important to appreciate that, although apparently weightless, a body still has mass and is just as difficult to push in space as it is on earth. So an astronaut 'floating' in a spacecraft could still be injured by hitting a hard but 'weightless' object.

Weightlessness has some interesting effects on people. Prolonged weightlessness in space can affect health. For example, red blood cells diminish, muscles lose their tone, and bones lose calcium and become brittle. Another effect of weightlessness is that a person could stand with arms outstretched without getting tired. This effect has many applications in athletics where people can experience temporary weightlessness in many free fall situations, for example, jumping, diving, trampolining or even between strides when running. In these situations, the limbs can be moved much more freely since only inertia, and not gravity, needs to be overcome, giving the person increased mobility.

Hazards Faced by Cosmonauts

The launching of a space craft and its re-entry into the earth's atmosphere both involve very large accelerations which are frequently expressed as multiples of g (9.8 m s^{-2}) and may be as high as 8g. Such accelerations can be dangerous to human beings, causing blood to be drawn from some parts of the body and to accumulate in others, causing loss of vision and even unconsciousness. Tests have shown that a person can withstand 15g for a few minutes when his/her body is at right angles to the direction of the acceleration, but only 6g when it is in the direction of the acceleration because, in this case, the brain can be deprived of blood. This is why at lift-off and re-entry an astronaut lies at right angles to the direction of travel.

When a spacecraft re-enters the earth's atmosphere at speeds of up to 30 000 km h^{-1}, very high temperatures are produced due to the force exerted on the spacecraft by the air. (What happens is that the force slows down the spacecraft, converting some of its kinetic energy into heat.) To protect it, the space craft is covered by a heat shield shaped like a honeycomb which disperses the heat.

Communications Satellites

The earth is encircled by a ring of man-made satellites that facilitate a continuous exchange of video, audio, and data signals. Most of these satellites inhabit a narrow band of outer space 35 786 km above the earth's equator. They complete one orbit of the earth every 24 hours (see p. 13), i.e. their angular velocity is the same as that of the earth's rotation on its axis, and so they appear to be stationary above a certain point of the earth's equator. They are said to be in geostationary orbit.

Congestion in the geostationary orbit is one of the problems being faced by satellite operators seeking preferential parking positions. Satellites must have an angular separation of about 2^{o} of longitude so as not to interfere with each other's signal. Each such satellite is allocated a position in this orbit by international agreement. For example, the Astra satellite, broadcasting TV and radio programmes (including Sky TV) to Europe is at the 13^{o} East position.

Note that the height of a satellite in geostationary orbit is between five and six times the radius of the earth, so that such a satellite is a long way above the earth's surface, and each satellite can therefore cover almost half the earth's surface. Three satellites can cover the whole of the earth, except for the polar regions. For example, there are three satellites which are used for international TV and telephone signals, one over each of the three major oceans, the Atlantic, the Pacific and the Indian, and these three are sufficient to cover all inhabited areas of the earth, and carry signals from continent to continent.

The vast majority of communications satellites, though not all, use geostationary orbits. Such orbits give poor coverage in high latitudes and polar regions. Systems that are required to cover these regions use highly eccentric elliptical orbits inclined at 63.4^o to the equatorial plane called a Molniya orbit (Molniya is Russian for lightning). These orbits typically have an apogee (furthest point from the earth) of approximately 40 000 km and a perigee of 1000 km. Because their orbits are so elliptical, the satellites move considerably more slowly at the apogee, and hence appears to remain quasi-stationary for 8-12 hours. Hence a full 24-hour service can be maintained by three such satellites with suitably phased orbits and hand-over facilities between them. Fig. 1.20(a) shows the view of the earth from a geostationary satellite, while Fig. 1.20(b) shows the view from a Molniya satellite.

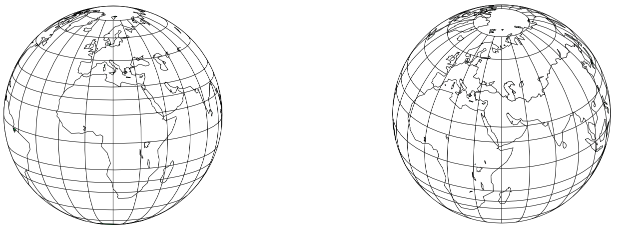


Fig. 1.20 *(a) Earth from a geostationary satellite, (b) earth from a Molniya satellite*

Satellite orbits are subject to perturbations due to a number of factors. The fact that the earth is not a perfect sphere (it is flatter at the poles - the equatorial radius is about 6378 km, and the polar radius about 21 km less) of uniform density means that the force of gravity at a point is not exactly towards the centre of the earth. This causes the orbital plane to rotate slowly about the earth's axis. Even at the equator the earth is not perfectly circular, and this causes slight instability in the period of satellites in geostationary orbit, which has to be corrected. The moon and sun both have disturbing effects on a satellite in geostationary orbit. Solar radiation increases orbital eccentricity. Signals from satellites in highly elliptical orbits are subject to significant Doppler shifts, and receiving apparatus must be capable of accommodating this.

Satellites receive their energy from the sun using solar panels to convert sunlight directly into electricity. However, for a period of about a month around the spring and autumn equinoxes, when the orbits of the earth, sun and geostationary satellite are in nearly the same plane, the satellite suffers eclipses when the earth is between it and the sun, lasting for up to 70 minutes per day. For this reason, satellites must carry secondary sources of energy, usually storage batteries, to avoid disruption of services.

For communications satellites stability is important so that their antennae remain directed towards the region of the earth where their signals are to be received. For stability, satellites use gyroscopes. In the simplest satellites, the body ('bus') of the satellite spins on its axis, and the antennae are counter rotated, thus effectively remaining stationary, directed towards the earth. In a 3-axis stabilised system, the spacecraft body contains a spinning gyroscope (called a momentum wheel) to act as an inertial reference.

1.6 Worked Examples

1. What is the weight of a 50 kg bag of cement?

$$W = mg$$

$$= 50 \times 9.8$$

$$= 490 \text{ N}$$

2. A stone of mass 180 g is whirled in a circle at constant speed in a horizontal plane at the end of a string 60 cm long. If the angular speed of the stone is 6.0 rad s^{-1} calculate (a) the linear speed of the stone and (b) the tension in the string.

(a) $v = r\omega$

$= 0.6 \times 6$

$= 3.6 \text{ m s}^{-1}$

(b) Tension in string = centripetal force

$T = mr\omega^2$

$= 0.18 \times 0.6 \times 6^2$

$= 3.9 \text{ N}$

3. If the stone in the previous example were swung in a vertical plane at the same speed, would it be travelling fast enough to complete a full circle? If so, calculate the tension in the string (a) at the top (b) at the bottom of the circle.

At the highest point in the circle the tension in the string plus the weight of the stone equals the centripetal force (see p. 19), i.e.

$$T + mg = mr\omega^2$$

If the stone is to complete the circle the string must not go slack, i.e.

$$T \geq 0$$

Therefore, $mg \leq mr\omega^2$

$mg = 0.18 \times 9.8$

$= 1.76 \text{ N}$

From the previous question $mr\omega^2 = 3.9$ N. Hence, the stone will complete the circle.

(a) At the top of the circle,

$$T + mg = mr\omega^2$$

$\Rightarrow$ $T = mr\omega^2 - mg$

$= 2.1 \text{ N}$

(b) At the bottom of the circle,

$$T - mg = mr\omega^2$$

$\Rightarrow$ $T = mr\omega^2 + mg$

$= 5.7 \text{ N}$

4. A body attached to a string 40 cm long is being swung in a vertical circle. What is the minimum speed that the body would require at the top of the circle in order to move in a complete circle?

The minimum speed required is such that, at the top of the circle, the tension in the string is just equal to zero, i.e.

$mg + 0 = mv^2/r$

$\Rightarrow$ $v^2 = rg$

$= 0.4 \times 9.8$

$\Rightarrow$ $v = 2.0 \text{ m s}^{-1}$

5. What is the gravitational force between two bodies each of mass 20 kg which are 50 cm apart (the distance between their centres of mass)? What acceleration would this force impart to each of the masses? ($G = 6.7 \times 10^{-11}$ N m^2 kg^{-2}.)

$$F = \frac{Gm_1m_2}{d^2}$$

$$= \frac{6.7 \times 10^{-11} \times 20 \times 20}{0.5^2}$$

$$= 1.07 \times 10^{-7}$$

$$= 1.1 \times 10^{-7} \text{ N}$$

$$F = ma$$

$$1.07 \times 10^{-7} = 20a$$

$$a = 5.4 \times 10^{-9}\ \text{m s}^{-2}$$

Note

To appreciate just how small this acceleration is consider that, starting from rest in a frictionless environment, it would take the body 32 minutes to accelerate through a distance of 1 cm.

6. Calculate the acceleration due to gravity on the surface of the moon. (Mass of moon = 7.4 x 10^{22} kg, radius of moon = 1.74 x 10^6 m, G = 6.7 x 10^{-11} N m^2 kg^{-2}.)

$$g = \frac{GM}{r^2}$$

$$= \frac{6.7 \times 10^{-11} \times 7.4 \times 10^{22}}{(1.74 \times 10^6)^2}$$

$$= 1.6\ \text{m s}^{-2}$$

7. What is the acceleration due to gravity at a height of 1000 km above the earth's surface? (Mass of earth = 6 x 10^{24} kg, radius of earth = 6.4 x 10^6 m, G = 6.7 x 10^{-11} N m^2 kg^{-2}.)

$$g = \frac{GM}{r^2}$$

$$= \frac{6.7 \times 10^{-11} \times 6 \times 10^{24}}{(6.4 \times 10^6 + 1 \times 10^6)^2}$$

$$= 7.3\ \text{m s}^{-2}$$

8. The mass of Jupiter is 316 times that of the earth, and its radius is 11 times that of the earth. Find the value of the acceleration due to gravity on the surface of Jupiter, given that its value on the earth is 9.8 m s^{-2}.

Note

In calculations of this kind, where comparisons are being made, simple formulae can often be derived to make the calculation easier.

$$g_E = \frac{GM_E}{r_E^2}$$

$$g_J = \frac{GM_J}{r_J^2}$$

Dividing, we get

$$\frac{g_J}{g_E} = \frac{M_J r_E^2}{M_E r_J^2}$$

$$\frac{g_J}{9.8} = \frac{316}{11^2}$$

$$g_J = 25.6\ \text{m s}^{-2}$$

9. Calculate (a) the height and (b) the speed, of a satellite in geostationary orbit around the earth. (Mass of earth = 6 x 10^{24} kg, radius of earth = 6.4 x 10^6 m, G = 6.7 x 10^{-11} N m^2 kg^{-2}.)

(a) In calculations involving satellites, moons or planets revolving about a central body, begin with:

gravitational force = centripetal force (see p. 20)

$$\frac{GMm}{r^2} = mr\omega^2 \qquad \text{(i)}$$

where M is the mass of the central body (the earth in this example) and m is the mass of the orbiting body (the satellite in this example). Since the period of a satellite is given by

$$T = \frac{2\pi}{\omega}$$

equation (i) becomes

$$\frac{GMm}{r^2} = \frac{4\pi^2 mr}{T^2}$$

$$r^3 = \frac{GMT^2}{4\pi^2}$$

$$= \frac{6.7 \times 10^{-11} \times 6 \times 10^{24} \times (24 \times 3600)^2}{4\pi^2}$$

$$= 7.6 \times 10^{22}$$

$$r = 4.2 \times 10^7\ \text{m}$$

This is the radius of the orbit of a geostationary satellite (T = 24 h). The height above the earth's surface, then, is

$$\begin{aligned}\text{height} &= 4.2 \times 10^7 - 6.4 \times 10^6 \\ &= 3.6 \times 10^7 \text{ m}\end{aligned}$$

(b) The speed of the satellite is calculated from

$$\begin{aligned} v &= r\omega \\ &= \frac{2\pi r}{T} \\ &= \frac{2\pi \times 4.2 \times 10^7}{24 \times 3600} \\ &= 3.1 \times 10^3 \text{ m s}^{-1}\end{aligned}$$

10. Calculate the escape velocity for the earth (i.e. the velocity needed for a body to escape from the earth's gravitational field). (Mass of earth = 6×10^{24} kg, radius of earth = 6.4×10^6 m, G = 6.7×10^{-11} N m^2 kg^{-2}.)

The escape velocity can be calculated by equating the kinetic energy necessary to escape with the gravitational potential energy at an infinite distance from the earth, i.e.

$$\frac{1}{2}mv^2 = \int_R^\infty \frac{GMm}{r^2}\mathrm{d}r$$

where v is the escape velocity, r is the distance from the centre of the earth, R is the radius of the earth and M is the mass of the earth.

Integrating, we get

$$\begin{aligned}\frac{1}{2}mv^2 &= \frac{GMm}{R} \\ v &= \sqrt{\frac{2GM}{R}} \\ &= 1.1 \times 10^4 \text{ m s}^{-1}\end{aligned}$$

Note

The escape velocity is equal to the (circular) orbital velocity x $\sqrt{2}$. Its value is approximately 40 000 km h^{-1}.

CHAPTER 2

SIMPLE HARMONIC MOTION

2.1 Background

Galileo and the Pendulum

According to legend, Galileo, in 1581, discovered the principle of the pendulum while observing a swinging lamp in the cathedral in Pisa. He realised that each complete cycle of the lamp took the same amount of time, compared to his own pulse, even though the amplitude of each successive swing was smaller than that of the previous one. He later showed that the period of oscillation of a simple pendulum is proportional to the square root of its length and does not depend on its mass. The discovery of the pendulum led directly to the invention of the first accurate mechanical clocks.

The period of a given pendulum varies with the gravitational field. Hence, a pendulum will swing faster at the poles than at the equator. Also, the period increases with increasing altitude, and so pendulums were commonly used as altimeters.

Pendulum Seismometer

The basic problem in measuring ground motions is to attain a point that remains steady when the ground moves. Various types of pendulum have been used for this purpose.

If the case containing a simple pendulum is vibrated horizontally, for example by an earth tremor, the bob of the pendulum will remain motionless through inertia, provided the period of the vibration applied to the case is much shorter than the period of the pendulum. Hence the movement of the ground relative to the pendulum can be recorded.

The Pendulum Clock

Galileo was the first to recognise that a pendulum could be used for measuring time. The first successful pendulum clock was invented in 1656 by the Dutch scientist, **Christiaan Huygens** (1629-1695), who solved the problem of making the period of the pendulum truly constant by devising a pivot that caused the bob to swing along the arc of a cycloid rather than that of a circle. Huygens' invention brought about a great increase in clockmaking. In 1670, **William Clement**, an English clockmaker, invented the long-case, or grandfather, clock.

For a clock which uses a pendulum of length 1 m, an increase in length of approximately 2.5×10^{-5} m causes it to lose one second per day. Altering the length of a pendulum is, therefore, a sensitive means of regulation. The alteration is usually carried out by allowing the bob to rest on a nut that can be screwed up or down the pendulum rod.

Changes of temperature, causing the rod of the pendulum to expand or contract, will obviously affect the timekeeping of a pendulum, e.g. a pendulum clock with a steel rod will lose one second a day for a rise in temperature of 2.2 °C. In order to keep the length of the pendulum constant as the temperature fluctuates, many clocks have pendulum rods made of invar, a special alloy which has such a small coefficient of expansion that small changes of temperature have a negligible effect.

The motion of the pendulum in a clock is maintained by transferring the energy in a spring, or a 'falling weight', to it to overcome the effects of friction. The mechanism by which this is achieved is called an escapement. The escapement also transfers the regular motion of the pendulum to a series of cog wheels which cause the hands of the clock to rotate.

The Metronome

A metronome is an instrument for marking musical tempo, invented in 1816 by the Dutchman, **Dietrich Nikolaus Winkel** (c.1776-1826) but patented by the German **Johann Nepomuk Maelzel** (1772-1838). It consists of a pendulum whose period is controlled by an adjustable mass that is slid up and down the pendulum. The metronome makes a ticking sound as the pendulum oscillates to and fro. By setting the metronome at the speed indicated on the music, a performer can be sure he/she is adopting the tempo intended by the composer. The conventional metronome is housed in a pyramidal case. Pocket and electric metronomes are also made.

2.2 Do You Know

What was Foucault's pendulum?

In 1851, the French physicist **Jean-Bernard-Léon Foucault** (1819-1868) suspended a pendulum consisting of a large iron ball by a long steel wire from the dome of the Panthéon in Paris. The pendulum was suspended so that its plane of swing was free to rotate and it was kept in motion by mechanical means, similar to a pendulum clock. He noticed that the plane of swing of the pendulum rotated in a clockwise direction (viewed from above). This was the first laboratory demonstration that the earth rotates on its own axis. What was really happening, of course, was that, while the earth rotated beneath it, the bob's inertial resistance to change of direction made it appear as if its plane of swing were rotating.

At the poles, the plane of swing of a pendulum rotates through 360° every 24 hours, with the rate of rotation decreasing with latitude until, at the equator, there is no rotation. In Paris, the period of rotation is approximately 32 hours.

In the southern hemisphere, the direction of rotation is counter-clockwise.

How are musical notes related to simple harmonic motion?

Musical sounds consist of a combination of many superimposed simple harmonic waves, the frequencies of which are multiples of a fundamental frequency. In fact, any regularly repetitive motion and any wave, no matter how complicated its form, can be treated as the sum of a series of simple harmonic motions or waves, a discovery first published in 1822 by the French mathematician **Baron Jean-Baptiste-Joseph Fourier** (1768-1830).

What is damped harmonic motion?

The description of simple harmonic motion usually disregards the friction which occurs in most oscillating systems and which causes them to decrease in amplitude and ultimately to stop vibrating entirely. For example, a real pendulum encounters resistance from the air through which it moves, causing its amplitude to decrease. Fig. 2.1 shows a graph of the displacement as a function of time for such a pendulum. Other examples of damped harmonic motion include the prongs of a vibrating tuning fork which radiate sound, resulting in a loss of energy in the tuning fork, and an oscillating electrical circuit which loses energy to its own circuit resistance and to whatever radiation is emitted. Almost the only cases of totally undamped oscillation occurring in nature are the vibrations of polyatomic molecules and the vibrations of the atoms of a medium through which radiation is passing without absorption.

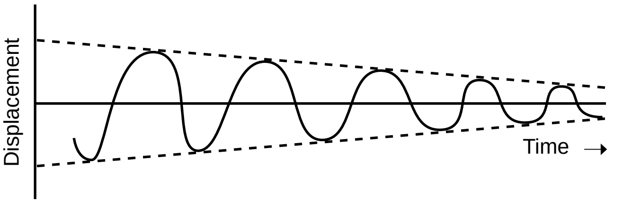


Fig. 2.1 *Damped harmonic motion*

Fig. 2.2 shows three different types of damped systems. Curve A shows a situation in which the system makes several oscillations before coming to rest. Such motion is said to be underdamped. In the case of Curve C, where the damping is so large that it takes a long time to reach equilibrium, the motion is said to be overdamped. Curve B represents critical damping: in this case equilibrium is reached most quickly.

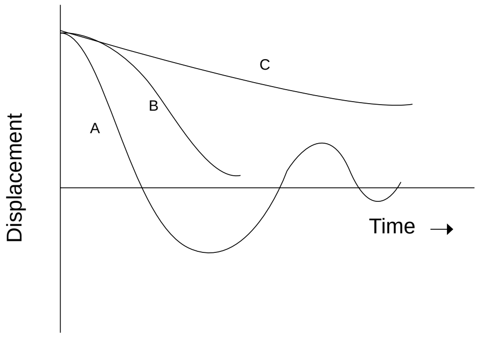


Fig. 2.2 *Different types of damped harmonic systems*

Damping is used in such systems as door closing mechanisms and shock absorbers in cars. Such devices are usually designed to give critical damping but, as they wear out, underdamping occurs - a door slams or a car bounces up and down several times when it hits a bump. Needles on analogue meters, e.g. voltmeters, ammeters, level indicators on tape recorders, are usually critically damped or slightly underdamped. If they were very underdamped they would swing back and forth excessively before arriving at the rest position, and if overdamped, they would take too long to reach equilibrium and any changes in the signal (say, recording level) would not be detected. The damping may be provided by viscous, frictional, electrical or magnetic means.

2.3 Conceptual Approach

Introduction

When we speak of a **vibration** or an **oscillation**, we mean the motion of an object that regularly repeats itself, back and forth, over the same path. That is, the motion is **periodic**. The simplest form of periodic motion is called simple harmonic motion.

The nearest approximation to simple harmonic motion with which most pupils are familiar is the motion of a pendulum. A simple pendulum is an ideal pendulum consisting of a particle (i.e. a body having mass but no size) on the end of a weightless inextensible support and moving without friction. In practice, the familiar spherical metal bob on the end of a light string approximates sufficiently closely to a simple pendulum and is usually referred to as such. Unfortunately, from a teaching point of view, the pendulum is not as easy to explain mathematically as an object oscillating on the end of a spiral spring, obeying Hooke's law. It is still, however, probably the best way to introduce simple harmonic motion to pupils.

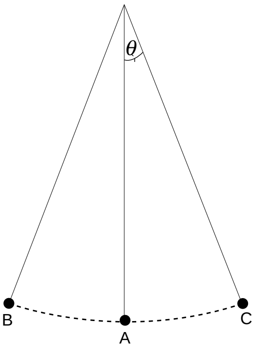


Fig. 2.3 *Simple pendulum*

At the outset, make sure that pupils understand that for a simple pendulum:

- the motion is periodic;
- the velocity is a maximum and the acceleration zero at A (Fig. 2.3);
- the velocity is zero and the acceleration is a maximum at B and C;
- the kinetic energy is a maximum at A and is zero at B and C, while the potential energy is maximum at B and C and a minimum at A;
- the total mechanical energy is conserved in the system.

It is worth pointing out at this time that:

- the motion of the pendulum only approximates to simple harmonic motion and that the approximation is acceptable when the angular displacement, θ, is small (in practice less than about 5°);

- the period of the pendulum is independent of the amplitude of the motion for small amplitudes (this is why pendulums are so useful in clocks);

- if a graph were drawn showing how the displacement of the pendulum varies with time, the graph would be sinusoidal. (It is worth showing a sinusoidal wave pattern on the oscilloscope, and also showing other wave patterns such as sawtooth or square wave to indicate that not all periodic motion is sinusoidal.) Note that the word 'harmonic' *means* that the motion is sinusoidal.

There are very many instances in nature where simple harmonic motion, or motion which is approximately simple harmonic, is encountered, e.g. the spring-and-mass oscillator, a vibrating tuning fork, a capacitor-inductor electrical circuit, a child's swing, a sounding organ pipe, a vibrating diatomic molecule, a bottle bobbing up and down in the water, etc. The range of phenomena which exhibit simple harmonic motion is very great, ranging from the vibrations of the surface of an atomic nucleus excited by the collision with a neutron where the frequency is of the order of 10^{20} Hz, to the oscillations of a Cepheid variable star which cause cyclical variations in the brightness of the star with a frequency of the order of 10^{-6} Hz.

An understanding of simple harmonic motion is also central to understanding wave motion. In water waves or sound waves, each oscillating molecule moves with simple harmonic motion. Since the same kind of description applies to electromagnetic waves and matter waves, we find that the concepts of simple harmonic motion are very important in our descriptions of nature.

Introducing the Mathematics of Simple Harmonic Motion

As is the case with circular motion, the mathematics of simple harmonic motion can appear abstract, and many pupils can find it difficult to relate the mathematics to the physical reality. It is important for the teacher to ensure that this relationship is not lost.

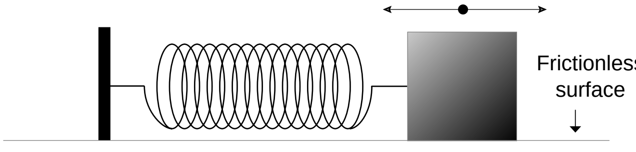


Fig. 2.4

Consider the case of the spring and mass oscillator, Fig. 2.4. It is obvious that the mass must be acted on by a special kind of force in order to vibrate. Pupils will understand that when it is in the equilibrium position there is no resultant force acting on it. After the body passes through the equilibrium position going left, the spring exerts a force on it pushing it to the right. This force gives it an acceleration to the right, first slowing down the body until it stops, and then speeding it up as it returns to the equilibrium position. After it passes through the equilibrium position again, this time going right, the force acting on it is towards the left. Hence the force is always directed towards the equilibrium point of the motion.

Hooke's law states that the magnitude of the restoring force is proportional to the displacement from the equilibrium position, i.e.

$$|F| \propto |s|$$

Taking into account the fact that when the mass is to the left of the equilibrium position the force acts towards the right, and vice versa, i.e. force and displacement are in opposite directions, this can be written as

$$F \propto -s$$

$$\Rightarrow \qquad ma \propto -s$$

If the mass at the end of the spring remains constant, then

$$a \propto -s$$

This relationship defines simple harmonic motion.

Definition

The motion of a particle is simple harmonic if its acceleration towards a fixed point is proportional to its displacement from that point.

The definition may be expressed in the form

$$a = -ks$$

where k is a positive constant of proportionality.

In the case of simple harmonic motion, the constant k is usually written as ω^2. So the definition becomes

$$a = -\omega^2 s$$

Note

In this context, ω^2 represents a positive constant. The square root of the constant, i.e. ω, is known as the angular frequency. The unit of ω is s^{-1} (from the defining equation) and it is inversely proportional to the periodic time. In fact, it can be shown (see below) that the periodic time of the motion is given by

$$T = \frac{2\pi}{\omega}$$

Period of a Particle in Simple Harmonic Motion

The derivation using calculus is included here, as the derivation without using calculus reverts to equating ω with angular speed, which can be confusing. The non-calculus derivation can be found in Appendix A.

The equation which defines simple harmonic motion,

$$a = -\omega^2 s$$

can be written in calculus notation as

$$\frac{d^2 s}{dt^2} = -\omega^2 s$$

which is a differential equation. One solution of this equation is

$$s = A \sin \omega t$$

That this is a solution may be verified by differentiating twice and substituting back into the original equation.

Thus, for a particle executing simple harmonic motion, the displacement varies sinusoidally with time. A is a constant which represents the maximum value of s (the maximum value of a sine function is 1), i.e. A is the amplitude of the motion. The periodic time of this sine function is

$$T = \frac{2\pi}{\omega}$$

Since the maximum value of s is A it follows from the defining equation that the maximum value of the acceleration is $\omega^2 A$, i.e.

$$a_{max} = \omega^2 A$$

Velocity of a Particle in Simple Harmonic Motion

Velocity is defined as the rate of change of displacement. In calculus notation this is

$$v = \frac{ds}{dt}$$

From the previous section the displacement of a particle in simple harmonic motion is

$$s = A \sin \omega t$$

Thus, the velocity is

$$v = \frac{ds}{dt}$$

$$= A\omega \cos \omega t$$

Since the maximum value of a cosine function is 1 the maximum velocity is

$$v_{\text{max}} = A\omega$$

and occurs when $t = 0$, i.e. when $s = 0$.

Period of a Spring and Mass Oscillator

The spring and mass oscillator system, Fig. 2.4 (p. 32), is subject to a force that obeys Hooke's law, i.e.

$$F \propto -s$$

$$\text{or} \qquad F = -ks$$

where s is the displacement from the equilibrium position, F is the resultant force and k is a constant, called the force constant, whose value depends only on the characteristics of the spring. Since $F = ma$ it follows that

$$a = -\frac{k}{m}s$$

Since k and m are constant for a particular spring and mass oscillating system, then this equation means that the motion of such a system is simple harmonic, where

$$\omega^2 = \frac{k}{m}$$

$$\omega = \sqrt{\frac{k}{m}}$$

Hence the periodic time, T, for the oscillator is

$$T = \frac{2\pi}{\omega}$$

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

The Motion of a Simple Pendulum

In practice, a simple pendulum consists of a small, heavy bob at the end of a light, inextensible string attached to a fixed point. Fig. 2.5 shows the forces acting on the bob of a pendulum. From the diagram, it can be seen that one component of the weight of the bob, $mg \cos\theta$, is in the opposite direction to the tension in the string, T.

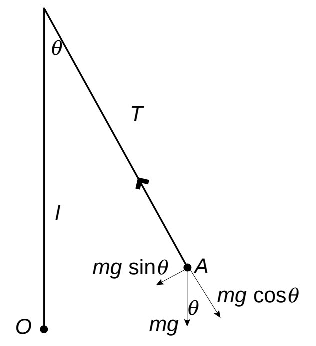


Fig. 2.5 *Simple pendulum*

The resultant of these two forces is the centripetal force. The other component of the weight, $mg \sin\theta$, is the restoring force, i.e. it tends to return the bob to the equilibrium position. The magnitude of the acceleration produced by this force is

$$|a| = g \sin\theta$$

For small values of θ, $\sin\theta \approx \theta$ when θ is measured in radians. So,

$$|a| = g\theta \qquad \text{when } \theta \text{ is small.}$$

But, since the bob is travelling in a circle of radius l

$$\theta = \frac{|\text{arc OA}|}{l}$$

$$= \frac{|s|}{l} \qquad \text{when } \theta \text{ is small}$$

where s is the displacement of A from O. Thus,

$$|a| = \frac{g}{l}|s|$$

Since the force is approximately towards the fixed point, O, while the displacement is from O this becomes

$$a = -\frac{g}{l}s$$

Since $\frac{g}{l}$ is a constant for any simple pendulum of length l, this equation means that the simple pendulum is executing simple harmonic motion with

$$a = -\frac{g}{l}s$$

The period, T, of the pendulum is given by

$$T = \frac{2\pi}{\omega}$$

$$= 2\pi\sqrt{\frac{l}{g}} \qquad (i)$$

This means that the period of the pendulum depends on the length of the string, but is independent of the mass of the bob or the angular displacement provided this does not go above approximately 5°. (It is a useful exercise for pupils to use a calculator to find sinθ, tanθ and θ in radians for values of θ between 1° and 10°). It shows also that the period of the swing is dependent on g, the acceleration due to gravity. This means that if the pendulum were used on the moon, where the acceleration due to gravity is smaller than on earth, the period of the pendulum would be greater.

Squaring both sides of equation (i), we get:

$$T^2 = \frac{4\pi^2}{g}l$$

This is the equation used in the experiment to measure the acceleration due to gravity using a simple pendulum. The variables are T^2 and l, and if a graph is drawn of T^2 against l, the result should be a straight line through the origin. The slope, m, of the line is given by

$$m = \frac{4\pi^2}{g}$$

$$\Rightarrow \quad g = \frac{4\pi^2}{m}$$

2.4 Student Experiment

Approximately simple harmonic motion occurs all around us. For example, a bottle bobbing in water, a cork being displaced by ripples, a marble rolling to and fro in a circular hollow, a vibrating guitar string, a mass vibrating at the end of a spring, a child's swing and the simple pendulum all exhibit approximately simple harmonic motion. The difference between simple harmonic motion and other periodic motions is well demonstrated using the cathode ray oscilloscope to show in slow motion examples of sinusoidal (which is simple harmonic), sawtooth and square waves.

Experiment 2.1

To Measure the Acceleration Due to Gravity using a Simple Pendulum

Apparatus

Pendulum bob, thread , split cork, retort stand and clamp, stop clock, metre stick.

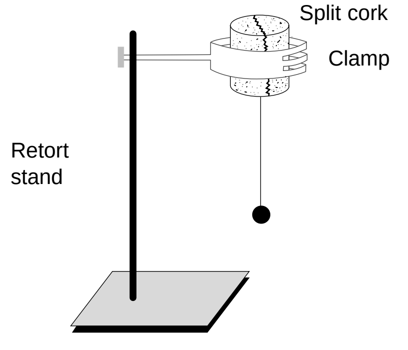


Fig. 2.6

Set up the apparatus as in Fig. 2.6. The slot in the split cork should be vertical so that there is a definite point of suspension. The length, l, of the pendulum is measured from this point to the centre of the bob. Set the length to 1.000 m. Set the pendulum swinging in a small arc (not more than 10°), and measure the time for 30 complete oscillations. Reduce the length of the pendulum to 0.900 m, and again measure the time for 30 complete oscillations. Continue this process using progressively shorter lengths, down to about 0.300 m (after this it becomes more difficult to time accurately).

l/m	Time for 30 oscillations	T/s	T^2/s^2
1.0			
0.9			
0.8			
0.7			

For each set of readings, calculate the time, T, for one swing, and get the square of this time. Tabulate your results as shown. Draw a graph of T^2 against l. The value for g, the acceleration due to gravity, is found from the equation:

$$g = \frac{4\pi^2}{m}*$$

where m is the slope of the graph.

**For derivation of this formula, see p. 35.*

Note

The above formula applies only when the angle of swing is small. In addition, if the swing becomes elliptical, i.e. the pendulum becomes a conical pendulum, the formula no longer holds, and the reading should be taken again.

2.5 Worked Examples

1. A particle in simple harmonic motion has a period of 3.0 s. What is its acceleration when its displacement is 20 cm from its equilibrium position?

$$\omega = \frac{2\pi}{T}$$

$$= \frac{2\pi}{3}$$

$$= 2.09$$

$$a = -\omega^2 s$$

$$= -(2.09)^2 \times 0.2$$

$$= -0.87 \text{ m s}^{-2}$$

The negative sign indicates that the direction of the acceleration is opposite to that of the displacement.

2. What is the period of a pendulum suspended from the top of a tall building on a light string 353 m long?

$$T = 2\pi\sqrt{\frac{l}{g}}$$

$$= 2\pi\sqrt{\frac{353}{9.8}}$$

$$= 37.7 \text{ s}$$

3. What is the period of a simple pendulum on the surface of the moon if its period on earth is 2.0 s, given that the acceleration due to gravity on the moon is one sixth that on earth?

In calculations of this kind, where comparisons are being made, simple formulae can often be derived to make the calculation easier. In this case,

$$T_M = 2\pi\sqrt{\frac{l}{g_M}}$$

$$T_E = 2\pi\sqrt{\frac{l}{g_E}}$$

Dividing, we get

$$\frac{T_M}{T_E} = \sqrt{\frac{g_E}{g_M}}$$

Substituting, we get

$$\frac{T_M}{2} = \sqrt{6}$$

$$T_M = 4.9 \text{ s}$$

4. What would the period of the pendulum in the previous example be if it were at a height above the earth equal to the radius of the earth?

From above,

$$\frac{T_H}{T_E} = \sqrt{\frac{g_E}{g_H}} \quad \text{(i)}$$

where T_H is the period at the required height, and g_H is the acceleration due to gravity at that position. We can find another comparison formula from the equation for g (p. 16) as follows.

$$g_E = \frac{GM_E}{r_E^2}$$

$$g_H = \frac{GM_E}{r_H^2}$$

Dividing, we get

$$\frac{g_E}{g_H} = \left(\frac{r_H}{r_E}\right)^2$$
$$= 4$$

Substituting in equation (i) we get

$$\frac{T_H}{T_E} = \sqrt{4}$$

$$T_H = 2\sqrt{4}$$
$$= 4 \text{ s}$$

So, the period of the pendulum at this height is twice what it is at the surface of the earth.

This result may also be obtained by reasoning as follows.

T is inversely proportional to the square root of g

g is inversely proportional to r squared

Therefore,

T is proportional to r

Thus, since r (the distance from the centre of the earth) is doubled, T is also doubled.

5. A load of mass 500 g is hanging vertically from a light spring for which the force constant is 10 N m^{-1}, Fig. 2.7. If the load is pulled down 15 cm from the equilibrium position and released, how long does it take to reach the equilibrium position again?

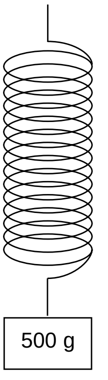


Fig. 2.7

$$\omega = \sqrt{\frac{k}{m}}$$

where k is the force constant and m is the mass of the load.

$$T = \frac{2\pi}{\omega}$$
$$= 2\pi\sqrt{\frac{m}{k}}$$
$$= 2\pi\sqrt{\frac{0.5}{10}}$$
$$= 1.4 \text{ s}$$

The period of the oscillation is 1.4 s. So the time taken to return to the equilibrium position from the maximum displacement is one quarter of a period, i.e. 0.35 s.

CHAPTER 3 WORK AND ENERGY

3.1 Background

The development of human civilisation has depended on the ability of mankind to find sources of energy and to harness the energy to assist in human endeavour. This process began when man first used fire to keep warm and to cook food. For this, he used mainly wood as fuel. He also developed another source of energy by using animals to pull ploughs and as a means of transport. The first energy saving measures began with the use of animal skins as clothing.

Charcoal fires were used in Egypt about 4000 BC to extract copper from mineral ores. From about 3500 BC copper was mixed with tin to form bronze and, from about 1000 BC, iron came into general use. These charcoal fires may represent the earliest harnessing of energy for 'industrial' use. The use of coal for heating purposes may have begun as early as 1000 BC in China, and **Theophrastus** (c.372 BC - c.287 BC) mentions its use by metal workers in Greece and Italy about 350 BC.

Sailing ships, using wind energy, were depicted in Egyptian rock carvings from 3300 BC, and the first windmills are thought to have been in Persia in the 7th century. Waterwheels were used to irrigate Babylonian fields over 3000 years ago. Both windmills and watermills were used for grinding corn and providing energy to drive other machinery of the time - for example, the Romans had water-powered sawmills.

Though **Hero of Alexandria** (c. 20 - ?) built a steam turbine in 100 AD, the development of steam as an alternative to wind and water as a form of power for industry began in the seventeenth century. In 1698 **Thomas Savery** (1650-1715) developed a steam pump. This was improved upon by **Thomas Newcomen** (1663-1729) who produced the first piston-operated steam engine. However, since Savery had the patent, the two went into business together producing pumps to remove water from coal mines. The engine produced in 1765 by **James Watt** (1736-1819) was much more powerful, and steam engines led to the huge growth in industrial activity and transport known as the industrial revolution. This in turn led to greatly increased demand for energy, mostly from coal.

Towards the end of the 18th century, **William Murdock** (1754-1839) produced gas from coal, and used it to light his office. The first gas street lights appeared in Soho, London in 1802. Around this time also, electricity was being developed as a source of power, largely due to the work of **Michael Faraday** (1791-1867) and **Joseph Priestley** (1733-1804). By the late 19th century, it began to be used for lighting and electric railways as well as for industrial uses and capital punishment ('the electric chair')! The energy required for generating the electricity came initially from coal. The first small hydroelectric station was built in Wisconsin in 1882, and ten years later water from Niagara Falls was used as the source of energy to provide electrical power for the city of Buffalo. (See module on electromagnetism.)

The emergence of oil as an important source of energy is dated to 1859 when **Colonel Edward Drake** drilled a well for oil at Titusville, Pennsylvania. Petroleum in the form of 'rock oil' had been known since ancient times. At various places around the world, petroleum materials occur at ground level in the form of oily liquids mixed with the soil, escapes of gas and deposits of asphalt, bitumen, tar sands and oil shale. These were used by ancient man in a variety of ways unconnected with energy. The Bible refers to the use of 'slime' for mortar by the builders of the Tower of Babel. This was asphalt, while the pitch used in the building of Noah's Ark was probably bitumen. **Nebuchadnezzar** (d. 562 BC), king of the Chaldean empire, had a road built from bricks laid in asphalt, and the ancient Sumerians used asphalt as cement for their mosaics. The Chinese burned natural gas to obtain salt from brine and by about 99 AD were transporting it through bamboo pipelines for lighting and heating.

Colonel Drake's well is generally seen as the beginning of the modern oil industry. The product most sought after from the early oil wells was paraffin, and other products were either burned off as waste, or simply allowed to run back into the ground. However, the development of the internal combustion engine in the 1880s changed all that, and petrol and diesel oil became the most important sources of energy for transport in the world. Oil now supplies 46% of the world's energy.

Nuclear energy technology, based on the nuclear fission process, has been developed this century, but it has proved expensive and has a number of hazards associated with it, not least of which is the nuclear waste produced in the reactors. Nevertheless, many countries use nuclear power stations to generate electricity, e.g. in France, some 70% of the electricity generated is obtained from nuclear power stations. Compared to nuclear fission, nuclear fusion offers the prospect of huge power output without many of the associated difficulties. However, while controlled nuclear fusion has been achieved for short periods, the power input has always greatly exceeded the power output.

Fig. 3.1 shows the current sources of the world's energy.

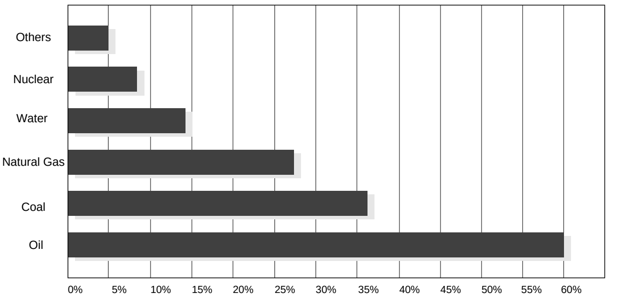


Fig. 3.1 *Percentage of world's energy usage obtained from different sources*

It is estimated that at current levels of usage, there is sufficient coal to last for about 300 years but the world's gas and oil supplies will run out around the middle of the 21st century. Hence there is a growing urgency in seeking new sources of energy, preferably renewable ones, and in promoting energy conservation.

Many of the world's pollution problems result from man's ever-increasing demand for energy. The pollution of the air by motor vehicles and industry, the fear of 'another Chernobyl' (where a very serious accident occurred in a nuclear power station in 1986), the problem of nuclear reprocessing, and the imminent shortage of the world's reserves of fossil fuels are all symptoms of our desire for more sources of cheap energy.

3.2 Do You Know

What is meant by the horsepower of a car?

The horsepower is a unit of power established in the late eighteenth century by James Watt, and is still used today, although less commonly than in the past. It is an arbitrary estimate for the power of a typical horse over a short period, and is equivalent to 746 W. An average man can work at rates between about $^1/_8$ h.p. and $^3/_4$ h.p., or approximately 100 - 600 W.

What is the kilowatt-hour?

The ESB charges consumers on the basis of the amount of electrical energy they use, i.e. convert from electrical energy to other forms. The 'unit' of electrical energy referred to in the ESB bill is the kilowatt-hour (kW h). 1 kW h is the amount of electrical energy used when an appliance rated 1 kW is used for 1 hour. So the amount of energy used is

Energy = power x time

1 kW h = 1000 x 3600 J

$= 3.6 \times 10^6$ J

Note

Because the word ‘watt’ appears in the term kilowatt-hour, pupils sometimes think of it as a unit of power. It is, in fact, a unit of energy.

One ‘unit’ (kilowatt-hour) of electricity could, typically:

- boil about 7 litres of water in an electric kettle;
- work a refrigerator for a day;
- run a two bar electric heater for half an hour;
- make 70 slices of toast in a toaster;
- light a 60 W bulb for 16.5 hours;
- give 1600 shaves with an electric shaver.

What is a calorie?

Like the horsepower, the calorie is a non-standard unit which is still in common usage. It is a unit of energy, and is defined as the amount of energy required to change the temperature of 1 g of water by 1 ^{o}C. One calorie is approximately equivalent to 4.2 joules.

How much energy does a person need?

Our bodies convert the food we eat into energy. An average man uses about 12 000 kJ of energy, and a woman 9000 kJ, per day.

How is energy conserved in satellite (and planetary) orbits?

One of the implications of Kepler’s second law (see p. 4) is that satellites in elliptical orbits travel more slowly the further they are from the earth. This is to conserve mechanical energy - the greater the distance from the earth, the greater the potential energy, and this is balanced by a reduction in kinetic energy. The same applies to planetary orbits about the sun.

The fact that satellites move more slowly at their apogee (furthest point) means that satellites in very elliptical Molniya orbits can be used for communications purposes (see p. 25).

What is the relationship between heating and temperature?

The heating effect of energy is detected by a rise in temperature or a change of state. So, what is the heating effect of 1 J of work? For example, if one joule of work were done on 1 g of water, what change in temperature would result?

Heat = mass x specific heat capacity x rise in temperature

1 J = 0.001 kg x 4180 J kg^{-1} K^{-1} x $\Delta\theta$

$\Delta\theta$ = 0.24 K

What is meant by the equivalence of mass and energy?

Einstein had shown in his theory of special relativity (1905) that the mass of a body increases as its speed increases according to the equation

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

where m is the mass of the body as judged by the observer, m_o the rest mass, i.e. its mass when at rest relative to the observer, v its velocity relative to the observer and c the speed of light in vacuum. Einstein reasoned that when a body is given kinetic energy, its mass must increase. He concluded that if an *increase* of mass corresponded to a *gain* in energy, the mass of a body even at rest must be equivalent to some energy.

The equivalence of mass and energy is basic to our understanding of nature. Apart from the release of energy in nuclear reactors and bombs, mass is transformed into radiant energy continuously in the sun and stars. In the sun, 5 x 10^{13} tonnes of hydrogen are converted into helium every second. However, the mass of helium produced is slightly less than that of the hydrogen, so that the mass of the sun is decreasing by 4 million tonnes every second. This mass is converted into radiant energy. The sun has been emitting radiant energy and losing mass at this rate for several billion years.

What is a perpetual motion machine?

Friction-free devices (e.g. atomic systems, superconducting systems, etc.) offer the possibility of providing 100% efficiency. The United States Patent Office has thousands of applications for patents on devices which are claimed to be perpetual motion machines, but it has a standing rule that perpetual motion patent applications must be accompanied by a working model! No one has yet presented a successful model of such a device.

What is the difference between elastic and inelastic collisions?

An elastic collision is one in which kinetic energy is conserved. At a molecular level, for example, it is an assumption of the kinetic theory of gases that collisions between the molecules themselves, and between molecules and the walls of the container are perfectly elastic. At a macroscopic level, collisions are never perfectly elastic. However, in some cases the amount of kinetic energy lost is so small as to be negligible, and these are said to be elastic. Examples include Newton's cradle and collisions between billiard balls.

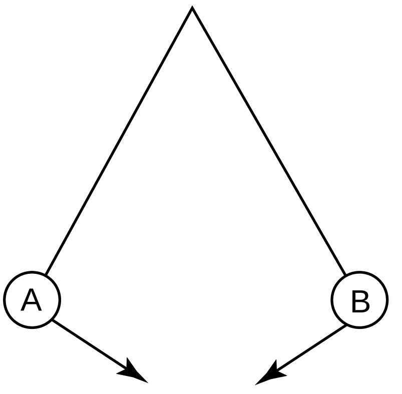


Fig. 3.2(a)

Fig. 3.2(a) shows an interesting example of a system where collisions are elastic. Two steel spheres, A and B, have masses of 0.3 kg and 0.1 kg, respectively. They are allowed to fall from the same height at the same instant, so that they have the same speed, say 1 m s^{-1} when they collide. Both momentum and energy are conserved in the collision, giving the equations

$$m_1u_1 + m_2u_2 = m_1v_1 + m_2v_2$$

$$\tfrac{1}{2}m_1u_1^2 + \tfrac{1}{2}m_2u_2^2 = \tfrac{1}{2}m_1v_1^2 + \tfrac{1}{2}m_2v_2^2$$

Solving these simultaneous equations gives the result that, after collision, A remains stationary while B bounces away with a speed of 2 m s^{-1}. (This is not a general result, but depends on the values of the masses. Another interesting case involves elastic collisions between bodies of equal mass, one of which is initially at rest, e.g. Newton's cradle.)

Momentum before collision:
$0.3 \times 1 + 0.1 \times (-1) = 0.2$ kg m s^{-1}

Momentum after collision:
$0.3 \times 0 + 0.1 \times 2 = 0.2$ kg m s^{-1}

Kinetic energy before collision:
$\tfrac{1}{2}(0.3)(1)^2 + \tfrac{1}{2}(0.1)(-1)^2 = 0.2$ J

Kinetic energy after collision:
$\tfrac{1}{2}(0.3)(0)^2 + \tfrac{1}{2}(0.1)(2)^2 = 0.2$ J

What happened to the two spheres on impact? At impact, the kinetic energy of both spheres was converted into elastic potential energy by the forces of impact. The molecules in the spheres were compressed closer together than normal. This elastic potential energy was very quickly transformed back into kinetic energy by equal and opposite forces which acted on the spheres as the compressed molecules returned to their normal positions.

Question: what happens when the smaller sphere returns and hits the (now stationary) larger sphere?

Answer: they move away from each other, each with a speed of 1 m s^{-1}.

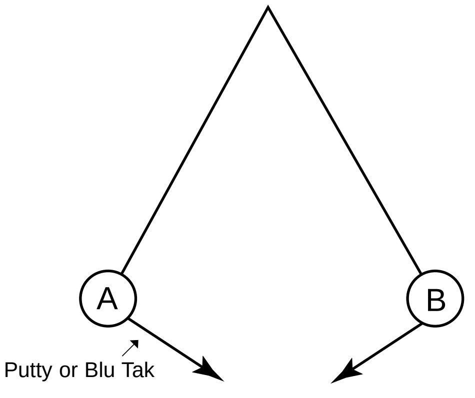


Fig. 3.2(b)

Consider, now, the situation in Fig. 3.2(b). In this case, when the spheres stick together after the collision, the collision is said to be perfectly inelastic. Momentum is conserved in this collision, and after collision, the spheres travel together with a speed of 0.5 m s^{-1}. However, the final kinetic energy in this case is

Kinetic energy after collision:

$$\tfrac{1}{2}(0.3)(0.5)^2 + \tfrac{1}{2}(0.1)(0.5)^2 = 0.05 \text{ J}$$

Loss of energy:

$$0.2 \text{ J} - 0.05 \text{ J} = 0.15 \text{ J}$$

In this case, since putty is a plastic rather than an elastic substance, it cannot develop elastic potential energy. The putty absorbs the kinetic energy by undergoing a distortion in its shape and a rise in temperature.

What is the greenhouse effect?

Apart from the small amount of heat which reaches the surface from the earth's centre, most of the warmth we experience comes from solar radiation. Ideally, over time, the earth radiates the same amount of energy into space as it absorbs from the sun. Certain gases, often referred to as 'greenhouse' gases, prevent some of this energy from escaping back into space, so keeping the earth warmer than it would be without these gases; this is somewhat analogous to the effect of the glass in a greenhouse. Increasing emission of these gases into the atmosphere may be expected to enhance the greenhouse effect and so cause the average global temperature to rise. The greenhouse gases include carbon dioxide, methane, water vapour, chlorofluorocarbons (CFCs) and nitrous oxide. Carbon dioxide resulting from the burning of fossil fuels accounts for two-thirds of the blocking of radiation of energy from the earth to space. These gases, in effect, act like an insulating blanket around the earth.

It is projected that the enhanced greenhouse warming will produce temperatures greater than at any time in the past 100 000 years. Even more importantly, the rate of temperature increase would be many times faster than past natural changes. The effects of global warming on such a scale are not predictable with any degree of certainty, but are likely to be very profound.

The increase in carbon dioxide levels in the atmosphere results from both the rate of burning of fossil fuels, and also from deforestation (plants absorb carbon dioxide in photosynthesis). The development of renewable energy sources, then, is becoming a matter of urgency and some people have suggested an international project to develop safe nuclear power.

What is meant by renewable sources of energy?

Most of the energy sources used by man are ultimately renewable but all are finite; they cannot be renewed indefinitely. Unfortunately, many of them such as coal, oil and natural gas will take millions of years to renew. Even some timber resources will take hundreds of years to renew. So the term renewable usually refers to sources of energy which are continually renewable, such as the sun's rays, wind, rainfall or tides, or those which can be replaced within a short time-span, such as biomass.

Only 2% of Ireland's total annual energy requirement of about 500 PJ is currently produced from renewable energy sources.

Solar energy

Most of the energy available on earth comes, or has come, from the sun. There are many ways in which this energy can be 'harvested'. These include fossil fuels, wind, waves, crops and biomass. However, the term solar energy generally applies only to energy from the sun which is 'harvested' directly using, for example, solar panels or solar cells.

Water energy

Hydro-electricity provides about 7% of the world's total energy demand. The use of water and gravity to produce electric power is very dependent on an adequate supply of rainfall or melting snow, and on a suitable landscape. Nearly all of Norway's electricity is generated in this way, over 50% in South America, about 20% in the United States but only about 2% in Ireland.

Wave energy

Waves on the sea possess very large amounts of energy. One method of 'harvesting' this energy is to use the motion of the waves rising and falling inside a vertical cylinder to push air through a valve and turn a turbine. Harnessing wave energy on a large scale, however, is proving very difficult.

Tidal energy

Tidal mills have been used in some European countries for hundreds of years. It is difficult, however, to find suitable sites to harness this energy for large-scale electricity generation. The only large-scale tidal power station in the world is on the Rance estuary in France. This produces up to 240 MW which, however, is less than 0.1% of France's total energy needs.

Wind energy

The power of the wind depends on the cube of the wind speed (v^3) so that even small changes in wind speed can cause large changes in power output.

In the eighteenth century, there were more than 20 000 windmills in use in the Netherlands, mostly used to pump water from low lying areas, thus reclaiming land from the sea. In 1975, it was estimated that there were 150 000, mostly small, windmills in use in the United States. In Ireland in early 1999 there were ten wind farms connected to the ESB network providing a total of c. 60 MW of installed capacity.

Geothermal energy

The temperature of the earth's core is about 2200 °C. Heat from the core seeps very slowly up through the earth's crust which is between 30 and 40 km deep in most places. At some places in this crust water trapped in porous rocks is gradually heated, resulting in a geothermal reservoir. Geysers and hot springs occur where cracks in the rocks allow this hot water to escape. This energy has been used for heating at least since Roman times when bath houses using hot springs were built. In modern times, geothermal energy is used in many parts of the world, most notably in Iceland where geothermal energy supplies most of the heat for the capital city, Reykjavik. In France, over 500 000 homes are heated by geothermal energy.

Geothermal energy was first used to generate electricity in Italy in 1904 when Prince Piero Conti used a steam engine driven by natural steam to drive a generator. Many countries now use geothermal energy to drive turbines. In most cases this involves drilling deep boreholes into the geothermal reservoir. The steam then rushes to the surface where it is used to generate electricity. The United States, New Zealand and Japan are just some of the countries that have large geothermal power stations feeding electricity into their national grids.

Although geothermal energy is not renewable in the sense described above, it is essentially a constant source of energy.

Bio-fuels

Fuels produced from quick-rotation crops or other biological sources (such as animal manure, waste cooking oil, etc.) are called 'bio-fuels'. Wood was the earliest and most common bio-fuel, but the energy needs in the world today are so great that wood cannot be grown quickly enough for fuel purposes, resulting in wide-scale deforestation. Many efforts are being made to find suitable biomass products. One of the problems is that many biomass products contain too much moisture to burn well. However, they can be converted into liquid or gas bio-fuels such as methane and ethanol. In Brazil, for instance, cars have been adapted to run on ethanol produced from specially grown sugar cane, and this now provides a large percentage of transport fuel.

In Ireland, less than 1% of energy demand is produced from biomass (it is 3% in the EU as a whole), but under the EU ALTENER programme, the amount of energy from renewable sources in Ireland is targeted to rise to between 7% and 10% by 2005, and biomass crops and residues are seen as the major components of this expansion. One major electricity generating plant (10-30 MW) is planned.

3.3 Conceptual Approach

Work

Work is one of those words which is used in everyday conversation, but which has a much more precise meaning when used in physics. In this context, 'work' is said to be done 'when a force moves a body'. So there are many aspects of work as understood in everyday conversation that would not count as work in physics - including studying physics!

Where the direction of the force is the same as the direction in which the object moves, then the work done is defined as

$$W = Fs$$

where W is the work done, F is the magnitude of the force acting on the body, and s is the magnitude of the displacement of the body. If the force does not have the same line of action as the resulting displacement, as in Fig. 3.3, then the work done is defined as the product of the component of the force in the direction of the motion and the distance through which it acts, i.e.

$$W = Fs \cos \theta$$

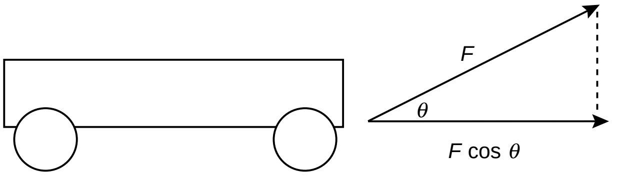


Fig. 3.3

Mathematically, work is the scalar (dot) product of force and displacement (see module on Mechanics I). Work is a scalar quantity and its unit is the joule. One joule is the work done when a force of one newton moves an object a distance of one metre in the same direction as the force; 1 J = 1 N m.

To examine the significance of the term cos θ, consider how much work is done by the force of gravity in keeping a satellite travelling in a circular orbit around the earth. In this case, the force always acts perpendicularly to the direction of travel, i.e. $\theta = 90°$, and since cos 90° = 0, the work done is also zero. Another example of the same principle is the conical pendulum, Fig. 3.4.

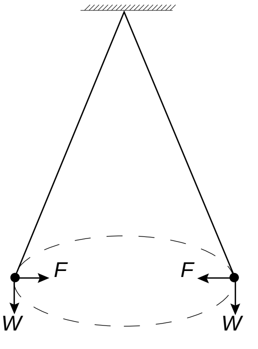


Fig. 3.4 *Conical pendulum*

Fig. 3.5 shows three paths to raise an object a height of 1 m. The work done is the same regardless of the path chosen (assuming that gravity is the only force acting on the object) because, even though the distance travelled along the path of least incline is greatest, in this case the component of the force in the vertical direction is least.

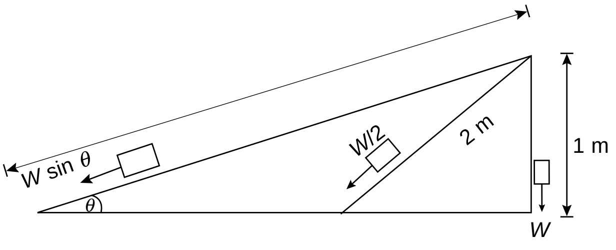


Fig. 3.5 *Work done is independent of path followed*

Energy

Energy is defined as the ability to do work. For example, a coiled spring can make the hands of a clock move, petrol can produce motion in the pistons in a car’s engine, a moving lorry can demolish a wall, water after falling from the top of a waterfall can cause turbines to rotate, and a battery can cause an ammeter needle to move when connected in a circuit. In other words, the coiled spring, the petrol, the moving lorry, the water near the bottom of the waterfall, and the battery can all cause work to be done. We say that each has energy. Like work, energy is measured in joules, and is a scalar quantity.

While there is only one energy, it is convenient to consider the energy involved in each of the above cases as being different forms of energy. In the case of the spring, the energy which is stored by virtue of the elasticity of the metal when it is coiled tightly is known as potential energy. The petrol has chemical energy which can be released when mixed with air and ignited. The lorry has kinetic energy by virtue of its movement - a stationary lorry will not damage the wall! The water near the bottom of the waterfall has kinetic energy since it is falling under gravity. The energy stored in the battery is, again, chemical energy which can be released as electrical energy. While these are referred to as different forms of energy it is important that pupils understand that the energy is identical in every case - the ability to do work.

This section of the physics course deals mainly with mechanical energy, i.e. kinetic energy and gravitational potential energy. (In some texts, all energy that is not kinetic is classed as potential.)

Kinetic Energy

The energy which a body has because of its motion is called kinetic energy. The amount of the kinetic energy of a body of mass m travelling with velocity v can be calculated by determining the amount of work done in accelerating it from rest to this speed. This is given by the equation

$$W = Fs$$

But $$F = ma$$

$$\Rightarrow \quad W = mas$$

and since, $$v^2 - u^2 = 2as$$

$$W = m.{}^1/_2(v^2 - u^2)$$

$$= {}^1/_2 mv^2 \qquad \text{since } u = 0$$

Note

This equation has been derived in the case of a body whose acceleration was uniform. It can be shown that it applies in other cases also.

Gravitational Potential Energy

If an object of mass m is lifted from ground level to a height h above the ground, then the potential energy it has gained is equal to the work done in lifting it to this height. This work is

$$W = Fh$$

But $$F = mg$$

$$\Rightarrow \quad W = mgh$$

This is the equation for gravitational potential energy based on the assumption that the potential energy at ground level is zero.

The use of the description of this equation as the ‘potential energy equation’ can be confusing for pupils because there are other forms of potential energy (e.g. in a wound up spring) which are not explained by this equation. To avoid confusion, it is advisable to use the term 'gravitational potential energy' when using this equation.

Law of Conservation of Energy

The concept that energy can neither be created nor destroyed represents one of the great generalisations of the nineteenth century, and was the result of the work of many scientists, in particular the English physicist **James Joule** (1818 - 1889). Before his time, it had been known that various forms of energy could be converted to heat and vice versa, but the quantitative equivalence of heat, or thermal energy, and other forms of energy had yet to be discovered. Joule showed experimentally that a given amount of heat is always equivalent to the same amount of energy. This law was understood in the nineteenth century to mean that 'the total amount of mechanical, thermal, chemical, electrical and other energy in an isolated system remains constant'. In 1905, Albert Einstein broadened the law still further to include the equivalence of mass and energy according to the equation

$$E = mc^2$$

An isolated system is one that does not have any interactions with objects outside the system, and the universe is, by definition, an isolated system. By extrapolation, the law of conservation of energy is assumed to apply to the entire universe. Hence, a modern statement of the law of conservation of energy would be: 'the total amount of energy in the universe is constant'.

Heat as Internal Energy

Which has more heat, a bucket full of water at 0 °C or a thimble full of water at 0 °C? To some pupils this might seem like a trick question - neither (they might say) has any heat. This is a good starting point for explaining the nature of heat as 'internal energy' and the distinction between heat and temperature.

The molecules in the water are moving and so have energy. As the temperature of the water is lowered, the molecules move more slowly (their energy decreases) until they stop moving completely. This occurs at absolute zero, 0 K or –273 °C. At all temperatures above this, the molecules of the water (or any other substance) have energy. This is called internal energy. Obviously, then, the bucket of water contains more moving molecules, and so has more internal energy than a thimble full at the same temperature. (See module on Heat and Temperature.)

Power

Power is defined as the rate at which work is done. It can also be defined as rate at which energy is converted from one form into another.

The unit of power is the watt. From the definition it follows that one watt is equivalent to one joule per second; $1 \text{ W} = 1 \text{ J s}^{-1}$.

The unit of power, the watt, is recognised, if not precisely understood in its physics meaning, in everyday conversation in relation to electric light bulbs. For example, a 100 W bulb converts 100 J of electrical energy into heat and radiant energy every second.

Note

In teaching this section of the course, it is interesting to read and interpret the power ratings from the back of electrical appliances.

Efficiency

The efficiency of a device, machine or engine is defined as the ratio of power output to power input. It is usually expressed as a percentage.

$$\text{Efficiency (\%)} = \frac{\text{useful power output}}{\text{power input}} \times 100$$

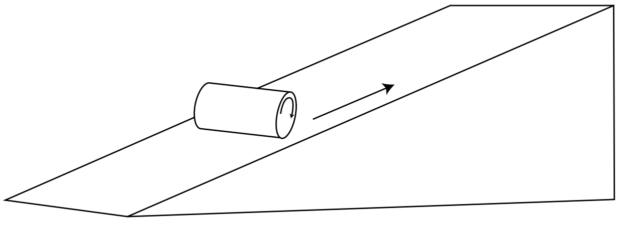


Fig. 3.6 *A barrel being rolled up an inclined plane*

For example, if a barrel is rolled up an inclined plane as shown in Fig. 3.6, its potential energy at the top will be less than the amount of energy required to roll it up the plane. The 'lost' energy appears as heat caused by the friction between the barrel and the plane which causes some energy to be changed into heat. This is an example of a situation where the efficiency of the device (i.e. the inclined plane) is less than 100%. In fact, any device changing kinetic to potential energy, or vice versa, is subject to frictional heating losses and has an efficiency of less than 100%. The more machinery that must be used to convert one form of energy into another, the more inefficient the process will be. For instance, the usual petrol engine has an efficiency of only about 20%.

One type of energy transfer which may be 100% efficient is the change of mechanical (or chemical or electrical) energy into heat. However, the reverse process can never be 100% efficient. The situation is analogous to pouring water from a bottle on to a floor (with 100% efficiency), and then trying to get it all back into the bottle.

If machines are so inefficient, why use them at all? Even simple machines, e.g. levers, pulleys, inclined planes, etc., yield less work that is put into them. We use them because other considerations are important. For example, a human can lift a load using a pulley or a lever which might be too heavy to lift otherwise. What the machine does, in fact, is to allow us to choose our relative values of F and s in performing work.

3.4 Worked Examples

1. What is the energy equivalent of 1 kg of mass?

   From Einstein's equation

$$E = mc^2$$

$$E = 9 \times 10^{16}\ \text{J}$$

2. A spacecraft of mass 4 tonnes travels with speed 3000 m s^{-1}. Calculate (a) its increase in mass compared with its rest mass, (b) its kinetic energy.

(a) The new mass is given by

$$m = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}}$$

$$m = \frac{4000}{\sqrt{1 - \frac{(3 \times 10^3)^2}{(3 \times 10^8)^2}}}$$

$$= 4000.0000002\ \text{kg}$$

Hence, the increase in mass is 0.0000002 kg, or 0.2 mg.

(b) The kinetic energy of the spacecraft is:

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

$$= \tfrac{1}{2} \times 4 \times 10^3 \times (3 \times 10^3)^2$$

$$= 1.8 \times 10^{10}\ \text{J}$$

Notes

Using the mass-energy conversion formula,

$$E = mc^2$$

1.8×10^{10} J of kinetic energy is the equivalent of a change in mass of 0.2 mg.

This example shows that the increase in mass in most cases is negligible. It is only when the speed of the body approaches that of light that the mass changes appreciably. It is because of this that it is not necessary to consider the effects of relativity in most cases involving energy changes, i.e. it is only necessary to consider mechanical, thermal, chemical or electrical energies, etc.

3. Calculate (a) the energy used, (b) the average power developed, by a man whose mass is 70 kg and who runs up stairs from the ground floor to the first floor of a building in 8 s, if there is a vertical height of 3 m between the floors.

(a) The energy used is equal to the increase in the man's gravitational potential energy.

$$E = mgh$$

$$= 2058 \text{ J}$$

(b) Average power used is

$$P = \frac{\Delta E}{t}$$

$$= \frac{2058}{8}$$

$$= 257 \text{ W}$$

4. A hoist raises an elevator with a total weight of 5000 N through a height of 30 m in 20 seconds. What is the average power developed by the motor?

$$P = \frac{W}{t}$$

$$= \frac{Fs}{t}$$

$$= \frac{5000 \times 30}{20}$$

$$= 7500 \text{ W}$$

$$\approx 10 \text{ h.p.}$$

5. A bullet of mass 40 g, travelling at a speed of 500 m s^{-1}, strikes and becomes embedded in a timber block of mass 12 kg which is at rest. If the block hangs by two threads as shown, and can swing freely, what is the greatest height to which it will rise?

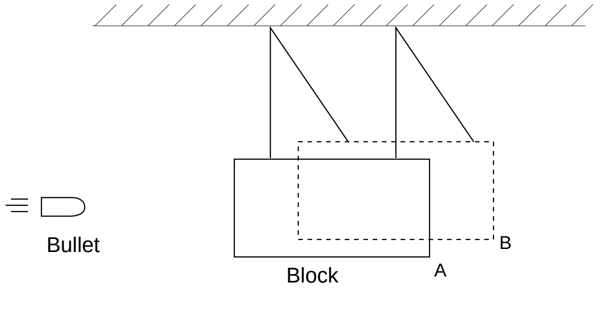


Fig. 3.7 *A bullet fired into a suspended block*

When the bullet strikes the block at A, Fig. 3.7, momentum is conserved. Thus, if u is the speed of the bullet before it strikes the block and v is the initial speed of the block and bullet after impact

$$mu = (M + m)v$$

$$0.04 \times 500 = 12.04v$$

$$v = 1.7 \text{ m s}^{-1}$$

If the potential energy of the block and bullet at A is taken as zero then their kinetic energy at A is equal to their potential energy at B, the highest point reached. So,

$$\tfrac{1}{2}(M + m)v^2 = (M + m)gh$$

$$\Rightarrow \quad \tfrac{1}{2} \times 12.04 \times (1.7)^2 = 12.04 \times 9.8 \times h$$

$$\Rightarrow \quad h = 0.15 \text{ m}$$

The latter part of this example, and Example 6, could be addressed using the equations of motion, but is much easier to solve using the concept of conservation of mechanical energy.

6. A pendulum has a length of 2.0 m and swings through an arc of 20°. Calculate the maximum speed of the bob.

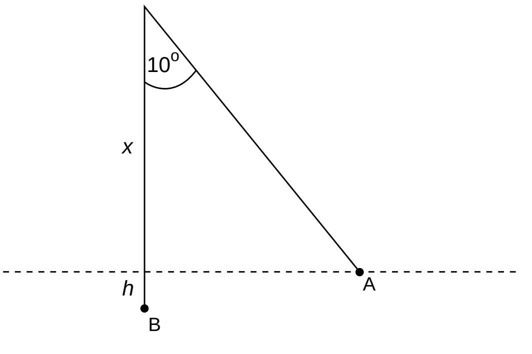


Fig. 3.8

The maximum speed occurs when the bob is at the central position, B, (Fig. 3.8). At this point, all its energy is kinetic, and this kinetic energy is equal to the potential energy it had at A.

$$E_k \text{ at B} = E_p \text{ at A}$$

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

$$\Rightarrow \quad v^2 = 2gh$$

$$v = \sqrt{2gh} \qquad \text{(i)}$$

From Fig. 3.8,

$$h = 2 - x$$

$$= 2 - 2\cos 10^0$$

$$\Rightarrow \quad h = 0.0304 \text{ m}$$

Substituting into (i) gives a value for v of

$$v = 0.77 \text{ m s}^{-1}$$

Note

Many problems in mechanics are much more easily solved using the equivalence of work and energy than by using the equations of motion. The following are some examples

7. A stone is dropped from a height of 10 m. What is its speed just before hitting the ground?

   Potential energy at point of release = kinetic energy just before hitting ground

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

$$v = \sqrt{2gh}$$

$$= 14 \text{ m s}^{-1}$$

8. A car of mass 800 kg travelling at 30 m s^{-1} is brought to rest in a distance of 40 m. What is the average force applied by the brakes?

   Kinetic energy of the car = work done by brakes in bringing it to rest

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

$$\tfrac{1}{2} \times 800 \times 30^2 = F \times 40$$

$$F = 9000 \text{ N}$$

9. Fig. 3.9 shows the path of a roller coaster travelling from A to D. If it is given an initial speed of 1.0 m s^{-1} at A, (a) what is its speed at B and at D, (b) what is the maximum radius of the loop in order that the passengers will not fall out of the car at C? (Ignore the effects of friction.)

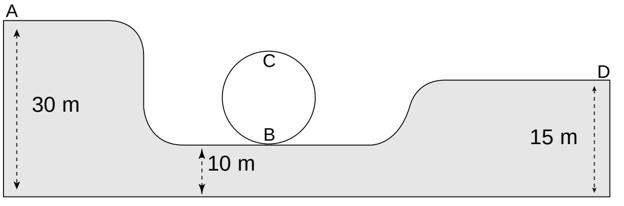


Fig. 3.9

(a) If we take the ground as zero potential energy, then the car has potential and kinetic energy at A, less potential and more kinetic energy at B, etc. The total mechanical energy is the same at A, B and C.

To find speed at B:

$$(E_p)_A + (E_k)_A = (E_p)_B + (E_k)_B$$

$$mgh_A + \tfrac{1}{2}mv^2_A = mgh_B + \tfrac{1}{2}mv^2_B$$

$$gh_A + \tfrac{1}{2}v^2_A = gh_B + \tfrac{1}{2}v^2_B$$

$$9.8 \times 30 + \tfrac{1}{2} \times 1^2 = 9.8 \times 10 + \tfrac{1}{2} \times v^2_B$$

$$v_B = 20 \text{ m s}^{-1}$$

Find speed at D similarly:

$$mgh_A + \tfrac{1}{2}mv^2_A = mgh_D + \tfrac{1}{2}mv^2_D$$

$$v_D = 17 \text{ m s}^{-1}$$

(b) In order for the passengers not to fall out of the car at the top of the loop, the speed at C must be

$$v_C = \sqrt{rg}$$ (see p.19)

So,

$$PE_A + KE_A = PE_C + KE_C$$

$$mgh_A + \tfrac{1}{2}mv^2_A = mgh_C + \tfrac{1}{2}mv^2_C$$

$$gh_A + \tfrac{1}{2}v^2_A = gh_C + \tfrac{1}{2}v^2_C$$

where $v_C = \sqrt{rg}$ and $h_C = 10 + 2r$, r being the radius of the loop

$$9.8 \times 30 + \tfrac{1}{2} \times 1^2 = 9.8 \times (10 + 2r) + \tfrac{1}{2} \times r \times 9.8$$

$$r = 8.0 \text{ m}$$

Note

This whole question could have been done more easily if B were taken as the position of zero potential energy.

Appendix A Derivations of Formulae

Alternative Derivation of Formula for Centripetal Acceleration Using Calculus

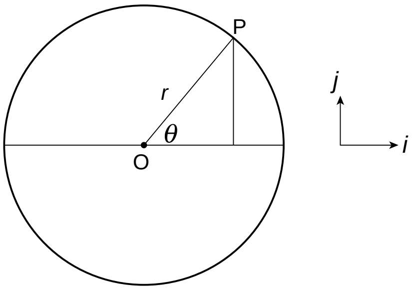


Fig. A.1

Consider a particle travelling with constant speed in a circular path of radius r, Fig. A.1. At a particular time, t, the displacement of the particle, relative to O, is

$$s = r\cos\theta i + r\sin\theta j \qquad \text{(a)}$$

where i and j are unit vectors in the directions shown.

Since the angular speed is constant

$$\theta = \omega t$$

and eqn. (a) becomes

$$s = r\cos\omega t\, i + r\sin\omega t\, j \qquad \text{(b)}$$

But, $v = \dfrac{\mathrm{d}s}{\mathrm{d}t}$

Therefore,

$$v = -r\omega\sin\omega t\, i + r\omega\cos\omega t\, j \qquad \text{(c)}$$

But, $a = \dfrac{\mathrm{d}v}{\mathrm{d}t}$

Therefore,

$$a = -r\omega^2\cos\omega t\, i - r\omega^2\sin\omega t\, j$$

$$= -\omega^2(r\cos\omega t\, i + r\sin\omega t\, j)$$

$$= -\omega^2 s$$

Thus, the acceleration of the particle is in the opposite direction to its displacement, i.e. is towards the centre, and is of magnitude

$$a = \omega^2 r$$

From eqn. (b), the displacement is along a line whose slope is

$$m_1 = \tan\omega t$$

From eqn. (c), the velocity is along a line whose slope is

$$m_2 = -\cot\omega t$$

Since

$$m_1 m_2 = -1$$

the velocity is at right angles to the displacement. Since the displacement at any point on the circle is along the radius at that point, the velocity must be along the tangent to the circle at that point.

Alternative Derivation of an Expression for the Period of Simple Harmonic Motion

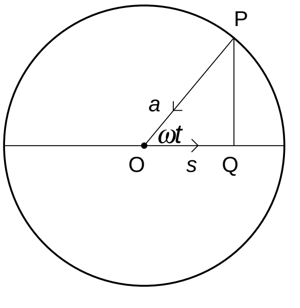


Fig. A.2

To derive an expression of the period of a particle executing simple harmonic motion, we will first show that if the particle P, Fig. A.2, is moving around O with constant angular speed, ω, then the projection of P on the x-axis, i.e. the point Q, moves back and forth with simple harmonic motion.

The acceleration of P is

$$a_p = r\omega^2 \quad \text{towards O}$$

where r is the radius

The acceleration of Q is the horizontal component of the acceleration of P, i.e.

$$a_Q = r\omega^2 \cos \omega t \text{ towards O}$$

But

$$\cos \omega t = \frac{|s|}{r}$$

$$\Rightarrow \quad a_Q = \frac{-r\omega^2 s}{r} \quad (a_Q \text{ and } s \text{ are in opposite directions})$$

$$= -\omega^2 s$$

Thus, the motion of Q is simple harmonic. Its periodic time, T, is the same as that of P, which is

$$T = \frac{2\pi}{\omega}$$

This derivation is for a particular case of simple harmonic motion, but holds for all cases.

Appendix B Data for Planets

It is useful to have figures of measurements for the planets, to enable new calculation questions to be made up. The values given for Pluto's mass, radius and acceleration due to gravity are of doubtful accuracy. [*Data for g calculated from* $g = GM/r^2$]

Name	Mass/kg	Radius/m	g/m s^{-2}	Orbital period	Radius of orbit/m
Mercury	3.28×10^{23}	2.44×10^{6}	3.68	88.0 d	5.79×10^{10}
Venus	4.87×10^{24}	6.05×10^{6}	8.87	225 d	1.08×10^{11}
Earth	5.98×10^{24}	6.38×10^{6}	9.81	365 d	1.50×10^{11}
Mars	6.42×10^{23}	3.39×10^{6}	3.71	687 d	2.28×10^{11}
Jupiter	1.90×10^{27}	7.14×10^{7}	24.9	11.9 y	7.78×10^{11}
Saturn	5.69×10^{26}	6.03×10^{7}	10.4	29.5 y	1.43×10^{12}
Uranus	8.68×10^{25}	2.56×10^{7}	8.84	84.1 y	2.87×10^{12}
Neptune	1.02×10^{26}	2.48×10^{7}	11.1	165 y	4.50×10^{12}
Pluto	1.29×10^{22}	1.15×10^{6}	0.65	248 y	5.91×10^{12}
Sun	1.99×10^{30}	6.96×10^{8}	274		
Moon	7.35×10^{22}	1.74×10^{6}	1.62	27.1 d	3.84×10^{8}

Table B.1 *Selected data for the planets and for the earth's sun and moon*

Name	Mass	Radius	g	Period	Radius of orbit
Mercury	0.0548	0.383	0.375	0.241	0.387
Venus	0.814	0.949	0.894	0.615	0.723
Earth	1	1	1	1	1
Mars	0.107	0.532	0.379	1.88	1.52
Jupiter	318	11.2	2.54	11.9	5.20
Saturn	95.2	9.41	1.07	29.5	9.54
Uranus	14.5	4.06	0.8	84.1	19.2
Neptune	17.1	3.89	1.2	165	30.1
Pluto	0.002	0.18	0.05	248	39.5

Table B.2 *Selected data for the planets expressed relative to the corresponding values for the earth*

MODULE 8

Modern Physics

CONTENTS

CHAPTER 1 X-RAYS

1.1 Background and History

Early Egyptians knew of nine elements - carbon, gold, iron, copper, lead, mercury, silver, sulphur and tin. However, it was not until 1661 that **Robert Boyle** (1627-1691)defined what exactly was meant by an element.

By 1803 the English chemist **John Dalton** (1766-1844) had put forward his theory about atoms. He envisaged them as tiny, indivisible particles that could neither be created nor destroyed, with all the atoms of a given element being exactly alike but different from those of other elements. When elements reacted to form compounds, their atoms combined together in simple whole-number ratios such as 1:1, 1:2, 2:3, etc. Dalton's theory held until near the end of the nineteenth century.

During the latter half of the nineteenth century, a significant number of far-reaching scientific discoveries were made - in many cases, one leading to the next, and eventually leading to a description of atomic structure.

Cathode Rays

Two inventions in particular, those of the induction coil in 1836 by **Revd Nicholas Callan** (1799-1864) of Maynooth, and of an efficient vacuum pump by **Henry Geissler** (1815-1879) in 1855, laid the foundations for many of the scientific discoveries at the end of the nineteenth century. Callan's coil provided the high voltages, and Geissler's pump the low pressures, that led to the investigation of electric discharges in gases. Geissler discovered that 'rays' of some kind were emitted from the cathode of evacuated tubes in which a discharge was occurring. The penetrating power of the rays was first demonstrated by the German physicist, **Heinrich Hertz** (1857-1894) and his assistant, **Philip Lenard** (1862-1947); this they did by passing the rays through thin aluminium foils. Another German physicist, **Eugen Goldstein** (1850-1931) introduced the name 'cathode rays' and began experiments that led to the discovery of the cathode rays' positive counterpart - *Kanalstrahlen* (channel or canal rays).

Two people particularly associated with the investigation of cathode rays were **Johann Hittorf** (1824-1914) of Germany and **William Crookes** (1832-1919) of England. Hittorf found that the rays were deflected by magnetic fields, and Crookes discovered, using a variety of discharge tubes, that the rays have momentum and energy, and are therefore streams of particles. **Jean Perrin** (1870-1942), who was French, discovered that the particles were negatively charged.

The Electron

It was **J. J. Thomson** (1856-1940) working in the Cavendish Laboratory in Cambridge in 1897 who identified these particles. He measured their *e/m* value (their specific charge) and found it to be about two thousand times greater than that of the hydrogen ion. He found that this *e/m* value was always the same, irrespective of both the gas in the tube and the metal of which the cathode was made. This meant that the particles had either a very large charge, or a very small mass, or a combination of both. On the assumption that they had a charge the same as that of the hydrogen ion, he concluded that their mass was about two thousand times less - therefore the particles were much smaller than the smallest known atom, i.e. they were sub-atomic particles. Thomson is hence accredited with discovering the electron. The name electron had already been used by **George Johnstone Stoney** (1826-1911) of Galway for the fundamental unit of electric charge in electrolysis.

In 1909, the American physicist **Robert Millikan** (1868-1953) measured the charge on the electron in a series of experiments based on oil drops. He always found that the charge on the drops was an integral multiple of 1.6×10^{-19} C and concluded that electric charge could never exist in fractions of this amount and that the magnitude of the electronic charge was therefore 1.6×10^{-19} C. Millikan was awarded the 1923 Nobel prize in physics for this work and also for his work on the photoelectric effect.

X-rays

The discovery of X-rays was one of those lucky scientific 'accidents' and was the result of experiments on cathode rays carried out by **Wilhelm Konrad von Roentgen** (1845-1923) of Germany. Roentgen was educated in Holland and Switzerland (where he studied mechanical engineering), started his scientific career as an assistant in the physics laboratory at Würtzburg, Germany, subsequently worked at the universities of Strasbourg (now France), Hohenheim (Germany), Strasbourg again, Giessen (Germany) and ended up in 1888 as Professor of Physics back at Würtzburg.

In England, at the time, cathode rays were considered to be particles but German scientists favoured some kind of ray. One way of detecting cathode rays is to make use of their property of causing fluorescence. On 8 November 1895 Roentgen was working with a cathode ray tube when he noticed an unexpected glow about two metres away from the tube, coming from some barium platinocyanide [$BaPt(CN)_4$], a fluorescent material that was often used to detect cathode rays. The effect could not have been caused by the cathode rays as it was known that they have a range in air of only a few centimetres. He knew that it had to be a different phenomenon - but what?

He spent the next seven weeks in his laboratory investigating this unknown radiation. He discovered that human flesh is transparent to it, but that bones are not. He produced a photograph showing the bones of the hand of his wife Bertha. Not yet knowing what kind of rays they were, he called them X-rays. He published the first of three papers in January 1896 and within a short time X-rays were being used for medical diagnosis. Roentgen showed that X-rays are not deviated by magnetic fields. He knew that they were similar to light rays because of the way they produced images on photographic emulsions. He tried, but was unable to show any of the properties of waves, such as diffraction, refraction, reflection and polarisation. In the end he postulated that X-rays are some kind of longitudinal waves similar to the transverse waves of visible light.

Roentgen's words (translated from the German) describe his experiments:

'If the discharge of a fairly large induction coil be made to pass through a Hittorf vacuum-tube or through a Lenard tube, a Crookes tube, or other similar apparatus, which has been sufficiently exhausted, the tube being covered with thin, black cardboard which fits it with tolerable closeness, and if the whole apparatus be placed in a completely darkened room there is observed at each discharge a bright illumination of a paper screen covered with barium platinocyanide, placed in the vicinity of the induction coil, the fluorescence thus produced being entirely independent of the fact whether the coated or the plain surface is turned towards the discharge tube…The most striking feature of this phenomenon is the fact that an active agent here passes through a black cardboard envelope which is opaque to the visible and the ultraviolet rays of the sun or of the electric arc; an agent too, which has the power of producing active fluorescence…We soon discover that all bodies are transparent to this agent, though to very different degrees. I proceed to give a few examples: paper is very transparent; behind a bound book of about one thousand pages I saw the fluorescent screen light up brightly, the printers' ink offering scarcely a noticeable hindrance. In the same way the fluorescence appeared behind a double pack of cards. A single sheet of tin foil is also scarcely perceptible; it is only after several layers have been placed over one another that their shadow is distinctly seen on the screen.'

The history of science has many examples of accidental discoveries and, of these, the discovery of X-rays is a prime example. Of all the discoveries made by man, possibly that of X-rays attracted the most public attention. The fact that the rays permitted one to 'see' through opaque objects made sensational tabloid material, and there was great consternation lest, by their use, fully-dressed people might be made to appear unclothed. 'Punch' and other similar magazines of the time made the most of cartoons of the phenomenon. When such speculation died down, there was widespread appreciation of the value of X-rays in setting broken bones, and the discovery was quickly put to this use.

For the discovery, Roentgen received the Rumford Medal of the Royal Society in 1896 and in 1901 the first Nobel prize in physics. He also conducted researches in light, heat and elasticity but none of these compares in importance with his discovery of X-rays. Roentgen died in 1923, from cancer of the intestines.

1.2 Thermionic Emission

In a metal, the electrons of the outer or valency shell are not held strongly to the atom. As a result, these electrons become free of their atoms and form an 'electron cloud' within the metal. They do not escape because of the attractive forces of the now positive ions within the metal. However, when the metal is heated, the surface electrons can gain enough kinetic energy to escape from the metal. This is known as thermionic emission and it was first observed by **Thomas Edison** (1847-1931) in 1883. He could not explain the effect, and it was left to the English physicist **Owen Richardson** (1879-1959) to do so in 1902. The resulting current is given by the following equation, now known as Richardson's equation.

$$I = AT^2e^{-\phi/kT}$$

where *A* and *W* are constants characteristic of the emitter, *k* is the Boltzmann constant, and *T* is the temperature in kelvin. *W* is the minimum amount of energy needed to remove an electron from the metal, i.e. the work function of the metal. Comparisons between the work functions found in thermionic emission and in the photoelectric effect show a close agreement (Table 1.1).

Metal	**Thermionic ϕ/eV**	**Photoelectric ϕ/eV**
Caesium	1.81	1.9
Tungsten	4 .52	4.49

Table 1.1 *Comparing thermionic and photoelectric work functions*

Thermionic emission provides the electron beams in modern X-ray tubes and other cathode ray tube devices such as oscilloscopes, TVs, and computer monitors. Cathode ray tubes are also used in the display devices of several types of medical equipment.

1.3 Generation of X-rays

X-rays are generated when fast-moving electrons hit a solid target, where they decelerate rapidly and some of their kinetic energy is converted to X-rays. Waves are produced in the wavelength range 10^{-8} m to 10^{-12} m. Less than 1% of the kinetic energy of the electrons is converted to X-rays, the rest goes to heat the target which therefore must be cooled in some way.

Early X-ray tubes were gas discharge tubes made essentially for producing cathode rays. From about 1896 focus tubes were manufactured, in which beams of cathode rays were concentrated onto platinum anodes. From about 1915, tungsten began to replace platinum because it could withstand higher temperatures, but often extra cooling was needed. These X-ray tubes gave off a lot of unwanted radiation, and the dangers of this were not fully realised. Some screening was included, but the precautions taken were often far from adequate.

In the modern X-ray tube (the Coolidge tube, named after its inventor, the American physicist **William Coolidge** (1873-1975)) the electrons are produced by thermionic emission (q.v.) from a hot cathode. A high-voltage transformer replaces the induction coil used to provide the high voltage for early X-ray tubes. The electrons are accelerated by voltages of the order of 25 kV to 400 kV, to collide with the metal target. Because most of the energy generated by the electrons on bombardment with the anode appears as heat, the anode must be able to withstand high temperatures; it should also have a high thermal conductivity in order to remove the heat as quickly as possible. Tungsten meets these criteria best (see Table 1.2) and is almost always used.

Metal	At. no	Thermal conduct-ivity/ $W\ m^{-1}\ K^{-1}$	Specific heat capacity/ $J\ kg^{-1}K^{-1}$	Density/ $kg\ m^{-3}$	Melting point/K
copper	29	4.00	386	8960	1356
molybdenum	42	1.38	252	16220	2880
lead	82	0.44	160	11350	600
tungsten	74	1.63	1332	19300	3650
iron	26	0.80	445	7870	1808

Table 1.2 *Comparison of metals for use in anode of X-ray tubes*

From the above table, it is seen that tungsten meets the requirements best; it has a high atomic number (74) so its nucleus has a large mass and a high positive charge. This increases the probability that collisions of electrons with it will lead to X-ray emission. It has a high melting point and also a high thermal conductivity, so it can withstand high temperatures and is an efficient conductor of the heat that is generated in the anode.

The actual metal target is usually part of a rotating metal disc that has a bevelled edge. This bevelled edge presents a large area to the electron beam but also produces a narrow X-ray beam. This also spreads the heating effect over a larger surface area and reduces damage to the metal target. A tungsten ring, about 6 cm wide, is set into the bevelled edge and the whole assembly is rotated at about 9000 revolutions per minute. The housing of the X-ray tube contains a bath of oil to carry away most of the heat generated. The glass used in the housing is heat-resistant. In standard radiography, a beam of X-rays is passed through the patient and on to a fluorescent material which emits light. This light then falls on a photographic film which, on development, produces the image.

Mobile X-ray systems, which are moved around hospitals, still use fixed anode systems. The actual anode part is made of a disc of molybdenum, which is about 10 cm in diameter and bevelled around the edge.

The X-radiation produced when electrons undergo rapid deceleration at the anode is known as *bremsstrahlung* or ‘braking radiation’. X-rays are also emitted as a result of electronic transitions in the atoms of the anode (see below).

X-rays may also be produced when energetic electrons or ions in the outer reaches of stars are deflected under the influence of very strong magnetic fields.

1.4 X-Ray Spectra

X-rays have a minimum wavelength that is determined by the energy of the electrons producing them, which in turn is determined by the tube voltage. The higher the tube voltage is the higher the energy of the electrons when they strike the target and therefore the higher the maximum energy of the X-ray photons produced. Since $E = hf$, the higher the energy of the photon is the higher the frequency and therefore the shorter wavelength. The minimum wavelength occurs when all the energy of an electron appears as a photon. Thus, the minimum wavelength is determined by the tube voltage.

X-ray photon energies are largely within the range 0.12 keV (1.9×10^{-17} J) to 1.2 MeV (1.9×10^{-13} J) at the high end. These energy values give wavelengths ranging from about 10 nm (frequency 3×10^{16} Hz) to about 0.001 nm (3×10^{20} Hz).

The spectrum of the rays from an X-ray tube, Fig. 1.1, is a continuous one, running from the minimum wavelength upwards without a break. However, the intensity of all the different wavelengths (or frequencies) is not constant, but shows peaks at some particular wavelengths which depend on the metal used for the anode; they do not depend on the tube voltage or other factors.

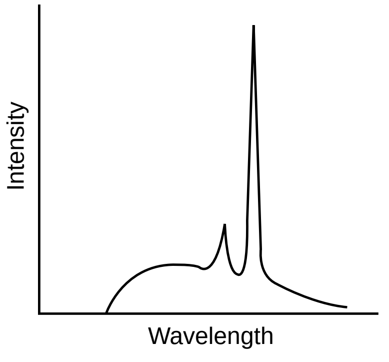


Fig. 1.1 *X-ray spectrum*

The characteristic peaks come from rapidly moving electrons knocking electrons out of lower energy levels of the target atoms. The spaces will then be filled by electrons moving down from higher energy levels, emitting X-rays as they do so. It was an investigation of these characteristic X-ray wavelengths that, in 1913, led the English physicist **Henry Moseley** (1887-1915) to the discovery of atomic number.

1.5 Detection of X-Rays

The two principal means of X-ray detection are photography and the GM tube. For accurate work in X-ray spectroscopy, a scintillation counter is sometimes used. The quality of an X-ray photograph improves with the intensity of the X-rays used. However, in medical applications a high-intensity beam can be dangerous to the subject, and a compromise must be made between safety and the quality of the picture. GM tubes are used when an estimate of X-ray intensity is needed, usually for safety monitoring. Accurate work is mostly carried out with scintillation counters and photomultiplier assemblies, which can both count the number of photons and also record their energies.

X-ray Imaging

When X-rays travel through matter they are attenuated (i.e. their intensity is reduced) because photons are deflected out of the beam and/or are absorbed. The amount of attenuation depends on the composition of the material through which the beam is passing. The atoms of soft tissue, such as muscle, have an average proton number of about 7, as muscle is composed mainly of carbon, hydrogen and oxygen. However, the atoms of bone have an average value of 14, as bone is mainly made up of calcium and phosphorus. Density is a second important factor; soft tissue has a relative density of about 1.0 but bone a relative density of 1.8.

These differences are the basis of X-ray imaging. In many cases X-rays can differentiate between tissues with very small differences in their chemical composition. If the differences are too small then they can be enhanced by the introduction of materials with high proton number like barium or iodine. Thus a ‘barium meal’ (a sort of slurry of barium sulphate and water) is often given to patients before an investigation of their digestive system.

Photographic film is sensitive to X-rays but much more so to light. To rely only on X-rays would require high exposure times. Owing to the ionising ability of X-rays this is not desirable. Exposure times must always be kept to a minimum (e.g. for a skull X-ray, about 0.40 s). Therefore an image-intensifying cassette is used. A fluorescent material such as zinc sulphide is used to coat the intensifying screens; this absorbs the X-rays and re-emits the energy as light. It is the resultant light that affects the photographic film.

1.6 Uses of X-Rays

The uses of X-rays are various. Medical uses are of two kinds - for diagnosis and for therapy. Diagnostic uses are primarily for the production of X-ray photographs of bones or lung tissue to identify fractures and other abnormalities. Diagnosis of a suspected bone fracture is possibly the first encounter with X-rays for most people and this is routine work in the accident and emergency departments of hospitals. Chest X-rays are also used for the diagnosis of tuberculosis and cancer. Women have mammograms, in which their breasts are X-rayed to search for early signs of breast cancer. Dentists use X-rays to locate problems of tooth decay and abscesses.

Both X-ray diagnosis and therapy benefit the patient but, as with the use of any form of radiation, the benefit must outweigh the risk for it to be used. X-ray tubes for diagnosis are run at voltages in the region of about 60 to 125 kilovolts; X-ray photons of this energy are largely absorbed in photoelectric events. For therapeutic purposes, X-ray photons are produced at a voltage of 1 MV or above, and are absorbed mostly in Compton scattering.

In the early days X-rays were used by people with little or no protection. However, in a very short time, doctors realised the harmful effects of over-exposure to the radiation. One of the first recorded deaths from X-ray exposure was that of **Clarence Dally**, who worked for **Thomas Edison**. He died in 1904 aged only 39 years.

In controlled environments, X-rays are used to kill tumours in the body. In a process known as radiotherapy, X-rays are directed at a tumour to destroy it. A way is found of giving high X-ray doses to malignant or other offending tissue without giving similar doses to nearby normal tissue, and in this way the offending tissue can be destroyed. The X-rays are produced by a linear accelerator generating energies up to 6 MeV (compared with energies of 120 keV for diagnostic purposes). Electrons move along a tube where the potential difference is repeatedly applied between neighbouring sections to promote electrons to higher energies. The X-ray machine is rotated in an arc around the patient. This ensures that the tumour receives a maximum dose but healthy tissue surrounding it does not receive too much. The tumour will probably have been discovered using a CT (see below) scanner and the radiotherapist will use these computer-generated images to target the X-ray beam.

Low-energy X-ray wavelengths happen to be of the same order of magnitude as the lengths of inter-atomic distances in molecules. Because of this, X-ray diffraction techniques are used to investigate molecular structures in a wide range of materials, from simple crystals to complex proteins. X-rays played a significant part in the discovery of the helical structure of DNA.

CT Scanning

The amount of information made available using X-rays only is quite limited. As the body is three-dimensional, a more-detailed way to look at it has obvious advantages and this came about with the advent of computer tomography (CT scanning) - a recent major development in the application of X-rays. This is a way of obtaining a sectional view of the body without shadows from other organs being imaged. A computer is used to collect the images from a large number of body ‘slices’, to produce a composite three-dimensional image. The method is particularly good in examining soft tissue.

In a CT scanner the patient lies on a horizontal movable couch which can slide in and out of the scanner. An X-ray source moves in an arc around the patient and produces short bursts of X-rays. A large number of X-ray detectors are arranged, also in an arc, on the opposite side of the patient from the source. The detectors consist of sodium iodide crystals to detect the X-rays and photodiodes to record the scintillations from the crystals. A computer is used to collect information from the photodiodes and to reconstruct a slice of the body on a TV screen in a few seconds. While the radiation dose is higher than from standard radiography, it can detect tumours, blood clots, etc., which a conventional X-ray machine would be unable to detect.

The CT scanner was invented by **Godfrey Hounsfield** who worked for the firm EMI. He was awarded the Nobel prize for medicine in 1979.

Fluoroscopy

In this technique, the X-ray beam passes through the patient into an image-intensifier, and the moving image is then displayed on a TV monitor. However this uses higher doses of X-rays than conventional imaging.

Examination	Effective dose/mSv	Equivalent period of natural background radiation
Limb (XR)	0.01	1.5 days
Chest (XR)	0.02	3 days
Lumbar Spine (CT)	2	1 year
Thoracic Spine (CT)	6	3 years
Chest (CT)	8	4 years
Barium enema (XR)	9	4.5 years

XR = conventional X-ray
CT = computed tomography

Table 1.3 *Radiation dose obtained during different types of X-ray examination*

1.7 Worked Examples

1. What is the kinetic energy of an electron accelerated through a potential difference of 50 kV in an X-ray tube? Calculate then the minimum wavelength of X-rays this electron will produce. ($h = 6.63 \times 10^{-34}$ J s; $c = 3.0 \times 10^{8}$ m s^{-1}.)

$$E = eV$$

$$= 1.6 \times 10^{-19} \times 50 \times 10^{3}$$

$$= 8.0 \times 10^{-15} \text{ J}$$

$$E = hf$$

$$= \frac{hc}{\lambda_{\min}}$$

$$\lambda_{\min} = \frac{hc}{E}$$

$$= \frac{6.63 \times 10^{-34} \times 3 \times 10^{8}}{8 \times 10^{-15}}$$

$$= 2.5 \times 10^{-11} \text{ m}$$

2. The potential difference across an X-ray tube is 60 kV and the current through it is 2.5 mA. Calculate (a) the number of electrons striking the anode per second, (b) the speed with which they strike it. ($e = 1.6 \times 10^{-19}$ C; m_e assumed constant $= 9.1 \times 10^{-31}$ kg.)

(a)

$$1 \text{ A} = 1 \text{ C s}^{-1}$$

$$2.5 \text{ mA} = 2.5 \times 10^{-3} \text{ C s}^{-1}$$

$$\text{No. of electrons/s} = \frac{2.5 \times 10^{-3}}{1.6 \times 10^{-19}}$$

$$= 1.6 \times 10^{16}$$

(b) $E = eV$

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

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

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

$$= \sqrt{\frac{2 \times 1.6 \times 10^{-19} \times 6 \times 10^{4}}{9.1 \times 10^{-31}}}$$

$$= 1.5 \times 10^{8} \text{ m s}^{-1}$$

3. An X-ray tube operates at a potential difference of 200 kV. Heat is generated at a rate of 630 W in the target where 99.9% of the energy of the electrons is converted to heat.

Calculate (a) the current in the tube, (b) the speed of the electrons striking the target. ($e = 1.6 \times 10^{-19}$ C; $m_e = 9.1 \times 10^{-31}$ kg.)

(a)

Heat represents 99.9% of total energy of electrons

Total energy per second, i.e. power is

$$P = 630/0.999$$

$$= 630.63 \text{ W}$$

$$P = VI$$

$$630.63 = 200 \times 10^3 \times I$$

$$I = \frac{630.63}{200 \times 10^3}$$

$$= 3.2 \times 10^{-3} \text{ A}$$

$$= 3.2 \text{ mA}$$

(b)

$$E = eV$$

$$= {}^1/_2\, mv^2$$

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

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

$$= \sqrt{\frac{2 \times 1.6 \times 10^{-19} \times 2 \times 10^5}{9.1 \times 10^{-31}}}$$

$$= 2.7 \times 10^8 \text{ m s}^{-1}$$

CHAPTER 2 PHOTOELECTRICITY

2.1 Photoelectric Effect

When electromagnetic radiation of sufficiently high frequency (short wavelength) falls on the surface of metals, electrons are ejected. This phenomenon is known as the photoelectric effect, and its discovery at the end of the nineteenth century caused scientists to re-evaluate the wave theory of light. In particular, the phenomenon was in conflict with the ideas of Maxwell, but gave support to the quantum theory as put forward by the German physicist **Max Planck** (1858-1947) to explain the facts of black body radiation.

It was **Isaac Newton** (1642-1727) who, in 1660, put forward the idea of the corpuscular theory of light, i.e. that light consists of moving particles or corpuscles. This theory could explain both reflection and refraction but not interference and diffraction. The Dutch scientist **Christiaan Huygens** (1629-1695) proposed a wave theory of light and he explained interference and diffraction by saying that light was a propagation of a wave in a medium. In 1802 the English physicist **Thomas Young** (1773-1829) successfully demonstrated interference of light and, as a result, Newton's ideas were subsequently abandoned.

One of the problems scientists were trying to understand towards the end of the 19th century was the way in which matter (atoms) interacts with radiation (light). The simplest way is to look at a hot object. A hot object radiates electromagnetic energy, and the hotter it is, the more energy it radiates at shorter wavelength (black body radiation curve). Scientists used the idea of a 'black body' because such a body absorbs all the radiation that falls on it. A black body could be a hollow sphere with a small hole in one side. Any radiation, such as light, which can get in through the hole will experience multiple reflections and be unlikely to escape. When the black body is heated, it first glows, becoming red-hot and eventually white-hot, giving out radiation. There is very little radiation at short wavelengths and again very little at long wavelengths but most of the energy is radiated in a middle band of wavelengths. A definite peak shifts towards the shorter wavelengths as the body gets hotter.

This is where the conflict arose. Classical theory says that a black body would have an infinite amount of energy at the shortest wavelengths. So all black body radiation should be able to produce a lot of high-frequency energy, particularly in the ultraviolet and beyond. This was known as the 'ultra-violet catastrophe'. A law, known as the Rayleigh-Jeans law, could explain the observations on the low-frequency side of the spectrum. **Wilhelm Wien** (1864-1928) developed a law that predicted the wavelength at which the peak of the curve occurred at any temperature.

From 1895 to 1900 Max Planck tried to solve this problem. Eventually in 1900 Planck proposed a solution that would encompass all of the curve. Unfortunately for Planck he had no physical basis for his solution, except that it worked. Planck proposed that energy be considered to be made up of discrete units E, given by $E = hf$, where h is a constant now known as the Planck constant and f is the frequency of the radiation. It took many years for physicists to fully realise the importance of his idea, and Planck was eventually awarded a Nobel prize in 1918.

Philipp Lenard (1862-1947) (Nobel prize 1905) had shown in 1899 that cathode rays can be produced by light shining onto a metal surface in a vacuum. His experiments involved monochromatic light which means that all the waves of the light had the same frequency. He showed that as the intensity of the light was increased, the number of electrons was increased, but the maximum velocity of the ejected electrons stayed the same. By increasing the frequency of the radiation (using ultraviolet radiation instead of blue light) the maximum velocity increased. There was no way to explain these effects using classical ideas.

In 1904 Albert Einstein applied the equation $E = hf$ to electromagnetic radiation, instead of to black-body radiation as Planck had suggested. He suggested that light was not a continuous wave but instead comes in definite packets of energy called quanta. All light of a particular frequency comes in packets which have the same energy E.

It was the American physicist **Robert Millikan** (1868-1953) who, in 1916, conducted a series of experiments that not only verified Einstein's equation for the photoelectric effect but also led to a value for h which agreed very closely with values obtained for black-body radiation.

2.2 Photoelectric Devices

Photovoltaic Cells

A photovoltaic cell consists of a p-n junction in which the p-layer is thin enough to allow light to pass through it onto the depletion layer. In the depletion layer the light produces pairs of free holes and electrons which move to the p-layer and the n-layer, respectively, under the influence of the junction voltage. If the two layers are connected externally a current will flow in the circuit. Photovoltaic cells, depending on the materials used, can produce 0.5 V per cell in full sunlight with a current of 35 mA per cm^2 of cell, and with an efficiency of 10-15%. This type of cell is used in solar cells.

Photoemissive Cells

When light of a frequency greater than the threshold frequency falls on the cathode of this type of cell, electrons are emitted. These electrons are attracted to the anode and hence a current can flow in an external circuit. The size of this current is proportional to the intensity of the light. If the light beam is broken or interrupted in any way, then the current ceases. This device is used in counting operations in industrial applications, and as part of burglar alarm circuits. It has also been used to 'read' the sound tracks of movie films.

Photoconductive Cells

This type of cell changes its resistance according to the intensity of the light falling on it. Such cells can have a resistance of up to 1 MΩ in the dark and as low as 100 Ω in bright light. One of the most common light-dependent resistor is the ORP12, which is made from cadmium sulphide (CdS). Unlike the photovoltaic cell, the photoconductive cell needs a battery in the circuit (a battery is needed to run a camera light meter). The major disadvantage is that these cells are slow (100 ms) to respond to changes in light intensity, compared with photodiodes (1 ms). Photoconductive cells are therefore not suitable for fast counting operations.

Photodiode/Phototransistor

A photodiode is a normal p-n junction diode but one in which light can enter through a small window in the casing. It is operated in reverse bias and the reverse current generated is in proportion to the amount of light entering the window. Photodiodes are used as optical fibre receivers, and as fast counters owing to their fast response time.

A phototransistor is usually used with the base disconnected and with a window allowing light to shine onto the collector-base junction. This allows more electron-hole pairs to be produced. As a transistor is a current amplifier, a larger collector current is produced. In this way a phototransistor is a photodiode plus an amplifier.

Photomultiplier Tube

When light from a scintillator hits a photocathode, it causes electrons to be emitted. These photoelectrons are then accelerated and multiplied by a series of electrodes called dynodes, held at increasingly positive potentials along the tube. Alloys such as beryllium-copper are used because they cause several secondary electrons to be emitted. The process is repeated a number of times, causing multiplication each time. The electrons are eventually collected by an anode and then outputted to a further amplifier. It is possible to amplify the original light by a factor of between 10^5 and 10^8. The tube can be used in an image-intensifying camera, to obtain pictures from low light levels (thermal-image camera) or in a scintillation counter to detect ionising events. The commonest scintillator is sodium iodide, activated with 0.5% thallium iodide. It is used in hospitals for the detection of X-rays and gamma rays.

2.3 Worked Examples

1. Find the energy of a photon which has a wavelength of 300 nm. (c = 3.0 x 10^8 m s^{-1}; h = 6.63 $\times$ 10^{-34} J s.)

$$E = hf$$
$$= \frac{hc}{\lambda}$$
$$= \frac{6.63 \times 10^{-34} \times 3 \times 10^{8}}{300 \times 10^{-9}}$$
$$= 6.6 \times 10^{-19} \text{ J}$$

2. Light of wavelength 650 nm falls on a metal whose work function is 1.75 eV. Find (a) the energy of the photons in eV, (b) the maximum energy of the emitted electrons in joules, (c) the stopping voltage. (c = 3.0 x 10^8 m s^{-1}; h = 6.63 x 10^{-34} J s; e = 1.6 x 10^{-19} C.)

(a)

$$E = hf$$
$$= \frac{hc}{\lambda}$$
$$= \frac{6.63 \times 10^{-34} \times 3 \times 10^{8}}{650 \times 10^{-9} \times 1.6 \times 10^{-19}}$$
$$= 1.91 \text{ eV}$$
$$= 1.9 \text{ eV}$$

(b)

$$\tfrac{1}{2}mv^2 = hf - f_0$$
$$= 1.91 - 1.75$$
$$= 0.16 \text{ eV}$$
$$= 0.16 \times 1.6 \times 10^{-19} \text{ J}$$
$$= 2.6 \times 10^{-20} \text{ J}$$

(c)

Energy of electrons = 0.16 eV, therefore stopping voltage = 0.16 V

3. If a metal has a work function of 2.4 eV, what is the threshold frequency of the light that will liberate electrons from it? (h = 6.63 $\times$ 10^{-34} J s; e = 1.6 $\times$ 10^{-19} C.)

$$\phi = hf_0$$
$$f_0 = \frac{\phi}{h}$$
$$= \frac{2.4 \times 1.6 \times 10^{-19}}{6.63 \times 10^{-34}}$$
$$= 5.8 \times 10^{14} \text{ Hz}$$

4. Monochromatic radiation falls on a photocell whose work function is 2.8 eV. If the stopping potential is measured at 1.4 V and the maximum current observed is 0.005 mA, find (a) the energy of each photon of radiation, (b) the frequency of the radiation, (c) the number of electrons emitted per second, (d) the threshold frequency for the cathode surface.
(h = 6.63 x 10^{-34} J s; e = 1.6 x 10^{-19} C.)

(a)

Photon energy = 1.4 + 2.8

$= 4.2$ eV

$= 4.2 \times 1.6 \times 10^{-19}$

$= 6.7 \times 10^{-19}$ J

(b)

$$E = hf$$

$$f = \frac{E}{h}$$

$$= \frac{6.7 \times 10^{-19}}{6.63 \times 10^{-34}}$$

$$= 1.0 \times 10^{15} \text{ Hz}$$

(c)

Since current = total charge per second, number of electrons per second = current divided by charge on electron

$$= \frac{I}{e}$$

$$= \frac{5 \times 10^{-6}}{1.6 \times 10^{-19}}$$

$$= 3.1 \times 10^{13}$$

(d)

$$\phi = hf_0$$

$$f_0 = \frac{\phi}{h}$$

$$= \frac{2.8 \times 1.6 \times 10^{-19}}{6.63 \times 10^{-34}}$$

$$= 6.8 \times 10^{14} \text{ Hz}$$

5. A metal surface has a threshold wavelength of 0.65 μm. If light of wavelength 0.45 μm shines on it, calculate the maximum speed of the electrons emitted from the metal surface. ($c = 3.0 \times 10^8$ m s^{-1}; $h = 6.63 \times 10^{-34}$ J s; $m_e = 9.1 \times 10^{-31}$ kg.)

$\lambda_0 = 0.65$ μm

$= 6.5 \times 10^{-7}$ m

$$f_0 = \frac{c}{\lambda_0}$$

$$= \frac{3 \times 10^8}{6.5 \times 10^{-7}}$$

$$= 4.6 \times 10^{14} \text{ Hz}$$

$\phi = h f_0$

$= 6.63 \times 10^{-34} \times 4.6 \times 10^{14}$

$= 3.05 \times 10^{-19}$ J

For light of wavelength 0.45 μm

$$f = \frac{c}{\lambda}$$

$$= \frac{3 \times 10^8}{4.5 \times 10^{-7}}$$

$$= 6.67 \times 10^{14} \text{ Hz}$$

$\frac{1}{2} mv^2 = h f - \phi$

$= (6.63 \times 10^{-34} \times 6.67 \times 10^{14}) - (3.05 \times 10^{-19})$

$= (4.422 \times 10^{-19}) - (3.05 \times 10^{-19})$

$= 1.372 \times 10^{-19}$ J

$$v^2 = \frac{2 \times 1.372 \times 10^{-19}}{m}$$

$$= \frac{2 \times 1.372 \times 10^{-19}}{9.1 \times 10^{-31}}$$

$$= 3.0 \times 10^{11}$$

$$v = 5.5 \times 10^5 \text{ m s}^{-1}$$

CHAPTER 3 RADIOACTIVITY

3.1 Background and History

We live in a radioactive world. So did Stone Age man. Radioactivity is nothing new; since the formation of the earth, radioactive elements in the earth's crust have been producing radioactive soils and seas and, as well, have been decaying to produce radioactive gases that have gone into the atmosphere. We cannot avoid radioactivity, no matter how hard we try. Our houses are radioactive, our schools, our food, and indeed we ourselves too. However, it is only since about the beginning of the twentieth century that an understanding of radioactivity has been acquired and that radioactive materials have been put to use.

About one hundred years ago, several remarkable and highly-important discoveries were crowded into the short space of ten years: X-rays in 1895, radioactivity the following year, the electron in 1897, quantum theory in 1900, and special relativity in 1905. Individually, each had enormous significance and collectively they heralded what is now known as 'modern physics'.

February 26 1896 might be regarded as one of the most important days in the history of physics. The Professor of Physics at what is today the National Museum of Natural History in Paris, **Antoine Henri Becquerel** (1852-1908), who was an expert on the luminescence of solids, and also skilled in the laboratory applications of photography, had become fascinated by the properties of uranium salts. Upon learning, in January 1896, of Roentgen's discovery of X-rays, he undertook to investigate whether there was some connection between this invisible radiation and the visible light that all luminescent materials, upon excitation, emitted. He was using sunlight to excite fluorescence in a salt of uranium and potassium (potassium uranyl sulphate) on a photographic plate. This plate in turn was wrapped in black paper. But it was raining in Paris at the time and he put the plate away in a drawer. Two days later he processed the plate - despite the fact that it had not been exposed to any sunlight. To his great surprise it showed a strong blackening. The blackening he traced to the uranium metal in the salt. Becquerel reported the results to the Academie des Sciences the next day. The full significance of his discovery did not become clear until some years later when Pierre and Marie Curie, also in Paris, made more systematic investigations. Seven years later in 1903, five years before he died, Becquerel was awarded the Nobel prize for physics, along with Pierre and Marie Curie.

Henri Becquerel was just one of four generations of Becquerels, who were at the forefront of science throughout the nineteenth century. The first scientific Becquerel was **Antoine César** (1788-1878). His scientific interests were wide: electrochemistry, of which he was one of the founders; optics; fertilisers; luminescence - the property of substances to emit light (fluorescence or phosphorescence) after stimulation with an excitation source.

When Antoine César died, both his son and grandson were already confirmed scientists. The second Becquerel, **Alexandre Edmond** (1820-1891) published numerous articles, together with his father, amongst others, on the chemical effects of light. He constructed a phosphoroscope, a device for measuring the emission of light by phosphorescent compounds and for the analysis of uranium salts.

The third and most famous Becquerel, Antoine Henri, had been interested in the absorption of light by crystals. He noticed that tetravalent uranium compounds were not phosphorescent, whereas uranyl salts exhibited a bright luminescence under the same conditions of excitation. Like his father, Antoine Henri was fascinated by the phenomenon of phosphorescence and at the time nobody suspected the secret hidden in the elements. His continued investigation of this problem finally led him to his discovery of radioactivity.

Jean Becquerel (1878-1953), the fourth and last member of the Becquerel dynasty, became interested in the modern physics of the twentieth century and also on the relation between music and physics. Although married twice, he died without children, and the Becquerel scientists came to an end.

For the next year or so after Becquerel's discovery, the investigation of uranic rays made little progress. There was more interest in the search for phosphorescent compounds, other than those of uranium, that might exhibit the same strange phenomenon. Even glow worms were investigated. In 1898, Pierre and Marie Curie entered the scene. When **Pierre Curie** (1859-1906) met **Maria Sklodowska** (1867-1934) in 1894, he was already famous for his achievements in physics. Together with his brother Jacques, he had discovered piezoelectricity. Becquerel's discovery had raised a puzzling problem. Uranium salts appeared to maintain a constant ability to blacken a photographic plate over months. The continuous emission of energy seemingly contradicted the first principle of thermodynamics. Did this energy arise from the uranium atom or from the surroundings?

Marie Curie who was thinking of a subject for her doctoral thesis, decided to investigate the new phenomenon. Her work began in December 1897 and her experiments are described in detail in three notebooks. The working conditions in what has been described as an old shed were miserable and one learns that on February 6, the laboratory temperature was 6 °C - the information being completed with 10 exclamation marks!

On April 12 is recorded 'radiation emitted by compounds of uranium and thorium'. In a first step, Marie Curie confirmed that all uranium compounds were active, and found that the intensity of the uranic rays was roughly proportional to the amount of uranium in the sample, and was independent of the chemical combination of the uranium. In a next step, she searched for other elements that exhibited the same phenomenon, and found that thorium and its compounds were active. Finally, minerals were investigated and the most important result of this search was that pitchblende, a variety of uraninite (UO_2), is nearly four times more active than metallic uranium. Several other minerals, too, were discovered to be more active than uranium. The notebook records: 'This fact is quite remarkable and suggests that these minerals may contain an element much more active than uranium itself'.

Because the phenomenon investigated was a property of the atoms of uranium or thorium, Marie Curie concluded that the excess of activity in pitchblende must arise from atoms of one or more unknown chemical elements present in the pitchblende, and more active than uranium itself. At this stage, further search for the identity of the presumed element became a matter of paramount importance. Pierre Curie abandoned his own research projects and joined his wife in the venture.

The second notebook, now signed by both, was published three months later, and announced the discovery of a new element. The word radioactive appears for the first time in the title: 'on a new radioactive substance contained in pitchblende'. Research had now turned from physics to chemistry; in a first trial, it was found that heating of the pitchblende in vacuum released sublimated products 30 times more active than uranium, but with a very low yield. The separations were performed with the ordinary procedures of chemical analysis. In time, they extracted several tons of pitchblende, to yield less than a gram of an element

300 times more active than uranium. This result was sufficient for the Curies to announce the discovery of the new element. They proposed that it be named polonium, in honour of Marie's native country.

At this point, the Curies were on the verge of another, much more important, discovery. Even before the announcement of the discovery of polonium, experiments had led them to suspect that the pitchblende might contain a second, more radioactive, element. In a note, dated December 26, came the announcement of a new strongly radioactive substance contained in pitchblende: 'We are led to think that this radioactive substance contains a new element, to which we propose to give the name radium…(it) contains very likely a high proportion of barium; nevertheless, the radioactivity of radium must be enormous'. The name radium, followed by a question mark, appears in the notebook on December 18. With the foregoing discovery of polonium, the Curies had, oddly enough, begun with the most difficult part of the work. In its own right, radium had outstanding advantages; the concentration of ^{226}Ra ($T_{1/2}$ = 1600 y) in the ore was about 5000 times greater than that of polonium.

Further work, lasting several years, led to the determination of the relative atomic mass of radium and eventually, in 1902, Marie Curie announced this to be 225 ± 1.

The discovery of radium gave the definitive impetus to the new field of radioactivity and in 1900 there were more published papers on radioactivity than on X-rays. The Nobel prize for physics in 1903 was awarded to Becquerel and the Curies for their 'joint researches on the radiation phenomena'. At this time, no mention was made of the discoveries of polonium and radium. In 1911, Marie Curie received the Nobel prize for chemistry for 'her services to the advancement of chemistry by the discovery of the elements radium and polonium, by the isolation of radium, and for the study of the nature and compounds of this remarkable element'.

Radioactivity Measurements

Radioactivity is measured in becquerel (Bq). *1 Bq is defined as 1 nuclear disintegration per second*. (The unit 'curie' was used in the past to measure radioactivity but is now obsolete; 1 curie = 3.7×10^{10} Bq.) Remember that the becquerel is a rate unit, and also that it is a very small unit. The radioactive potassium in every adult has an activity of several thousand becquerel; the average amount of radon in every cubic metre of air in houses in the UK has an activity of about 20 becquerel.

Radiation dose is measured in sievert (Sv). *1 Sv is equal to 1 J of energy deposited in 1 kg of tissue × quality factor*. The quality factor is a constant which takes into account the fact that alpha particles and neutrons have greater effects on biological cells than beta particles and gamma rays of the same energy, Table 3.1.

Radiation	Quality factor
Beta	1
Gamma	1
Alpha	20
Fast neutrons	10

Table 3.1 *Quality factors for various types of radiation*

A microsievert (1 μSv) is approximately:

- one tenth of the dose that would be incurred by flying from Ireland to Spain in a jet aircraft;
- the difference in annual dose from cosmic rays that would be received by moving from a first floor flat to one 20 metres higher (about a seventh floor);
- one twentieth of the average dose from a single chest X-ray;
- one tenth of the annual dose from radioactive fallout in Ireland and the UK in the 1980s;
- two-thirds of the average dose to the UK population due to discharges from existing nuclear installations.

The amount of energy deposited in unit mass of tissue is called the absorbed dose. The unit of the absorbed dose is the gray (Gy). 1 Gy is 1 J kg^{-1}.

Radiation dose limitation

There are three general principles governing radiological protection.

(a) Practices involving the use of sources of ionising radiation should produce sufficient benefit to offset the radiation detriment they cause **(justification)**.

(b) In relation to any particular source within a practice, all exposures must be kept as low as reasonably achievable (ALARA). This is the principle of **optimisation**.

(c) Doses to individuals must not exceed the limits specified in the regulations **(dose limitation)**.

The annual whole body dose limits specified in the European Communities (Ionising Radiation) Regulations, 1991 (Statutory Instrument No. 43 of 1991) are 20 mSv for occupationally exposed workers (over 18 years of age) and 1 mSv for members of the public.

It should be noted, however, that S I No 43 of 1991 will be revoked and replaced, from 13 May 2000, by regulations transposing into Irish law the Council Directive 96/29/Euratom of 13 May 1996 laying down the revised Basic Safety Standards for the protection of the health of workers and the general public against the dangers arising from ionising radiation. The above dose limits will nevertheless remain the same.

Pupils and teachers using sources of ionising radiation for demonstrations and/or experiments in science in Irish post-primary schools are considered, for radiation protection purposes, as members of the public. Provided that good laboratory radiation safety procedures are adhered to, the dose to any teacher or pupil over the course of a school year, will be unlikely to exceed a few tens of microsieverts. Exposures from this practice will, therefore, amount to only a small proportion of the annual dose limit of 1 mSv for a member of the public.

It is possible, therefore, for pregnant teachers to conduct experiments and demonstrations involving sources of ionising radiation without health concern. Furthermore, doses to other staff and pupils in the school and to visiting members of the public will be negligible. It must be emphasised, however, that pupils under 16 years of age must not be allowed to participate in demonstrations or experiments involving ionising radiation.

Note
For information on the types of sources of ionising radiation suitable for use in post-primary schools for teaching purposes the Radiological Protection Institute of Ireland should be contacted. The Institute is the competent authority in this country for sources of ionising radiation.

3.2 Do You Know

What famous Austrian physicist was a Senior Professor in the Institute for Advanced Studies in Dublin?

Erwin Schrödinger (1887-1961) did research in both physics and the history and philosophy of science at the Institute from 1941 to 1956. Schrödinger is best known for his work in quantum mechanics for which he shared the Nobel prize in physics with the English physicist, **Paul Dirac** (1902-1984) in 1933. During the war years he arranged for Dirac and other famous physicists, including Max Born and Wolfgang Pauli, to lecture in Dublin.

Radioactive contamination of the Irish Sea does not pose a hazard to food safety.

According to a report from the Radiological Protection Institute of Ireland the level of radioactive contamination of the Irish Sea is now so low that eating fish is quite safe. The report states that the estimated annual radiation dose to a heavy consumer of seafood, attributable to discharges from Sellafield, fell from 70 μSv in the early 1980s to just 2 μSv in 1995. By comparison, the annual average dose to a person in Ireland is estimated at 3000 μSv due to background radiation and 350 μSv due to medical procedures.

The weather in Paris contributed greatly to the discovery of radioactivity.

In February 1896 Henri Becquerel wrapped a photographic plate in a double thickness of black paper, coated the paper with 'bisulphate of uranium and potassium', and exposed it to the sun for several hours. When the plate was developed it showed the outline of the coating of the chemicals. Two days later, he prepared another plate in the same way to repeat the experiment, but the sky was cloudy for the next two days and the plate stayed in a drawer. On 1 March he developed the plate and again found the outline of the uranium salt.

Who won the first Nobel prize in physics?

The first Nobel prize in physics was awarded in 1901 to **Wilhelm Roentgen** for the discovery of X-rays in 1895.

What is the difference between a radiologist and a radiographer?

The radiologist is a qualified doctor who has specialised in radiation medicine. The radiologist will interpret the images from an X-ray machine. The radiographer works in the radiology department of a hospital. The radiographer will have training to degree level and will be responsible for the images produced in the department.

Who took the first clinical radiograph in Ireland?

The Lancet of February 1896 has a report of a lecture by Dr Cecil Shaw of Belfast, in which he unveiled a number of X-ray photographs. Other sources credit it to a Professor Barrett from the Royal College of Surgeons who, on March 16 of that year, used X-rays to locate a needle that had lodged in a girl's hand two years before.

Who was the first person to 'split the atom'.

Ernest Thomas Sinton Walton (1903-1995), Nobel Laureate in 1951, was a graduate of Trinity College Dublin, where he had taken the BA and MSc degrees. He obtained his PhD degree in 1931 under the supervision of Ernest Rutherford, staying on in the Cavendish Laboratory as the Clerk Maxwell Scholar. In 1932, he and John Cockcroft performed a crucial experiment in Cambridge. It was the first transmutation of an element by artificial means. This realised the age-old dream of alchemists, as the newspaper headlines were quick to point out. It was also the birth of the modern era of experimental nuclear physics using particle accelerators, now grown to gargantuan proportions at CERN and elsewhere. Walton returned to Trinity College Dublin not long after his epoch-making experiment, where he stayed as Professor of Experimental Physics for the rest of his career, until his retirement in 1974.

Are particles other than alpha, beta and gamma, ever emitted by radioactive elements?

Certain radio-nuclei can undergo radioactive decay by emitting complete carbon-14 nuclei. Such decay events are very rare but they have been detected experimentally. An example is ^{223}Ra ($T_{1/2}$ = 11.2 days) which decays to ^{209}Pb by emitting a ^{14}C nucleus with a branching ratio relative to alpha decay of approximately 10^{-9}.

Why is radio-caesium more dangerous than plutonium?

If you consumed food contaminated with plutonium, less than 0.01% of the plutonium is incorporated (i.e. crosses the gut), with 99.99% being excreted. On the other hand, if the food were contaminated with radio-caesium, virtually all of the caesium is incorporated. Minute traces of fallout plutonium are present in every human being's body. Typical levels are in the range 0.1 - 0.2 Bq per person.

3.3 Radon

Radon-222 is a naturally-occurring radioactive gas produced from the natural decay of uranium and thorium. It is a colourless, odourless gaseous element and it is the heaviest of the noble gases listed in the periodic table. It was discovered in 1900 by the German physicist **Friedrich Ernst Dorn** (1848-1916). It has a half-life of 3.8 days, decaying by the emission of alpha particles into an isotope of the element polonium.

Most people receive their greatest radiation exposure from radon, a naturally occurring radioactive gas that comes from the uranium in rocks and soils. The distribution of natural radioactivity varies largely with the underlying geology. Although all rocks and soils contain radioactive elements, the proportion they contain varies. Granite, for example, is high in radioactivity because it contains more uranium than most rock types. Building materials are another source of radioactivity. Old buildings constructed from granite, limestone or sandstone blocks are more radioactive than others

Radon is the biggest single source of natural background radiation, accounting for about half of most people's annual radiation dose. Present in all buildings, it is not considered hazardous unless its concentration builds up through inadequate ventilation. It is easily dispersed in the open air, but its ability to cause lung cancer over a long exposure period, when concentrated, can be an occupational hazard for miners.

When radon is produced it diffuses to the surface where it easily disperses in the atmosphere. However, a house tends to trap radon, resulting in elevated levels indoors. Levels rise in the winter as there is a lower air pressure indoors due to higher levels of heating. Houses that have very good draught proofing also have higher levels of radon. Radon can be removed by ventilation or by the installation of a sump under the floor of the house.

The presence of radon is measured in terms of its radioactivity per unit volume, that is in becquerel per unit volume. Surveys of the levels of radiation indoors have been done in many countries and the results of these are shown in Table 3.2.

Country	**Average conc./ Bq m^{-3}**
Canada	33
Denmark	68
Finland	90
France	76
Germany	49
Ireland	68
Italy	55
Japan	10
Netherlands	31
UK	20
USA	61

Table 3.2 *Average radon concentrations in various countries*

The International Commission for Radiation Protection, ICRP, has recommended a radiation exposure limit for nuclear workers of 20 mSv per year. In radon terms this equates to about 400 Bq m^{-3}. Ireland, from the above table, has a relatively high concentration level and in recent years much research has been carried out to determine the areas of highest concentration.

In 1992 the Radiological Protection Institute of Ireland initiated a national geographically-based radon survey of the country. Two radon detectors were sent to each of a percentage of householders in each of the counties of the country. One of these was to be placed in a main living room and the other to be put in a bedroom. After one year, the detectors were returned and analysed. The results of this survey are shown in Table 3.3.

County	No. of houses surveyed	Mean radon concentration/Bq m^{-3}	Maximum radon concentration/Bq m^{-3}	No. of houses >200 Bq m^{-3}
Carlow	194	123	1562	30
Cavan	180	67	780	5
Clare	742	88	1489	66
Cork	1211	76	1502	71
Donegal	487	69	512	18
Dublin	155	73	260	6
Galway	1213	112	1881	181
Kerry	932	70	1924	52
Kildare	480	90	1114	29
Kilkenny	181	100	717	16
Laois	334	83	565	17
Leitrim	145	60	433	6
Limerick	524	77	1102	41
Longford	132	75	450	8
Louth	124	112	751	14
Mayo	1184	100	1214	152
Meath	233	102	671	18
Monaghan	120	68	365	4
Offaly	286	68	495	7
Roscommon	235	91	1387	17
Sligo	270	145	969	54
Tipperary	852	79	1318	63
Waterford	162	119	1359	20
Westmeath	289	91	699	20
Wexford	469	99	1124	54
Wicklow	185	131	1032	24

Table 3.3 *Radon concentration in homes in Ireland by county*

Radon and the Human Body

Most of the knowledge of the effects of radon has come from the experience of uranium workers. These men have a long history of contracting lung disease as a result of exposure to radon at high concentration levels. It is this experience that leads to the assumption that increased levels in the home could be a problem. However the picture, even at this stage, is still not clear.

Computer modelling has shown that the lungs receive the highest dose from indoor radon, followed by the sensitive cells just below the surface of the skin. Inside the body the kidney receives a significant dose from radon daughters as it acts as a filter for these nuclei. Bone marrow also receives a high dose, owing to the presence of fat cells in the marrow. Their significance is that radon is 16 times more soluble in fat than in surrounding red marrow.

The fat cells act as sponges soaking up radon. When radon decays, some of the alpha energy can be lost from the fat cell, and thus irradiate the red marrow.

Radiation in general will bring about changes in the DNA of an individual. It can bring about a single-strand break in the DNA chain, a double-strand or even a compound break. In severe cases, the cell will be sterilised and will be unable to reproduce itself.

High doses can kill a large number of cells, resulting in severe and lethal changes to the body. At lower doses the DNA can be damaged and then repaired in such a way that the cell continues to divide but with an altered message in the DNA. This altered message may eventually result in a cell turning into a cancer. Thus any dose of radiation increases the risk of cancer. The main danger from high radon exposure is the increased risk of lung cancer. However it must also be borne in mind that natural background radiation can bring about a change in the DNA as well. However, the human body has a remarkable ability to heal itself.

3.4 Radioactivity Detectors

Radioactivity can be detected with a variety of different devices. Geiger-Müller counters are perhaps the most familiar, crackling when radioactivity is detected. Radioactivity can also be measured with Cherenkov counters, scintillation counters and solid-state detectors (semiconductor nuclear devices).

Geiger-Müller Counters

When radiation enters a Geiger-Müller tube, the gas inside it becomes ionised. An electric circuit within the counter to which the tube is attached generates an electric pulse every time ions are formed. The pulses generate sound in a loudspeaker or a reading on a meter. The first version of the device was invented in 1908 by the German physicist **Hans Geiger** (1882-1945), and in 1928 Geiger and **Walther Müller** produced an improved design - accounting for the name the Geiger-Müller counter that is in use today.

Cherenkov Counters

These counters are used to detect charged particles that are travelling at very high speeds. Cherenkov radiation is electromagnetic radiation with a maximum wavelength in the blue end of the visible spectrum which is produced when the speed of a particle in a particular medium is greater than the speed of light in that medium. For example, the most energetic beta particles emitted during the decay of ^{60}Co travel faster in water than does light and as a result a bluish glow is observed when these beta particles pass through a tank of water. The direction of maximum intensity of the emitted light can be used to calculate the speed of the beta particles.

Scintillation Counters

Scintillation counters are used to detect and count high-energy particles and radiation. They depend on the emission of flashes of light (scintillations) by certain substances on being excited by radiation. The light is allowed to fall on a sensitive photocell where electrons are emitted. These electrons are accelerated to an anode where each electron knocks out several ‘secondary electrons’ from the anode. These electrons are accelerated to a second anode where further electrons are produced. This process is repeated until there are enough electrons to constitute a current which can be amplified and thus operate a counting device. This arrangement of a photocell and a series of anodes at successively higher voltages is known as a photomultiplier. These counters are used to detect charged particles such as protons and cosmic rays, and ionising radiations such as gamma rays. A variety of scintillators have been investigated, the more important ones now in use being sodium iodide activated with thallium, various plastics, and certain organic liquids. The great advantage of scintillation counters over some other particle detectors is that they can count at the rate of the order of a million particles per second. They are useful for counting and measuring the energy of gamma rays, protons, alpha particles and others, since the brightness of the scintillation depends on the energy of the charged particle or the gamma ray photon.

Solid-state Detectors

These devices (also known as semiconductor nuclear detectors) consist of a reverse-biased p-n junction. Particles striking the device create electron-hole pairs in the depletion layer. These then produce pulses of current in a circuit. These pulses are amplified and measured by standard methods. Solid-state detectors may be kept at the temperature of liquid nitrogen (73 K, or –200 °C) to reduce the effect of thermal noise (interference from the effects of heat) and so improve their accuracy.

3.5 Nuclear Power

All around the world people use different sources of energy, and their requirements vary. Subsistence farmers struggling for a living in poor countries need energy to heat and light their homes and to cook their food, whereas in wealthy countries people use energy to run their cars and to chill their drinks.

Just as energy use varies dramatically between developed and developing countries, so too do the sources of energy. Fossil fuels produce about 90% of all the energy used in the world but almost all of this is produced and used by the industrialised nations. In developing countries, where each person uses only a fraction of the energy that we do, the story is very different. Firewood, charcoal, renewables, and animal and crop residues are the main energy sources for about half of the world's population - over 2.5 billion people. More people in the world depend on wood for energy than any other source.

Where they can afford it, developing countries are using oil and gas (especially where large gas reserves have recently been discovered) but they are also turning increasingly to nuclear power, renewables and coal. As the poorer nations of the world industrialise, and as world population grows, the amount of energy consumed will rise rapidly. The World Energy Conference estimates that the world will need at least half as much energy again by the year 2020, the bulk of the increase being in Asia, Africa and South America.

There is no perfect answer, no perfect energy source. Each source has advantages and disadvantages. So most governments build up a balanced energy programme that does not rely too heavily on any one source. Nuclear power is in use in many of the countries of the world, and often provides a significant percentage of the country's electricity.

Country	**Number of reactors in use**	**Percentage of country's electricity supplied**
Argentina	2	11
Belgium	7	60
Brazil	1	1
Bulgaria	6	45
Canada	16	14
China	3	<1
Czech Republic	4	19
Finland	4	30
France	59	78
Germany	20	32
Hungary	4	40
India	10	2
Japan	54	35
Korea	12	34
Lithuania	2	81
Netherlands	1	3
Romania	1	10
Russia	29	14
Slovak Republic	4	44
Slovenia	1	40
Spain	9	29
Sweden	12	46
Switzerland	5	41
UK	35	27
Ukraine	16	47
USA	107	20

Table 3.4 *Nuclear power reactors in operation and under construction, December 1997*

Nuclear Reactors

There are three main types of nuclear reactor in use today - Magnox, AGR (Advanced Gas-Cooled), and PWR (Pressurised Water Reactor). The fuel in Magnox gas-cooled reactors is natural uranium metal encased in a magnesium alloy can, and in this type of reactor the coolant reaches a maximum temperature of 400 °C. In the AGR, the fuel is uranium dioxide pellets in stainless steel tubes assembled in graphite. Because of the design of the fuel, the coolant reaches a temperature of 660 °C. This higher fuel temperature increases the amount of energy that can be extracted from the uranium. In the PWR, the fuel is uranium dioxide encased in zirconium alloy cans. This type of reactor is very compact, making it ideal for marine use. The maximum temperature reached is 324 °C.

Nuclear power stations emit few of the gases associated with acid rain and neither do they contribute to the enhanced greenhouse effect (except in small amounts during power station construction). Since 1973, the use of nuclear power has saved about 2.2 billion tonnes of coal. However, nuclear power is not without its environmental impact. The main issue faced by the nuclear industry is the disposal of radioactive waste.

Nuclear Waste

The main environmental issue faced by the nuclear industry is the safe disposal and storage of radioactive waste. Each year a nuclear power station produces around 100 m^3 of solid radioactive waste. Most of this waste is only slightly radioactive and can be stored without too many special arrangements. Very stringent rules apply though to the disposal of the other waste. Current government policy in European countries and the USA, is to dispose of most nuclear waste underground. Safety standards in underground repositories are extremely high and there are several barriers between the radioactive material and the outside world. The waste is contained within multi-layered metal and concrete vessels. Repositories are built in areas where local rocks allow little ground water to pass through them.

All waste has an effect on the environment, and most of it is unpleasant and/or dangerous. Domestic waste tips are a haven for rats and diseases, and many are potentially explosive. Industrial wastes contain substances (e.g. asbestos, heavy metal compounds) that can remain toxic forever. The volume of radioactive waste produced by the UK nuclear industry, for example, is very small compared with the volume produced from other sources.

Type of waste	Volume/10^3 m^3
Industrial solid waste	40 000
Industrial gaseous effluent	5500
Domestic rubbish	40 000
Coal mining waste	25 000
Radioactive waste	40

Table 3.5 *Volume of different types of waste produced in the UK*

Classification of Waste

Radioactive waste can vary enormously in the amount of radiation that it is giving out, and so it is categorised into low-level, intermediate-level, and high-level waste. In the UK, the organisation that processes and stores radioactive waste is British Nuclear Fuels Ltd (BNFL).

Low-level waste (LLW) includes solids and liquids like glassware, building materials, used protective clothing and air filters that might be contaminated with traces of radioactive materials. Brazil nuts contain as much natural radioactivity as a lot of low-level waste. Over 84% of radioactive waste is low-level. Most low-level waste is packed into 200 litre steel drums and stored in purpose-built concrete-lined vaults. When full, the vaults are covered with two metres of top soil and the site restored to nature.

Intermediate-level waste (ILW) consists of solid and liquid materials from power stations, fuel reprocessing and defence establishments and includes items such as reactor components, cladding from fuel rods, solidified sludges and resins from effluent-treatment plants. Intermediate-level waste is potentially dangerous. About 14% of radioactive waste is intermediate-level. BNFL encapsulates this waste inside stainless steel drums. At present these are stored in a silo at Sellafield (Cumbria, England), but they will eventually be stored in a purpose-built repository cut into hard rock deep underground.

High-level waste is concentrated waste that is produced when nuclear fuel is reprocessed. Even though it represents only about 1.5% of the total volume of waste, it contains about 99% of the radioactivity in the irradiated fuel. Most high-level waste is ‘vitrified’ (turned into glass blocks). This reduces its volume by about two thirds and aids its safe storage. High-level waste is initially in the form of a liquid which is hot and very radioactive. In the UK it is stored initially at Sellafield in stainless steel tanks cooled by cold water. Vitrification involves driving the water from the waste, and then turning it into glass blocks which will be stored in deep underground purpose-built vaults, where they will stay for 50 years.

A Natural Fission Reactor

For thirty years it was assumed that the first nuclear chain reaction to occur on Earth was that set up by **Enrico Fermi** (1901-1954) in Chicago in 1942. However, it has now been established that a natural reactor operated in a natural uranium deposit in west Africa 1.8 billion years ago.

Evidence for this came in an interesting way. Natural uranium from Gabon was exported to France; an examination of the isotopic content showed that the proportion of uranium-235 was slightly lower than normally found (0.717 per cent, compared to 0.720 per cent). This small difference was enough to arouse the interest of some of the scientists working on fuel processing, and they looked more closely into the composition of the ore. They found traces of the fission products of uranium in higher proportions than in normal uranium ore. Their suspicion was that, at some time in the geological history of the uranium, some of it had undergone a fission reaction. But how could a chain reaction have been established in natural uranium?

The seam of ore which was being extracted was unusually rich in uranium - up to 10 per cent. Geological conditions had conspired to accumulate large quantities in a small area. In addition, the proportion of uranium-235 would have been higher at an earlier date. The water of crystallisation of the minerals in the ore might have acted as a moderator.

It is now believed that a natural fission chain reaction must have taken place in the ore approximately 1800 million years ago. It may have run for 10^5 to 10^6 years, emitting a thermal power of tens of kilowatts (any greater power would have led to the evaporation of the water required as a moderator). In the course of its lifetime, it would have consumed a similar amount of uranium as a present-day power reactor consumes in a year.

Nuclear Accidents

No part of the power industry is free of risk. Coal mines, drilling rigs, gas mains, oil tankers and dams are vulnerable to accident. Nuclear power is no different and the nuclear industry’s safety record stands with the best.

The reactor in a nuclear power station contains a large amount of radioactive material. Properly maintained, there is no danger to anyone. However, in the event of a major accident, large amounts of radioactive material could escape - with consequent damage to the environment, and to animals, plants and people that come into contact with it. However, for radioactive material to escape from a modern reactor on a major scale, a whole chain of failures would have to take place. The water or gases used to cool the system would have to escape or fail, the fuel would have to overheat, the cans in which the fuel is held would have to melt and the fuel would have to break down to release radioactive products. After that, the radioactivity would still have to escape from the pressure vessel and the additional surrounding barriers.

In the UK, before a nuclear power station is licensed, its owners must show that it is safe and prove that the likelihood of uncontrolled radioactivity escaping is literally one in a million for every year of the reactor’s life. To achieve these results, everything, from earthquakes and fire to crashing aircraft, must be allowed for. Designers also assume that human operators can make mistakes, so all protective systems are duplicated or triplicated. Some systems can detect unsafe conditions and restore the plant to a safe condition automatically.

Chernobyl

The world's most devastating nuclear accident happened at Chernobyl, in the Ukraine, in 1986. In the early hours of the morning of 26 April of that year, there were two loud explosions that blew the roof off and completely destroyed the No. 4 reactor, releasing during the course of the following days, 6 to 7 tonnes of radioactive material, with a total activity of about 10^{18} becquerels, into the atmosphere. The discharge included over a hundred, mostly short-lived, radioisotopes, but iodine and caesium isotopes were of main relevance from a human health and environmental point of view. Contamination in the surrounding areas was widespread, with the half-life of some of the materials measured in thousands of years. Radioactive material from the plant was subsequently detected over practically the entire Northern Hemisphere.

For the next two weeks tonnes of material were dropped into the reactor to put out the fires, after which the reactor was entombed in concrete.

An 'exclusion zone', initially some 30 kilometres in radius was established around the site and about 116 000 people within it were evacuated to less-contaminated areas in the months following the accident. The exclusion zone was later extended to cover an area of 4300 km^2.

Large numbers of people involved in the initial clean up of the plant received average total body radiation doses of about 100 mSv - about five times the maximum dose permitted for workers in nuclear facilities. Average worldwide total body radiation dose from natural 'background' radiation is about 2.4 mSv annually.

During, and soon after the accident and the initial clean-up, at least 30 plant personnel and firefighters died from burns and radiation. In the eight years following the accident, a further 300 suffered radiation sickness, and there are possible links between the accident and increased numbers of thyroid cancers in neighbouring regions.

After ten years of detailed analysis by international experts, the principal causes of the accident are well understood. The accident occurred because of severe deficiencies in the design of the reactor, compounded by violations of operating and safety procedures.

The Chernobyl reactor was of the type known as an RBMK reactor, a type used only in the Soviet Union at the time (and one that would never have been licensed for use in many other countries in the world). It was a graphite-moderated water-cooled reactor in which the zirconium-niobium fuel rods were held in groups of 18 in an assembly which could be removed for refuelling through tubes from the top. The cooling water was circulated through 1680 tubes which passed into the reactor containment vessel, then through the graphite before passing out again to the steam separator and boiler that powered the turbines. Helium and nitrogen were used to prevent graphite oxidation and the containment building did not have to withstand high pressures. The fuel used was uranium dioxide enriched to 2% with uranium-235.

It had been decided that before the next planned shutdown, a test would be made to find out whether the remaining kinetic energy in a turbine and generator could be used to power the cooling pumps for the 40 seconds needed to start the emergency generators. Similar experiments had been done successfully on two previous occasions but some electrical changes had been made in the system. In order to prevent the emergency cooling system from interfering with the test, it was switched off in violation of safety regulations.

When the test on the generator started there were four pumps connected to the national grid and four to the remaining generator. After 35 seconds the voltage from the slowing generator fell sufficiently for its safety cutout to trip and disconnect the power to the four pumps. It is almost certain that the voltage did not hold up for the full 40 seconds because of the fact that the test began with the generator already running at much reduced power. Immediately the other four pumps started to vibrate, probably due to cavitation as the high temperature water from the steam condensate accelerated through the pumps. Cavitation would cause the pumps to fail and only froth to arrive at the reactor core. These pumps were then also switched off.

At this point a catastrophic surge in the heat output from 6% to 50% occurred in 10 seconds. A scram was ordered but the fully-withdrawn control rods could not be replaced in time. The fuel rods burst and the fuel mixed with the water which, turning to steam, lost any cooling ability. A thermal explosion resulted, rupturing the containment vessel and the reactor building, and dispersing hot radioactive material around the power site causing more than 30 fires, most of which were under control after four hours. Unsuccessful attempts were made to flood the reactor compartment and stop the graphite core fire but it was not until 10 days later, and after placing 5000 tonnes of lead and dolomite on the reactor by military helicopter, that the fire was brought under control.

The study of the accident by the experts from all over the world will endeavour to ensure that one like it will never happen again. The reactor's remains are now enclosed within a large concrete structure or 'sarcophagus' that was built in the months following the accident. One of the four original reactors at the site is still in operation. Between 1987 and 1991, a first stage of upgrading was performed on all other RBMK reactors to eliminate the design deficiencies, to improve shutdown mechanisms and to heighten safety awareness amongst staff.

Compared with other nuclear events, the Chernobyl explosion put 400 times more radioactive material into the earth's atmosphere that the atomic bomb dropped on Hiroshima. It is horrific though to note that the atomic weapons tests conducted during the 1950s and 60s all together are estimated to have put some 100 to 1000 times more radioactive material into the atmosphere than the Chernobyl accident.

3.6 Radioactivity in Medicine

Ionising radiation has two very different uses in medicine: diagnosis and therapy. Both benefit the patient but, as with the use of any form of radiation, the benefit must outweigh the risk to justify its use. Many people, including school pupils, have had an X-ray examination at some stage to aid diagnosis; however, the use of ionising radiation solely for that purpose is much less common. Physicians use this method if they cannot make a diagnosis without it. Radiation doses for this purpose are generally, but not always, low.

Much higher doses are required to treat malignant diseases or malfunctioning organs, sometimes in combination with other forms of treatment. A beam of radiation may be used to irradiate the affected part of the body or a fairly high activity radionuclide may be administered to the patient.

The use of X-rays for examining patients is called diagnostic radiology and the use of radionuclides for diagnosis or therapy is called nuclear medicine. When radiation beams are used to treat patients, the procedure is called radiotherapy.

Nuclear Medicine

For a diagnostic procedure in nuclear medicine, the patient is given a radionuclide in a carrying substance, such as a pharmaceutical, that is preferentially taken up by the tissue or organ under study; administration may be by injection, ingestion or inhalation. The radionuclide emits gamma rays.

Most of the diagnostic procedures make use of the radionuclide technetium-99; this has a half-life of 6 hours, gives off gamma rays with an energy of 0.14 MeV, can be conveniently prepared in a hospital, and readily labels a variety of carrying substances. A special detector called a gamma ray camera is used to observe how the organs or tissue behave or how quickly the radionuclide moves.

Detectors and Imagers

Gamma camera

The gamma camera was invented in 1957 by Anger. It consists of a large disc of sodium iodide which emits flashes of light when struck by gamma ray photons. The disc is typically 40 cm in diameter and 1 cm thick. To the disc are attached an array of photomultiplier (PM) tubes arranged in a geometric pattern. These tubes receive signals from each flash of light in the crystal. Because they are at different distances from it, the intensity will differ. This difference enables the coordinates of each flash to be determined and, with the aid of many flashes from different parts of the organ under study, permits an image to be reconstructed. Various types of lead collimators are also employed to improve image definition. The isotope most commonly used is ^{99}Tc with a photon energy of 140 keV.

Thermoluminescent dosimeter

Thermoluminescent dosimeters (TLDs) are devices that are replacing film badges for monitoring personnel, and are also used for monitoring patients' doses during therapy and for calibrating radiation sources in hospitals. The principle of the TLD is electron trapping. Certain materials, such as lithium fluoride (LiF) and calcium fluoride (CaF_2) can, when exposed to radiation, store part of the energy in electrons trapped in the isolated levels resulting from impurities in the material. These electrons cannot migrate and conduct, but if the crystal is heated they can gain thermal energy which then enables them to lose all their additional energy as light. The amount of light emitted is a measure of the number of trapped electrons, and hence the original exposure. TLDs are inexpensive, requiring only small quantities of LiF, which can be reused. They can be made very small to measure exposure within the body at sites such as tumours.

Fluoroscope

Fluoroscopes make use of fluorescent substances - ones in which the energy of radiation moves electrons into the excitation band. They then immediately re-emit this absorbed energy as light. Dynamic imaging of the body with X-rays was carried out for many years by use of a glass screen coated with zinc-cadmium sulphide, the fluoroscope. Direct viewing is no longer the practice because of the high exposure involved. Modern fluoroscopy employs an image intensifier tube instead. Computer image-enhancement techniques are also used.

X-ray tomography

X-ray tomography (CT scanning) is the major recent development in X-radiography. It enables detailed pictures to be produced of particular planes of the body. Those produced with conventional radiography are only longitudinal. The method is particularly good in examining soft tissue, as it can detect small differences in attenuation of X-rays. The tomographic technique removes the interference from structures which are out of the chosen plane. As it is measuring the amount of X-rays absorbed quantitatively, this information can be stored digitally by computer, recalled and processed to give a selection of reconstructed images.

Magnetic resonance imaging

Magnetic resonance imaging (MRI) is a completely new method of imaging, being based on the phenomenon of nuclear magnetic resonance (NMR), a process in which protons interact with a strong magnetic field and radio waves to generate electrical pulses that can be processed in a similar way to the computed tomography for X-rays. The images produced are in some ways similar to those of the CT scanner but without the radiation hazard. For certain parts of the body, including brain, spinal cord and heart, and in certain diseases, MRI is superior to CT scanning. It offers an early and easy diagnosis of many conditions, including cancer, and is a help in planning the most effective treatment. The latest instruments have an additional capability: using the technique of magnetic resonance spectroscopy (MRS), they can be used to detect and measure chemicals within the body without having to take any samples. Questions that can be answered include: has the heart muscle been irreversibly damaged after a heart attack, is a cancer responding to radiotherapy or chemotherapy, etc.?

3.7 Other Applications of Radioisotopes

Heart pacemakers often contain a nuclear-powered battery. In a typical arrangement, a small quantity of plutonium decays by alpha emission and the heat generated is converted to electricity by a semiconductor thermopile.

Gamma cameras are used in hospitals to detect the position of radioactive tracers (e.g. ^{99}Tc) injected in to the patient's bloodstream, helping in the diagnosis of a wide range of illnesses and lesions.

Radioisotopes find wide industrial application, e.g. in the measurement of the thickness of metal tyre cord using a ^{90}Sr source, in the inspection of solid objects (e.g. jet engines) to find possible flaws in a similar way to medical X-ray examination, etc.

Radioisotopes are used in the irradiation of food to kill parasites, bacteria and pest insects, and to inhibit germination, sprouting or premature ripening. The food itself does not become radioactive.

Strong gamma sources (e.g. ^{60}Co) are employed for the sterilisation of medical equipment.

Radioisotopes are used to label biochemicals in order to trace chemical processes in living organisms. These techniques have applications in important medical fields such as immunology.

The decay of ^{14}C is used in radiocarbon dating, where the measurement of the ratio of ^{14}C to ^{12}C permits the age of carbon-bearing materials to be determined. In Ireland, the technique has found application in the dating of the important Céide Fields Neolithic settlement, which is now known to be about 5500 years old.

The relative proportions of different radioactive isotopes are used in geology to determine the age of rocks. A similar approach using different isotopes enables sedimentologists and marine scientists to establish the chronology of sedimentary layers and reconstruct the deposition history of sediments in lakes, coastal zones and deep oceans.

3.8 The Irish Radium Institute (1914-1952)

It was in Dublin that radium was first divided into small quantities in capillary tubes, placed in needles, and used to treat cancer growths by the implantation or interstitial method. The Royal Dublin Society was instrumental in the formation of an organisation called the Irish Radium Institute in 1914. In 1900, **John A. McClelland** had been appointed Professor of Physics in University College, Dublin. McClelland was interested in the recently-discovered element radium, and in 1903 the RDS purchased 60 mg of radium bromide to enable him to continue his researches on the radiation and emanations from radium. By 1910, Dublin's physicists and chemists were well alerted to the properties of radium and its potential as a source of radioactivity. **John Joly** (1857-1933), Professor of Geology in TCD, recorded the following: 'A case of facial rodent ulcer was brought to my notice by one of the Junior Surgeons - Dr. Walter C. Stevenson - as intractable to treatment. I had at that time in my possession 4 mg of radium bromide sealed in two glass tubes. As the result of exposure to the rays emitted by these tubes, the rodent ulcer rapidly healed.'

Early in 1914 the Committee of Science and its Industrial Applications of the Royal Dublin Society had under consideration, at the suggestion of Prof. Joly, the desirability of approaching the Council of the Society with a view to the establishment of a Radium Institute in Ireland. As a result, the Science Committee passed the following Resolution:

'The Committee unanimously recommend the Council to vote the sum of £1,000 to purchase radium in addition to the Society's present stock. The Committee further unanimously recommend the Council to take steps to found a Radium Institute, and appoint the following Committee to draw up a report, formulating a scheme for a Fête, to be held in the year 1915: Prof. Joly, Prof. McClelland, Prof. Purser, Dr Adeney, Dr McWeeney, Prof. A.F. Dixon, Sir Howard Grubb, and the Honorary Officers.'

'For almost 40 years thousands of patients in Dublin hospitals benefited from the radon produced by the Society. One can only realise the immense value of the work of the Irish Radium Institute from a study of the many reports submitted by the medical practitioners concerned and published in the Scientific Proceedings of the RDS.'

3.9 Worked Examples

1. Complete the following nuclear reactions.

(a) ${}^{238}_{92}U + {}^{1}_{0}n \rightarrow ?$

(b) ${}^{27}_{13}Al + {}^{1}_{0}n \rightarrow {}^{4}_{2}He + ?$

(c) ${}^{53}_{24}Cr + {}^{4}_{2}He \rightarrow {}^{1}_{1}H + ?$

(d) ${}^{27}_{12}Mg + {}^{1}_{1}H \rightarrow {}^{1}_{0}n + ?$

(a) ${}^{238}_{92}U + {}^{1}_{0}n \rightarrow {}^{239}_{92}U$

(b) ${}^{27}_{13}Al + {}^{1}_{0}n \rightarrow {}^{4}_{2}He + {}^{24}_{11}Na$

(c) ${}^{53}_{24}Cr + {}^{4}_{2}He \rightarrow {}^{1}_{1}H + {}^{56}_{25}Mn$

(d) ${}^{27}_{12}Mg + {}^{1}_{1}H \rightarrow {}^{1}_{0}n + {}^{27}_{13}Al$

2. Calculate the number of alpha decays and beta decays involved in each of the following transformations.

(a) ${}^{232}_{90}Th \rightarrow {}^{208}_{82}Pb$

(b) ${}^{237}_{93}Np \rightarrow {}^{209}_{83}Bi$

(c) ${}^{238}_{92}U \rightarrow {}^{206}_{82}Pb$

(d) ${}^{235}_{92}U \rightarrow {}^{207}_{82}Pb$

Mass number has decreased by 24 (232 - 208) $\Rightarrow$ 6 α-emissions

$\Rightarrow$ decrease in atomic number of 12

But, atomic number has decreased by 8 (90 – 82)

$\Rightarrow$ 4 β-emissions (each of which causes an increase of 1 in atomic number)

(b) 7 α-emissions + 4 β-emissions

(c) 8 α-emissions + 6 β-emissions

(d) 7 α-emissions + 4 β-emissions

3. A radioactive isotope ${}^{x}X$ decays by alpha decay followed by two beta decays. What is the resulting isotope?

$$ {}^{x}_{y}X \rightarrow {}^{x\text{-}4}_{y\text{-}2}Y \rightarrow {}^{x\text{-}4}_{y\text{-}1}Z \rightarrow {}^{x\text{-}4}_{y}X $$

The isotope formed has the same atomic number as the original, i.e. it is the same element, but with a mass number smaller by 4.

4. A radioactive source has an activity of 8×10^4 Bq and a half-life of 20 minutes. Calculate the activity of the source after one hour.

1 hour = 3 half-lives,

$\Rightarrow$ activity is one-eighth $[(1/2)^3]$ of original, i.e. 1×10^4 Bq

5. The energy per fission of ${}^{235}U$ is 175 MeV. Calculate the energy released during the fission of 1 g of uranium-235. (Avogadro constant = 6.0×10^{23} mol^{-1}; 1 mol uranium-235 = 235 g; $e = 1.6 \times 10^{-19}$ C.)

235 g of uranium-235 contains 6×10^{23} atoms

$\Rightarrow$ 1 g contains $6 \times 10^{23}/235$ atoms = 2.55×10^{21} atoms

Total energy released is

$$E = 2.55 \times 10^{21} \times 175 \text{ MeV}$$

$$= 4.5 \times 10^{23} \text{ MeV}$$

$$= 4.5 \times 10^{23} \times 10^{6} \times 1.6 \times 10^{-19} \text{ J}$$

$$= 7.2 \times 10^{10} \text{ J}$$

Appendix I Chronology of Some Modern Physics Discoveries

Year	Discoverer	Nationality	Invention/discovery
1836	Callan	Irish	Induction coil
1855	Geissler	German	Improved vacuum pump
1859	Geissler	German	‘Rays’ emitted from cathode in discharge tube
1869	Hittorf	German	Effect of magnet on rays
1876	Goldstein	German	Investigations of electric discharges in gases
1888	Hertz	German	Photoelectric effect
1891	Stoney	Irish	Introduced the name ‘electron’
1893	Lenard	German	Investigated cathode rays in air
1895	Perrin	French	Cathode rays are negatively charged
1895	Thomson	English	Electrons in all atoms, leading to the ‘plum pudding’ model of the atom
1895	Roentgen	German	X-rays
1896	Becquerel	French	Radioactivity of uranium
1897	Thomson	English	Measured *e*/*m* for cathode rays
1897	Rutherford	British	Distinguished different types of radiation from uranium
1898	Curies	French	Radium and polonium
1899	Becquerel	French	Magnetic deflection of alpha and beta rays
1899	Lenard	German	Photoelectric emission due to electrons
1900	Planck	German	Introduced quantum theory of radiation
1900	Becquerel	French	Showed beta rays to be identical to cathode ray particles
1900	Villard	French	Gamma rays
1903	Rutherford	British	Radioactive processes involve transmutation of elements
1903	Crookes	English	Principle of scintillation counter
1905	Einstein	German	Quantum explanation of photoelectric effect
1906	Richardson	English	Thermionic emission
1909	Rutherford	British	Structure of alpha particles
1911	Millikan	American	Electron charge
1911	Rutherford	British	Nuclear model of the atom
1913	Geiger	German	Geiger counter
1913	Bohr	Danish	Electronic structure of atoms
1914	Moseley	English	Atomic number
1915	Einstein	German	General theory of relativity
1919	Rutherford	British	Transmutation of elements
1923	Compton	American	The photon has momentum
1932	Cockcroft & Walton	English & Irish	Split lithium atom
1932	Chadwick	English	Neutron discovered
1934	Joliot-Curie	French	Induced (artificial) radioactivity
1938	Hahn	German	Nuclear fission
1942	Fermi	Italian	Nuclear reactor

* The countries listed are not necessarily those in which the work was carried out.

MODULE 9

Particle Physics

CONTENTS

CHAPTER 1 BACKGROUND

1.1 Introduction

The history of modern particle physics is fascinating and has followed a pattern which, in a sense, can be traced back to the time of the ancient Greeks. Examination of this pattern is very interesting in the light of the more recent history of the subject.

The search for the ultimate structure of matter is a noble and antique discipline. By the sixth century BC, Greek philosophers had proposed four basic elements - earth, air, fire and water - to be the ultimate constituents of matter. This hypothesis was not based on proper scientific methods, as we now know them, but on philosophical methods of speculation and logical reasoning. It could be said that the first step on the road to modern elementary particle physics was the discovery of fire. The first cosmological speculations arose simply from the observation of the regular patterns of day and night.

The concept of four basic elements was followed by an idea originated by the Greek philosopher **Empedocles** (c. 490 BC - c. 430 BC), and further developed by another Greek philosopher **Democritus** (c. 470 BC - c. 380 BC), that the four basic elements were composed of atoms (from the Greek *atomos*, meaning indivisible). These atoms were thought to be endowed with shapes characteristic of the particular element and to be in continuous random motion in a vacuum. This very fundamental idea of the indivisibility of some small constituent of matter was not further advanced for a considerable period of time. This happened because the possibility of the transmutation of one substance into another became apparent and considerable time and effort were invested in the pursuit of alchemy, the shady relative of chemistry with ingredients of philosophy, mysticism and chemistry. One of the main aims of the alchemists was the conversion of base metals into gold. It is not surprising that it was studied with such fervour!

One thing that became very clear as time progressed and elements such as gold, silver and mercury were discovered was that their existence could not be explained in terms of the four classical elements. However, chemistry and alchemy remained confused disciplines until the time of the French chemist **Antoine Laurent Lavoisier** (1743-1794). He did more than any other scientist to give a precise meaning to the Greek notion of the chemical element. The English physicist **John Dalton** (1766-1844), a founding father of atomic theory, wrote in 1810, ‘We have endeavoured to show that matter, although divisible in an extreme degree, is not infinitely divisible.’ More cautious remarks of this nature were to be echoed by the Russian physicist **George Gamow** (1904-1968) in his book *One, Two, Three, Infinity,* which was published in 1947.

1.2 The Cycle of Development

In the last two centuries, our progress has followed a remarkably cyclical pattern. Four times scientists have arrived at the ‘ultimate’ description of matter, only to find themselves proved wrong or, at the very least, their theories found to be incomplete. It is not yet clear whether the current theory has finally succeeded in uncovering the basic constituents of matter, but it is illuminating to look briefly at the history of the search before considering it in more detail. Each time the search followed a logically similar pattern which can be summarised as a series of steps.

Step 1

A new level of structure is revealed or deduced and candidates emerge as the ultimate constituents of matter.

Step 2

A theoretical system evolves to explain the nature of the particles and their interactions - the 'elementary particle physics' of its day.

Step 3

There is a veritable 'population explosion' of the supposed elementary particles - there seem to be too many of them for all to be truly fundamental.

Step 4

Scientists uncover a regular pattern or symmetry among the elementary particles and they are grouped into patterns or tables. They display systematic variations of chemical or physical properties.

Step 5

Experiments hint at further structure in the elementary particles themselves. They display the behaviour of composite systems made up of even simpler things. When eventually the existence of the simpler entities, the new 'elementary' particles, is established, it is time to return to Step 1.

1.3 The First Cycle

The science of chemistry evolved very rapidly to explain the interactions amongst atoms, the second generation of fundamental species (if the early Greek elements of earth, air, fire and water are included). Chemical techniques became more sophisticated and revealed an increasing number of atomic species. Lavoisier's list of 30 chemical elements was doubled by the mid-nineteenth century. Could nature employ so many fundamental building blocks of matter?

There then followed the emergence of a pattern; in 1869, the Russian chemist **Dmitri Mendeleev** (1834-1907) conceived the idea of a periodic table of chemical elements in which there was a clear relationship between atomic mass and periodically recurring chemical and physical properties. On the basis of his periodic table Mendeleev accurately predicted the properties of elements not even discovered at that time. Many scientists realised that the atom could not be a fundamental entity. Each one of this large family of atoms had its own individual characteristics such as mass, size and chemical affinity. Elements when heated produced characteristic spectra of light and this was to provide a very powerful tool, that of atomic spectroscopy. Atoms, far from being fundamental and indivisible, were in general made from many parts and structured in a very ordered manner.

The discovery of the first sub-atomic particle, the electron, arose as a result of the study of electrical discharges in low pressure gases. In 1858, streams of faint blue light were seen to emerge from the cathode of a discharge tube as the gas pressure was reduced. These 'cathode rays' were investigated in great detail from 1894-1897 by the English physicist **J.J. Thomson** (1856-1940). He confirmed an earlier suggestion by another English physicist **William Crookes** (1832-1919) in 1879 that they were a stream of negatively charged particles and measured the ratio of their charge to mass in 1897. This he found to be a quantity which did not vary with experimental conditions. He concluded that the particles, named electrons, were present in all matter and that the charge to mass ratio was a characteristic of the electron itself. (The name electron was coined in 1891 by the Irish physicist **George Johnstone Stoney** (1826-1911), then Vice-President of the Royal Dublin Society.)

In 1910, the British physicist **Ernest Rutherford** (1871-1937) conducted experiments in which alpha particles were fired at a thin gold leaf. They showed that, while almost all of the alphas passed straight through, a few of them actually recoiled back from the target. In the words commonly attributed to Rutherford, ‘It was almost as incredible as if you fired a 15-inch shell at a piece of paper and it came back and hit you.’ This result showed that the mass of the atom was concentrated in a tiny nucleus and the rest was largely empty space. Rutherford was able to calculate that the diameter of this positively charged nucleus, which contains almost all of the mass, was less than 1/10000 of the atomic diameter. The remaining mass is made up by the negatively charged electrons which were thought to circle the nucleus as planets circle the sun.

Once more, events had set the proceedings back to Step 1.

1.4 The Second Cycle

The structure of atoms was explained by the new science of atomic physics. The term orbiting, as applied to atomic electrons, gave rise to problems in a classical model of the atom because it implied that the electron was accelerating. Classical electromagnetic theory shows that accelerating charges emit electromagnetic energy, so that the electrons would lose energy, be pulled towards the nucleus and the atom would collapse.

Atoms were known to be neutral and stable. Rutherford had found his alpha particle scattering measurements to be consistent with the existence of a tiny positively charged atomic nucleus containing almost all of the mass of the atom. All these facts were consistent with some model of the atom in which the negatively charged electrons were bound to the positively charged nucleus by the electromagnetic force. Then, there were the characteristic emission spectra of atoms such as the well-known Balmer series of lines in the spectrum of atomic hydrogen…?

The intensity of the electromagnetic radiation from a cavity radiator, such as a small hole in an incandescent tungsten tube, depends only on the Kelvin temperature and not on the material of the cavity walls. By the early 1900s, it had become clear that the intensity of the emitted radiation varied with wavelength in a way that was inexplicable using classical theory. The German physicist **Max Planck** (1858-1947) introduced the startling concept that electromagnetic radiation is emitted from, and absorbed by, the cavity walls in concentrated bundles or quanta, in a successful attempt to formulate a theory which would explain the form of this radiation spectrum. He proposed that the size of a quantum was equal to a constant, now known as the Planck constant, *h*, multiplied by the frequency of the wave. He adjusted the value of this constant to give the best fit to the experimental data. In 1905, **Albert Einstein** (1879-1955) proposed that the energy of the radiation itself was quantised and this bold concept was successfully applied to explain the photoelectric effect.

A novel, if only temporary, solution to the problem of the structure of the atom was provided by the Danish physicist **Niels Bohr** (1885-1962) in his atomic model of 1913. Essentially, it said that electrons could move only in certain ‘orbits’ without radiating energy - neatly side-stepping the problem! The emission of radiant energy of discrete wavelengths corresponded to the transition of an atomic electron from one allowed orbit to another of lower energy and the frequency of the emitted radiation, multiplied by the Planck constant, is simply equal to the difference in the energies of the two levels. This, in essence, was the proposal that atomic energy levels should be quantised. Bohr then used a simple classical calculation to obtain the energy values corresponding to a given orbital radius for an electron which circles the nucleus of a hydrogen atom.

It proved necessary to assume also that the orbital angular momentum associated with the circling electron was a variable which could only assume values that were integral multiples of $h/2\pi$ ($\hbar$), i.e. that angular momentum was quantised. From his calculations, Bohr was able to obtain discrete values for the energies of the ground and excited states of the hydrogen atom in terms of known constants. Thus, not only was the origin of the discrete line spectra explained, but Bohr was able to obtain values for the wavelengths of the spectral lines of atomic hydrogen which accurately fitted the experimental observations. His model was not however successful when applied to other atoms and a more comprehensive theory was needed.

A new and revolutionary methodology, quantum mechanics, was developed in the mid to late 1920s. Applied to the structure of atoms, quantum mechanics was highly successful but not everyone embraced it wholeheartedly by any means. An excerpt from a letter sent from Einstein to the German physicist **Max Born** (1882-1970) in December 1926 reads, 'Quantum mechanics is very impressive. But an inner voice tells me that it is not yet the real thing. The theory produces a good deal but hardly brings us closer to the secrets of the Old One. I am at all events convinced that He does not play dice.' The last sentence is a reference to the fact that quantum mechanics deals not with certainties but with probabilities.

In 1925, a startling suggestion was put forward. It was proposed that the electron itself has an intrinsic angular momentum, which has the same value for all electrons, independent of their environment or state of motion. It is as much a characteristic property of the electron as are its charge and mass. An electron with an intrinsic angular momentum has a corresponding intrinsic magnetic moment and behaves like a minute bar magnet in a magnetic field, behaviour which was first observed in the Stern-Gerlach experiment. This property was also deduced from observations of some very closely spaced pairs of lines in atomic spectra.

The evidence pointed very strongly to a value of $^1/_2\,\hbar$ for the electron intrinsic angular momentum. The value of $\hbar$ is precisely the unit or 'quantum' of orbital angular momentum which Bohr had used in formulating his model for the hydrogen atom. As already seen, orbital angular momentum is an integral multiple (1,2,3…) of the fundamental unit $\hbar$. The fact that intrinsic angular momentum is $^1/_2\hbar$ and not an integral multiple of $\hbar$ facilitated a clear distinction between effects due to orbital and intrinsic angular momenta.

To reflect this additional intrinsic property of the electron, it was assigned a new quantum number called spin (see below). In terms of the basic unit or quantum of angular momentum $\hbar$, the electron spin, s = $^1/_2$. Quantum mechanically this means that the electron spin can have only two orientations relative to the axis in space which is defined when a magnetic field exists.

A basic law of atomic physics, the Pauli exclusion principle formulated by the Austrian physicist **Wolfgang Pauli** (1900-1958) in 1924, says that no two electrons in an atom can exist in the same quantum state. They can not, for example, have the same spin, the same value of orbital angular momentum and the same energy and, since all these variables are quantised, we can simply say that they cannot have the precisely same quantum numbers. Electrons and other subatomic particles which behave in this way are called 'fermions' and this will have some far-reaching consequences in the exotic world of the quark. More to the point here, it determines the electronic structure of atoms. In the simplest instance, it means that the ground state of an atom, where the total angular momentum is zero, can hold only two electrons, one 'spin up' and the other 'spin down'. The electronic structure of all atoms can be determined using the Pauli principle.

In 1928, the English physicist **Paul Dirac** (1902-1984) derived a relativistic wave equation and succeeded in predicting all of the observed electron properties, including its spin. Spin angular momentum is a purely quantum mechanical phenomenon and has no classical analogy. In the context of the stationary electron, there cannot be an angular momentum, in the classical sense, because the electron is a point particle and does not have a moment of inertia. In the years which have elapsed since 1925, it has become clear that all subatomic particles have spin.

Many of the properties of matter could now be explained. The value of the mass number A was used to define each specific element in the periodic table, where A was defined, at that time, as the integer nearest to the atomic mass measured in unified atomic mass units, u (where 1u = 1.661×10^{-27} kg). Also, each atomic nucleus was found to have a definite value of electric charge which was always an exact multiple of the charge on the electron. This multiple, the atomic number Z, was used together with A to classify all existing chemical elements. Z could be measured by alpha particle scattering experiments or found from measurements of X-ray spectra. There still remained the embarrassing problem of the large number of atomic species and as far back as 1913 Thomson discovered two types of neon, chemically identical but differing in mass. More of these so-called 'isotopes' were discovered and with them came the realisation that the number of different isotopes (i.e. nuclei) far exceeded the number of chemical elements. Each isotope had a unique pair of A and Z values. Isotopes of a given element characterised by its Z value had different A values. A systematic pattern was once again emerging.

Further indications of the complex nature of nuclei were seen in the phenomenon of radioactivity. Evidence of the occurrence of natural radioactivity, the spontaneous transformation of certain naturally occurring nuclei into different nuclei, was first seen by the French physicist, **Henri Becquerel** (1852-1908) in 1896. In one such transformation, alpha decay, the A value of the parent nucleus is reduced by 4 and the Z value by 2 as a helium nucleus is emitted.

$$^{A}_{Z}\mathrm{X} \rightarrow {}^{A-4}_{Z-2}\mathrm{Y} + {}^{4}_{2}\mathrm{He}$$

X is termed the 'parent' nucleus and Y the 'daughter' nucleus.

Perhaps the most significant hint was to be found in gamma decay, which can accompany alpha decay. Neither A nor Z is changed in this process, which is the emission of very high energy electromagnetic radiation from the nucleus. The frequency of the radiation is characteristic of the daughter nucleus, which hints at the possibility of internal structure within the nucleus. Has the nucleus, like the atom, energy levels and is it, too, a composite object, structured in an ordered manner according to the rules of quantum mechanics and whose constituents obey the Pauli principle?

In 1932, the discovery of the neutron by the English physicist **James Chadwick** (1891-1974) marked another return to Step 1 of this cyclical history. Neutrons and protons were evidently the constituents of all atomic nuclei and so the science of nuclear physics was born. A certain complacency set in. Surely the basic constituents of matter had finally been discovered. They were the proton, the neutron, and the electron. (A is now defined as the total number of protons and neutrons in a nucleus and Z is defined as the number of protons.)

1.5 The Third Cycle

The Conservation of Energy and Momentum in Nuclear Reactions

In classical physics, mass and energy are separately conserved. In an explosion, if all the debris including even gaseous material were to be recoverable, it would be found that the number of individual atoms had not changed. In 1905 however, Einstein in his theory of Special Relativity had predicted that mass itself is a form of energy and that a stationary mass m has a 'rest' energy of mc^2 where c is the velocity of light in a vacuum. In nuclear reactions such as radioactive decay, where one nucleus is transformed into another, it is the conservation of total (i.e. rest + kinetic) energy together with the conservation of momentum that governs the process.

It is customary in atomic, nuclear and particle physics to use the electron volt as a unit of energy, because the joule is a very large unit in this context.

The energy equivalent of a mass of 1u can be obtained as follows.

$E = mc^2$

$= 1.661 \times 10^{-27} \times (2.998 \times 10^8)^2$

$= 1.493 \times 10^{-10}$ J

Even this is a large amount of energy in the present context as the following example shows.

Example 1.1

Calculate the energy released in the alpha decay of uranium-238 to thorium-234.

$$^{238}_{92}\text{U} \rightarrow {}^{234}_{90}\text{Th} + {}^{4}_{2}\text{He}$$

Atomic masses are 238.050784 u, 234.043593 u and 4.002603, respectively.

Mass difference, Δm = 238.050784 – 238.046196

= 0.004588 u

Energy released, $\Delta E = \Delta mc^2$

$= 4.588 \times 10^{-3} \times 1.493 \times 10^{-10}$

$= 6.85 \times 10^{-13}$ J

To present this in its more customary units, we can express it in electron volts (eV), where 1 eV is the change in potential energy of 1 electron when it moves through a potential difference of 1 volt. Since work = charge x potential difference and the charge on an electron is 1.60×10^{-19} C

1 eV = $1.60 \times 10^{-19} \times 1.0$

$= 1.60 \times 10^{-19}$ J

Thus the energy released in the above reaction is

$$\Delta E = \frac{6.85 \times 10^{-13}}{1.60 \times 10^{-19}}$$

$= 4.28 \times 10^6$ eV

= 4.28 MeV

This energy, sometimes called the disintegration energy, is converted to kinetic energy of the decay fragments.

To give a convenient conversion from mass, measured in atomic mass units, to energy, we can use the fact that a mass of 1 u is equivalent to 931.5 MeV of energy. This can easily be verified.

The decay process used in Example 1.1 is a good example of the conservation of charge, one of the universally obeyed conservation laws of nature.

Z = charge in units of the electron charge e.

Z(U) = 92

Z(Th) + Z(He) = 90 + 2 = 92. Charge is conserved.

An elementary particle carrying a charge which is a fraction of the electronic charge has not, so far, been detected experimentally.

Note

Assuming that the uranium nucleus decays at rest, the thorium nucleus must recoil in the direction opposite to the outgoing alpha particle so that momentum is conserved in the reaction. Almost all of the energy is taken up by the alpha particle because the thorium is so much more massive.

Example 1.2

Show, in the above decay, that the alpha particle takes up almost all the available energy.

At an energy of 4.28 MeV, an alpha particle behaves like a classical particle, with momentum = mv and kinetic energy = $\frac{1}{2}mv^2$.

Let M and V be the mass and velocity of the thorium nucleus.

The uranium decays at rest so that the two decay products must leave in opposite directions to conserve momentum.

$$U$$
$$Th \bullet \leftarrow \circ \rightarrow \bullet \alpha$$
$$MV = mv$$

$$v = \frac{MV}{m}$$

$$\frac{{}^1/_2 mv^2}{{}^1/_2 MV^2} = \frac{M}{m}$$
$$= \frac{234}{4}$$
$$= 58.5$$

The kinetic energy of the alpha particle is 58.5 times that of the thorium nucleus.

Alpha particles are seen to be emitted with a single energy, characteristic of the particular decay mode. The only exceptions are cases where the daughter nucleus may be formed in one or more excited states as well as in the ground state and in that case alpha particles of different energies are seen. Gamma radiation is subsequently emitted as the nucleus returns to a lower energy state. The wavelength of the gamma radiation can be calculated from the difference in the energies of the two states.

This unique emission energy always occurs when the emitted particle is one of only two end products of a reaction (i.e. when there is a *two-body final state*) because of the conservation of both momentum and energy.

Nuclear Beta Decay and the Discovery of the Neutrino

Early observations of so-called beta decay, in which an electron was emitted from a nucleus, showed that the electron could have a wide range of energies up to a maximum value characteristic of the decay. In the well-known decay of carbon-14, this maximum value is 0.156 MeV.

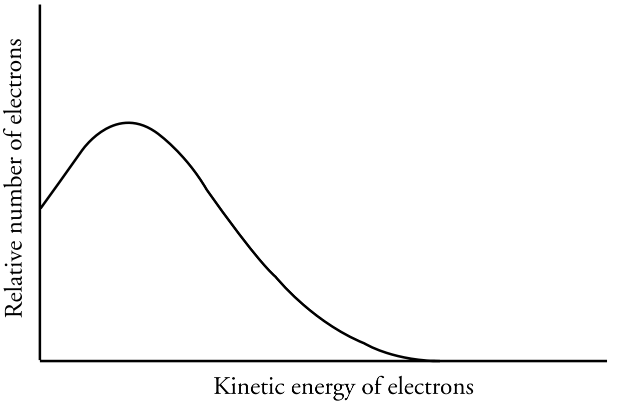


Fig. 1.1 *Typical beta decay spectrum*

Example 1.3

If the beta decay of carbon-14 takes place by the process shown below, calculate the kinetic energy of the emitted electron.

$$^{14}_{6}C \rightarrow {}^{14}_{7}N + {}^{0}_{-1}e$$

Atomic masses of the neutral atoms are 14.003241 u and 14.003074 u, respectively.

$$\Delta m = 1.67 \times 10^{-4} \text{ u}$$
$$1 \text{ u} = 1.493 \times 10^{-10} \text{ J}$$

Disintegration energy

$$\Delta E = 1.67 \times 10^{-4} \times 1.493 \times 10^{-10}$$
$$= 2.49 \times 10^{-14} \text{ J}$$

Or, expressed in MeV

$$\Delta E = 1.67 \times 10^{-4} \times 931.5$$
$$= 0.156 \text{ MeV}$$

The disintegration energy appears almost entirely as the kinetic energy of the electron.

This is precisely the value expected if the electron is one of just two end products of the decay. However, experimentally, the emitted electron is seen to have a whole range of energies up to this value, so that this cannot be true. In fact two things emerge from these observations.

a) Since conservation of energy and momentum permits only a single way of sharing the available energy between two decay products, either the laws are incorrect or there are more than two particles involved.

b) Any additional particle must have a mass of, or close to, zero because the measured maximum electron energy agrees within experimental error with the value predicted by assuming only two decay products.

In 1931, Pauli made the inspired suggestion that an additional unseen particle was also emitted by a beta-decaying nucleus. The particle had to be uncharged in order to conserve charge and of extremely small, if not zero, mass to explain the form of the electron energy spectrum. The particle was later named the neutrino (ν) by the Italian physicist **Enrico Fermi** (1901-1954), who by 1934 had formulated a theory of beta decay. As for the neutrino ν, it was indeed so weakly interacting that it was not observed until 1956 by Cowan and Reines who looked for interactions initiated by the very large flux of neutrinos from the beta decay of nuclear reactor fission products. The full equation for the beta decay of carbon-14 is shown below.

$$^{14}_{6}C \rightarrow {}^{14}_{7}N + {}^{0}_{-1}e + \overline{\nu}$$

The anti-neutrino ($\overline{\nu}$) is an example of an anti-particle (see p. 22).

The basic nuclear process in this carbon to nitrogen transformation is the decay of a neutron to a proton within the nucleus. The neutron mass exceeds the proton mass so that the disintegration energy is positive and this means that free neutrons will beta decay to protons, i.e. the neutron is not a stable particle. The average lifetime of a free neutron is about 16 minutes. There is another form of beta decay in which a proton is converted into a neutron within the nucleus but this obviously cannot occur in the free state because the proton is the less massive of the two. Here, energy must be supplied from within the nucleus. An example of this process is the decay of oxygen to nitrogen as follows.

$$^{15}_{8}O \rightarrow {}^{15}_{7}N + {}^{0}_{+1}e + \nu$$

The positron (${}^{0}_{+1}e$) is the anti-particle of the electron (see p. 22).

Example 1.4

A nuclide X with a total of 9 protons and 11 neutrons has an atomic mass of 19.999982 u. What element is this?

Ans: Fluorine

What element would be created if it underwent a) alpha decay and b) beta decay, with the emission of an electron?

Ans:
a) Nitrogen (${}^{16}_{7}N$, mass = 16.006100 u)
b) Neon (${}^{20}_{10}Ne$, mass = 19.992435 u)

What factor determines the possibility of a decay?
Ans:
Whether or not the total mass beforehand exceeds that afterwards. If not, a spontaneous decay cannot occur.

If the mass of the alpha particle is 4.00260 u, can alpha decay occur?

Ans: No.

Can beta decay occur? If so, calculate the maximum energy of the emitted electron.

Ans: Yes.

The maximum energy of the electron = 7.03 MeV

Now is an opportune time to introduce the early 1930s table of known particles. It has become customary to express the masses of sub-nuclear particles in units of energy.

Particle	Mass/MeV	Electric charge/e	Mean lifetime
Electron	0.511	−1	Stable
Proton	938.3	+1	Stable
Neutron	939.6	0	~16 minutes
Neutrino	0	0	Stable
Photon	0	0	Stable

Table 1.1 *Elementary particles of the 1930s*

The magnetic moments of the neutron and proton were measured in the 1930s and surprisingly it was found that the neutron, despite being uncharged, has a magnetic moment. Here, in retrospect, was the first hint of internal structure within the nucleons themselves.

The neutron, proton and the electron all have spin = 1/2. In the beta decay of a free neutron, a spin 1/2 particle decays to three others, of which two have spin 1/2 – so that the neutrino must also have a spin of 1/2 since spin angular momentum is a conserved quantity. The first four of the particles listed are then, like the electron, fermions.

The Transmutation of Elements

Radioactivity played another important role in nuclear physics by providing the alpha particle, a very useful projectile which was the first probe of nuclear structure. It was the earliest tool with which a nucleus was transformed into a second with different mass number and atomic number. The first artificial transmutation was observed by Rutherford in 1919 when he bombarded nitrogen gas with alpha particles from a naturally-occurring radioactive isotope and produced protons and an isotope of oxygen in the reaction

$$^{4}_{2}\mathrm{He} + ^{14}_{7}\mathrm{N} \rightarrow ^{17}_{8}\mathrm{O} + ^{1}_{1}\mathrm{H}$$

Thus, in a sense, the medieval alchemists' dream had come true.

In nuclear reactions, the energy transformed from rest mass energy to kinetic energy is called the *Q*-value of the reaction. If *Q* is positive, kinetic energy is released and the reaction is referred to, in this context, as exothermic. Including the *Q*-value, Rutherford's equation can be written as shown in Example 1.5

Example 1.5

Calculate the *Q*-value of the above reaction.

$$^{4}_{2}\mathrm{He} + ^{14}_{7}\mathrm{N} \rightarrow ^{17}_{8}\mathrm{O} + ^{1}_{1}\mathrm{H} + Q$$

m_{He} = 4.00260 u; m_{N} = 14.00307 u

m_{O} = 16.99913 u; m_{H} = 1.00783 u

$$Q = (m_{\mathrm{He}} + m_{\mathrm{N}} - m_{\mathrm{O}} - m_{\mathrm{H}})\, c^2 = -\,1.20 \text{ MeV}$$

The fact that *Q* is negative means that the minimum kinetic energy of the projectile alpha particle must be 1.2 MeV greater than the final kinetic energy of the system in order to satisfy the conservation laws.

The Acceleration of Protons

In the period from 1928-1932, at the Cavendish Laboratory in Cambridge, the English physicist **John Cockcroft** (1897-1967) and the Irish physicist **Ernest Walton** (1903-1995) who was the Professor of Physics at Trinity College, Dublin from 1946 to 1973, designed and constructed the first so-called 'atom-smasher', a machine that was capable of accelerating protons up to energies of more than 0.7 MeV. These protons were used in the first artificial splitting of the nucleus. The 1951 Nobel prize for physics was awarded to Cockcroft and Walton for their 'pioneer work on the transmutation of atomic nuclei by artificially accelerated atomic particles'. Fig. 1.2 shows the Cockcroft and Walton apparatus which is termed an electrostatic generator. In the photograph, it is Walton who is in the lead-shielded 'hut'.

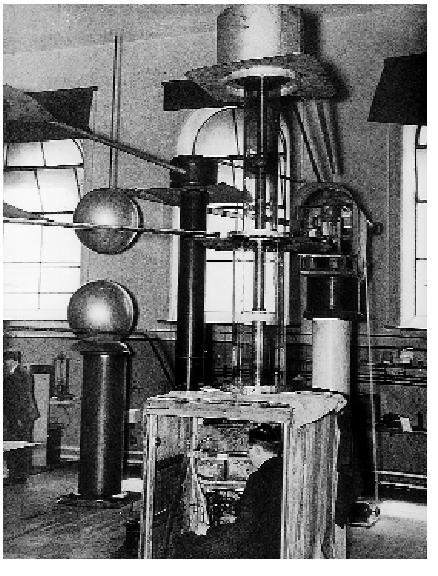


Fig. 1.2 *The Cockcroft and Walton accelerator (courtesy of the Walton family)*

A large amount of electric charge was accumulated on a spherical capacitor and the protons were accelerated in a single-step process in the large electric field created by the charge. The proton energy was determined from the electric potential through which they were accelerated.

Using protons with a kinetic energy of approximately 0.7 MeV to bombard a lithium target, Cockcroft and Walton observed the reaction

$$^{1}_{1}\mathrm{H} + ^{7}_{3}\mathrm{Li} \rightarrow ^{4}_{2}\mathrm{He} + ^{4}_{2}\mathrm{He}$$

Example 1.6

Calculate the *Q*-value of this reaction.

m_{Li} = 7.016005 u. m_H and m_{He} as given in Example 1.5

$$Q = (m_H + m_{Li} - 2m_{He}) = 17.36 \text{ MeV}$$

Q is not only positive but is large. Thus, although the incoming protons had a kinetic energy of only 0.7 MeV, the two outgoing alpha particles had a combined kinetic energy in excess of 17 MeV. This reaction served as the first experimental check of the mass-energy relationship $E = \Delta mc^2$. The experiment was a significant milestone in nuclear physics because the incoming particle energy was controlled.

Converting Energy into Matter

Einstein's mass-energy relationship ($E = mc^2$) introduces the possibility of creating matter from energy in the collision of two sub-nuclear particles. It should then be possible to produce new particles, either in addition to or instead of, the original particles. The greater the energy of the colliding particles, the more energy is available for new particle production. The development of particle accelerators will now be briefly described.

In 1929, the American physicist **Ernest Lawrence** (1901-1958) had the idea of accelerating protons repeatedly through a relatively small potential difference instead of in one giant step. In his cyclotron, which he developed in association with Livingstone at Berkeley, a uniform magnetic field holds protons in a circular orbit whilst an electric field acts on them periodically in such a way as to gradually increase their energy. The fundamental principle lies in the fact that a particle of fixed mass moving in a plane perpendicular to a uniform magnetic field executes circular motion at a fixed frequency, independent of its speed (Example 1.7). In the cyclotron, Fig 1.3, an electric field with a frequency equal to the 'cyclotron' frequency can be used to gradually accelerate the protons because they remain 'in phase' with the electric field each time they enter it, and are accelerated each time they cross the gap between the 'dees'.

This was a major advance in the development of particle accelerators since it made possible the acceleration of protons to much higher energies than those obtained with other accelerators. The first cyclotron built by Lawrence had a diameter of a few centimetres but, by 1939, he had built a 200 tonne cyclotron with a diameter of 1.5 metres.

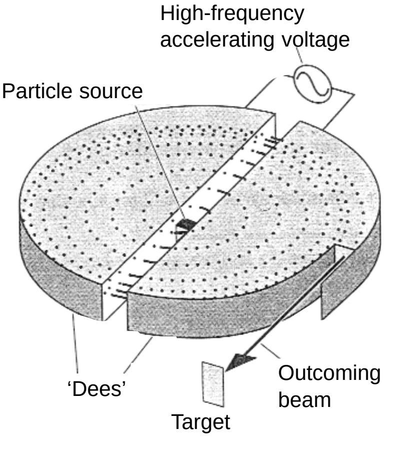


Fig. 1.3 *The basic principle of the cyclotron*

Example 1.7

Calculate the radius of curvature *r* of the path of a particle of charge *Q* moving in a plane perpendicular to a uniform magnetic field *B* and show that the cyclotron frequency *f* is constant.

Since the force on the moving charge is always at right angles to its velocity the particle follows a circular path. The resultant force on the particle is

$$F = QvB$$

Therefore

$$QvB = \frac{mv^2}{r}$$

$$r = \frac{mv}{QB}$$

$$\frac{v}{r} = \frac{QB}{m}$$

The period of revolution is

$$T = \frac{2\pi r}{v}$$

The frequency of revolution is

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

$$= \frac{QB}{2\pi m}$$

Since *Q, B* and *m* are constant the frequency of revolution is constant.

The principle of using magnetic fields to confine and manipulate the particle beams and electric fields to accelerate the particles is the basis of modern cyclic accelerators called proton synchrotrons. To attain greater particle energies, it was necessary to modify Lawrence's basic design, as is done in proton synchrotrons.

As the particle energy increases, the orbital radius increases (Example 1.7). The radius is maintained at a fixed value by increasing the magnetic field as the energy rises. With increasing energy, there is also an increase in particle mass (a relativistic effect), and the orbital frequency decreases (Example 1.7). In the proton synchrotron, individual bunches of particles move within a ring-shaped vacuum pipe of fixed radius under the influence of magnetic fields. At a number of points around the ring there are cavities containing variable frequency electric fields which are tuned to accelerate discrete bunches of protons on arrival. As in the cyclotron, the protons complete many orbits. Fig. 1.4 shows the basic set-up of a proton synchrotron. The focusing and bending magnets are not shown. Prior to injection into such a machine, the protons are accelerated to a kinetic energy of about 1 MeV by, for example, a Cockcroft and Walton accelerator.

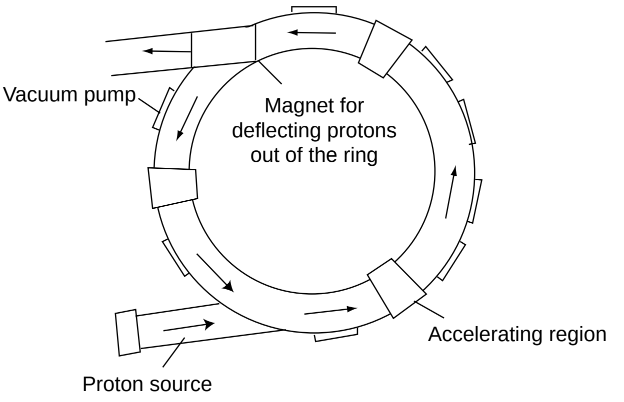


Fig. 1.4 *Proton synchrotron*

At energies of about 1 MeV, proton collisions with nuclei in a target may fragment the nuclei into their constituent protons and neutrons, but the number of protons and neutrons remains constant. By the mid 1950s, protons were being accelerated to energies in excess of 1 GeV (1 GeV = 1000 MeV), where impacts of accelerated protons on stationary protons reveal yet another phenomenon. The protons themselves are 'unbreakable' and additional particles are produced as a result of the conversion of some of the incident protons' kinetic energy into mass.

An example is the reaction

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

which is a very common process at energies of about 2 GeV. The π^+ and π^- are positively and negatively charged 'pions', sub-nuclear particles which have a mass of about $^1/7$ of the proton mass. The formal introductions will be made later.

In 1955, the Russians pushed accelerator energies upwards with the surprise unveiling of a 10 GeV machine. The next major leap forward was at CERN (Centre Européen pour la Recherche Nucleaire) where, as a result of the co-operative effort of twelve European nations, a 28 GeV proton synchrotron was built. Progress has continued apace since that time and with it the physical size of accelerators. As already mentioned, charged particles radiate energy in the form of electromagnetic radiation (called 'synchrotron radiation') when accelerated. The power radiated in this way is inversely proportional to the fourth power of the ring radius R so that an increase in R reduces the energy lost in this manner. An even more critical factor in accelerator sizes is the size of the magnetic field needed to hold the protons in their orbit. It increases with increasing energy but is smaller when R is larger. The modern very high energy accelerators use superconducting magnets. The Tevatron accelerator at the Fermi National Accelerator Laboratory (FNAL), near Chicago, which began to produce 1000 GeV (1 TeV) protons in 1984 has a diameter of 1000 metres, a far cry from Lawrence's prototype cyclotron. Fig. 1.5(a) is a photograph of the accelerator centre at CERN. The SPS accelerator - the smaller ring, circumference 7 km - was originally a proton synchrotron, later adapted to accelerate counter-rotating protons and antiprotons. The LEP machine, circumference 27 km, accelerates electrons and positrons. (The tunnels for both machines are underground.) The scale of such machines is well illustrated by the inclusion of Geneva airport in the foreground of the picture. Fig 1.5(b) is a photograph taken inside the LEP tunnel.

Fig. 1.5(a) *An aerial view of CERN (Photo CERN)*

Fig. 1.5(b) *Inside the* LEP *tunnel (Photo CERN)*

High energy accelerators do not accelerate particles to such energies from rest. The particles are pre-accelerated and bunched in smaller machines. It is amazing to think that the primary accelerator for the Tevatron accelerator, where the latest quark find was recently made, is an 800 MeV Cockcroft and Walton electrostatic generator.

CHAPTER 2

THE FUNDAMENTAL FORCES OF NATURE

2.1 Gravitational Force

We know of four fundamental forces which exist in nature. The first of these to be put on a theoretical footing was gravitation.

The full theory of gravitation was not published until 1687, although the English physicist **Isaac Newton** (1642-1727) recalled later that he had conceived the basic ideas some twenty years earlier. He had realised that the forces which keep the moon in its orbit about the earth and the planets in their orbits about the sun have the same origin. It is a force which children experience at a very early age and yet, paradoxically, it is the force about which the least is known at the most fundamental level.

2.2 Electromagnetic Force

Next there is the electromagnetic force of which a part is the electrostatic force between two charges. The law governing the electrostatic force was formulated by the French physicist **Charles Augustin de Coulomb** (1736-1806), about a century after Newton's gravitational law.

The size of both these forces is known to depend inversely on the square of the separation of the interacting objects. It is interesting to estimate the relative strength of these two forces at the atomic level by working out the ratio of the sizes of the gravitational and electrostatic forces between a proton and an electron in a vacuum.

$$F_G = \frac{Gm_p m_e}{r^2}$$

$$F_E = \frac{e^2}{4\pi\varepsilon_0 r^2}$$

Substituting gives

$$\frac{F_G}{F_E} \approx 10^{-40}$$

Both forces depend on $1/r^2$ so that they fall off fairly rapidly as r increases but do not go to zero until $r \to \infty$. In principle, these forces have an infinite sphere of influence or 'range'.

Gravitational force is always attractive and acts on all particles but, as the calculation shows, is not significant at the atomic level. The electrostatic force acts between all charges and may be attractive or repulsive depending on the nature of the charge. The magnetic force on moving charges constitutes the second part of the total electromagnetic force, and it is this electromagnetic force which, in some way or another, affects us most as we go about our daily lives because it binds atoms together. This force can be friend or foe, on the one hand keeping nails in walls but on the other, causing wear and tear, to quote just two of the innumerable instances of its effects.

To explain how, for example, two charges can exert forces on one another without being in contact, classical physics uses the concept of a field related to the force and permeating the space around the charges. The equations of electromagnetism, developed in the 1860s by the Scottish physicist

James Clerk Maxwell (1831-1879), predicted the existence of electromagnetic radiation, as the result of oscillating electric charges or currents. These produce oscillating electric fields, which induce oscillating magnetic fields and these, in turn, induce new electric fields and so on. Thus, electromagnetic radiation is emitted by an accelerated electric charge.

With the advent of the theory of the quantisation of radiant energy with its quantum, the photon, it seems a natural progression to think in terms of the quantisation of the electric field between charges. It was Dirac who put forward the idea that the electromagnetic field itself was quantised. The force due to that field is transmitted by the exchange of photons, just as a soccer ball is exchanged by two players on a pitch. To create a better understanding of the process, Fig. 2.1(a) shows a ball exchanged by two ice skaters and Fig. 2.1(b), a photon exchanged by two electrons.

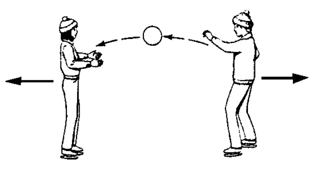


Fig. 2.1(a) *Ice skaters*

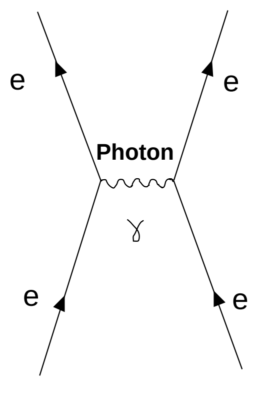


Fig. 2.1(b) *Electrons*

There are, however, difficulties. There is an electrostatic force between two electrons at rest, but photons from a charge at rest have never been detected experimentally. Also, there is no energy available to transmit these photons from one electron to the other.

However, salvation was at hand in the shape of the uncertainty principle proposed by the German physicist **Werner Heisenberg** (1901-1976), which allowed the existence of 'virtual' photons (see below) which do not have to obey the conservation laws.

The Uncertainty Principle

Heisenberg reasoned that it is only realistic to think in terms of measurable quantities and concluded that, in dealing with small particles, some pairs of variables are in a sense inseparable, in that it is impossible to make accurate measurements of the values of both at the same time. One such pair is the position and momentum of a particle. A simple explanation of this particular example is made possible by remembering that sub-atomic particles exhibit wave properties. This is very powerfully illustrated in Fig. 2.2 by comparing the diffraction patterns typical of (a) X-rays and (b) electrons, after traversing aluminium foil.

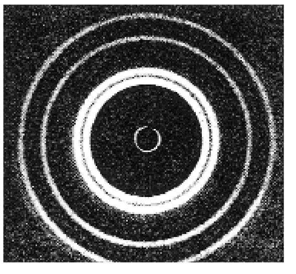


Fig. 2.2(a)

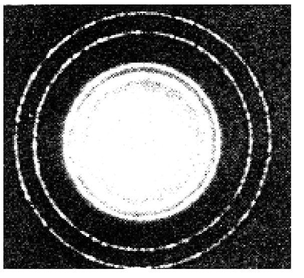


Fig. 2.2(b)

A sub-atomic particle may be represented by a wave packet. Since a wave packet must spread over a finite distance in space, this implies that there is an uncertainty Δx in the position of the particle. In Fig. 2.3(a) the particle is more precisely located in space than in Fig. 2.3(b), where the wave packet contains more whole wavelengths.

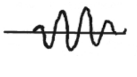


Fig. 2.3(a)

Fig. 2.3(b)

We can express the momentum p of a particle in terms of its wavelength λ.

$$p = \frac{h}{\lambda}$$

Therefore, an accurate estimate of the momentum relies on the presence of a large number of whole wavelengths (Fig. 2.3(b)). It is thus possible to get a good estimate of position on the basis of Fig. 2.3(a) and a good estimate of momentum from Fig. 2.3(b) but we cannot simultaneously get a good estimate of both quantities.

The fundamental uncertainty in the simultaneous measurement of the position x and the momentum p of a particle is usually stated to be

$$\Delta x \, \Delta p \approx \hbar$$

where Δx and Δp are the uncertainties in the measurements.

A similar formula links energy and time and this permits a single photon of energy E, emitted by an electron, to exist for a maximum time $\Delta t \approx \hbar / E$, without violating the law of conservation of energy (such a photon is called a virtual photon). The fact that this photon must be absorbed by the second electron within the time interval Δt has implications for the mass of the photon. Any particle which has a mass m has a minimum or rest energy of $E_0 = mc^2$ and this implies that there is a minimum time of $\Delta t \approx \hbar / E_0$ for which it can exist without violating the conservation laws. If the photon has zero mass, then it can exist for a time unlimited by its mass. Infinitely low energy virtual photons may be emitted by a charge, and in that case, $E \to 0$, and $\Delta t \to \infty$. Photons travel at a constant speed and these low-energy photons may therefore travel an infinite distance. The intensity of the electromagnetic field is seen to decline steadily as the distance from a charged particle increases but at no point is it seen to drop sharply to zero. The field is said to have infinite range and its carrier, the photon has zero mass.

2.3 Strong Force

The strong or nuclear force is an attractive force which acts between, for example, protons and neutrons. It is the strongest of the known forces and has a very short range, R, of about 10^{-15}m. Its size and form are independent of the electric charge of the interacting particles. The time scale of an interaction is the time during which the particles are close enough to interact. If a proton travelling at a speed of the order of the speed of light passes a second proton, the duration of the interaction will be given by

$$\Delta t \approx 10^{-15}/c \approx 10^{-23} \text{ s}$$

Processes such as the decay of particles via the strong interaction will also take place over times of this size.

A classical analogy is that a passenger travelling in a car at 30 m s^{-1} past a letter box of length 0.30 m will have a time interval of only 0.01 s during which he can post a letter.

The presence of a strong or nuclear force is something which is undeniably required by the fact that nuclei of stable elements contain protons and neutrons. The positively charged protons mutually repel one another and, in the absence of this attractive force, nuclei would not be stable. There is, however, a limit to the size of stable nuclei. Stable nuclides with an atomic number larger than 83 are not found and this observation is in broad agreement with the fact that the strength of the nuclear force is approximately two orders of magnitude larger than that of the electromagnetic force.

Nucleons are arranged within the nucleus in a manner similar to that of the electrons in an atom. They, like electrons, have spin and obey the Pauli exclusion principle. The strong force is spin-dependent and a good example of its complex nature can be seen by comparing the energies required to fragment nuclei. It takes 2 MeV to split up a deuteron but a massive 30 MeV is needed to totally split up a helium nucleus (alpha particle)! It is only possible to fully describe this force in terms of the forces acting between the constituents of individual nucleons.

In 1935, the Japanese physicist **Hideki Yukawa** (1907-1981) put forward a mathematical expression for the size of the strong force. It is similar to Coulomb's formula for the force between electric charges, but also contains a factor that makes the force fall off rapidly as the inter-nucleon separation increases. He then made the remarkable proposal that the field due to this force is transmitted by particles, in a manner similar to that in which the electric field is carried by photons. The finite range or abrupt fall in intensity of the field due to the strong force implies that these particles have a non-zero rest mass. This mass can be calculated by using Heisenberg's uncertainty principle in the form $\Delta E\ \Delta t \approx \hbar$ and interpreting ΔE as the rest mass of the particle.

Assuming the range of the force, R, to be $\approx 10^{-15}$ m and the speed of the particle to be $\approx c$

$R \approx c\Delta t$ and hence $\Delta t \approx 10^{-23}$ s

$\Delta E\ \Delta t \approx \hbar$ giving $\Delta E \approx \hbar / \Delta t$

$\Delta E \approx 100$ MeV

2.4 Weak Force

Finally there is the weak force which has the extremely short range of about 10^{-18} m. The neutrino is the only known particle which interacts with matter solely via the weak interaction. An example of the action of the weak force is the interaction between an anti-neutrino and a proton in which a neutron and a positron are produced.

$$\bar{\nu} + p \rightarrow n + e^{+}$$

It is responsible for, amongst other things, the conversion of neutrons to protons, and vice versa, in nuclear beta decay. In fact the above process is known as inverse beta decay. The weak force is thus responsible for one form of natural radioactivity, the familiar process which causes the decay of certain forms of elements within the earth's crust. The radioactive decay of certain long-lived elements permits us to estimate the age of artefacts by radiocarbon dating and even the age of the earth itself.

Weak interactions in general have a low probability of occurring - neutrinos can pass right through the earth - and this is reflected in the range and size of the weak force compared to the strong and electromagnetic forces.

It is possible to make an estimate of the typical lifetime of a particle which decays via the weak interaction, using the same sort of reasoning which was used to estimate the lifetime of strongly interacting particles.

In this case, the range R is approximately 10^{-18} m but the weak interaction is a factor of approximately 10^{-13} times weaker than the strong interaction, implying that the particles will live longer.

$$\Delta t \approx 10^{13}\ (10^{-18}/c) \approx 10^{-13}\ \text{s}$$

This lifetime is not characteristic of all particles which decay in this manner, as is seen in the enormous range of lifetimes of beta decaying radioisotopes. One of the factors which influences the lifetime is the mass difference between the decaying particle and its decay products.

If the weak force is also transmitted by a particle or mediator then the fact that its range is so short means that the mediator must be much more massive than the proton. A theory which unified weak and electromagnetic interactions as different aspects of the same basic (electroweak) interaction was proposed by the American physicist **Steven Weinberg** (1933-), the Pakistani physicist **Abdus Salam** (1926-1996) and the American physicist **Sheldon Glashow** (1932-), independently, in 1967 and 1968. (This is not by any means a unique example of the coincidence of ideas among physicists.) They proposed that the carriers of the electroweak field are the familiar mass-less photon and three massive particles - the charged W^{+} and W^{-} and the uncharged Z^{0}. According to their theory, the masses of the W and Z were predicted to lie in the range 80 to 100 GeV. The weak and electromagnetic processes experienced in laboratory experiments have very different properties and can still be classified separately.

Table 2.1 below shows the relative strengths of the four forces. The figure for the ratio of the gravitational to electromagnetic interaction is based on the interaction between two protons to give a better comparison with the size of the strong force.

Force	Strong	Electro-magnetic	Weak	Gravity
Relative strength	1	10^{-2}	10^{-13}	10^{-39}

Table 2.1 *The four fundamental forces*

The figure for the weak interaction is such as would have been measured in the 1960s and its extreme weakness is a reflection of its very short range of about 10^{-18} m (10^{-4} times the diameter of a typical nucleus). As the energy increases, the strength of the so-called ‘weak’ interaction approaches that of the electromagnetic interaction, as predicted by the unified theory of Weinberg, Salam and Glashow.

It seems ironic that gravity, the only force known to affect all particles, does not play a significant role in nuclear and sub-nuclear phenomena. It is currently believed, as mentioned earlier, that there is also a particle which transmits the gravitational force. This particle, the graviton, must have zero mass as the force has infinite range. Gravity waves, which are emitted by all masses, are so weak that it has not yet proved possible to detect them experimentally. (It is estimated that the total power emitted as gravity waves by the earth whilst orbiting the sun is a paltry 10 watts - not even adequate to illuminate the average room in a house.) There are however events which signal the victory of this seemingly feeble force over the much more powerful electromagnetic force in circumstances where the amount of matter involved is very large. In the collapse of a massive star at the end of its normal life as it exhausts its supply of nuclear fuel to produce heat from fusion, its atoms are compressed, forcing the electrons into the nucleus where they combine with protons to make neutrons and neutrinos. The escaping neutrinos from a ‘supernova’, arising from such a collapse, were detected at the earth’s surface in 1987. The remaining neutrons are compressed into a ‘nucleus’ with a radius of about 10 km and a colossal density of the order of several megatonnes per cubic metre. A very massive star will undergo further gravity-driven collapse to end up as a black hole. There is good experimental evidence for the existence of both neutron stars and black holes.

CHAPTER 3 **MORE PARTICLES**

3.1 Introduction

In the years that followed 1932, particles with masses between those of the electron and the proton were discovered. The energetic particles needed to create them were provided by nature, in this case in the shape of high-energy cosmic ray protons which continually bombard the earth. These protons interact with nuclei in the upper atmosphere, and produce sub-nuclear particles. In the years following 1937, a particle seen in numerous cloud chamber photographs taken at various altitudes was identified as the much sought-after ‘Yukawa’ particle, which had been proposed as the mediator of the strong interaction (see p. 15). It was found to have an average mass of 100 MeV, as had been predicted. It was called the mu-meson, μ (meson after the Greek word for intermediate). However, the recorded fluxes of mu-mesons on mountain tops, at sea level and even at the bottom of lakes were essentially the same. If these particles were really those responsible for binding nuclei and therefore being susceptible to the strong interaction, nuclei of the atmosphere would absorb them very readily. Thus their numbers would diminish noticeably as the particles travelled through the atmosphere or any other material. We now know that they do not even belong to the group of particles classified as mesons (see p. 21) and they are now known as muons (μ).

Yukawa’s particle, the pi-meson or pion, eventually made its appearance in 1947 in cloud chamber photographs from unmanned balloon experiments. Pions travel a much shorter distance in the atmosphere than do muons because there is a higher probability of interactions between them and atmospheric nuclei as a result of the strong force. Only a year later, artificially accelerated protons from the 200 MeV accelerator at Berkeley, California were seen to produce pions and the mass of this particle was fixed at 140 MeV.

Fig. 3.1 shows some examples of the decay of pions which come to rest in nuclear photographic emulsion. The pions decay into muons, which subsequently decay into electrons. In each instance, the length of the muon track is the same (approximately 2/3 mm), implying that the muons are produced with precisely the same energy in each case. This means that the decay of a pion to a muon involves just one other particle which is, in fact, a second type of neutrino. If the pion is stationary when it decays, the muon and the neutrino will be emitted with equal and opposite momenta. We now know that both the electron and the now correctly classified muon have their own associated neutrino.

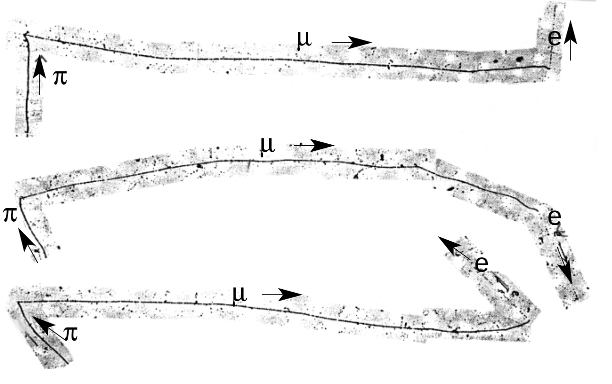


Fig. 3.1 *Pion decay*

3.2 A New Quantum Number

The K-meson or kaon (K) was the first of a group of 'strange' particles to be observed. It can be produced in an interaction between, for example, a negatively charged pion and a proton, which is a strong interaction. However, the kaon does not decay in the same manner as it is produced, even though there are lighter mesons (see Table 3.1, p. 21) to which it could in principle decay, via the strong interaction, without violating any known conservation laws…stranger and stranger! The earliest observed example of the production of a kaon was

$$\pi^- + p \rightarrow \Lambda^0 + K^0$$

This interaction was originally seen in cloud chamber photographs taken by Rochester and Butler in 1947. The lifetime of the K^0 is typical of particles decaying via the weak interaction (see p. 16). To explain the manner in which the kaon is seen to decay, it was assigned a new quantum number called strangeness (*S*), which is conserved in strong interactions but not conserved in weak interactions. The kaon was arbitrarily assigned a strangeness, $S = 1$. The kaon is the lightest strange meson and cannot therefore decay via the strong interaction.

In the interaction shown above, a second particle, the Λ^0 (lambda) is seen to be produced. To conserve strangeness the Λ^0 is assigned a strangeness, $S = -1$. The two particles have zero total strangeness, as is required, since the pion and proton are not strange particles. This is called 'associated production' of strange particles.

Other strange particles such as the Σ (sigma) and Ξ (cascade), which behave similarly, were discovered. The values of the strangeness of such particles were fixed on the basis of the arbitrary value assigned to the K^0. The complications seemed to increase!

A schematic representation of the associated production of a K^0 and a Λ^0 is shown in Fig. 3.2. A negative pion collides with a proton, in an interaction which is frequently seen in hydrogen-filled bubble chambers. The pion track is shown to terminate at point A, where it interacts with the proton. The two uncharged particles created in the interaction do not leave tracks but their decays are clearly seen at points B (Λ^0) and C (K^0).

The K^0 decays weakly into a pair of oppositely charged pions

$$K^0 \rightarrow \pi^+ + \pi^-$$

The Λ^0 also decays weakly

$$\Lambda^0 \rightarrow p + \pi^-$$

The lifetimes of each of the two decaying particles can be inferred from the distance between the point at which they are produced and the point at which they decay.

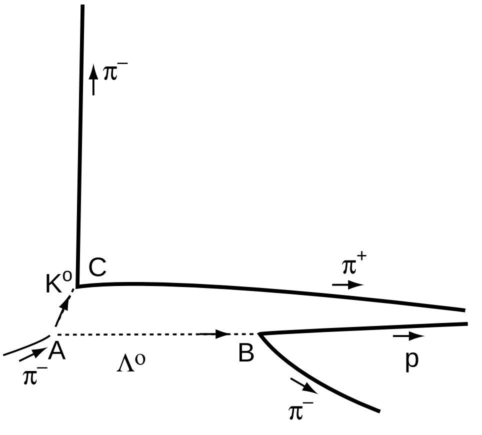


Fig. 3.2 *The associated production of a pair of strange particles*

With the growth in accelerator energies described earlier, the number of particle species known seemed to be limited only by the energy available to create them. By the early 1960s, some hundreds of particles had been identified - a veritable population explosion! Some form of classification of these particles was necessary so that both similarities and differences might become apparent.

3.3 Classification of Particles

Particles can be classified according to whether or not they are affected by the strong force.

Leptons

These are particles which are unaffected by the strong force. There is a quantum number, the lepton number (*L*) which is given a value of 1 for leptons and 0 for all other particles. The electron and the muon are examples of leptons. Lepton number was seen to be conserved in all interactions and further, electron and muon lepton numbers were seen to be separately conserved. This is an important pointer to the fundamental nature of leptons, to become apparent later.

Hadrons

All hadrons are affected by the strong force. There is a further sub-division within this group, according to the spin of the particle.

Mesons are strongly interacting particles, which have zero intrinsic angular momentum (s = 0). The pi-meson or pion as it is now known, is a meson.

Baryons are strongly interacting particles which have half integral values of intrinsic angular momentum (s = 1/2, 3/2). Baryons are assigned a quantum number, baryon number B, which is 1 for baryons and zero for all other particles.

Particle	Mass/MeV	Electric charge/e*	Spin/ $\hbar$	Baryon no.	Lepton no.	Strange-ness	Lifetime/s
Electron e^-	0.511	–1	$^1/_2$	0	1	0	Stable
Neutrino ν_e	0	0	$^1/_2$	0	1	0	Stable
Muon μ^+, μ^-	105.7	1, –1	$^1/_2$	0	1	0	≈ 2 x10^{-6}
Neutrino υ_μ	0	0	$^1/_2$	0	1	0	Stable
Pions π^+,π^-	139.6	1, –1	0	0	0	0	≈ 3 x 10^{-8}
π^0	135.0	0	0	0	0	0	≈ 10^{-16}
Kaon K^+, K^0	493.6	1, –1	0	0	0	1	≈ 10^{-8}
K^- $\bar{K}^0$	497.7	0	0	0	0	–1	≈ 10^{-10**}
Proton p	938.3	1	$^1/_2$	1	0	0	Stable
Neutron n	939.6	0	$^1/_2$	1	0	0	≈ 10^3
Lambda Λ^0	1116	0	$^1/_2$	1	0	–1	≈ 3 x 10^{-10}
Sigma Σ^+	1189	1	$^1/_2$	1	0	–1	≈ 10^{-10}
Σ^0	1193	0	$^1/_2$	1	0	–1	≈ 10^{-10}
Σ^-	1197	–1	$^1/_2$	1	0	–1	≈ 10^{-10}
Cascade Ξ^0	1315	0	$^1/_2$	1	0	–2	≈ 3 x 10^{-10}
Ξ^-	1321	–1	$^1/_2$	1	0	–2	≈ 2 x 10^{-10}

Table 3.1 *Some elementary particles - early 1960s (******e = 1.6 x 10^{-19} C)*

(**There are in fact two neutral kaons, the K-short with the lifetime shown and the K-long with a lifetime of about 5 x 10^{-8} s. Each of these particle states is a mixture of two quantum mechanical base states.)

It became apparent from experiments that not only things like energy, momentum, electric charge and intrinsic spin, but also baryon and lepton numbers are conserved in all interactions. It was understandable why the lightest baryon, the proton, would be stable.

Table 3.1 shows a sample of the elementary particles known or discovered by the early 1960s. Only longer-lived ones are shown. Even this small sample of the known family of elementary particles illustrates some very striking features.

Example 3.1

Give a reason why the following reactions or decays should not be allowed

a) $p \rightarrow n + e^+ + \upsilon_e$

Ans: The proton mass is less than the neutron mass.

b) $n \rightarrow \pi^+ + \pi^- + \pi^0$

Ans: Baryon number is not conserved.

c) $\pi^+ + p \rightarrow \Lambda^0 + K^0$

Ans: Charge is not conserved.

d) $\Sigma^- + p \rightarrow K^0 + n$

Ans: Baryon number is not conserved.
Strangeness is not conserved.

Example 3.2

Two protons of equal energy collide head-on. Calculate the minimum kinetic energy of each if the following reaction is to occur.

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

Ans: $^1/_2(m_{\pi^+} + m_{\pi^-}) \approx 140$ MeV (the π^+ and π^- have equal mass, they are a particle anti-particle pair – see below).

If each proton has an energy E of 1000 GeV, calculate the maximum number of pions which could, in principle, be produced.

Ans: The total energy available = 2000 GeV and $m_\pi \approx 140$ MeV

Maximum number = 14

3.4 Anti-particles

Dirac set out to combine the theories of Special Relativity and Quantum Mechanics and in 1928 had his first major achievement in the area, the equation of motion for free electrons. The immediate difficulty facing him was that the equation was quadratic in energy with a negative root, implying the existence of both positive and negative energy states.

Instead of merely dismissing the negative energy solutions as non-physical, Dirac proposed that the negative energy states were present but were normally filled with electrons, arranged according to the Pauli principle. In this way, there is no net energy so that their presence would not normally be detectable. Dirac further postulated that any unoccupied negative energy state or ‘hole’ would present itself as a positively charged particle. The hole would behave in the same way as a particle with the same mass as the electron, but with opposite charge.

Any natural system tries to achieve its minimum possible energy and free electrons would radiate away their energy as they dropped (rather like apples!) into the so-called ‘sea’ of vacant negative energy states. Such processes are called ‘pair annihilation’ (see p. 23). On the other hand, an electron could be liberated from the sea by giving it energy and the result would be a pair of particles, the liberated electron and a hole in the sea, in a process referred to as ‘pair production’ (see p. 23).

The Discovery Of The Positron

A particle with the same mass as the electron but with opposite charge was seen in 1932 by the American physicist **Carl Anderson** (1923-) in cloud chamber photographs and was called the positron (e^+). Charged particles, moving in the plane perpendicular to a uniform magnetic field, travel in a circle at constant speed and their charge can be inferred from their direction of motion. To ascertain this, a lead plate is inserted across the centre of the cloud chamber. The positrons emerge from the plate with reduced energy and momentum and since the radius of curvature is proportional to momentum (see Example 1.7 on p. 10) the direction in which the positron is travelling can be determined (Fig. 3.3). Here, the positron is travelling in the direction indicated by the arrow.

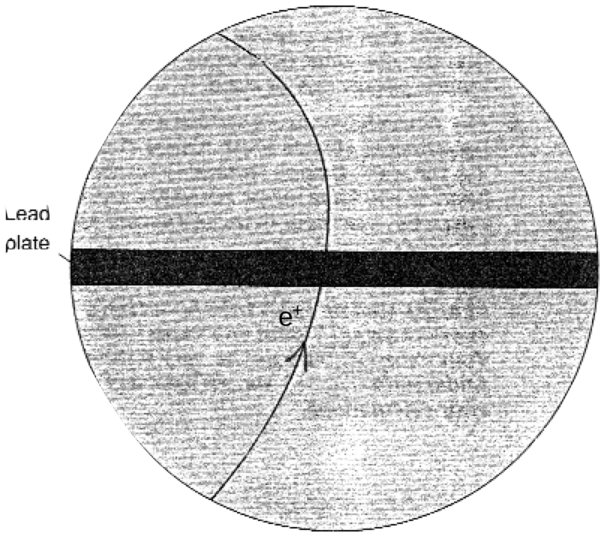


Fig. 3.3 *The discovery of the positron (from C. D. Anderson, Physical Review, 1932)*

Later theories reinterpret the negative energy solutions as positive energy states of the positron and remove the rather uncomfortable notion of an electron sea. They further show that for every particle there must exist an antiparticle with exactly the same mass and with opposite values of electric charge, baryon number, lepton number, etc. – in a certain sense, the mirror image or alter ego. The first anti-protons were produced in the laboratory in 1955.

In Table 3.1, the π^+ and π^-, for example, are a particle antiparticle pair, whereas the γ, for example, is its own antiparticle.

Pair Production

Dirac's theory implies that, for example, a photon with an energy of 1.022 MeV, twice the mass of the electron, could produce an electron-positron pair. In this case, the electron and positron would be at rest, since all the energy of the photon would be converted into mass. This implies that both particles have zero momentum and that the momentum of the incoming photon has somehow been transferred to another particle. Therefore, this process can only take place in the vicinity of a nucleus which can recoil such that the total momentum is conserved.

Pair production: $\gamma + N \rightarrow e^+ + e^- + N$ where N = nucleus.

Pair Annihilation

The reverse process, in which an electron and a positron annihilate to form a pair of photons, is also possible. Once again, conservation of momentum prohibits the production of a single photon and further demands that the photons which are produced travel in opposite directions.

Pair annihilation: $e^+ + e^- \rightarrow \gamma + \gamma$

Note the conservation of charge in both processes.

The processes of pair production and pair annihilation illustrate the equivalence of mass and energy in a very impressive way.

Pair annihilation has proved to be a very useful tool in diagnostic medicine, for example in the study of brain abnormalities. The technique is called Positron Emission Tomography (PET). Some radio-nuclides such as ^{15}O decay by positron emission (as seen earlier). If injected into the body, they collect at specific sites and there the emitted positrons annihilate almost immediately with electrons from body tissue. If both emitted photons strike opposite sides of a ring of detectors this defines the line along which their source is located. Such information leads to a computer-generated image of areas of possible abnormality.

CHAPTER 4

NEW FUNDAMENTAL PARTICLES

4.1 The Emergence of a Pattern

Just as Mendeleev had conceived a periodic table of atomic elements almost one hundred years earlier, Gell-Mann and Ne'eman independently did the same thing in 1960-1961 for the embarrassingly large and increasing numbers of hadrons, some of which have been listed in Table 3.1. The results were quite striking - the particles could be grouped into regular patterns according to their electric charge and strangeness. Baryons and mesons were to be found in separate arrangements. Mesons formed either singlets or octets and baryons either octets or decuplets. This pattern was called the 'eight-fold way'.

An example of a meson octet, involving particles drawn from Table 3.1, is shown in Fig. 4.1(a). All known particles could be accommodated in these new arrangements. There was however a missing member in one of the baryon decuplets shown in Fig. 4.1(b). Gell-Mann was able to estimate its mass, charge and strangeness and, one year later, in 1963 the predicted Ω^- particle was found, complete with all of its expected attributes. It was reminiscent of Mendeleev's prediction of the elements gallium, germanium and scandium. (Note: The * symbols on some of the particles in Fig. 4.1(b) just mean that these particles are higher mass states of the corresponding particles in Table 3.1. The Δ particles are examples of the very numerous short-lived particles which are not shown in Table 3.1.

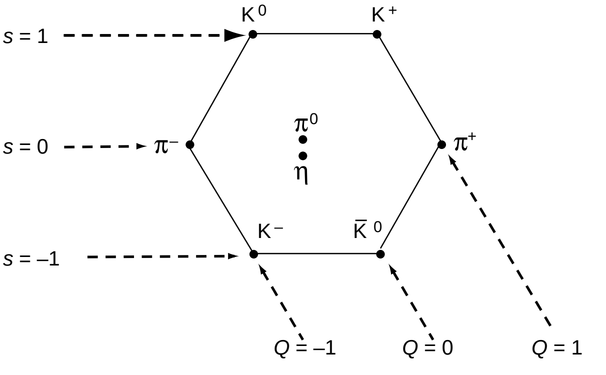


Fig. 4.1(a)

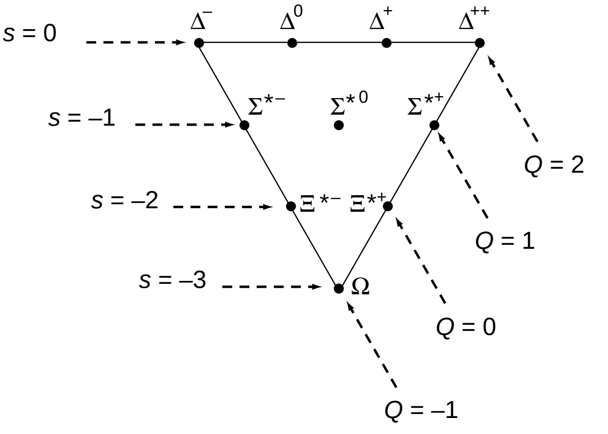


Fig. 4.1(b)

However, this is only Step 4 (see p. 2). The really basic question of why the patterns exist still remained unanswered.

4.2 Quarks and Leptons

In 1964, independently, Gell-Mann and Zweig published a model which proposed the hadron to be a composite structure. The constituents of hadrons were named 'quarks' by Gell-Mann (from the quotation, 'Three quarks for Muster Mark', in *Finnegan's Wake* by James Joyce). Coming hot on the heels of the eight-fold way with its prediction of the Ω^-, it aroused tremendous interest amongst physicists.

Quark Properties

Gell-Mann's paper set out the properties of the three types of quark needed to describe all known hadrons. They are rather unimaginatively labelled up (u), down (d) and strange (s). The most immediately noticeable difference between them and the experimentally observed particles is that quarks were assigned charges which were not integral multiples of the electronic charge.

Quark	Charge/*e*	Spin/ $\hbar$	Baryon no.	Strangeness
u	2/3	1/2	1/3	0
d	−1/3	1/2	1/3	0
s	−1/3	1/2	1/3	−1
$\bar{u}$	−2/3	1/2	−1/3	0
$\bar{d}$	1/3	1/2	−1/3	0
$\bar{s}$	1/3	1/2	−1/3	1

Table 4.1 *Quark properties*

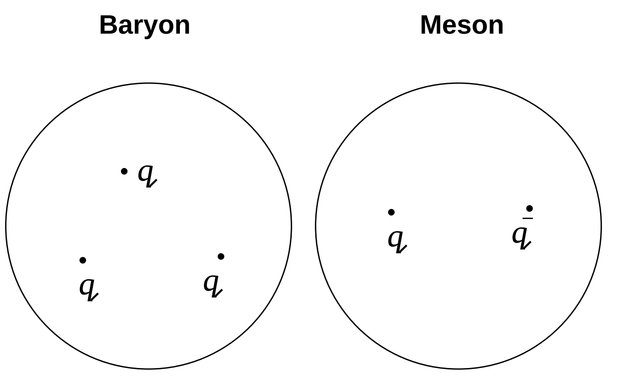
$\bar{u}$ is the notation for the u anti-quark and similarly for the other anti-quarks.

Baryons, Mesons and Quarks

The baryon-meson division within the hadrons is readily accommodated in the quark theory. The basic rules of hadron composition are easily stated as follows.

A baryon contains three quarks (and an anti-baryon three anti-quarks). It will have a baryon number of 1 and a spin of 1/2, 3/2, 5/2, etc., depending on the orientation of quark spins.

A meson (or its antiparticle) contains a quark and an anti-quark. Its baryon number is zero and its spin 0, 1, 2, etc.



(q = quark and $\bar{q}$ = anti-quark)

Fig. 4.2 *Pictorial representation of the quark composition of baryons and mesons*

Example 4.1

Using Tables 3.1 (p. 21) and 4.1, identify the particles which correspond to the following combinations of quarks.

a) u u d Ans: p
b) d $\bar{u}$ Ans: π^-
c) $\bar{u}$ s Ans: K^-
d) u d s Ans: Λ^0 or Σ^0

Experiments in the late 1960s, in which fast electrons were scattered off protons, showed results reminiscent of Rutherford's alpha particle scattering experiments which had probed the atom and revealed it to consist mostly of empty space except for a very small dense core. Analysis of these later experiments showed the charge of the proton to be concentrated in very small charged 'nuclei' within the proton, in agreement with the quark model.

Leptons, on the other hand, show no internal structure and can also be regarded as elementary particles. So, there are now two types of elementary particle, quarks and leptons, both of which have a spin of 1/2. This is then Step 1 once again.

Particle	Mass/MeV	Electric charge/*e*	Spin/ $\hbar$	Baryon no.	Strangeness	Quark make-up
Pion π^+	140	1	0	0	0	$u\bar{d}$
π^0	135	0	0	0	0	$u\,\bar{u}$ or $d\,\bar{d}$
Kaon K^+	494	1	0	0	1	$u\,\bar{s}$
K^0	498	0	0	0	1	$d\,\bar{s}$
Proton p	938.3	1	$^1/2$	1	0	u u d
Neutron n	939.6	0	$^1/2$	1	0	u d d
Lambda Λ^0	1116	0	$^1/2$	1	1	u d s

Table 4.2 *Quark composition of some known particles*

One More Quark

In the very year following the publication of the 3-quark model, Glashow and Bjorken suggested that a fourth quark existed. On purely aesthetic grounds, this fourth elementary particle would equate the numbers of quarks and leptons (the known leptons at the time were the electron, the muon and their respective neutrinos). This symmetry is also required by the unified theory of weak and electromagnetic interactions. By the early 1970s the properties of the fourth quark were predicted. It would have a spin of 1/2, charge of 2/3, strangeness of zero and be more massive than the other three. In addition, it was endowed with the brand new quantum number of charm.

Growth in the Numbers of Quarks

The first experimental evidence for the fourth quark came in late 1974 from two experiments using rather different methods. The first experiment which was technically more difficult involved analysis of proton collisions with a beryllium target - a strong interaction. The second experiment studied high energy electron-positron collisions, an electromagnetic interaction for which the theory was well understood. Both detected a new meson with a mass in excess of 3 GeV. This in itself was not unusual, but what was very surprising was that the particle did not decay very rapidly (i.e. strongly) to lighter particles in a whole variety of possible ways. In fact the lifetime of the new meson, called the J/ψ (J/psi) because of the joint discovery, indicated an electromagnetic decay. This was correctly attributed to the fact that its composition was cc (c is the symbol for a charm quark). The strong interaction must also conserve charm (remember strangeness and the behaviour of the kaons, p. 20) and if there is no particle which has a mass less than half of the mass of the J/ψ and contains at least one charmed quark, a strong decay is forbidden. In the immediate aftermath of the appearance of J/ψ, a number of more massive excited states were uncovered by the second experiment which was capable of searching at higher energies with greater sensitivity than the first. The third of these had a lifetime characteristic of a strong decay. This put a lower limit on the total mass of a pair of charmed mesons.

In the sense that the J/ψ is composed of a charmed quark and an anti-quark, the total value of its charm quantum number is zero, and the J/ψ is said to contain 'hidden charm'. The hunt for a meson containing a single charm particle intensified. By the end of the 1970s, both charmed mesons such as the D^+ (quark composition, $c\bar{d}$) and charmed baryons such as the Λ_c^+ (udc) had been identified. A research group from the Experimental Physics Department at University College Dublin participated in experiments which first observed the Λ_c^+.

In the meantime, in 1977, there was an announcement of the detection of a bound state of a new type of quark, the b (bottom) quark. Its existence was inferred once again from the discovery of mesons with masses greater than 9 GeV and with lifetimes far in excess of those expected unless they, too, contained yet another type of quark. Ironically, this surprising discovery was not made by an 'e^+e^-' experiment, but by an experiment studying the strong interactions of protons on nuclei, because the e^+e^- accelerators were limited to a maximum energy of 8 GeV at that time. It remained for the e^+e^- experiments of 1980 to observe the lightest strongly decaying $b\bar{b}$ meson and hence fix the mass of the lightest bottom mesons, as had been done for charm. A typical example of these is the B^+ ($c\bar{b}$).

New Leptons

In the meantime, during this period of rapid growing weather for quarks, a new lepton made its appearance in the blossoming particle garden. It was discovered in 1975 and was the first new lepton to be found in the forty years since the muon had created such unwarranted joy by masquerading as the Yukawa particle. The tau, as it is called, 'weighs in' at a hefty 1.78 GeV, almost twice the mass of the proton, uncharacteristically heavy for a lepton (the family name is derived from the Greek for light !). It was assumed that the new lepton would have its own neutrino as have the electron and muon and indirect indications leave no doubt about this. Direct evidence of its existence will take some time to obtain, as is customary with these very weakly interacting particles. The enlarged lepton family now numbered six.

The Search for a Sixth Quark

Quarks and leptons can be arranged in families of increasing mass. Only the first and least massive family or generation is needed to explain all phenomena within our everyday experience. Higher mass hadrons, created at higher energies, contain heavier quarks and, together with the heavier leptons, form two more generations.

Until recently, the third and most massive generation was incomplete. An additional quark, already named the t (top) quark had not been found experimentally. Evidence for its existence was finally announced in 1995 by the CDF collaboration at the Tevatron at FNAL, close to 20 years after the discovery of bottom. Their discovery was a result of the analysis of proton anti-proton collisions.

4.3 Quarks and Leptons - the New Fundamental Particles

Summary

The six quarks and six leptons, together with their antiparticles are currently considered to be the fundamental constituents of all matter in the universe. The span of time taken to isolate them, from the realisation of the importance of the electron, revealed by Thomson's experiments, has been about one hundred years. Recent measurements on decays of the Z^0, one of the carriers of the weak interaction, have restricted the total number of generations to three.

Generation	1		2		3	
Quarks	u	d	s	c	b	t
Lepton	e	ν_e	μ	ν_μ	τ	ν_τ

Table 4.3 *Quark and lepton generations*

So Step 1 is again complete. We have not yet seen evidence of any internal structure within the new elementary particles. However, current theory does not explain why there should be three and only three generations, nor can it predict the masses of the quarks and leptons, which make up those generations. Only time will reveal whether or not these elementary particles are really the fundamental constituents of matter.

As already mentioned, the first generation is all that is needed to explain our everyday world of atoms and their interactions. We only begin to see evidence for higher mass generations at the higher energies such as are achieved in accelerators built on the earth or provided by nature at one of the many distant cosmic sources which are currently being studied with such tremendous interest. Such sources have been referred to as 'cosmic powerhouses' because of their truly phenomenal output of energy.

Conclusion

The most fundamental of sciences are those which probe the ultimate structure of matter, the very nature of space and time and the origin and fate of the universe. It is interesting that the very knowledge gained in the search for the ultimate constituents of matter (looking through a microscope, so to speak) is now being used in attempts to explain how these cosmic powerhouses (viewed through various kinds of telescope) produce their energy.

The traditionally rather diverse fields of particle physics and cosmology have recently spawned the exciting field of high energy particle astrophysics. It has become apparent that the ultimate constituents of matter must have been present in the early and very hot stages of the universe, just after the big bang, and that fundamental particle interactions determine the nature of the cosmic radiation which falls on the earth, a minor planet of a very ordinary middle-aged star. Research in the area of particle astrophysics has advanced beyond all recognition from the earliest speculations mentioned at the very beginning of the introduction.

MODULE 10

Some Irish Contributions to Physics, Mathematics and Technology

SOME IRISH CONTRIBUTIONS TO PHYSICS, MATHEMATICS AND TECHNOLOGY

by Charles Mollan

Introduction

Ireland had a remarkable flowering of talent in the physical sciences, particularly in – though not limited to – the century between the 1830s (when Thomas Grubb manufactured his first large telescope and Revd Nicholas Callan invented the induction coil) and the 1930s when Ernest Walton 'split the atom'. It seems such a pity that we do not adequately recognise this thread of ability clearly resident in our Irish genes. Knowledge of the past might encourage more of our very able young people – today's beneficiaries of these creative genes – to look to the physical sciences both to provide for them valuable and rewarding careers, and to secure Ireland's future as a rapidly developing centre for technical excellence.

Robert Boyle (1627-1691)

It is Robert Boyle who can claim to be the most influential Irish-born scientist ever, so he makes a good starting point in this select roll-call of Ireland's achievers in the physical sciences. He played a key role in the history of science because of his part in establishing the experimental method, on which all modern science is based. By using carefully devised experiments, Robert demonstrated the power of practical science, and knowledge took a giant leap forward.

Although referred to as 'The Son of the Earl of Cork and the Father of Chemistry', Robert's contributions to what is now called physics were also of seminal importance. He was born in Lismore, Co. Waterford, the second youngest of fifteen children of Richard Boyle, first Earl of Cork. Richard, the father, came to Ireland from England to make his fortune in 1588, and got a job dealing with properties reverting to the Crown in the absence of legal heirs. He was thus in a good position to direct some of these his own way, and he had the good sense to marry an heiress. Although imprisoned for embezzlement and theft, he managed to receive a Royal pardon, and went on to accumulate a massive fortune and to advance his social standing and his political influence (see Canny 1982).

On the death of his first wife, he married the well-connected Katherine Fenton (when she was 17 and he was 36), and Robert was their youngest son. Robert was different from the other children in the family. In a brief autobiography of his early life (Boyle 1648-49), he paints himself as a rather self-righteous swot, preferring study to normal boyish pursuits. But his father, perhaps rather surprisingly given his own energetic career, clearly approved, for Robert wrote that he was very much his father's favourite. He went to school at the famous Eton College for a while, and then was sent at the age of eleven on a grand tour in Europe, which lasted no less than six years. He got a broad education in Europe, for he reports that, while in Florence, he was allowed to visit the famous Bordellos, though he claims he went to them out of 'bare [sic] curiosity', retaining his 'unblemished chastity' (Maddison 1969,40). He was no more attracted to other diversions, for he also reported that he was somewhat rudely pressed by what he called the 'preposterous courtship' of a couple of Friars, but he managed to escape.

His father, remembered as ‘The Great Earl of Corke’, had tried to arrange for Robert to marry and live in Mallow in Co. Cork. But the bride chosen for him opted to marry her cousin, much to the relief of Robert, who was thus free to devote himself entirely to his studies.

In 1654 he moved to Oxford, where he worked on pneumatics. In 1658-59 he had an air pump built for him by Robert Hooke (1635-1703) based on that invented by Otto von Guericke (1602-1686) in 1654. He showed that air was essential for the transmission of sound, for respiration and combustion (Daintith 1994,109). He used a U-shaped tube with a shorter closed end and a longer open end into which he poured mercury, so that there was a given volume of air in the closed end. With this device he found that, at constant temperature, volume is inversely proportional to pressure – the famous ‘Boyle’s Law’. This work was published in his book *New experiments physico-mechanicall, touching the spring of the air and its effects* (1660, second edition 1662).

Robert was a founder of the Royal Society in London in 1660, and the next year he published the most famous of his many books, *The sceptical chymist* (Boyle 1661). In this, he questioned the early belief that materials were made up of four elements – earth, air, fire, and water, instead anticipating modern atomic theory. He introduced many chemical tests, including the use of vegetable dyes as acid-base indicators, and flame tests to detect metals.

Robert Boyle died in London on 30 December 1691.

Thomas Romney Robinson (1793-1882)

On top of today’s bright orange conical road hazard bollards, you will often see a little horizontal three-cup windmill, complete with reflectors. This helps to draw attention to the bollard and thus to whatever it is you are supposed to avoid. They have contributed to the saving of many lives on our roads. You will see the same gadget in the context of weather stations and in other settings where the speed of the wind is of relevance. Next time you see one, remember its inventor, Irishman Thomas Romney Robinson.

It was in 1850 that the Reverend Robinson, at the time Director of Armagh Observatory, reported details of the ‘Robinson anemometer’ for measuring wind speed to a meeting of the Royal Irish Academy (published in Volume 22 of their *Transactions*). His actually had four cups, not three as is more common today, but otherwise the instrument remains the same. The hemispherical cups, on horizontal rotating arms, are blown by the wind, and Thomas worked out the relationship between the rate of rotation and the speed of the wind. All you have to do, then, is measure the number of rotations over a given period, and this will give you the average wind speed over that period. The original idea was suggested to him many years earlier by Richard Lovell Edgeworth (1744-1817), father of novelist Maria Edgeworth (1767-1849). Robinson was a close friend for many years of the Edgeworth family, and his second wife, whom he married in 1843, was Lucy Edgeworth, half-sister of Maria (Bennett 1990,141).

Thomas Romney Robinson was born in Dublin, the first son of a fashionable portrait painter, who had studied under George Romney (hence his second name). The family moved to Belfast in 1801, and Thomas attended Belfast Academy, where he gained a reputation as a child prodigy. He entered Trinity College Dublin at the tender age of twelve, graduated with a gold medal in 1810, became a Fellow in 1814, an MA in 1817, and a Bachelor of Divinity in 1822. He acted as a deputy Professor of Natural Philosophy in Trinity and, in 1820, published a text-book for his students: *A system of mechanics, for the use of students in the University of Dublin*, before accepting the College living of Enniskillen in 1823. He was Director of Armagh Observatory for nearly sixty years, from 1823 to1882 (Bennett 1990,59-64).

He played an important role in getting the Dublin firm of telescope makers, the Grubbs, ‘off the ground’ (*appropriate term that*!) by commissioning in 1833 from Thomas Grubb (see below) a reflecting telescope with a 15 inch diameter mirror. He was a great supporter of the astronomical work of William Parsons, third Earl of Rosse (*see below*). Robinson is also remembered in astronomy for his book

Places of 5,345 stars (a catalogue of approximate star positions) published in 1859, firmly establishing the reputation of the Observatory, which continues to carry out important astronomical work today. The 'Armagh catalogue' won for him the Royal Medal of the Royal Society, of which he had become a Fellow in 1856. He was president of the Royal Irish Academy from 1851 to 1856, and he played a central part in acquiring the use of 19 Dawson Street from the Government for the use of the Academy (Bennett 1999). Recently refurbished, this remains the headquarters of the Academy today.

The Revd Thomas Romney Robinson died in Armagh on 28 February 1882.

Nicholas Callan (1799-1864)

If you have assets you should surely use them. Such presumably was the thinking of one of Ireland's most ingenious physicists. He was the Reverend Nicholas Callan, who was born near Ardee in Co. Louth on 20 December 1799. He was Professor of Natural Philosophy at St Patrick's College, Maynooth, from 1826 until his death.

In a Pontifical University – and St Patrick's is the only one in these islands – you will find seminarians. In the mid nineteenth century, these were encouraged to bow to the authority of their superiors without demur. If you are a pioneer in electrical research, you need a sensitive means of checking your currents and voltages. Add the opportunity to the need and what do you get – a new use for seminarians!

Nicholas Callan was the inventor in 1836 of the induction coil. This was the essential forerunner of today's step-up and step-down transformers without which we wouldn't have limitless electricity at our fingertips. He sent an example of his coil to William Sturgeon (1783-1850) in London in 1837, where it was exhibited to members of the Electrical Society to their great amazement. Nicholas also invented new kinds of batteries, and he produced almighty voltages using his experimental coils. What better way to test the intensity of his electricity than to pass it through a 'voluntary' seminarian or two. And he did.

One of his best known 'voltmeters' was William Walsh, later to be the famous Archbishop of Dublin. The unfortunate young Walsh was rendered unconscious by Callan. Another seminarian who received similar treatment was Charles Russell, who, like William Walsh, became President of Maynooth. Russell apparently wasn't knocked out like Walsh, but he had to 'spend time on the infirmary after doses of Callan's high-tension electricity' (McLaughlin 1965,35).

A young student at Maynooth, called Lawrence Johnson, wrote home to his folks in 1855, telling them that:

We have a priest here from Co. Louth, Dr. Callan, the Professor of Science, and many are afraid he will blow up the College...but he is a very holy priest (McLaughlin 1965,37).

He was also something of a showman. He got the local blacksmith in Maynooth, James Briody, to make him an enormous U-shaped electromagnet of soft iron. When current was passed through the coils wound around the ends, they became magnetic, and they held tight to a keeper of iron placed across them. Callan challenged a team of his experimental seminarians to a tug-of-war. They were to try to separate the keeper from the electromagnet. No go.

Then the professor plays a little trick. He cuts the current as the team makes a mighty heave: the magnet is no longer active and the members all fall on a heap on the floor, much to the amusement and applause of the onlookers (McLaughlin 1965,70).

To further entertain his audience, the Revd Nicholas carried out the impressive experiment of electrocuting a turkey – an experiment apparently no longer recommended in today's student laboratories, even coming up to the Christmas break!

But it was not Callan who cashed in on his discovery. This distinction went to his contemporary, Heinrich Ruhmkorff (1803-1877), who was born in Hanover, but set up his own instrument-making business in Paris in 1839. He improved the induction coil in the 1850s, and it became one of the most important electrical instruments, so much so that examples are still usually called ‘Ruhmkorff coils’ rather than ‘Callan coils’. By 1878 (just after his death) Heinrich’s firm had a turnover of around 200,000 francs, and his apparatus was so well known, appearing in all the text books, that he apparently didn’t even need to advertise. Certainly no printed catalogue has survived (Brenni 1994,4-8).

We are probably not aware of it, but many of us own a Callan coil of our own – that is if we drive a petrol-powered car. The internal combustion engine was first demonstrated by Etienne Lenoir (1822-1900) in 1860, and taken up around 1908 by Henry Ford (1863-1947). (Henry’s father had emigrated from Ballinacarty, Co. Cork, in 1847.) The spark in our internal combustion engines is produced by an induction coil (Cawley 1998).

Maynooth University still maintains an impressive scientific tradition, and St Patrick’s College has a wonderful collection of old scientific instruments displayed in its newly-renovated ‘National Science Museum’ (*see* Mollan 1994). It includes original induction coils, batteries and other apparatus invented or used by Callan, including his ‘Maynooth Battery’ which was produced commercially. This had iron and zinc electrodes and an electrolyte of sulphuric and nitric acids. The Museum also displays the impressive Patent granted to Callan by Queen Victoria in 1853 for his method for preventing iron from rusting.

The Revd Professor Nicholas Callan, holy priest and showman, died from natural causes – and not from electrocution – at Maynooth on 10 January 1864.

Humphrey Lloyd (1800-1881)

The son of a previous Provost of Trinity College Dublin also became Provost – and indeed there was a remarkable similarity in the careers of father and son. Bartholomew Lloyd (1772-1837), the father, had been Professor of Mathematics, Greek, and Natural and Experimental Philosophy, and had lectured in divinity, before being elected to the position of Provost – a post he held from 1831 to 1837. He was President of the Royal Irish Academy from 1835 to 1837.

His son, Humphrey, was born in Dublin on 16 April 1800, and he took first place in the entrance examination to Trinity in 1815, graduating with a gold medal in science in 1819. He became Professor of Natural and Experimental Philosophy in 1831 (in succession to his father), Vice-Provost in 1862, and was Provost from 1867 to 1881. He also followed in his father’s footsteps by becoming President of the Royal Irish Academy, a post which he held from 1846 to 1851.

He carried out important scientific work on light, and received international notice when, in 1832, he experimentally verified the theoretical prediction – *conical refraction* – made by William Rowan Hamilton (*see below*) about the way light is bent in travelling through a special type of crystal – a *biaxial* crystal. He is remembered by physicists for the *Lloyd Mirror Experiment* of 1834, in which he demonstrated interference patterns between a ray of light and its reflected image. This showed that a change of phase occurs when light is reflected by a mirror. He was also the main force behind the foundation of the Engineering School in Trinity in 1841 (Mollan 1995a,18).

The earth’s magnetic field, as we measure it at the surface, is a complex one, arising from eddy currents in the earth’s core of molten metal, modified by the ever-changing solar wind which streams past the earth in its orbit around the sun (McConnell 1986,26). The use of terrestrial magnetism is of great importance both to science and to navigation, but is complicated by the fact that magnetic north (as determined with a horizontal magnetic needle) and geographical north are not usually the same, and that the direction of magnetic north varies in both the short and long term. So also does the vertical component of the earth’s magnetic field, measured by a magnetic needle which departs from the vertical to greatly differing extents at different points on the earth’s surface.

In order to study the direction and intensity of terrestrial magnetism, Humphrey in 1837-1838 set up a Magnetical Observatory in the Fellow's Garden in Trinity College (the building was removed to the UCD Belfield Campus in 1974 when the new Arts Block was built at Trinity). Lloyd used instruments made in Paris and London, and himself devised several new or modified instruments. In 1838, the British Association for the Advancement of Science, with government support, undertook the setting up of permanent stations to make simultaneous magnetic measurements in different parts of the world, and Lloyd's observatory was used as a model for the other stations (*DSB* VIII,426).

Humphrey Lloyd died in the Provost's House in Trinity College Dublin on 17 January 1881.

William Parsons, third Earl of Rosse (1800-1867)

The Parsons family has lived at Birr, in Co. Offaly, since 1620 (Moore 1971,2). Its most famous member was William Parsons, although he was actually born at York, England, on 17 June 1800, the eldest son of the second Earl, Sir Lawrence Parsons (1758-1841), and his wife Alice (née Lloyd), from Gloster in Co. Offaly. He succeeded to the family title as third Earl of Rosse in 1841.

His father didn't approve of boarding schools, so William was educated at home in Birr until the age of 18, when he went up to Trinity College Dublin. He transferred to Magdalen College, Oxford, where he graduated first class in mathematics in 1822 (*DNB* XV,425). Although evidently a talented mathematician, his great love was for engineering. If he had gone to school, his passion for practical pursuits would probably have been beaten out of him (Scaife 1988,104-5). At that time, it wasn't considered proper for a gentleman, much less an aristocrat, to get his hands dirty with practical engineering projects.

Some time earlier, a famous German-born astronomer – William Herschel (1738-1822) – had built large reflecting telescopes and had discovered the planet Uranus in 1781 – the first planet to be discovered since ancient times. Herschel had died without passing on the secrets of his trade, so no-one had equalled his success in making large telescopes, much less improved on it (Mollan 1996a,8-9). The engineering challenge appealed to William Parsons, and of course he appreciated correctly that larger and better telescopes would improve our understanding of our universe. So he got to work.

He trained local craftsmen and with them carried out amazing experiments to cast, grind and polish large metal mirrors made of speculum metal, a very reflective alloy of copper and tin. This alloy, though, was very brittle, and the castings kept fracturing. But, after much effort, he succeeded in casting enormous mirrors six feet in diameter, and he invented a steam driven machine to grind and polish them to the right curvature. A key innovative breakthrough was to make a mould for the castings having a base of steel strips through which gases could escape. This prevented pock-marking on the surface of the mirror (it was cast reflective side down) (Mollan 1996a,16).

The casting of the large mirror in 1842 was observed by Thomas Romney Robinson (*see above*) – the Director of Armagh Observatory (Parsons 1926, 21):

On this occasion, besides the engrossing importance of the operation, its singular and sublime beauty can never be forgotten by those who were so fortunate as to be present. Above, the sky, crowded with stars and illuminated by a most brilliant moon, seemed to look down auspiciously on their work. Below, the furnaces poured out huge columns of nearly monochromatic yellow flame, and the ignited crucibles during their passage through the air were fountains of red light, producing on the towers of the castle and the foliage of the trees, such accidents of colour and shade as might almost transport fancy to the planets of a contrasted double star. Nor was the perfect order and arrangement of every thing less striking: each possible contingency had been foreseen, each detail carefully rehearsed; and the workmen executed their orders with a silent and unerring obedience worthy of the calm and provident self-possession in which they were given.

In the meanwhile, William had the good sense to marry an heiress – Mary Field – and so had enough money to build a massive telescope to house his mirrors. The tube was 57 feet long and up to eight feet in diameter. It was suspended between mighty masonry walls 70 feet long, 56 feet high, and 24 feet apart. The whole instrument was reported to have cost between £20,000 and £30,000 (Scaife 1999), a truly astronomical sum for an amazing astronomical instrument.

The telescope was a considerable success, seeing further into space than anything which had gone before, and it remained the largest telescope in the world from 1845 until well into this century. It attracted as visitors to Birr the most eminent scientists and others who wished to experience the wonders seen through it. Its most famous discovery was the spiral shape of some nebulae, now known to be neighbouring galaxies.

Between 1997 and 1999, the telescope, nicknamed *The Leviathan of Parsonstown* – the Leviathan being a legendary monster, and Parsonstown being an old name for Birr – was restored, and is now a magnificent sight at Birr Castle Demesne. One of the two original six foot mirrors – the second is lost – is in the Science Museum in London, but attempts to get it back have proved unsuccessful so far.

The third Earl and his Countess had eleven children, but only four boys survived to adulthood. His eldest son Laurence (1840-1908) continued the astronomical work of his father, and carried out important studies on the temperature of the moon. His youngest son, Charles (1854-1931 – *see below*), became Ireland’s finest engineer, inventing the steam turbine engine, which revolutionised electricity generation and marine transport.

Sir William Parsons, third Earl of Rosse, who received many honours in his day, including Presidency of the Royal Society from 1848-1854, died from cancer in Monkstown, Co. Dublin, on 31 October 1867.

Thomas Grubb (1800-1878) and Howard Grubb (1844-1931)

In the eighteenth and nineteenth centuries, Ireland had many scientific instrument makers (see Burnett 1989 and Mollan 1995b). But one firm stands out as the most successful and internationally respected of all, the Dublin firm founded by Thomas Grubb and continued by his youngest son, Howard.

Thomas Grubb was born in Waterford on 4 August 1800, but moved first to Kilkenny and then to Dublin. Around 1832, he set up in business supplying specialised machinery and tools, including cast iron beds for billiard tables (Glass 1997,10). He later constructed printing machines for bank notes, and he became Engineer to the Bank of Ireland in 1842 (Wayman 1989,[5]). He seems to have taken an early interest in astronomy, and he constructed an observatory with a small reflecting telescope. Soon he began to make larger instruments.

One of his first contracts was for a mounting of what was then the most powerful astronomical lens in the world – a 13.3 inch diameter lens manufactured in France between 1830 and 1834 (Wayman 1987,96) – which had been bought by Edward Cooper for his observatory at Markree Castle in Co. Sligo. Another was for the reflecting telescope ordered by Thomas Romney Robinson (*see above*), the Dublin-born director of Armagh Observatory.

The telescope supplied to Armagh in 1835 introduced important design features of great significance to the future development of the telescope (Butler 1990,54-5). These included the viewing position of the user who observed the image at the back of the instrument through a small hole in the centre of the main mirror (a ‘Cassegrain’ arrangement). Before that, observers usually viewed images from the top of the telescope, which could be both cold and uncomfortable. He also mounted the telescope so that it could revolve around an axis pointing to the North Pole, and this made it easier to keep stars in sight as the earth revolved. And he used a system of ‘equilibrated levers’ to support the mirror in such a way that the weight was distributed evenly over the back surface, thus minimising distortion of the image due to flexure. This system was later used to support the giant 72 inch diameter mirror of the *Leviathan of Parsonstown*, built at Birr by the third Earl of Rosse (*see above*).

With this successful start, the firm began to build large telescopes which were exported all over the world, to Madras, Madrid, Mecca, and Mississippi – to list a just few places beginning with the letter 'M' (Mollan 1995b,381-9). Particularly large and important telescopes were built for Melbourne in Australia (a reflecting telescope with a 48 inch diameter mirror) and for Vienna in Austria (a refracting telescope with a 27 inch diameter lens).

Howard Grubb, who was born in Dublin on 28 July 1844, joined his father in 1865, after studying at Trinity College Dublin, and the firm went from strength to strength. Their factory in Rathmines in Dublin is still remembered in the name 'Observatory Lane', now leading to the Leinster Cricket Club grounds. Unusually for instrument makers, both Thomas and Howard were elected to Fellowship of the Royal Society (in 1864 and 1883 respectively). Thomas died in Dublin on 19 September 1878. Howard was knighted in 1887 and received the Boyle Medal of the Royal Dublin Society in 1912. He could not receive the latter honour sooner, since he was an Officer of the Society, serving as Honorary Secretary from 1889 to 1893 and Vice-President from 1893 to 1922. The rules were changed in 1911 to allow Officers to be awarded the Medal if their work merited it.

The firm made other optical instruments, including periscopes, gun sights, and range finders during the First World War, and they had to re-locate to St Albans in England, because crossing the Irish Sea with this vital cargo had become so hazardous, due to the U-boat threat. By this time, Howard was 74 years old. In 1925, the firm was taken over by Charles Parsons, (see below), and Grubb-Parsons went on to make many large and famous telescopes. Howard returned to Ireland where he spent his last years in Dun Laoghaire and Monkstown. He died at Monkstown, Co. Dublin, on 16 September 1931.

Grubb telescopes can still be seen in Ireland at Armagh Observatory, at Dunsink Observatory in Castleknock, Dublin, and in the Crawford Observatory on the campus of University College Cork.

William Rowan Hamilton (1805-1865)

In 1863, the new American National Academy of Sciences appointed an Irishman, Dublin-born William Rowan Hamilton, the first of fourteen Foreign Associates (*DSB* Vol.6,86). He was considered by the Academy to be the greatest of living scientists. William is generally regarded as Ireland's top mathematician of all time, although the country has produced quite a few others distinguished in this discipline as will be seen later.

William claimed to have been born in Dublin on the stroke of midnight on 3-4 August 1805. Such precision may perhaps be expected in the mind of a mathematical genius, but one wonders whether his mother, Sarah Hutton, acknowledged such an instant birth? It doesn't seem to allow for what, presumably, were hours of painful labour! Whatever about that, young William, who was tutored by his Uncle, the Revd James Hamilton in Trim, was an infant prodigy. His father, a Dublin Solicitor called Archibald, boasted that his son knew Hebrew, Persian, Arabic, Sanscrit, Chaldee, Syriac, Hindustanee, Malay, Mahratta, and Bengali, in addition to Latin, Greek, and the modern European languages, by the age of ten (Hankins 1980,13)! One of the problems encountered was the difficulty of acquiring Chinese books in Dublin.

His uncle James had a systematic way of teaching the youngster English. William was required to master all the monosyllabic words beginning with a, then those beginning with b, and so on through the alphabet. After that, he progressed to two syllables, and so on. 'From the beginning William was working on obscure words that most adults had never heard or seen' (Hankins 1980,13).

As well as his classical and linguistic gifts he was brilliant at sums. At the age of 13, however, he was pitted against a 14-year-old American calculating wizard, called Zerah Colburn, and, to his great consternation and amazement, Zerah beat him (Hankins 1980,15). He wasn't used to this. But then Zerah could, for example, instantly calculate the number of minutes in 1811 years (remember you have to allow for leap years). Impelled to greater efforts by this setback, he entered Trinity College Dublin and came first in every classics, mathematics and science exam which he sat.

Anxious to retain his amazing gifts, Trinity College appointed him Andrews' Professor of Astronomy and Royal Astronomer of Ireland in 1827, before he had even graduated, although they knew that astronomy, especially the tedious night-time observational astronomy, was not his forte. But this adventurous move for a sometimes conservative institution was a sound one, for he will be evermore remembered as one of Trinity's greatest academic jewels.

He sprang to international academic fame by predicting in 1832 a special kind of refraction in crystals, a prediction which was experimentally verified by his colleague, Humphrey Lloyd (*see above*), Trinity Professor of Natural Philosophy. This remarkable achievement led to him being knighted by the Lord Lieutenant of Ireland in 1835, during the first meeting in Dublin of the British Association for the Advancement of Science (which had been founded in 1831). He also won the Royal Medal of the Royal Society in 1836, and was President of the Royal Irish Academy from 1837 to 1846.

His work is still widely used today. He introduced the familiar mathematical terms vector and scalar, and devised the method of *quaternions*, involving the concept of the imaginary square route of minus one. *Hamilton quaternions* are in regular use today in computer graphics and in the guidance systems of spacecraft. *Hamilton graphs* are commonly employed in modern discrete mathematics.

A habitual scribbler, William is reputed to have carried out some of his calculations on his finger nails and the shell of his breakfast egg. In 1843, on the way to a meeting of the Royal Irish Academy, of which he was then President, a flash of inspiration came to him. He inscribed his basic *quaternion* formula on Brougham Bridge, on the Royal Canal, lest he should forget it.

While blossoming as a mathematician, he wrote poetry, and was a friend of William Wordsworth and Samuel Taylor Coleridge. It is the former who is credited with encouraging him to stick to the day job, judging his poetry to be less creative than his sums (Wayman 1987,57).

Not everything went perfectly for the calculating young genius. At the age of 20 he fell deeply in love with Catherine Disney, the sister of one of his College friends. Her parents didn't approve, and she turned him down, but he retained this unrequited love for the rest of his life. Meanwhile, in 1833, he married Helen Bayly, and they had three children. Helen suffered from continued ill health, although she still managed to outlive her husband by four years.

William himself was reputed to have died from overwork, but it seems that alcohol helped him into his grave. He died at Dunsink Observatory on 2 September 1865. His statue, and that of Robert Boyle, today grace the entrance to Government Buildings in Merrion Street (formerly the College of Science) where, presumably, they inspire our legislators.

James MacCullagh (1809-1847)

One of the great treasures of the National Museum in Dublin is the Cross of Cong. It is not widely known that it was presented to the Royal Irish Academy (and subsequently transferred by the Academy to the Museum) by one of Ireland's great mathematicians James MacCullagh. It cost him 100 guineas, no mean sum for an academic at the time (Spearman 1990a,23).

James was born in Landahussy, Co. Tyrone in 1909, and graduated from Trinity College Dublin in 1828. He was elected to Fellowship in Trinity in 1832 and became Professor, first of Mathematics (1835-1843) and then of Natural and Experimental Philosophy (1843-1847).

A little younger than his Trinity colleague William

Rowan Hamilton, James was of the opinion that William's famous discovery of conical refraction (*see above*) was 'an obvious and immediate consequence of his own published work' and he evidently resented the glory which Hamilton received (Spearman 1985,213). Indeed, when going through his papers after his death, Hamilton confessed to George Salmon (*see below*) that MacCullagh 'seems to have been very near finding the theory [of conical refraction] for himself' (Hankins 1980,94).

James is particularly remembered for his work in the development of mathematical models for the aether, work which was further developed to great effect by other Irish mathematicians and physicists (*see under Joseph Larmor below*).

A sufferer from periodic bouts of depression, James took his own life in Dublin on 24 October 1847.

Robert Mallet (1810-1881)

If you refer to that multi-volume Bible of the English Language *The Oxford English Dictionary*, you will find no less than nine words starting with 'seism' – seismic, seismograph, seismological, seismologist, seismologue, seismology, seismometric, seismometry and seismoscope – which acknowledge 'Mallet' as their instigator. It was Robert Mallet, born in Dublin on 3 June 1810, who could be said to have 'invented' the science of seismology. The summary of a recent important historical study (Dean 1991,39) records:

Though the name of Robert Mallet was once inevitably associated with the scientific study of earthquakes, it is less well known today. As part of an overdue reappraisal, this essay examines Mallet's major seismological projects and publications, emphasizing his theoretical contributions. Mallet's own claim to be a founder of modern seismology is upheld. Beyond that, however, he is also seen to be an important precursor of plate tectonics.

Indeed, modern seismology can be said to have begun on Killiney beach in south Co. Dublin:

The first series of experiments began at Killiney Bay on 18 July 1849 with the laying out of a mile's distance. Gunpowder explosions would take place at one end, and their wave energies would be measured at the other by a seismoscope. Elaborately calibrated chronographs, started (like a stop-watch) by a signal from the explosion end, would record the elapsed time. All of this took weeks to prepare, however, and when the first real attempts were made, on 25 October, they failed. Mallet then reduced the range of the experiments to half a mile, adjusted his power and chronographs, and tried again. He obtained three excellent series of results on 30 and 31 October and 2 November 1849 (Dean 1991,49).

One can only speculate on the reaction of the locals or the day trippers who travelled on the train from Dublin to enjoy sunshine and bathing at the beach. Gunpowder and picnics do not seem to be compatible companions. Perhaps it was the inevitable reaction which resulted in the transfer of the experiments to offshore Dalkey Island the next year. The resident goats were likely to prove less effective complainants.

Robert's day job was as a most distinguished civil engineer, who built up his father's iron, brass and copper founding business into a practice which 'ultimately absorbed all the engineering works of note in Ireland' (*DNB*). He raised the roof of St George's Church in Dublin, bored an artesian well for Guinness & Co, built several swivel bridges over the Shannon, built also terminal railway stations, the Fastnet Rock lighthouse, and monster mortars for the Crimean War. In 1852 he patented the buckled plate, used to strengthen flooring, combining the maximum of strength with the minimum of weight, and this was used, for example, in Westminster Bridge in London.

His lasting scientific reputation, though, remains with earthquakes. With his son, John W. Mallet, he compiled an extensive two-volume earthquake catalogue of the world (1850-1858), and was the first to measure an earthquake's focus or epicentre – in Naples in 1858 (Cox 1985,42). A graduate of Trinity College Dublin (1830), he became a Fellow of the Royal Society in 1854, and received the Cunningham Medal of the Royal Irish Academy in 1862. He died in Enmore in Surrey on 5 November 1881.

George Boole 1815-1864

Born in Lincoln, England, on 2 November 1815, the son of a shoemaker and a lady's maid, George Boole left school early to support his family. He earned money from teaching and had set up his own school by 1834. He taught himself advanced mathematics and gained a reputation from his published papers. In 1841 he introduced Invariant Theory, a new branch of mathematics, and the concept of invariance later became part of the inspiration for Einstein's theory of relativity. George was awarded the first ever gold medal for mathematics by the Royal Society in 1844, and was appointed the first Professor of Mathematics at Queen's College (now University College) Cork in 1849. His sister, Mary Ann, was tutor to George Francis FitzGerald (*see below*).

While at Cork he published *An investigation of the laws of thought* (1854), which introduced the mathematics used in today's computers. He has been variously described as 'the founder of pure mathematics, father of computer science, and the discoverer of symbolic logic' (MacHale 1985) – not a bad bag for an early school leaver! Concepts such as set theory, binary numbers and probability can all be traced to his work.

Sadly, he died in Cork on 8 December 1864, aged only 49, at the height of his intellectual powers, leaving 'virtually destitute' his wife Mary (née Everest) and their five daughters aged six months to eight years (MacHale 1985,227,252). Mary's uncle, Lieutenant-Colonel Sir George Everest, was the Surveyor-General of India, after whom the highest mountain in the World is called. Mary was clearly a very resourceful woman, and several of her daughters achieved distinction in their own right – Alicia (Alice) in mathematics, Lucy in chemistry, and Ethel (Lily) in literature, the latter being particularly remembered (under her married name of Ethel Lilian Voynich or E.L.V.) for her first novel *The gadfly* (1897) (MacHale 1985,261-276).

George Salmon (1819-1904)

'Salmon's *Conic Sections* – the first 150 years of a mathematical best-seller' is the title given to a recent article by a Professor of Mathematics at University College Dublin (Gow 1998,26). That this book, written in 1848, is still in print (in the sixth reprint of its sixth edition), and that it is still useful to teachers of mathematics in today's Leaving Certificate, is a remarkable tribute to the Revd George Salmon:

Conic Sections *was the first of four highly influential textbooks written by Salmon. The other three were* Higher Plane Curves (1852), Modern Higher Algebra (1859), *and* Analytic Geometry of Three Dimensions (1862). *Historians of mathematics all attest to the important role that these textbooks played in disseminating knowledge of new developments in geometry and algebra in the second half of the nineteenth century. Their influence in Germany was particularly strong, where translations of all four books went through several editions, and they were given warm praise in the writings of the German mathematician, Felix Klein.*

George Salmon was born in either Dublin or Cork on 25 September 1819. Most references give Cork, and he certainly received his early education there, but he himself believed he was born in Dublin (Gow 1998,26), and presumably he should know! Whatever about this, he was a bright lad, for he entered Trinity College in 1833, at the age of thirteen or fourteen, won a classical scholarship in 1837, and took first place in his mathematics degree (Senior Moderatorship) in 1838. He became a Fellow in 1841, was ordained in 1844, and was Regius Professor of Divinity from 1866 to 1888, when he became Provost – a post he held until his death in Trinity on 22 January 1904. He received the Cunningham Medal of the Royal Irish Academy in 1858. He was elected Fellow of the Royal Society in 1863, and was awarded its Royal Medal in 1868 and its Copley Medal in 1889.

While he is recognised for his contributions to mathematics – his researches on surfaces being particularly noteworthy (Gow 1997,36) – he is chiefly remembered for his textbooks. The review of the third edition of Conic Sections in the *Philosophical Magazine* (written by English mathematician T.A. Hirst) noted:

The treatment throughout is admirably clear, strict and elegant – in fact such as can be achieved only by one who, besides that perfect mastery of the subject which can only be acquired by original research, possesses also that unacquirable talent of lucid exposition, and is guided by that knowledge of the difficulties usually encountered by students, which experience only can give (quoted in Gow 1998,26).

In keeping with his calling as a cleric, George also wrote extensively on religious matters. His book *The infallibility of the church* (1889, second edition 1890) was 'a trenchant criticism of papal infallibility, which reflected the author's deep distrust of any authority over the individual conscience' (Spearman 1990c,35). He played a central and constructive part in the reorganisation of the Church of Ireland after its disestablishment in 1870.

As Provost he was conservative. The admission of women to university degrees at Trinity College, which was finally permitted in the last year of his provostship, was introduced against his better judgement (*DNB*). The views of the elderly male Board on female access to Trinity at the time have been summarised as:

Once a woman was in College it would be impossible to know where she went and how long she stayed, or, indeed, whether she ever came out at all…. An occasional scandal was bound to occur, and in any case many a susceptible boy would become entangled in an imprudent marriage. From such liaisons dangereuses and fortune-hunters the Board has a duty to protect its students (McDowell 1982,347).

George Gabriel Stokes (1819-1903)

A baby called Victoria was born in London on 24 May 1819, the same year which saw the birth, on 13 August, of the seventh child of a Church of Ireland Rector in the distant village of Skreen, Co. Sligo. In 1837 Victoria ascended to the throne, the same year in which George Gabriel Stokes entered Pembroke College, Cambridge. George became one of many Irishmen who added distinction to the Empire of Queen Victoria in a variety of fields of endeavour. He was elected President of the Royal Society in Victoria's London, the highest scientific honour in the world at the time, and he held this position from 1885 to 1890. He had, before that, been Secretary to the Royal Society for no less than 31 years – from 1854 to 1885 – so he had definitely earned his elevation. He was followed in this exalted post by another Irish-born innovator – William Thomson, Lord Kelvin (*see below*), who was President for the next five years.

The very first President of the Royal Society (from 1662 to 1677) was mathematician William, second Viscount Brouncker (c.1620-1684) 'of Castle Lyons in the Irish peerage' (DNB). Presumably this is Castlelyons in Co. Cork. Robert Boyle, from neighbouring Co. Waterford, one of the founders of the Society, was invited to be President in 1680, but he declined the honour due to scruples about taking oaths. Other Irishmen, though, have held this distinction since then: Sir Robert Southwell from Kinsale Co. Cork was President from 1690-1695; Sir Hans Sloane (Founder of the British Museum) from Killyleagh Co. Down from 1727-1741, and William Parsons, third Earl of Rosse (*see above*), from Birr Co. Offaly from 1849-1854. The first woman to be admitted to Fellowship of the Royal Society was the tenth child of the postmaster of Newbridge, Co. Kildare, X-ray crystallographer, Kathleen Lonsdale (1903-1971), who achieved this distinction in 1945, 'only' some 285 years after the foundation of the Society. (Biochemist Marjory Stephenson was also admitted on the same day – 17 May 1945.) Kathleen was later (1956) created a Dame Commander of the Order of the British Empire.

Due to certain differences of opinion between Ireland and Britain over the years, there has been a tendency to ignore the Irish background of such distinguished people. Perhaps we have now reached the position where we can be proud of them and, indeed, acknowledge the opportunities which Britain granted them – to the long-term advantage of the pursuit of knowledge.

Following a distinguished undergraduate career, culminating in first place in mathematics in Cambridge University in 1841 (the student who achieves this distinction is called the 'Senior Wrangler'), George Stokes remained on at Cambridge, being appointed in 1849 to the Lucasian Chair of Mathematics – a post held earlier by Isaac Newton and later by Stephen Hawking (Wood 1995,49). He held this position for 53 years. In 1859 he married Irish girl Mary Susannah Robinson, daughter of Thomas Romney Robinson of Armagh (*see above*).

There is no doubt that George Gabriel was greatly inspired by his upbringing in the West of Ireland, and he returned regularly for the summer vacation, a non-trivial exercise in the pre-railway era, while a student in England. Even after the death of his parents he continued to visit his brother John Whitley, then a clergyman in Tyrone, and his sister, Elizabeth Mary, to whom he was greatly attached, in Malahide, almost annually until his death (Wood 1995,52).

A cousin of G.G. – Dublin-born William Stokes (1804-1878) – is remembered in medicine by *Cheyne-Stokes Respiration*, and by the *Stokes-Adams Syndrome* in cardiology. George Gabriel is remembered in mathematics in *Stokes' Law* (dealing with viscosity), *Stokes Layer* (a boundary between a fluid and solid), *Stokes' Theorem* (relating to the surface integral of the curl of a vector), *Stokes' Line* (in spectral studies), *Stokes' Conjecture* (about waves of greatest height), the *Navier-Stokes equations* (dealing with fluid motion), and the *Campbell-Stokes Sunshine Recorder* (Wood 1999, Walker 1995,1051, Mollan 1995b,88). The CGS unit of kinematic viscosity is the stokes.

George's early research was in hydrodynamics, in which he introduced the concept of 'internal friction' of an incompressible fluid, a concept which could also be applied to other continuous media like elastic solids. He studied oscillatory waves in water, and fluid motion. His major advance was in the theory of light, in which he examined mathematically the properties of the aether, treating it as an incompressible elastic medium. This enabled him to obtain results on the mathematical theory of diffraction, which he confirmed by experiment, and on fluorescence (a term which he introduced), leading him into the field of spectrum analysis (Wood 1999). He made contributions to both theoretical and practical science.

He became a Fellow of the Royal Society in 1851, Member of Parliament for Cambridge University in 1887, and was knighted by Queen Victoria in 1889. He died at Cambridge on 1 February 1903, and his obituary in The Times of London, two days later, recorded:

It is sometimes supposed…that minds conversant with the higher mathematics are unfit to deal with the ordinary affairs of life. Sir George Stokes was a living proof that if the mathematician is only big enough, his intellect will handle practical questions so easily and as well as mathematical formulas (quoted in Wood 1995,57).

John Tyndall (1820-1893)

John Tyndall was one of the great public figures of nineteenth-century British science. As a researcher, an educator, a lecturer and a controversialist, he played a major role in the professionalization and popularization of science (Attis 1999c).

Born in Leighlinbridge, Co. Carlow, on 2 August 1820, John was educated locally and, at the age of 18, joined the Ordnance Survey in Ireland. After three years, he was in 1842 chosen to transfer to the English Survey, but he was summarily dismissed in 1843 following a protest about working conditions of Irish assistants (Burchfield 1981,2). After returning home for a spell, he worked in England as a surveyor for the railways, and then, in 1847, became a mathematics teacher in Queenwood College in Hampshire. His interest in science stimulated there, he transferred to the laboratory of the famous Robert Bunsen (1811-1899) in Marburg, Germany, where he obtained his PhD within two years. After a further year in Germany, he returned to England in 1851, but was unsuccessful in applications for academic posts in universities in Toronto, Sydney, Cork and Galway.

His abilities as a lecturer, however, led to his appointment as Professor of Natural Philosophy at the Royal Institution in 1853. He was to remain there for a further 34 years, succeeding Michael Faraday (1791-1867) as Director in 1867.

Almost from the moment he arrived in London, Tyndall became an evangelist for the cause of science. His experience in Germany had convinced him that traditional British education, with its emphasis on the classics and rote mathematics was hopelessly outdated and detrimental to the good of the country. Thus, his lectures at the Royal Institution were not merely to entertain or even to instruct his audience, but to awaken them to the beauty and importance of science (Burchfield 1981,6).

John had wide interests, studying diamagnetism and the magneto-optical properties of crystals. He studied glaciers, solar heat and radiation. He was the first to explain why the sky is blue – a result of the scattering of light of different wavelengths to different extents in the atmosphere. He made the first pollution measurements of the London atmosphere, and was involved in early work on bacteriology. Several scientific phenomena are named after him – including the *Tyndall Effect* and *Tyndall Scattering* (both referring to the scattering of light), *Tyndallimetry* (dealing with the concentration of suspended material in a liquid) and the related *Tyndall Cone*, and *Tyndallisation* (used in bacteriology) (Walker 1995,1143). His scientific reputation was such that he was awarded honorary degrees from Cambridge (1865), Edinburgh (1866), Oxford (1873), Tubingen (1877), and Trinity College Dublin (1886) (Burchfield 1981,5).

He helped found the famous journal *Nature* in 1869, was a prolific author of books, and played an important part in advancing the cause of science in America (Sopka 1981,193-203).

John remained a bachelor until the age of 56, when he married Louisa Hamilton. It was a very happy, though childless, marriage – but it ended in tragedy at Hindhead in Surrey on 4 December 1893, when Louisa mistakenly administered to him a lethal dose of chloral (Burchfield 1981,12).

Samuel Haughton (1821-1897)

Mathematics has been called 'the Queen of the sciences', and there is no doubt that mathematical analysis and calculation are fundamental to the advance of scientific progress and social order. Mathematicians can be called upon to answer many practical as well as theoretical questions. For example, in the days (fortunately now past in this country) when capital punishment was not unusual, a humanitarian problem was to ensure that the unfortunate condemned criminal did indeed meet his final end when subject to the hangman's noose. It was a Trinity College Dublin cleric and polymath who was called upon to provide a formula which determined how far the condemned should fall, according to his weight, to ensure instantaneous death. The result became known as *Haughton's drop* (O'Brien 1984,141).

Samuel Haughton was born in Carlow on 21 December 1821, was educated locally, and entered Trinity College at the age of seventeen, graduating with first place in mathematics (Senior Moderatorship) in 1843. After obtaining Fellowship in 1844, he took holy orders the following year. Initially he followed in the footsteps of his teacher, James MacCullagh (*see above*) and, indeed, later co-edited MacCullagh's papers for publication (1880). He worked on elasticity – what would now be called continuum mechanics. He received the Cunningham Medal of the Royal Irish Academy in 1848 for his memoir *On the equilibrium and motion of solid and fluid bodies* (Spearman 1990b,36). He shared MacCullagh's interest in the refraction of polarised light by crystalline media, and turned his interest to mineralogy. In 1851 he was appointed Professor of Geology and Mineralogy in Trinity, a post he held for 30 years. He carried out pioneering work on the chemical composition of rocks, and applied his mathematical skills to topics such as tidal motion, planetary equilibrium and the age of the earth (*see under John Joly below*).

Acquiring an interest in anatomy, Samuel, while still holding his Professorship, and already a Fellow of the Royal Society (1858), entered the Trinity Medical School in 1859, graduating with a Bachelor in Medicine degree in 1862: the reaction of his fellow students to this relatively elderly Professor in their midst is not recorded, but can be imagined! He worked on muscular action and the mechanism of joints, and published his results in his 1872 book *The principles of animal mechanics* (Spearman 1985,224).

If you visit the Zoo in Dublin, you may eat in the restaurant, which is called 'Haughton House' as it was built in his memory by public subscription. His connection with the Royal Zoological Society lasted many years: he became a member of Council in 1860, Honorary Secretary in 1864 and President in 1883. 'But for his energy in grappling with the financial difficulties with which the Society was beset during his period of office as Secretary, it would probably have ceased to exist' (*DNB*). As already mentioned, mathematicians can be very useful!

The Revd Samuel Haughton died in Dublin on 31 October 1897, and was buried in Carlow.

William Thomson, Lord Kelvin (1824-1907)

One of the greatest of all nineteenth century physicists was William Thomson who was born in Belfast on 26 June 1824. His Ulster family background played a very important part in his outlook on life and his approach to science:

Sir William Thomson's elevation to the peerage as Baron Kelvin of Largs symbolized the social summit of a remarkable life lived in the context of Victorian Britain. Yet that ultimate acclaim did not flow from scientific and technical achievement alone, but also from his direct involvement in the political cause of Liberal Unionism during the 1880s. That involvement derived from the Irish context into which he had been born, a context of cultural and social liberalism upon which his enduring personal values were founded (Smith 1989,3).

He was no ivory tower theorist, but a man of remarkable intellect who addressed and resolved many industrial challenges.

Of his numerous contributions to science and industry, he is particularly remembered as being the brains behind the successful transatlantic telegraph cable of 1866. And the 'absolute' scale of temperature is called after him, the freezing point of water on this scale being 273.15 kelvin.

William's father was James Thomson, Professor of Mathematics at the Academical Institution in Belfast – known as 'Inst' to its students. James was appointed Professor of Mathematics at Glasgow University, and the family moved there in 1832. William studied in Glasgow and then at Cambridge, where he graduated in 1845. While in Cambridge, he was a champion oarsman, and one of the founders of the University Music Society. It is often the case that ability in science and in music go together. After a brief period in Paris, William returned to Glasgow University as the Professor of Natural Philosophy, at the early age of 22. He retired from this position in 1899, 53 years later, having resisted invitations to move elsewhere. He was knighted in 1866, created a peer, Baron Kelvin of Largs, in 1892, and was conferred with the Order of Merit in 1902. He was President of the Royal Society from 1890 to 1895. Although married twice – to Margaret Crum in 1852 (she died in 1870) and to Frances Blandy in 1874 – he had no children.

In Glasgow, William carried out pioneering work on mechanical energy and heat, the basis of modern low-temperature engineering. He helped to develop the second law of thermodynamics, which sets limits to the efficiency of engines (Attis 1999b). He demonstrated the application of the reversible heat engine to refrigeration (Garvin 1993,19).

The first transatlantic cable, laid in 1858, did not work for long, largely due to poor insulation and to the large voltages which were used. Thomson invented a very sensitive mirror galvanometer which could detect extremely feeble signals, and this was the essential element in the successful cable of 1866, which ran from Valentia Island off Co. Kerry to Trinity Bay, Newfoundland, in Canada. His mathematical work had provided a basic understanding of how signals were sent, and his laboratory tested the purity of the copper in the cable (Attis 1999b). He received his knighthood for this work.

William was actually aboard the *Great Eastern*, the ship which laid the cable. The first officer and key navigator was another Irishman, Robert Halpin (1836-1894 – *see* Rees 1992) from Wicklow Town. Halpin was appointed captain of the ship in 1869, and went on to earn further distinction for cable laying, being responsible for connecting Portugal and Brazil, Suez and Aden, Aden and India, India and Singapore, Singapore and Australia (Rees 1994,37). He earned the nickname 'Mr Cable'.

William Thomson has another claim to nautical fame. He improved compasses for navigation, and his new compass was adopted by all British Navy vessels.

Ships' compasses used before 1880 consisted of a pivoted magnet under a heavy card. They were very sluggish and inclined to stick, and were useless in stormy weather or on a battleship during gunfire. With the increase in the number of iron ships, compasses were proving unsatisfactory because they were affected by the metal. Thomson designed a new improved compass which consisted of eight narrow steel needles suspended from a light aluminium ring under a thin circular card (Garvin 1993,20).

William formed a working relationship, and later a company, with the Glasgow instrument maker James White, and together they produced and sold a new generation of precise electrical and other instruments patented by Thomson (Mollan 1995b,496). He accumulated great wealth through his patents and business interests.

William didn't always get it right. Based on the cooling of the earth, he estimated in 1846 that the age of the earth was about 100 million years. Another Irishman, John Joly (*see below*), using radioactive decay in minerals, played a key role in establishing that it was much older than this – about 4500 million years old.

Lord Kelvin died in Glasgow on 17 December 1907, and is buried in Westminster Abbey. A commemorative statue in the Botanic Gardens, Belfast, is inscribed 'He elucidated the laws of nature for the service of man'.

George Johnstone Stoney (1826-1911)

Co. Offaly has been an important breeding ground for Irish scientists. While the Parsons family made the greatest contributions to science and technology (see William Parsons – *above*, and Charles Parsons – *below*), others have confirmed places in the history of science. These include The Hon. Mary Ward and John Joly. All of these have merited mention in the present select list of Ireland's innovators in the physical sciences.

One of the most eminent of Offaly's sons was George Johnstone Stoney, born at Oakley Park, near Birr, on 15 February 1826, who is best remembered for introducing the term 'electron' to science in the *Transactions of the Royal Dublin Society* (Volume 4, p.583) in 1891. This was an important conceptual step leading to the eventual physical discovery of the electron by J.J. Thomson at Cambridge in 1897-1899. George was studying the theoretical framework which gave rise to spectral lines. He tried, without ultimate success, to explain the lines by vibrations analogous to those observed in sound waves, with the lines being related to a fundamental frequency and its harmonics (O'Hara 1993,11-12). Although this concept had to be abandoned, it had a profound positive influence on the further development of spectral studies. In the course of this work, he introduced the electron as a quantum of electrical charge associated with each chemical bond. He considered his electrons in terms of a solar-system type of molecule, with the electrons moving in perturbed elliptical orbits. After Thomson's 'discovery' of what he called 'corpuscles', it was George Francis FitzGerald (Stoney's nephew – *see below*) who first suggested that free electrons could exist, and Thomson's corpuscles were renamed electrons (Mollan 1992,57-9). The term has survived for one of the most important entities in the universe as we know it.

After taking second place in mathematics in Trinity College Dublin in 1847, George decided to study for the Trinity Fellowship – a stepping stone to an academic career. While working for this, he was employed by the third Earl of Rosse (*see above*) as astronomical assistant at the Birr Observatory between 1848 and 1850. Unfortunately he was unsuccessful in the Fellowship process, a bitter disappointment to him, but also, in retrospect, a great loss to Trinity for, had he been able to stay there, he would undoubtedly have had an even more distinguished scientific career. In the event, after another stint at Birr in 1852, and then appointment as Professor of Natural Philosophy at Queen's College, Galway, he became Secretary of the Queen's University in 1857. This was the administrative body for the colleges at Belfast, Cork and Galway. He thus had to carry out much of his scientific work on an out-of-hours basis, and he made use of the facilities of the Royal Dublin Society (RDS) for this purpose. The volume and quality of his scientific work was impressive in spite of his difficult circumstances, which were compounded by the untimely death of his wife in 1872, and two severe illnesses – small-pox in 1875 and typhoid in 1877 (*DNB*). He was Honorary Secretary of the RDS from 1871 to 1881, Vice-President from 1881 to 1911, and received the Society's first Boyle Medal in 1899. He was elected to Fellowship of the Royal Society in 1861, and served as Vice-President between 1898 and 1899, and on Council from 1898 to 1900.

Although primarily known for his work in theoretical physics, Stoney had a practical turn of mind also, and he published papers: *On the energy expended in propelling a bicycle; On gearing for bicycles and tricycles; On the possibility of prolonging the tones of a pianoforte so as to produce an instrument with the quality of tone of the piano, and the continuity of tone of the organ;* and *On a mounting for the specula of reflecting telescopes, designed to remove the impediment of their being used for celestial photography and spectroscopy*, all in the scientific journals of the Royal Dublin Society (Mollan 1987,109-110). His interest in music was also demonstrated by his role in the introduction of recitals to the cultural programme of the RDS. He devised a new form of spectroscope, which incorporated a large-sized silvered prism to provide greater dispersion (instead of the more-usual train of smaller prisms used for this purpose), together with a large convex mirror (Mollan 1995a,27).

Concerned at the expense of heliostats – mirrors turned by a clockwork mechanism to compensate for the movement of the sun and thus obtain steady beams of light for optical experiments – he designed, around 1867, a relatively inexpensive version, which was subsequently produced commercially in Ireland and England. He reported on this to the 1869 Exeter meeting of the British Association for the Advancement of Science:

This heliostat was planned throughout with a view to cheapness. It costs only five guineas and yet, in the opinion of the author, who has used the first of them for a year and a half, is quite as efficient as the more expensive instruments. It....has the adjustments of the mirror under easy control, and is adapted for use at any station within a range of latitude of five or six degrees (Stoney 1869,53).

In 1893 George left Dublin for London so that his daughters could have the opportunity of a university education, something not possible at the time in Dublin (*DNB*). He died at Notting Hill Gate on 5 July 1911, though his ashes were brought back home to Dublin for burial (*DNB*).

The practical genes in the Stoney family are demonstrated by other members. His younger brother, Bindon Blood Stoney (1828-1909), who had also worked as assistant to the third Earl of Rosse, became a distinguished engineer, holding the post of Chief Engineer to the Dublin Port and Docks Board from 1862 to 1898. He published the important work *The theory of stresses in girders and similar structures* (1866 and subsequent editions) and was elected Fellow of the Royal Society in 1881 (*DNB*).

George's eldest child, George Gerald Stoney (1863-1942), also distinguished himself in the field of engineering. After graduation in 1886, he was employed briefly by his uncle Bindon Blood, but soon joined Charles Algernon Parsons (*see below*), who was then working in the Firm of Clarke, Chapman & Co in Gateshead in England. As George Johnstone Stoney had been tutor to Charles Parsons, the two young men were well known to each other. There followed a long and productive collaboration between these Irishmen which lasted until 1912. In particular, George Gerald was a key player in the development of the Parsons' steam turbine engine (*see below*) in the firm of C.A. Parsons & Co at Heaton, Newcastle, which had been founded in 1889. He was elected to Fellowship of the Royal Society in 1911 (*DNB*).

Mary Ward (1827-1869)

Irish women scientists have not received the credit they deserve. For many years, it was generally considered not quite proper for women to engage in such activities. But a number of Irish women did make useful contributions to the development of science. One of these was Mary King, born near Ferbane, Co. Offaly, on 27 April 1927. She was a first cousin of the third Earl of Rosse (*see above*), although 27 years his junior, as their mothers were sisters.

A constant visitor to Birr Castle, she took an early interest in microscopy and astronomy, and persuaded her parents to buy her a fine microscope in 1845 (McKenna-Lawlor 1998,32). She used this to study butterflies and other biological specimens, and she privately published her first book *Sketches with the microscope* in 1857, including her own illustrations (she was an accomplished artist), the first book on microscopy to be written and published by a woman.

In the meantime, she had married Henry Ward of Castle Ward in Co. Down in 1854. 'The dashing second son of the third Viscount Bangor' (McKenna-Lawlor 1998,34), Henry was something of a waster. He resigned his commission in the 43rd Regiment with the rank of Captain in 1855, and 'pursued no formal employment but enjoyed, instead, a life of hunting, fishing, playing cricket and attending social gatherings'. As the family finances dwindled, Mary had a total of eleven pregnancies, and eight live infants, between 1855 and 1867. All this was not exactly helpful to her scientific pursuits, but she continued to write and publish books: *The world of wonders revealed through the microscope* (1858), *Telescope teachings* (1859), and *Microscope teachings* (1864). Her book on the telescope included valuable observations and drawings of the comet named after G.B. Donati of the Museum of Florence, which was a 'spectacular feature' of the night skies between September and October 1858 (McKenna-Lawlor 1998, 43-49).

These books on the microscope and telescope were re-issued in several editions, and they established Mary's reputation as an outstanding woman writer on popular scientific topics. As an example of the esteem in which she was regarded, she became one of only three women entitled to receive the *Monthly Notices of the Royal Astronomical Society*. Women, of course, could not become Members or Fellows, or even university graduates, at that time. Her companions in receipt of the *Notices* were Queen Victoria and Mary Somerville (McKenna-Lawlor 1998,50).

Unfortunately her connection with Birr Castle, with its passion for engineering, led to her untimely death. She was travelling in a steam carriage designed and built by the third Earl. Going around a sharp corner, she was thrown from the vehicle, and was crushed by one of its heavy wheels, dying instantly at the early age of 42, leaving behind her large family.

Robert Stawell Ball (1840-1913)

It is not generally well-known that Irish verbal ability is not confined to the world of fiction. There have been generations of distinguished Irish science writers, particularly of text books, as will be evident from the series of brief biographies in this article. (A good example is Mary Ward just considered above.)

Of all the Irish popularisers of science in the second half of the nineteenth century, Robert Stawell Ball probably ranks in the top three, following in the footsteps (or, rather, loquacity!) of his somewhat older contemporaries John Tyndall (*see above*) and Monsignor Gerald Molloy (1834-1906) – the latter being the Rector of the Catholic University in Dublin from 1884. They were all superb lecturers, and they all wrote books.

Robert Ball was born in Dublin on 1 July 1840, and attended Trinity College, where he had a distinguished career. Like George Johnstone Stoney (*see above*), he did not succeed in obtaining Fellowship. Also like Stoney, he worked for the third Earl of Rosse, as tutor to his children and astronomical assistant, from 1865 to 1867 (Wayman 1987,107-125).

He was then appointed Professor of Applied Mathematics at the Royal College of Science in Dublin and, in 1874, he became Royal Astronomer of Ireland and Andrews' Professor of Astronomy at Trinity, residing at Dunsink Observatory. Finally, he became Lowndean Professor of Astronomy at Cambridge in 1892, a position he retained until his death at Cambridge on 25 November 1913. He became a Fellow of the Royal Society in 1873, and was Knighted in 1886. He carried out important work on the theory of screws, and was considered to be 'one of the two or three greatest British mathematicians of his generation' (*DNB*).

However he is best remembered for his astronomy books and his lectures. Between 1877 and 1906, he published no less than thirteen popular works, including: *The story of the heavens* (1886 and subsequent editions); *An atlas of astronomy* (1892); *The story of the sun* (1893); *Great astronomers* (1895); *A popular guide to the heavens* (1905 and subsequent editions); and *Comets and shooting stars* (1910) (Wayman 1987,123-4). He delivered courses of Christmas lectures at the Royal Institution in London, and travelled to America on lecturing tours in 1884, 1887 and 1901.

John Boyd Dunlop (1840-1921)

Transport museums usually display a range of old bicycles – including one nicknamed 'the bone rattler'. When cobble-stones were common, and when tyres were of solid rubber, such a nickname was well deserved. The smooth ride we get today is due to John Dunlop.

John was born at Dreghorn in Ayrshire in Scotland on 5 February 1840, but he moved to Belfast in his late 20s, where he set up a successful veterinary practice. His son, also John, was a keen cyclist and, at the end of 1887, John senior made for him a rubber tube, fixed it to a wooden disc, and inflated it with a football pump. After some successful tests, he applied for a patent on the idea, which was granted in 1888 (Garvin 1993,26-9):

An Improvement in Tyres or Wheels for Bicycles, Tricycles or other Road Cars…A hollow tyre or tube made of india-rubber and cloth, or other suitable material, said tube or tyre to contain air under pressure or otherwise and to be attached to the wheel or wheels in such a method as may be found most suitable (Patent 10607).

At the Queen's College Sports in Belfast in 1889, a bicycle fitted with Dunlop's pneumatic tyres, and ridden by the captain of a local cycling club – W. Hume – easily won the cycle race, 'beating superior riders mounted on solid-tyred cycles' (*DNB*), an event which heralded a revolution in road transport. Everyone is now familiar with Dunlop tyres.

John himself moved to Dublin in 1892. Perhaps unsurprisingly, the residents of Westland Row in Dublin, where the tyres were manufactured, objected to the smell, and this led to the transfer of the company to England – first to Coventry and then to Birmingham. 'Thus was lost to Dublin and Ireland an industry which grew to enormous proportions' (*The Irish Times* 30 March 1998). John remained on in Dublin, where he died on 23 October 1921.

John Philip Holland (1841-1914)

You can pit sword against sword, rifle against rifle, cannon against cannon, iron-clad against iron-clad. You can send torpedo-boat destroyers against torpedo-boats, and destroyers against destroyers. But you can send nothing against the submarine boat, not even itself…..

You cannot see under water. Hence, you cannot fight under water. Hence, you cannot defend yourself against an attack under water, except by running away. If you cannot run away, you are *doomed. Wharves, shipping at anchor, the buildings in seaport towns cannot run away. Therefore, the sending of a submarine against them means their inevitable destruction….*

It may be that the tacticians can solve the problem. To me it is the most profound puzzle. To me there seems to be but one solution, and that is too Utopian for serious consideration. Nations with sea ports will have to refrain from making war.

Unfortunately it was too Utopian. But it was an interesting philosophy, as penned by John Philip Holland (Holland 1900,894). He was a small bespectacled man who was born in Liscannor in Co. Clare around 24 February 1841. His boyhood ambition was to go to sea, but his sight was too bad. It has been suggested that his eye condition was due to a diet of Indian Meal provided for the starving Irish during the years of the famine (Rynne 1987).

Whatever about that, he took the initial vows of the Christian Brothers, and perhaps his first real taste of warfare was when he attempted to drum knowledge into the students of North Mon school in Cork, his first teaching assignment! With his background, a nationalist ideology is hardly surprising. His views were reputably reflected in his belief that:

We can't beat the English on land, we can't beat them in the air, and we can't beat the world's greatest navy on the seas, so let's beat them underwater (Rynne 1987).

That this is apocryphal is confirmed in that air warfare wasn't invented at the time, but it likely reflected his sentiments nonetheless. Anyway, his thoughts apparently turned to the submarine after the abortive Fenian rising of 1867 and he experimented with a clockwork boat in his bath-tub. Lack of funds meant that he couldn't build a real one-man submarine as he wished.

He solved his financial problem by obtaining dispensation from the Christian Brothers in 1873, before taking his final vows, emigrating to America, and receiving sponsorship from Clan na Gael. There he constructed an increasingly sophisticated range of submarines, one of the most famous being the aptly-named *Fenian Ram*. Unfortunately, internal feuding between Clan na Gael and the Fenian Brotherhood led to the theft of the *Ram* from his boatyard, but he persevered. He stoutly defended the safety of the submarine, writing (Holland 1900,901):

For twenty-one years, I have been experimenting with submarine craft. I have travelled in submerged boats under all sorts of conditions and with all sorts of crews. All my work has been experimental, the most dangerous stage of any mode of travel. Yet I have never had an accident.

His perseverance paid off. His 1897 design, and variants of it, were purchased by the US Navy and by the navies of not only the Japanese and Russian Empires, but even by that of the old enemy itself – Britain. The Royal Navy A1 – its first submarine – built at Barrow in October 1900 was constructed to his design. His submarines, which were propelled by electricity underwater, stabilised by hydroplanes, relied on buoyancy control for diving and staying down, came into service around the world, and remain the basis of the modern submarine (Murphy 1989,41).

And he, himself, didn't end up a millionaire. The John P. Holland Torpedo Boat Company, founded in 1893, was taken over by the Electric Boat Company, and he was forced to resign in 1904 over legal disputes about patents (Collins 1999). Subsequent fighting through the courts cost him all his money.

He died from pneumonia in Newark, New Jersey, on 12 August 1914, just as submarine warfare, making full use of his best ideas, really began. Sadly it was a German U-boat which, in 1915, sank one of the finest non-military vessels – the *Lusitania* – fitted with the turbine engine invented by Ireland's greatest engineer, Sir Charles Parsons (*see below*). Holland's dream of nations with sea-ports refraining from making war was dramatically unfulfilled.

Agnes Mary Clerke (1842-1907)

It was a remarkably audacious thing to do. A woman with no formal scientific education dared to criticise professional male astronomers for neglecting important questions in their subject. She hadn't even gone to school and, of course, university education wasn't available to girls then. And, although initially treated with not a little sarcasm – Sir David Gill (1843-1914) commented 'no woman could do justice to [this] noble science' – she soon earned the respect of these conservative men, and has merited a lasting place in the history of astronomy.

Agnes Clerke was born in Skibbereen in Co. Cork on 10 February 1842 into an talented family. Her father, John William Clerke, mounted a telescope in their garden, and by the age of 15 Agnes had decided that she would write a history of astronomy (McKenna-Lawlor 1998,58). Due to her delicate health, she spent time in Italy, and became a talented linguist, and this meant that she could read the astronomical literature in the original texts.

She fulfilled her ambition with her book *A popular history of astronomy during the nineteenth century,* the first edition of which was published in 1885. It was a triumph, made her reputation, and ran to four editions. Following this success, she got the opportunity for the first time actually to take part in astronomical observations – at the Royal Observatory at the Cape in South Africa. She then proceeded to write more books, including *The system of the stars* (1890, second edition 1905), and her most influential book *Problems in astrophysics* (1903). Referring to the latter, the same Sir David Gill commented:

I do not believe that there is a man living who knew before hand all the facts that you have brought together, and brought together so well in their proper places (quoted in McKenna-Lawlor 1998,65).

There was a convert indeed! And she soon had the signal honour of being invited to become an Honorary Member of the Royal Astronomical Society (1903). Agnes has been described as 'the chief astronomical writer of the English speaking world' (Osterbrook 1984). She died in London on 20 January 1907.

Margaret Lindsay Huggins (née Murray) (1848-1915)

Everyone, at least everyone after Noah, had observed the rainbow, but it was Isaac Newton (1642-1727) who, in 1666, 'procured me a Triangular glasse-Prisme, to try therewith the celebrated Phaenomena of Colours' (Bennett 1984,1). He reported how he could split sunlight into a spectrum of colours, and recombine these colours back into white light.

It took the combined talents of a very able German chemist Robert Bunsen (1811-1899 – best remembered for the introduction of the 'Bunsen burner' in 1855), and his equally able physicist colleague, Gustav Kirchhoff (1824-1887), to demonstrate the revolutionary significance of the analysis of spectra in 1860.

But now a new and wonderful mode of inquiry has been devised, and has rapidly taken its place as the most important of all the methods of research which science has as yet placed in the hands of her servants. I refer to spectroscopic analysis, or the analysis of light by means of the prism….

The whole power of the new method of research depends on the emission of light from an object. It matters not whether the object be in the laboratory of the chemist, or half a mile off, or a hundred millions of miles off, or in fine as far off as the most distant star; if we can only obtain light enough from it to form a distinct spectrum, we can tell what is its nature (Proctor 1899,68-69 – originally written in 1869).

In essence, different elements give rise to characteristic spectra and, by looking at spectra, chemical composition can be established. This was an electrifying technique for astronomers particularly, for they could, for the first time, tell what the heavenly bodies were made of, long before anyone could go there to find out (it will be a long time before we get to other stars!). It was rapidly established that the Moon was not, after all, made of green cheese.

Kirchhoff and Bunsen soon found sets of discrete spectral lines which they could attribute to no known element, and thereby discovered two new elements, caesium and rubidium, in the mineral waters of Baden and Dürkheim (Roscoe 1869,63).

An astronomer who took up the study of stellar spectra in a particularly successful way was William Huggins (1824-1910), working at his private observatory at Tulse Hill in London. He found, for example, that some nebulae had continuous spectra (like that of our sun), while others had line spectra, characteristic of a luminous gas. Many had thought previously that, if only a good enough telescope were obtained, then all nebulae would be resolvable into discrete stars and, indeed, this was a major motivation for the building by the third Earl of Rosse of his enormous *Leviathan of Parsonstow*n (*see above*). The fact that he was not successful in resolving many of the nebulae was due not to deficiencies in the instrument, but to the fact that they were not after all made up of discrete stars.

As well as making this dramatic discovery, Huggins made an even more momentous one – the red shift. He noticed that the characteristic spectra of stars were shifted towards the red end of the spectrum, and realised that this was a result of the 'Doppler effect' in which light emitted from objects moving away from the observer increases in wavelength (the same way that the sound from a speeding ambulance changes as it moves closer or further away). His observation, therefore, was evidence that the Universe is expanding. Cosmology was changed for all time, and his results are still leading to dramatic advances in the understanding of our Universe.

Back in Ireland, in Monkstown in Co. Dublin, a young woman read about the work of William Huggins, whom she had previously met socially in London, and she even built for herself a small spectroscope (McKenna-Lawlor 1998,75). Meanwhile William had ordered astronomical instruments from Sir Howard Grubb in Dublin (*see above*), and had come to Ireland to oversee progress. Howard introduced the 51 year old astronomer to his distant admirer, the 27 year old Margaret….They married in 1875 in the parish church at Monkstown.

This began a remarkable scientific collaboration, for Margaret did not become a conventional housewife, but was worked hard by her husband in the progressing of his work, and soon was recognised by him and their peers as a worthy collaborator, rather than just an assistant. Between them William and Margaret are acknowledged as the founders of the science of astrophysics. When, during the Diamond Jubilee of Queen Victoria in 1897, William was created a Knight Commander of the Order of the Bath, his citation read:

for the great contributions which, with the collaboration of his gifted wife, he had made to the new science of astrophysics.

William served as President of the Royal Society from 1900 to 1906. So an Irish connection – through Lady Huggins – was continued in this Presidency, following the filling of this distinguished office between 1885 and 1895 by two Irishmen – George Gabriel Stokes and Lord Kelvin (*see above*). The President between 1895 and 1900 was Joseph, Lord Lister (1827-1912), the famous surgeon.

Meanwhile Margaret had, together with Agnes Clerke (*see above*), been invited to be an Honorary Member of the Royal Astronomical Society in 1903. That not one but two Irish women should be granted this rare honour at the same time is remarkable.

In spite of being considerably his junior in age, Margaret did not survive her husband for too long, for she died in London on 24 March 1915. They had been a devoted couple, and had no children.

George Francis FitzGerald (1851-1901)

Described as 'the most talented and versatile Irish theoretical physicist of the time' (Spearman 1985,229), George Francis FitzGerald, the son of George Johnstone Stoney's sister Anne Frances, played an important role in the dynamic development of physics which was a characteristic of the last quarter of the nineteenth century. Had he not died prematurely, he would have certainly added to the continuation of that dynamism into the new century.

Born in Dublin on 3 August 1851, the son of the Rector of St Anne's Church in Dublin who later became a Church of Ireland Bishop, George had the good fortune to have as tutor the sister – Mary Ann – of mathematician George Boole (*see above*) (MacHale 1985,209). It was clearly a successful combination, for he entered Trinity College at the age of sixteen and took top place, with large gold medals, in both mathematics and experimental science in his degree examinations in 1871. He spent all of his working life in Trinity, holding the Erasmus Smith Professorship of Natural and Experimental Philosophy there from 1881 until his death. He became a Fellow of the Royal Society in 1883 and received its Royal Medal in 1899 for his contributions to theoretical physics. He was Honorary Secretary of the Royal Dublin Society from 1881 to 1889.

His major work was in the development of the theory of electromagnetic radiation, which had been introduced by Scotsman James Clerk Maxwell (1831-1879). He investigated the possibility of radiation from alternating currents and, in 1882, he suggested 'the principle of the method of production of "electric waves" which Hertz used in 1887, and he contributed much to our knowledge of their properties' (*DNB*). When Heinrich Hertz (1857-1894) actually demonstrated that detectable electromagnetic waves could be produced by alternating currents, it was FitzGerald, in his role as President in 1888 of the Mathematical and Physical Section of the British Association for the Advancement of Science, who enthusiastically endorsed Hertz's work and ensured that its importance was realised.

Albert Michelson (1852-1931) and Edward Morley (1838-1923) had tried to prove the existence of the aether by a light interference experiment, but failed. George, together with Hendrik Lorentz (1853-1928) are remembered in the FitzGerald-Lorentz contraction which attempted to explain away the negative result by suggesting that a moving body contracts in the direction of its motion (Landy 1985,62). This marked one of the first departures from classical physics and it had an important influence on Einstein and his theory of relativity (MacHale 1985,209), which, ironically, specifically rejects the aether concept (Spearman 1985,229-230).

FitzGerald does not receive the recognition he deserves because his name is not remembered by laws and theories. In a letter to Oliver Heaviside (1850-1925), dated 1889, he writes:

I admire from a distance those who contain themselves till they worked to the bottom of their results, but as I am not in the very least sensitive to having made mistakes, I rush out with all sorts of crude notions in hope that they may set others thinking and lead to some advance (quoted in DSB Vol.5,15).

He is being unfair to himself here, particularly in the word 'crude' because, as is evident from a fine collection of letters (now in the possession of the Royal Dublin Society) written to him from many eminent physicists of the time, his ideas could be extremely useful. 'He possessed extraordinary versatility…throwing out luminous suggestions with splendid prodigality and rejoicing if they were absorbed and utilised by others' (*DNB*). He inspired a generation of Irish students – people like Thomas Preston (1860-1900) and Sir John Townsend (1868-1957) (*see below*). Another inspired student in FitzGerald's time who, like Preston and Townsend, took a double 'Senior Moderatorship' (i.e. first class degree) in both mathematics and experimental science was Charles Smith, who attained top place in both subjects in the degree examinations in 1889. (Preston took third place in maths and second in science in 1885, and Townsend took first place in maths and second in science in 1890.) However, Charles was beaten in the Fellowship examination and, disheartened, retired from academic work to take holy orders, ending up in a remote country parish in Co. Donegal. He is the grand-father of the present writer!

FitzGerald took a great interest in technical education and it was largely due to his efforts that Kevin Street Technical School (now part of the Dublin Institute of Technology) was founded by Dublin Corporation in 1887 (Landy 1985,63).

He is remembered in Trinity for building and trying to fly an aeroplane in College Park – not very successfully though – leading to his nickname 'flightless Fitzgerald'!

George Francis FitzGerald died in Dublin on 22 February 1901, at the early age of 49.

Charles Parsons (1854-1931)

Ireland's most eminent engineer was the youngest son of William Parsons, the third Earl of Rosse (*see above*), from Birr Castle in Co. Offaly. Although he spent most of his childhood in Birr, Charles Parsons was actually born in London on 13 June 1854. Charles invented the steam turbine engine in 1884, and this invention made dramatic contributions to electricity generation on land, and to transport at sea. Although the diesel engine has now taken the place of the steam turbine for ships, the modern turbine for generating electricity still embodies features which were introduced by him.

As in previous generations, the Parsons children were educated at home and not at public schools. This fact must account for the very untypical interest in science and technology which the family shared. Not only would public schools of the day almost certainly not have kindled such interests, but they would more likely have generated a dismissive attitude to any such avocation…

[Charles Parsons] *inherited from his father much more than a modest fortune. More precious to the future inventor were the childhood years spent with his father absorbing the scientific approach, the knowledge of workshop skills and above all else his father's example of determination to see a practical undertaking brought to a conclusion in the face of great discouragements* (Scaife 1988,104-5).

So, instead of going to school, Charles was educated by tutors at Birr. Some of these were assistant astronomers to his father, who had, in the 1840s, built the largest telescope in the world, the 72 inch reflecting telescope nicknamed the *Leviathan of Parsonstown*, which has now been restored at Birr.

After his private schooling, Charles went to Trinity College Dublin, and then on to Cambridge University, where he graduated in mathematics in 1877. He was already making model engines, as well as rowing in the St John's College first boat. A great need in the 1880s was for an engine which could drive a dynamo directly to generate electricity efficiently. High rotational speeds were necessary, and these were beyond the range of normal piston engines. Charles's breakthrough was to pass the steam through a series of bladed wheels, alternately rotating and stationary. As it went through each pair, the steam expanded by a small amount, giving up some of its energy to rotate the shaft. In his first steam turbine of 1884, which incorporated several other brilliant innovations, he used 15 pairs of these wheels, and obtained an engine with a rotation speed of 18 000 revolutions per minute (Scaife 1985,102). He developed a dynamo to cope with this speed, and the age of the steam turbine was born. It was used first to light ships, as it was small and portable, but then was developed for use in power stations. By the end of his life, his turbines were powering stations all over the world, each producing up to 50 000 kilowatts (Mollan 1996b,24) .

Realising its potential for marine propulsion, Charles built a small experimental boat, the *Turbinia*, and in her tried out different engines and propellers. By 1897, he had a vessel which could travel at an unheard of 34 knots. Having trouble convincing a conservative British Navy of its importance, he hit on an audacious sales pitch. He joined the Naval review organised to celebrate the Diamond Jubilee of Queen Victoria. The *Turbinia* could not be caught as it careered amidst Her Majesty's Ships (Appleyard 1933,104). The sales pitch worked, and before long his turbine was commissioned for all British warships, and for merchant ships too. The latter included the sister ships, the *Lusitania*, torpedoed in 1915, and the *Mauretania*, which held the Blue Riband of the Atlantic from 1907 to 1929. They were both 785 feet long, with 44 000 tons displacement, a mighty step up from the 100 foot 44 ton *Turbinia*. Charles was knighted in 1911.

In 1925, he bought the telescope-making firm set up by the Grubb family of Dublin (*see above*), and Grubb-Parsons built many successful large telescopes for the world, a very appropriate activity for the man who had been brought up in the shadow of the *Leviathan of Parsonstown*.

As well as his work on turbines, Charles invented an apparatus, based on compressed air, to amplify sound – called the auxetophone (Appleyard 1933,208-9). The humorous magazine *Punch* reported on an (imaginary) demonstration of the auxetophone in the Albert Hall in its issue of 29 March 1905:

The Hon. Charles Parsons remarked that the best way to promote British music was to secure for it the widest hearing. He had recently patented an improved gramophone which, on a calm, windless day, could be distinctly heard at a distance of three miles (Cheers). But the instrument was only in its infancy, and he was not without hopes that in a short time he would be able to make it heard across the Channel, and perhaps even in Leipsic, Munich, Berlin, and Beyreuth.

Charles also spent a great deal of money trying to make artificial diamonds. In one attempt, he tried to shoot a bullet from a gun into a small chamber drilled in a metal block, and charged with graphite (Parsons 1934,219). Although his idea was right, he was unsuccessful, and it wasn't until 1955 that artificial diamonds were obtained by the General Electric Company, Schenectady, New York, using pressures of about 100 000 atmospheres and temperatures above 3000 °C (*EB CD-ROM*).

He died – fittingly enough on board ship – on the *Duchess of Richmond*, as the sun was setting on Kingston Harbour in Jamaica on 11 February 1931.

John Joly (1857-1933)

Some scientists gain fame for one major discovery, and never reach the same heights of achievement again. There are others who keep on making important and useful discoveries. One of the most fertile of Irish scientific minds was surely that of John Joly. He was born in 1857 in Hollywood House – the Church of Ireland Rectory – at Bracknagh in Co. Offaly on 1 November 1857. He entered Trinity College Dublin in 1876 and graduated in engineering in 1882. He worked in the engineering and physics departments before becoming Professor of Geology and Mineralogy, a post he held for 36 years, from 1897 to 1933. He spent all his working life in Trinity and did not marry. He was elected to Fellowship of the Royal Society in 1892, and awarded its Royal Medal in 1910. He was Honorary Secretary of the Royal Dublin Society from 1897 to 1909, President from 1929 to 1932, and received the Society's Boyle Medal in 1911.

John had a wide range of interests and made many contributions to science, writing about 270 scientific papers and several books. He invented quite a few scientific instruments, the best known being his meldometer for measuring the melting points of minerals, his steam calorimeter for measuring specific heats, and his photometer for measuring light intensity (Mollan 1995b,260,251,359).

One of his claims to fame was the making of an accurate estimate of the age of a geological period – an essential step in estimating the age of the earth. Indeed, several Trinity academics made contributions of varying accuracy to the estimation of the age of the earth. Archbishop James Ussher (1581-1686), Fellow of Trinity in 1600, estimated the date of Creation as 4004 BC using chronological information from the Bible. John Phillips, the first Professor of Geology and Mineralogy (from 1843 to 1845), gave an estimate of 96 million years, based on the rate of sedimentation in the oceans. Samuel Haughton, Trinity's third Professor of Geology (*see above*), made several calculations, with widely different estimates ranging from 2,298 million down to 200 million years. William Sollas, the fifth Professor of Geology (from 1883-1897), made more estimates, with results between 17 and 80 million years. John Joly also carried out different calculations. Based on the volume of sodium in the oceans, he came up with 97 million years. Based on the accumulation of sediment, he estimated 80 million years. However, using the decay in radioactivity in minerals, he finally (in 1913, with Sir Ernest Rutherford – then working in Manchester) estimated the beginning of the Devonian period (the geological period between the Silurian and Carboniferous) as about 400 million years, an age which is in line with that accepted today (Wyse Jackson 1992,93-94). The origin of the earth is now estimated as around 4500 million years.

Another practical achievement was Joly's successful collaboration with Dr Walter Stevenson of Dr Steeven's Hospital in Dublin – of which John was a Governor – in the use of radiation for the treatment of cancer. He had used radium bromide sealed in glass tubes to treat successfully an otherwise intractable 'facial rodent ulcer', brought to his attention by Dr Stevenson, at that time a junior surgeon in the hospital. In 1914, he persuaded the Royal Dublin Society to set up the Radium Institute, and this provided capillary tubes containing radium emanation – the gas now known as radon – to hospitals for many years for the treatment of tumours (Mollan 1995a,34).

In 1894, John patented his method for colour photography, the first successful method of producing colour photographs from a single plate. What he did was effectively to place three filters on the one glass screen by ruling fine lines – about 200 per inch – successively in red/orange, yellow/green, and blue/violet. The resulting transparency, when viewed through a similar screen, 'appears in vivid colour and with all the realism and relief conferred by colour and colour perspective'. As the method could not be used for printing on paper, it did not receive widespread use at the time (although acclaimed as a 'veritable triumph of scientific research'): but his ideas have been developed in modern methods of colour photography (Mollan 1995a,34).

He collaborated with his close friend and Trinity colleague, Henry Horatio Dixon (1869-1953) in explaining how sap rises in plants, the first time this had been done. They found that the motive force was largely due to evaporation from leaves. Nobody believed them at first, but they had the satisfaction of proving their point and demonstrating that they were right and their critics wrong.

John Joly, one of Ireland's great innovators, died in Dublin on 8 December 1933.

Joseph Larmor (1857-1942)

Born at Magheragall, Co. Antrim, on 11 July 1857, Joseph Larmor was educated at the Royal Belfast Academical Institution ('Inst') and Queen's College Belfast, before going up to Cambridge University, where he graduated with first place in the mathematical tripos ('Senior Wrangler') in 1880. The Second Wrangler was another Joe, Joseph John (J.J.) Thomson (1856-1940) 'who with Larmor laid the foundations of the electromagnetic theory of matter' (*DNB*). Larmor returned to Ireland as Professor of Natural Philosophy at Queen's College Galway in 1880, but went back to Cambridge in 1885, where, in 1903, he was appointed Lucasian Professor of Mathematics in succession to Sir G.G. Stokes (*see above*), a position he retained until 1932. This meant that there was an Irishman in this post, previously held by Sir Isaac Newton and now held by Stephen Hawking, for over 80 years, during a key period in the development of mathematics and physics in which Cambridge University played such a central role.

Joseph is best remembered for his contributions to the theory of electromagnetism, in particular the electron theory of matter:

When James Clerk Maxwell showed in the 1870s that light is an electromagnetic wave, the search was on for a theory of the aether that would explain not only the phenomena of light, but also those of electricity and magnetism. Irish scientists showed a particular interest in aether theories. In the 1880s George Francis FitzGerald….adapted a model of the aether created in the 1830s by James MacCullagh…FitzGerald showed that the model *could mathematically account for much of optics and electromagnetism, though its properties were difficult to visualise mechanically. Larmor took up this model with a view towards understanding the relationship between electromagnetic fields and charged bodies. He soon found that small point-like strains or twists in the aether would act like charged particles. He called these strains 'electrons', a term coined by Irish physicist George Johnstone Stoney* (Attis 1999a).

It is not surprising that Larmor's theory was described by one commentator as 'a curiously Irish creation' (quoted in O'Hara 1993,20)!

As we will see (under Thomas Preston below), Joseph's model could explain the Zeeman effect in spectroscopy, but could not explain Preston's anomalous Zeeman effect. As we now know, there is no aether, and the full explanation of the relevant phenomena had to wait for the theory of relativity and the introduction of quantum mechanics. Nevertheless these studies by Irish mathematicians and physicists were of great importance in the development of today's interpretation of the fundamental forces of nature:

Seen in its place in the history of physics, Larmor's work marks the end of the attempt to express everything in terms of the Newtonian mechanics of matter and the beginning of the electromagnetic theory of matter. But it led on immediately to the more revolutionary theory of relativity. For one of the main problems which Larmor attacked was the failure to find definite evidence of motion of the earth through the aether. Of this he was able to give a partial and approximate explanation….This explanation led up directly to the more radical outlook of Einstein which, while completing the discussion, shook the rigid framework of Newtonian conceptions of absolute time and space.

… Although radical in his natural philosophy, he was conservative in temperament, questioning modern trends even in such matters as the installation of baths in [St John's] college (1920). 'We have done without them for 400 years, why begin now'?…Yet once the innovation was made he was a regular user (DNB).

Joseph crowned his distinguished contributions in his famous book *Aether and matter* (1900). He had been elected Fellow of the Royal Society in 1892, and served as one of its Secretaries from 1901 to 1912, continuing the Irish tradition of service to that Society. He received the Royal Society Royal Medal in 1915, and its Copley Medal in 1921. He was knighted in 1909, and from 1911 to 1922 (following the footsteps of G.G. Stokes) he represented Cambridge University in Westminster. He edited the collected works of fellow Irishmen, Stokes, Kelvin and FitzGerald. After his retirement in 1932, he returned to Ireland, where he died, at Holywood, Co. Down, on 19 May 1942 (DNB).

Thomas Preston (1860-1900)

Born at Ballyhagan, Kilmore, Co. Armagh, on 23 May 1860 and educated at Armagh Royal School, Thomas Preston entered Trinity College Dublin in 1881, the year in which George Francis FitzGerald (*see above*) became Professor of Natural and Experimental Philosophy. He graduated in mathematics and experimental science in 1885, and became Professor of Natural Philosophy in University College Dublin in 1891 (Weaire 1987,618-9). He was elected to Fellowship of the Royal Society in 1898, and was awarded the second Boyle Medal of the Royal Dublin Society in 1899.

He is particularly remembered in two ways. First as the author of physics textbooks; and second as the discoverer of the *anomalous Zeeman effect*.

[*Preston's books*] The theory of light *(1890) and* The theory of heat *(1894) are exceptionally thorough and professional treatments of their subjects. A wealth of informative detail is carried along by a magnificent prose style and beautiful illustrations. Preston produced an expanded second edition of Light in 1895 and the world wide popularity of both books was such that they survived as recommended texts throughout half of the present century, with further editions revised by other hands* (Weaire 1990,50).

Spectral studies played a key role in elucidating the structure of the atom. Several Irish mathematicians and physicists, including James MacCullagh, Sir George Gabriel Stokes, Lord Kelvin, George Johnstone Stoney, George Francis FitzGerald and Sir Joseph Larmor (*see above*) made important contributions to the discussion, particularly in trying to explain the aether in mathematical and mechanical terms.

Pieter Zeeman (1865-1943) had found in 1896 that the spectral lines of certain elements were split into three when the sample was placed in a strong magnetic field perpendicular to the light path. If the field was parallel to the light path, the lines were split in two. This Zeeman effect could be explained by Joseph Larmor's theory of electromagnetism (*see above*).

When leading spectroscopists in Europe and America were engaged, during 1897, in exploring the recently-discovered Zeeman Effect, they were overtaken by a relatively obscure physicist working in Dublin. Thomas Preston had previously been known only for his excellent textbooks. His achievement in discovering the Anomalous Zeeman Effect was immedately recognised (Weaire 1987,617).

What Thomas had found was that the splitting of the lines of the spectrum could depart from the simple triplet form observed by Zeeman. This demonstrated the inadequacy of Larmor's theory and other theories current at the time – notably that of Heindrik Lorentz (1853-1928).

Thomas went on to establish empirical rules for spectral lines, which are still associated with his name, but he unfortunately died from a perforated duodenal ulcer in Rathgar, Dublin, on 7 March 1900, at the early age of 39 (Weaire 1990,51). He had already established an impressive reputation, and would undoubtedly have gone on to greater things had he been spared.

His widow, Katherine (née McEwen), became a pioneer of education for women in Ireland as Principal of Alexandra College, Dublin (*DNB*).

John Sealy Edward Townsend (1868-1957)

Born in Galway on 7 June 1868, the son of the Professor of Civil Engineering in Queen's College there, John Townsend attended Trinity College, where he studied under George Francis FitzGerald. He obtained a Senior Moderatorship (first class degree) in both mathematics and experimental science in 1890. After a few years' teaching, he had the distinction of being one of the first two non-Cambridge University graduates to be admitted as research students to the Cavendish Laboratory, then under the direction of the famous J.J. Thomson (1856-1940). The other research student, likewise admitted, was Ernest Rutherford (1871-1937), a graduate of Canterbury College, Christchurch, New Zealand. Rutherford succeeded Thomson as the Cavendish Professor of Physics in 1919, where he dominated the newly emerging field of nuclear physics. It was in Rutherford's Cavendish Laboratory that Ernest Walton and John Cockcroft 'split the atom' in 1932 (*see below*).

Following J.J. Thomson's success in the 1897 physical discovery of his 'corpuscles', later re-christened 'electrons' – the name introduced by George Johnstone Stoney (*see above*) – Townsend in 1898 measured the charge on the electron, the fundamental unit of electrical charge. He did this, appropriately enough, by using Stokes Law (named after fellow Irishman George Gabriel Stokes – *see above*). He measured the rate of fall of a cloud which had condensed on an electrified gas, liberated in electrolysis, and then bubbled through water. He proved that the fundamental constant of electrolysis was equivalent to the charge carried by a gaseous ion whatever its mode of production (*DSB Vol.13,446*).

The first measurement of the elementary ionic charge [was] an outstanding example of elegance and simplicity, using but a laboratory balance, an electrometer, and a photographic camera (DNB).

He was appointed in 1900 as the first Wykeham Professor of Experimental Physics in Oxford University – a post he held until his retirement in 1941 – where he enhanced his reputation through his studies of electricity in gases. He formulated a theory of ionisation by collision, showing that the motion of electrons in an electric field would release more electrons by collision. These would in turn release more electrons in a multiplication of charges known as an 'avalanche', and this allowed him to explain the passage of currents through gases where the electric field was thought to be too weak (Daintith 1994,887-8). The collision principle was the basis of the 1908 particle detector developed by Rutherford and Hans Geiger (1882-1945), subsequently improved as the famous *Geiger-Müller Counter*.

Townsend was elected to fellowship of the Royal Society in 1903, published his *Electricity in gases* in 1915, and was knighted in 1941. The founder of the kinetic theory of ions and electrons in gases (Williams 1994,486), John died in Oxford on 16 February 1957.

His wife, Mary Georgiana (née Lambert) from Castle Ellen in Co. Galway was twice Mayor and became an Honorary Freeman of the City of Oxford (*DNB*).

Guglielmo Marconi (1874-1937)

His name doesn't sound very Irish, right enough, but Guglielmo Marconi's mother was Annie Jameson from Enniscorthy in Co. Wexford, a daughter of the famous distilling family – a family which has brought pleasure to many Irish and Italian palates. Annie had been sent by her parents to Italy to study singing (*DNB*) and, in a nice complementarity between the arts and sciences, it is recorded that it was she who encouraged her son's practical scientific work, as his father wanted him to give this up to promote the family businesses (Barry 1994,24). Guglielmo's Irish background presumably influenced his 1905 choice of wife, for she was the Hon. Beatrice O'Brien, daughter of Lord and Lady Inchiquin of Dromoland Castle in Co. Clare (O'Loughlin 1994).

Guglielmo was born in Bologna on 25 April 1874, and studied at the Leghorn lyceum. He then began, in 1894, to experiment with the electromagnetic waves which Hertz had discovered (*see under George Francis FitzGerald above*) with the assistance of Augusto Righi (1850-1920), Professor of the Institute of Physics at Bologna University (*DSB* 9,98). He found that if he used a coherer (an evacuated tube of filings which 'cohered' and created a circuit when a burst of electromagnetic waves went through) at the receiving end, and large elevated antennae, he could substantially increase the distance between transmitter and receiver. A quick worker, he had, by 1895, achieved a transmission distance of almost a kilometre. As the Italian government didn't show much interest, he moved to London in February 1896, where he demonstrated and patented his invention. He set up the Wireless Telegraph and Signal Company in 1897, which became Marconi's Wireless Telegraph Company in 1900.

Meanwhile he was steadily improving his techniques. He achieved a notable first in Kingstown Harbour (now Dun Laoghaire, Co Dublin) in July 1898, when the results of the Royal St George Yacht Club Regatta were transmitted from a yacht in Dublin Bay to the Dublin *Daily Express*. Some of the apparatus used in this transmission is preserved in the National Museum of Science in St Patrick's College, Maynooth (Mollan 1994,117). Various Marconi stations were built around the coast of Ireland – the first at Crookhaven in Co. Cork being started in 1901. A major breakthrough came on 12 December 1901, when he successfully sent a signal (the letter 'S' in Morse – three dots) across the Atlantic – from Cornwall to Newfoundland. Construction of the massive Clifden station in Co. Galway started in July 1905 to handle transatlantic commercial traffic, and a station at Ballybunion was the site of the first transatlantic telephone message in 1919 (Barry 1994,26).

Guglielmo received the Nobel Prize for Physics in 1909. He died in Rome on 20 July 1937 and was accorded a State Funeral by the Italian Government.

Harry George Ferguson (1884-1960)

Agriculture has been, and remains, a vitally important industry in Ireland. It is appropriate then that it is an Irish engineer who is credited with revolutionising not only Irish but world agriculture.

Harry Ferguson didn't invent the tractor, but he dramatically changed its design, and the way it is used with accessories.

Harry was born at Growell, near Hillsborough, in Co. Down on 4 November 1884, the fourth son in a family of eleven. He left school at 14 to work on his father's farm, but soon became an apprentice to his brother's car and cycle repair business. While there, he became involved in motor cycle and car racing, and fortunately lived to tell the tale. He was the first man in Ireland to design and build his own aeroplane, which he flew on 31 December 1909 (Garvin 1993,34-5).

He had set up his own garage business in 1911, and in 1914 he began to sell an American tractor, called the 'Overtime' – which weighed over two tons (Rae 1980,1). But heavy tractors were limited in use and beyond the means of most farmers. Lighter tractors of the day were liable to tip over or rear backwards if the implement they were pulling caught on an obstruction, like a root or a stone, sometimes with fatal results. Harry designed and built a new style plough, noted for its simplicity and lightness, but most of all for a brilliant innovation. The plough was coupled to the tractor in a three point linkage, so that tractor and plough formed a single unit. He patented this 'Ferguson System' in 1926.

The plough had no wheels of its own, and the coupling transferred the weight, so that, when an obstacle was encountered, the front end did not rise, and the plough could be raised easily using a lever beside the driver. Other implements could replace the plough using the same linkage. He later added an hydraulic system. His 'unit principle' was defined as:

A system for the mechanization of agricultural operations with the aim and objective of enabling increased production of food and other farm commodities easily, economically and profitably, so that eventually farm products will reach all the people at prices they can pay (quoted in Rae 1980,2).

In 1938, he agreed a deal with Henry Ford (1863-1947) in Michigan, USA, to sell a tractor of Harry's own design (Rae 1980,12). Henry described Harry as:

an inventive genius whose name will go down in history with those of Alexander Graham Bell, the Wright Brothers, and Thomas Edison (DNB).

Son of a famine emigrant from near Bandon, in Co. Cork, Henry became the richest man in the world through mass-producing cars. By 1947, 300 000 Ford Ferguson tractors had been built. But Harry and the Fords fell out, and Harry designed a new model, the TE-20 or the 'Wee Fergie', which was built by the Standard Motor Company in Coventry. Between 1946 and October 1956, 517 649 of these lightweight, inexpensive, and highly manoeuvrable tractors had been built (Massey Ferguson 1993,28), and it was claimed that the name Ferguson was as familiar to farmers the world over as their own names. Indeed the Welsh word for tractor is 'fergie'. One of his employees is reputed to have said: 'We were not employees: we were converts. Joining the Ferguson organisation was like joining the Church'(*DNB*)!

Harry's company later merged with a large Canadian Company, Massey Harris, to form Massey Ferguson, which continues to supply tractors and agricultural machinery world wide.

A full-scale replica of Harry's aeroplane, and an early tractor and plough, are on display at the Ulster Folk and Transport Museum at Cultra, Co, Down.

When he retired in 1954, he turned his energies again to cars. He built a four-wheel drive racing car. He also invented a device for preventing the locking of wheels in a skid. Both these innovations are now in common use in production cars (Garvin 1993,37). Harry refused a knighthood which had been offered to him for his services to the Allies in the Second World War.

Harry Ferguson died in Stow-on-the-Wold, Gloucestershire, on 25 October 1960.

Erwin Rudolf Josef Alexander Schrödinger (1887-1961)

As his name would suggest, Erwin Schrödinger was not of Irish birth, but he became an Irish citizen in 1948, and lived and worked in Dublin from 1939 to 1956, during which time he published seven books and about 75 articles on the natural sciences and the philosophy of science (McConnell 1988,8).

Born in Vienna on 12 August 1887, he attended the University of Vienna from 1906 to 1910, and held University posts at Vienna, Jena, Stuttgart, Breslau, Zürich, Berlin, Oxford and Graz before arriving in Dublin (McConnell 1990,106). Although not a Jew – his father was a Roman Catholic and his mother a convert to an evangelical Lutheran (protestant) church (Moore 1989,11) – he was an opponent of nazism, and his life was changed on the advent of Hitler to power in 1933. By co-incidence, this was also the year in which he was awarded the Nobel Prize for Physics, with Paul Dirac (1902-1984) of Cambridge University, 'for the discovery of new forms of atomic theory and applications of them' (Moore 1989,289). It was during his time in Zürich (from 1921 to 1927) that Erwin had proposed what became known as the 'Schrödinger Equation':

He wrote a set of papers on wave mechanics which revolutionised scientific thought and had a profound effect on physics, chemistry and biology, as well as on the philosophy of science…Schrödinger applied his equation to the hydrogen atom and obtained discrete, that is discontinuous, energy levels….which agreed with experimental results. This success more than anything else established the position of Schrödinger in the history of physics, and his equation soon became the basis for the quantum-mechanical treatment of problems in atomic and molecular physics and in chemistry (McConnell1988,2-3).

It was Eamon de Valera who attracted Erwin to Dublin to a position as Senior Professor at the newly established Dublin Institute for Advanced Studies. Afterwards, Erwin referred to his time in Dublin as:

a *very, very beautiful time. Otherwise I would never have gotten to know Ireland and learned to love this beautiful island of Ireland. It is impossible to imagine what would have happened if, instead, I had remained in Graz for these 17 years* (DSB Vol.12,p.222).

One of Erwin's many books *What is life?* (1944) has been particularly influential. For example:

What is Life? *had a determining influence on the career of James Watson. He read the book in the spring of 1946, while he was an undergraduate at Chicago and undecided what to do and where to go for graduate work. 'From the moment I read Schrödinger's* What is Life? *I became polarized towards finding out the secret of the gene'* (Moore 1989,403).

And, indeed, James Watson (born 1928) and Francis Crick (born 1916) did just this in what has been called 'the most significant discovery of the century' – the discovery of the double helical structure of DNA (Daintith 1994,188).

A colourful character, Erwin lived in a *menage à trois* in Clontarf in Dublin, and was especially attracted to young women, with whom he had a succession of love affairs (*see* Moore 1989). Returning to Vienna, Erwin Schrödinger died there on 4 January 1961.

Robert (Roy) Charles Geary (1896-1983)

Apart from a few years in Paris, Cambridge and New York, Ireland's greatest statistician spent the whole of his working life in Dublin, where he was born on 11 April 1896. After graduating from University College Dublin, and further study at the Sorbonne in Paris, he joined the Statistics Branch of the Department of Industry and Commerce in 1923. In 1949, he became Director of the newly established Central Statistics Office and, in 1960, Director of the Economic Research Institute (Spencer 1990,66-7).

Roy Geary was a great Irish scientist whose research in mathematical statistics, demography, national accounting and economics is world famous. His prolific output was written during a career, not as an academic, but as a statistician *working in the Irish Civil Service. Alongside his theoretical achievements, he contributed enormously to the building of official statistical records during the first 60 years of the State* (Conniffe 1997).

Roy worked on the sampling theory of ratios and ‘his 1930 derivation of the density of the ratio of two normal variates remains one of the few essential references in the field’ (Spencer 1990,66). Another major theme was testing for normality and enquiring into the robustness of inferential methods which depended formally on normality. He worked on the estimation of relationships where the variables are subject to errors of measurement. He is considered one of the leading pioneers of the now standard technique of instrumental variables.

An interesting insight into his personality is provided in a letter Roy wrote to the journal *Nature* (Vol. 151, 24 April 1943, p.476 – remembered in Spencer 1997,71). It was claimed in the *Dictionary of National Biography* that Robert Boyle (1627-1691) – *see above* – had ‘but loosely demonstrated’ the principle of Boyle’s law. Referring to Boyle’s original work, Roy noted that ‘Boyle’s observations were so accurate and so conclusive’ that statistical calculations were actually unnecessary. The reference to looseness was obviously less than fair to the work of a great man of science. Spencer notes: ‘This was Roy’s patriotism – nobody was hurt and an Irish achievement was honoured’.

A recipient of Honorary Doctorates from the National University of Ireland (1961), The Queen’s University of Belfast (1968), and Trinity College Dublin (1973), and of the Boyle Medal of the Royal Dublin Society in 1981, Roy died in Dublin on 8 February 1983.

Another Irish connection in the historical development of statistics is the fact that the well-known Student’s t-test was developed by William Sealy Gosset (1876-1937), an English-born Brewer who worked at the Guinness Brewery at St James’s Gate in Dublin from 1899 to 1937 (Somerfield 1990,58-9).

John Lighton Synge (1897-1995)

While John Millington Synge (1871-1909) is well known in Ireland for his dramatic works, particularly *The playboy of the western world*, which caused a riot at the old Abbey Theatre in Dublin in 1907, his nephew and namesake is not so well known. John Lighton was the son of Edward, brother to John Millington, and is recognised by those who do know his work as ‘undoubtedly the most talented and distinguished Irish mathematician of his generation’ (Spearman 1999).

Born in Dublin on 23 March 1897, he obtained top place in his degree examinations (Senior Moderatorship) in both mathematics and experimental science in Trinity College in 1919, and was appointed Professor of Natural Philosophy (i.e. applied mathematics) there in 1925. He had further Professorships in Toronto, Columbus (Ohio), and Pittsburgh, before returning to Ireland as Senior Professor in the School of Theoretical Physics at the Dublin Institute for Advanced Studies in 1948. He remained in Dublin for the rest of his life (Spearman 1985,234).

[Synge] *showed great originality and versatility in the application of mathematics to a wide variety of problems; his natural curiosity and his fascination with analysis and mathematical description and solution of problems remained with him and were still evident right up to his death at the age of 98* (Spearman 1999).

He is best known for his work on relativity. Among his books, those on *Relativity – the special theory* (1956) and *Relativity – the general theory* (1960) ‘profoundly influenced and shaped a whole generation of students of relativity and cosmology’ (Spearman 1999). In his introduction to his 1956 book he writes:

Splitting hairs in an ivory tower is not to everyone’s liking and no doubt many a relativist looks forward to the day when governments will ask his advice on important questions.

Synge’s researches also covered classical mechanics, theory of elasticity, fluid mechanics and classical electrodynamics. His 1937 book *The relativistic gas* ‘laid the foundations of a systematic treatment of relativistic thermodynamics’, and that on *The hypercircle in mathematical physics* (1957) provided ‘a major contribution to numerical analysis’ (Boyle Medal citation). He was co-editor of *The mathematical papers of Sir William Rowan Hamilton* (1931).

John became a Fellow of the Royal Society in 1943, was President of the Royal Irish Academy from 1961 to 1964, received the Boyle Medal of the Royal Dublin Society in 1972, and honorary doctorates from the University of St Andrews (1966), The Queen’s University of Belfast (1969) and the National University of Ireland (1970). He died in Dublin on 30 March 1995.

One of the three daughters of John and his wife, Elizabeth (née Allen), Cathleen Morawetz, is a distinguished mathematician who has served as President of the American Mathematical Society, and was awarded the US National Medal of Honor (Spearman 1999).

Ernest Walton (1903-1995)

The only Irish-born scientist (so far) to win a scientific Nobel Prize has been Ernest Walton. As a young man in Cambridge University, he collaborated with John Cockcroft in the building of a linear accelerator which could accelerate protons (hydrogen nuclei) to energies of 700 000 electron-volts. With this apparatus, he and Cockcroft ‘split the atom’ in 1932. For this work, they jointly received the Nobel Prize for Physics in 1951. It was the beginning of accelerator-based experimental nuclear physics, which continues to teach us so much about the nature of matter.

Ernest was born in Dungarvan, Co. Waterford, on 6 October 1903, the son of a Methodist Minister. He attended Methodist College, Belfast – known affectionately as ‘Methody’ – and entered Trinity College Dublin in 1922, graduating with double first class honours in experimental science and mathematics in 1926. Following a strong recommendation from John Joly (*see above*), he obtained a place at the Cavendish Laboratory in Cambridge, working under the direction of Ernest Rutherford.

What he and Cockcroft actually did in 1932 was to bombard the element lithium with their accelerated protons. These were energetic enough to shatter the target lithium atoms and produce alpha particles, or helium nuclei. They had achieved a transmutation of the elements – the conversion of one element into another – by entirely artificial means – long the goal of the alchemists. While the conversion of lithium to helium was not quite the conversion of a base metal into gold, the dream of the alchemists, it was a dramatic step forward in science. And in carrying out this transformation, they were able to verify Einstein’s famous equation $E = mc^2$. The mass of the two alpha particles was less than that of the lithium and proton, the missing mass being converted into kinetic energy. This achievement was one of the great landmarks in physics, and was a formidable feat with the very limited experimental equipment available. The apparatus was made using bits of petrol pumps and car batteries, and the fact that it worked at all was a direct result of Walton’s great manual ability (Mollan 1995a,40).

Ernest soon returned to Trinity College, being elected to Fellowship in 1934, and he became Erasmus Smith Professor of Natural and Experimental Philosophy – a rather grand title whose origins date back to 1723 – in 1946. There, he established a lasting reputation as a devoted teacher who inspired generations of Trinity physicists. He died in Belfast on 25 June 1995.

While he remains the only Irish-born scientific Nobel Laureate, contrasting with four in literature – W.B. Yeats in 1923, George Bernard Shaw in 1925, Samuel Beckett in 1969, and Seamus Heaney in 1995, there are quite a few other scientific and medical Nobel prizewinners with Irish connections. The most famous of these is the inventor of wireless telegraphy, Guglielmo Marconi (*see above*), who won the physics prize in 1909. Other winners with Irish genes include Frederick Banting (physiology or medicine 1923), Maurice Wilkins (physiology or medicine 1962), Luis Alvarez (physics 1968), Martin

Ryle (physics 1974), Dudley Herschbach (chemistry 1986), George Hitchings (physiology or medicine 1988), Joseph Murray (physiology or medicine 1990), Richard Taylor (physics 1990), and Peter Doherty (physiology or medicine 1996). Quite a few others have obviously Irish surnames, but their origins have still to be investigated (Mollan 1997,5).

Kay McNulty Mauchly Antonelli (born 1921)

It was in 1839 that Charles Babbage (1792-1871) designed the first mechanical digital computer – his 'difference engine', although he didn't receive the necessary funding to build it. He was assisted in his endeavour by mathematician Ada Countess Lovelace, daughter of Lord Byron.

The first electronic computer (the Electronic Numerical Integrator and Computer – or ENIAC – which was ten feet tall, eighty feet wide and weighed thirty tons) was built by John Mauchly at the University of Pennsylvania. A young Irish woman was one of six recruited in 1945 to program it. These women were actually called 'computers'. One of them stole a march on the others, for she married John Mauchly in 1948.

Kay McNulty was born in the Creeslough Gaeltacht in Co. Donegal on 12 February 1921. On the night she was born, her father, James, who was an IRA Training Officer at the time, was arrested and imprisoned. After he got out of jail, the family emigrated to the USA in 1924, settling in Pennsylvania. After graduating from Chestnut Hill College in 1942 – one of only three women maths majors – Kay joined the US Army Women's Corps, and she was employed as a human computer to calculate the trajectories of artillery. She recalls:

*We were preparing a firing table for each gun, with maybe 1,800 simple trajectories. To hand-compute just one of these trajectories took 30 or 40 hours of sitting at a desk with paper and a calculator (*quoted in http://www.ohkaycomputer.com – from which the information in this biography is taken: see also Sharkey 1998).

In February 1946 the US Army unveiled ENIAC, but the vital role of the women computers was not mentioned in the records of the demonstrations nor in the press releases. It is only recently that this injustice is being reversed, and these pioneering women mathematicians are being accorded due recognition. Kay's husband went on to devise, with Presper Eckert, the first commercially available computer – the UNIVAC, and again Kay worked on its programming. John died in 1980, and Kay married photographer Severo Antonelli in 1985. Kay, and fellow computer Jean Bartik, took part in a lecture tour in Ireland in April 1999.

John Stuart Bell (1928-1990)

'The man who proved Einstein was wrong' is the provocative title of an article by the well-known science writer John Gribbin in *New Scientist* on 24 November 1990 (pages 43-45):

Albert Einstein never accepted the indeterminism of quantum mechanics. But a quiet Irishman, John Bell, devised a test that revealed the alarming truth behind quantum reality.

Born into a poor family in Belfast on 28 July 1928, John Bell left school at 16. But, after a year as a laboratory assistant in The Queen's University of Belfast, and encouraged by his mother, he enrolled as a student in the Physics Department and graduated in 1949. He immediately began employment at the Atomic Energy Research Establishment at Harwell in England, where he worked on the design of CERN's first accelerator, the Proton Synchroton. He transferred to CERN in Geneva in 1960 and remained there until his premature death on 1 October 1990 (Daintith 1994,67).

It was in 1964 that John published what has been described as 'the single most important theoretical paper in physics to appear since 1945' – 'On the Einstein Podolsky Rosen paradox'. When he died:

Colleagues referred to him as the only physicist of his generation to rank with the pioneers of quantum physics, such as Max Born and Niels Bohr, for the depth of his philosophical understanding of the implications of the theory. Had he lived a little longer, he would surely have been noticed by the

Nobel Committee (Gribbin 1990,43).

In essence he anticipated a test which would establish whether or not there was such a thing as ‘action at a distance’. Let Gribbin explain:

You cannot, even in principle, measure precisely both the momentum and the position of a particle at the same time. This is the basis of quantum uncertainty. It is often explained in physical terms, that the act of measuring the position of a particle disturbs its momentum, and vice versa. But this physical explanation obscures the fact that uncertainty is intrinsic to the quantum equations – in quantum mechanics, a particle does not possess both a precise position and a precise momentum.

Einstein and his colleagues pointed out, however, that if two particles A and B interacted with one another and then flew apart, an experimenter could measure the combined momentum of the pair at the time of the interaction, and later the momentum of particle A and the position of particle B. It is then simple to calculate the momentum of particle B, whose position we already know, from the measurement made on particle A.

At least it would be simple if quantum particles behaved like snooker balls. The only way to save quantum uncertainty….is to allow what Einstein referred to as a ‘spooky action at a distance’, a communication that operates instantaneously between two particles even when they are far apart, so that particle A can be disturbed by the measurement made on particle B – a communication that Einstein, the originator of relativity theory, could never accept, since it seemed to travel faster than light…..To Einstein it was ‘obvious’ that there could be no action at a distance, and therefore that quantum uncertainty was an illusion.

It was Belfast man John Bell, who, in his 1964 paper, made the conceptual breakthrough which led to a practical experiment to measure this bizarre effect. The actual experiments, involving the measurement of the polarisation of photons leaving an atom, were carried out by Alain Aspect working at the Institute of Optics at the University of Paris in the 1980s. This established that the quantum theory ‘including the spooky action at a distance which Einstein hated so much’ is correct.

This has profound implications as, according to the big bang theory, all particles originate from a common set of interactions at the birth of the universe. ‘It implies that there is a cosmic web linking all particles in the Universe’. It may imply that absolutely everything is predetermined – and where does that leave our assumption of free will?

‘For me’, said Bell, ‘it’s a dilemma. I think it’s a deep dilemma, and the resolution of it will not be trivial; it will require a substantial change in the way we look at things’ (quoted in Gribbin 1990,45).

Anne Kernan (born 1933)

In June 1995, the National University of Ireland awarded the Honorary Degree of Doctor of Science to one of its own, Dubliner Anne Kernan, who had graduated in Physics in University College Dublin in 1952, obtained a PhD there in 1957, and lectured in the Department of Experimental Physics from 1958 to 1962. She was born in Dublin on 15 January 1933.

Judging from her citation, attainment of doctoral status did not pass without suitable acknowledgement:

The details are probably best glossed over at a serious academic occasion such as this, not least because the celebrations involved, in addition to today’s distinguished graduand, five future professors of this university and the future chief executive of one of Ireland’s largest companies. There may well be little truth in contemporary descriptions of a future distinguished professor in the Science Faculty arrayed in turban pantaloons being ejected from the Metropole Ballroom in the course of the evening; let us just say that the celebrations were memorable!

It was the beginning of a distinguished academic career which brought her to the Lawrence Laboratory at Berkeley, through Stanford University, the University of California at Riverside, CERN in Switzerland, and back to Riverside, where she was Professor of Physics from 1972 until 1991, and Vice-Chancellor for Research from 1991 until 1994 (Mollan 1996c,9).

Anne led the US team in the multinational experiment at CERN which, in 1983, found the theoretically predicted, long-sought, carriers of the weak force. This involved the discovery of two new fundamental particles, christened ‘W’ and ‘Z’. The importance of that achievement lay in the fact that the existence of these two particles confirmed physicists’ understanding of ‘weak nuclear interaction’, one of the ways in which fundamental particles interact, thus increasing our knowledge of basic physics. Its value to science was acknowledged by the award of the 1984 Nobel Prize for physics to the overall leader of the group, Italian Carlo Rubbia.

Her honorary doctorate was a proper Irish acknowledgement of Anne’s major contribution to this work, and celebrated also her abilities as a teacher and an administrator. In the latter category, she held a succession of national appointments in the USA, including membership of the Advisory Committee for Physics to the National Science Foundation, the Council of the American Physical Society, and the High Energy Physics Advisory Panel to the Department of Energy.

Jocelyn Bell Burnell (born 1943)

It must be the ultimate dream of many human scientists to discover the little green men (women?) who inhabit our universe, but have yet to reveal themselves to us. An Irish-born woman scientist nearly got there.

Jocelyn Bell was working at the Cavendish Laboratory at Cambridge in 1967 with a massive radio telescope which she had helped to build. When studying the charts produced from this telescope, she noticed a regular pattern of pulses coming from the skies. It was established that these didn’t come from man-made radio interference, so she and her supervisor, Anthony Hewish, christened their project LGM 1 (little green men 1!). It transpired, though, that they didn’t come from intelligent life out there – friendly or otherwise. They came from rapidly rotating neutron stars. Jocelyn had discovered ‘pulsars’ (short for pulsating radio sources). A neutron star is one which has collapsed to a very small size under its own gravity, and the pulsars are the remnants of such ‘supernova’ explosions of large stars. Many more have been discovered since, but it was Jocelyn who discovered the first four. In 1974, Anthony Hewish was awarded the Nobel Prize in Physics for the pulsar work, and many have felt that Jocelyn should have been a joint winner.

Susan Jocelyn Bell, from Lurgan in Co. Armagh, was born on 15 July 1943, and started her education in Lurgan College. She graduated in physics from Glasgow University in 1965, and was awarded a PhD from the University of Cambridge in 1968. While in Glasgow, she had been the only woman in the class of 50 sitting the finals. Starting with her discovery of the first pulsars, she has had a very distinguished career. She worked at the University of Southampton, the Mullard Space Science Laboratory London, and the Royal Observatory Edinburgh. Her studies included gamma ray, X-ray, and infra-red astronomy. In 1991 she was appointed Professor of Physics in the largest University in the UK, the Open University, thus doubling the number of women professors of physics in the UK at the time. Her interests include the astrophysics of neutron stars, the physics of synchrotron nebulae around pulsars, and the surface temperature of pulsars. She has won many international medals and awards, and has several honorary doctorates (Mollan 1996c,9).

She lists among her interests: knitting and sewing, the public understanding of science, and promoting women in science.

Bibliography

Note: ***DNB*** stands for *Dictionary of national biography*, and ***DSB*** for *Dictionary of scientific biography* – see entries under **Dictionary** below. ***EB*** stands for *Encyclopaedia Britannica* – in this case details of the reference are given in the text.

Appleyard 1933: *Charles Parsons, his life and work*, by Rollo Appleyard, Constable & Co., London, pp. xiii + 333.

Attis 1999a: 'Joseph Larmor', by David Attis, in Mollan 1999.

Attis 1999b: 'William Thomson, Lord Kelvin', by David Attis in Mollan 1999.

Attis 1999c: 'John Tyndall', by David Attis, in Mollan 1999.

Barry 1994: 'Irish contributions to a century of wireless communications', by Noel Barry, *Technology Ireland*, May 1994, pp. 23-26.

Bennett 1984: *The Celebrated Phaenomena of Colours: the Early History of the Spectroscope*, by James Bennett, Whipple Museum of the History of Science, Cambridge, pp. 21.

Bennett 1990: *Church, state and astronomy in Ireland – 200 years of Armagh Observatory*, by James Bennett, Armagh Observatory , Armagh, pp. vi +277.

Bennett 1999: 'Thomas Romney Robinson', by James Bennett, in Mollan 1999.

Berry 1915: *A history of the Royal Dublin Society*, by Henry F. Berry, Longmans, Green & Co., London, pp. xv + 460.

Boylan 1978: *A dictionary of Irish biography*, by Henry Boylan, Gill & Macmillan, Dublin, pp. xi + 385.

Boyle 1648-49: *An account of Philaretus* [i.e. Robert Boyle himself] *during his minority*, by Robert Boyle, manuscript reproduced in Maddison 1969, pp. 2-56.

Boyle 1661: *The sceptical chymist*, by Robert Boyle, London pp. (18) + 442.

Brenni 1994: Heinrich Daniel Ruhmkorff (1803-1877), by Paolo Brenni, *Bulletin of the Scientific Instrument Society*, No. 41, June 1994, pp. 4-8.

Brock 1981: *John Tyndall – essays on a natural philosopher*, Edited by William. H. Brock, Norman D. McMillan and R. Charles Mollan, Royal Dublin Society, Dublin, pp. xii +219.

Burchfield 1981: 'John Tyndall – a biographical sketch', by Joe Burchfield, pp. 1-13 in Brock 1981.

Burnett 1989: *"Vulgar and mechanick" – the scientific instrument trade in Ireland 1650-1921*, by John Burnett and Alison Morrison-Low, National Museums of Scotland and Royal Dublin Society, Edinburgh and Dublin, pp. ix + 166.

Butler 1990: *Seeing Stars – Two Hundred Years of Astronomy in Armagh*, by John Butler, (Exhibition Catalogue), pp.102.

Canny 1982: *The upstart Earl – a study of the social and mental world of Richard Boyle (1566-1643)*, by Nicholas Canny, Cambridge University Press, Cambridge, pp. xii + 211.

Cawley 1998: *The work of Nicholas Callan and its relevance to today's technology*, by Michael Cawley, address given at the unveiling of the Institute of Physics plaque to Callan at St Patrick's College, Maynooth, 17 June 1998.

Collins 1999: 'John Philip Holland', by Timothy Collins, in Mollan 1999.

Conniffe 1997: *Roy Geary 1896-1983: Irish statistician*, Edited by Denis Conniffe, Oak Tree Press, Dublin, pp. xvii + 205.

Cox 1985: 'Robert Mallet', by Ron Cox, pp. 42-3 in Mollan 1985.

Daintith 1994: *Biographical encyclopedia of scientists* (second edition), Edited by John Daintith, Sarah Mitchell, Elizabeth Tootill and Derek Gjertsen, Institute of Physics Publishing, Bristol, two volumes, pp. 1-506 & 507-1075.

Dean 1991: 'Robert Mallet and the founding of seismology', by Dennis R. Dean, *Annals of Science*, 48, pp. 39-67.

Dictionary 1921-2: *Dictionary of national biography*, abbreviated to *DNB*, Oxford University Press. These volumes, with later additions, are now conveniently available as a CD-ROM (1995). If a Volume number is not given, then the CD-ROM version has been used.

Dictionary 1981: *Dictionary of scientific biography*, abbreviated to *DSB*, Editor-in-Chief, Charles Gillispie, sixteen volumes bound in eight, Charles Scribner's Sons, New York.

Garvin 1993: *Northern Ireland scientists and inventors*, by Wilbert Garvin and Des O'Rawe, Blackstaff Press, Belfast, pp. 50.

Glass 1997: *Victorian telescope makers – the lives and letters of Thomas and Howard Grubb*, by Ian Glass, Institute of Physics, Bristol, xiii + 279.

Gow 1997: 'George Salmon 1819-1904 – his mathematical work and influence' by Rod Gow, *Bulletin of the Irish Mathematical Society*, No.30, Christmas 1997, pp. 26-76.

Gow 1998: 'Salmon's Conic sections – the first 150 years of a mathematical best-seller', by Rod Gow, in *The Irish Scientist 1998 Year Book*, Samton Ltd, Dublin 1998, p. 26.

Gribbin 1990: 'The man who proved Einstein was wrong', by John Gribbin, *New Scientist*, 24 November 1990, pp. 43-5.

Hankins 1980: *Sir William Rowan Hamilton*, by Thomas Hankins, Johns Hopkins University Press, Baltimore, pp. xxi + 474.

Holland 1900: 'The submarine boat and its future', by John Philip Holland, *North American Review*, 172, pp. 894-903.

Landy 1985: 'George Francis FitzGerald', by Sheila Landy, pp. 62-3 in Mollan 1985.

McConnell 1986: *Geophysics and geomagnetism – catalogue of the Science Museum collection*, by Anita McConnell, Science Museum, London, pp. v + 45.

McConnell 1988: 'Erwin Schrödinger (1887-1961) – Austro-Irish Nobel laureate', by James McConnell, *Occasional Papers in Irish Science and Technology*, No. 5, Royal Dublin Society, pp. 13.

McConnell 1990: 'Erwin Schrödinger', by James McConnell, p. 106 in Mollan 1990.

McDowell 1982: *Trinity College Dublin 1592-1952 – an academic history*, by R.B. McDowell and D.A. Webb, Cambridge University Press, pp. xxiii +580.

MacHale 1985: *George Boole – his life and work,* by Desmond MacHale, Boole Press, Dublin, pp. xiii + 304.

McKenna-Lawlor 1998: *Whatever shines should be observed*, by Susan McKenna-Lawlor, Samton Limited, Dublin, pp. 136.

McLaughlin 1965: *Nicholas Callan, priest-scientist 1799-1864*, by P.J. McLaughlin, Clonmore & Reynolds, Dublin, pp. 128.

Maddison 1969: *The life of the honourable Robert Boyle F.R.S.*, by R.E.W. Maddison, Taylor & Francis, London, pp. xxii +332.

Massey Ferguson 1993: *Harry Ferguson – a brief history of his life and tractors*, published by Massey Ferguson Tractors Limited, Coventry, pp. 43.

Meenan 1981: *The Royal Dublin Society 1731-1981*, Edited by James Meenan and Desmond Clarke, Gill & Macmillan, Dublin, pp. 288.

Mollan 1985: *Some people and places in Irish science and technology,* Edited by Charles Mollan, William Davis and Brendan Finucane, Royal Irish Academy, Dublin, pp. 107.

Mollan 1987: *Nostri plena laboris – an author index to the RDS scientific journals 1800-1985*, by Charles Mollan, Royal Dublin Society, Dublin, pp. 120.

Mollan 1990: *More people and places in Irish science and technology*, Edited by Charles Mollan, William Davis and Brendan Finucane, Royal Irish Academy, Dublin, pp. 107.

Mollan 1992: 'Who discovered the electron?', by Charles Mollan, *Technology Ireland*, October 1992, pp. 57-59 – see also O'Hara 1993.

Mollan 1994: *The scientific apparatus of Nicholas Callan and other historic instruments*, by Charles Mollan and John Upton, Maynooth College and Samton Limited, Maynooth and Dublin, pp. 304.

Mollan 1995a: *The mind and the hand – instruments of science 1685-1932*, by Charles Mollan, Samton Limited, Dublin, pp. 64.

Mollan 1995b: *Irish national inventory of historic scientific instruments*, by Charles Mollan, Samton Limited, Dublin, pp. 501.

Mollan 1996a: *The Birr astronomy story* – briefing document prepared for the Birr Scientific Heritage Trust by Charles Mollan, pp. 61.

Mollan 1996b: *Charles Parsons (1854-1931)* – briefing document prepared for the Birr Scientific Heritage Trust by Charles Mollan, pp. 47.

Mollan 1996c: 'Irish Nobel women – the sky's the limit!', by Charles Mollan, *The Irish Scientist 1996 Year Book*, Samton Limited, Dublin pp. 8-9.

Mollan 1997: 'Nobel genes', by Charles Mollan, in *The Irish Scientist 1997 Year Book*, Samton Limited, Dublin, p. 5.

Mollan 1999: *Irish innovators in science and technology*, Edited by Charles Mollan, William Davis and Brendan Finucane, Royal Irish Academy and Enterprise Ireland, Dublin, in press.

Moore 1971: *The astronomy of Birr Castle*, by Patrick Moore, Michael Beazley, London, pp. xii + 81.

Moore 1989: *Schrödinger life and thought*, by Walter Moore, Cambridge University Press, pp. xi +513.

Murphy 1989: 'Dive! dive! dive!' by William Murphy, *New Scientist*, 7 January 1989, pp. 36-41.

O'Brien 1984: *A portrait of Irish medicine – an illustrated history of medicine in Ireland*, by Eoin O'Brien, Anne Crookshank and Gordon Wolstenholme, Ward River Press, Dublin, pp. xv + 307.

O'Hara 1993: 'George Johnstone Stoney and the conceptual discovery of the electron' by James O'Hara, pp. 5-28 in 'Stoney and the electron', *Occasional Papers in Irish Science and Technology*, 8, Royal Dublin Society, pp. 39.

O'Loughlin 1994: 'Seminar commemorates 100th anniversary of Marconi's radio', by Edward O'Loughlin, *The Irish Times*, 10 June 1994.

Ó Raifeartaigh 1985: *The Royal Irish Academy – a bicentennial history 1785-1985*, Edited by T. Ó Raifeartaigh, Royal Irish Academy, Dublin, pp. vvi + 351.

Osterbrook 1984: *James E. Keeler, pioneer American astrophysicist*, by B.E. Osterbrook, Cambridge University Press.

Parsons 1926: *The scientific papers of William Parsons, third Earl of Rosse 1800-1867, collected and republished by The Hon. Sir Charles Parsons, K.C.B., F.R.S.*, London, pp. 221. This reference also contains papers by Thomas Romney Robinson describing the Birr telescopes.

Parsons 1934: *Scientific papers and addresses of The Hon. Sir Charles Parsons*, Edited by The Hon. G.L. Parsons, Cambridge University Press, pp. xxviii + 260.

Proctor 1899: *The orbs around us*, by Richard A. Proctor, Longmans, Green & Co., London – the quoted passage was first published in Fraser's Magazine for December 1869.

Rae 1980: *Harry Ferguson and Henry Ford*, by John Rae, Ulster Historical Foundation, pp. x + 32.

Rees 1992: *The life of Capt. Robert Halpin*, by Jim Rees, Dee-Jay Publications, Arklow, pp. 169.

Rees 1994: 'Captain Robert Halpin, cable king', by Jim Rees, *Technology Ireland*, March 1994, pp. 34-37.

Roscoe 1869: *Spectrum Analysis*, by Henry E. Roscoe, Macmillan, London, pp. xv + 348.

Rynne 1987: *John Philip Holland*, address by Etienne Rynne at the opening of the Holland Theatre in the University of Limerick in June 1987.

Scaife 1985: 'The Parsons steam turbine', by Garrett Scaife, *Scientific American*, April 1985, pp. 100-106.

Scaife 1988: 'Science and engineering from galaxies to turbines', by Garrett Scaife, *Physics Bulletin*, 39, pp. 103-105.

Scaife 1999: 'William Parsons', by Garrett Scaife, in Mollan 1999.

Scott 1992: *Treasures of the mind*, Edited by David Scott, (Trinity College Dublin Colonnades Gallery Exhibition Catalogue), Sotheby, London, pp. xvi + 176.

Sharkey 1998: 'Oh Kay, computer', by Patricia Sharkey, *The Irish Times*, 30 November 1998.

Smith 1989: *Energy and empire – a biographical study of Lord Kelvin*, by Crosbie Smith and M. Norton Wise, Cambridge University Press, Cambridge, pp. xxv + 866.

Somerfield 1990: 'William Sealy Gosset ("Student")' by Adrian Somerfield, pp. 58-9 in Mollan 1990.

Sopka 1981: 'John Tyndall: international populariser of science', by Katherine Sopka, pp. 193-203 in Burchfield 1981.

Spearman 1985: 'Mathematics and theoretical physics' by David Spearman, pp. 198-239 in Ó Raifeartaigh 1985.

Spearman 1990a: 'James MacCullagh', by David Spearman, pp. 22-23 in Mollan 1990.

Spearman 1990b: 'Samuel Haughton', by David Spearman, pp. 36-5 in Mollan 1990.

Spearman 1990c: 'George Salmon', by David Spearman, pp. 34-5 in Mollan 1990.

Spearman 1999: 'John Lighton Synge' by David Spearman, in Mollan 1999.

Spencer 1990: 'Robert (Roy) Charles Geary', by John Spencer, pp. 66-7 in Mollan 1990.

Spencer 1997: 'R.C. Geary: his life and work', by John Spencer, pp. 3-78 in Conniffe 1997.

Stoney 1869: 'On a cheap form of heliostat', by G.J. Stoney, *Report of the 39th Meeting of the British Association for the Advancement of Science,* Exeter.

University of Ulster 1989: *The history of technology, science and society 1750-1914*, papers from a Symposium held in September 1989 (29 papers, unpaginated).

Walker 1995: *Larousse dictionary of science and technology,* General Editor, Peter Walker, Edinburgh, pp. xii +1236.

Wayman 1987: *Dunsink Observatory, 1785-1985, a bicentennial history*, by Patrick Wayman, Dublin Institute for Advanced Studies and Royal Dublin Society, pp. xiii + 353.

Wayman 1989: ‘Thomas Grubb, telescope maker and banknote printer’, by Patrick Wayman, eleventh paper in University of Ulster 1989.

Weaire 1987: ‘Unfulfilled renown – Thomas Preston (1860-1900) and the anomalous Zeeman effect’, by Denis Weaire and Shea O’Connor, *Annals of Science*, 44, pp. 617-644.

Weaire 1990: ‘Thomas Preston’, by Denis Weaire and Shea O’Connor, pp. 50-51 in Mollan 1990.

Williams 1994: *Biographical dictionary of scientists*, Edited by Trevor Williams, Harper Collins, Glasgow, pp. xviii + 602.

Wood 1995: ‘George Gabriel Stokes 1819-1903, an Irish mathematical physicist’ by Alastair Wood, *Bulletin of the Irish Mathematical Society,* Number 35, Christmas 1995, pp. 49-58.

Wood 1999: ‘George Gabriel Stokes’, by Alastair Wood, in Mollan 1999.

Wyse Jackson 1992: ‘A man of invention: John Joly (1857-1933), engineer, physicist and geologist’, by Patrick Wyse Jackson, pp. 89-96 in Scott 1992.

Acknowledgement

The author is most grateful to David Attis of the Program in the History of Science, Princeton University, New Jersey, for reading an earlier version of the text, for pointing out errors, and for suggesting amendments which have been incorporated into the final version. Mistakes which remain are, of course, the responsibility of the author. If you spot any, please inform Charles Mollan at Phone 01-289-6186, Fax 01-289-7970 or E-mail cmol@iol.ie

Bibliography

100 Years of Radium, ASE, 1998.

A Brief History of Time, S. Hawking, Bantam Books, 1988.

A Laboratory Manual of Physics, F. Tyler, Edward Arnold, Ltd, 1967.

A Laboratory Manual of Physics, F. Tyler, Hodder and Stoughton, 1977.

A New Physics, D. Bryant and D. G. Kershaw, Hodder and Stoughton, 1978.

A Science Foundation Course, Unit 32 The search for fundamental particles, Open University, 1989.

Access to Advanced Level Physics (Second Edition), Dave Bush and Bob Drumgoole, Stanley Thornes (Publishers) Ltd, 1996.

Advanced Level Physics, Nelkon and Parker, Heinemann Educational Books, 1979.

Advanced Level Practical Physics, M. Nelkon and J. Ogborn, Heinemann Educational Books, 1982.

Advanced Physics (Fourth Edition), Tom Duncan, John Murray, 1994.

Advanced Physics Revision Handbook, Stephen Pople, Oxford University Press, 1996.

Advanced Physics, Patrick Fullick, Heinemann, 1994.

Advanced Questions on Everyday Physics, Susan Williams, Nelson Blackie, 1993.

Adventures with Micro-Electronics, Tom Duncan, John Murray, London, 1979.

ASE Guide to Secondary Science Education, Mary Ratcliffe (ed.), ASE, 1998.

Asimov's Biographical Encyclopedia of Science and Technology, I. Asimov, Pan Books Ltd, 1975.

Becquerel's legacy: a century of radioactivity, *Radiation Protection Dosimetry* **68**(1/2), 1996.

Beyond the Visible - One Hundred Years of X-Rays, ASE, 1995.

Big Bang Science, Exploring the Origins of the Universe, Science & Engineering Research Council,1989.

Bubbles, Voids & Bumps in Time: the new cosmology, J. Cornell (ed.), Cambridge University Press, 1989.

Changing the Subject - Innovations in Science, Mathematics and Technology Education, P. Black and J. M. Atkins (Eds), Routledge, 1996.

Concepts in Physics, R.K. Adair, Academic Press, 1969.

Core Physics, Cackett, Kennedy and Steven, Oxford University Press, 1979.

Curie in 90 minutes, John and Mary Gribbin, Constable, 1997.

Darwin in 90 minutes, John and Mary Gribbin, Constable, 1997.

Directory of Physics Resource Materials, The Association for Science Education, 1985.

Einstein in 90 minutes, John and Mary Gribbin, Constable, 1997.

Electricity and Modern Physics, G.A.G. Bennett, Edward Arnold, Ltd, 1968.

Electricity, C. A. Coulson, Oliver and Boyd, 1965.

Essential Principles of Physics, P. M. Whelan & M. J. Hodgson, John Murray, 1977.

Everyday Things and How They Work, Stephen Parker, Kingfisher Books, 1991.

Fundamentals of Physics, D. Halliday and R. Resnick, John Wiley and Sons, NY, 1986.

Fundamentals of Physics, D. Halliday, R. Resnick and J. Walker, Wiley, 1997.

Fundamentals of Physics, M. Nelkon, Granada Publishing (Hart-Davis Ed Ltd), 1977.

Galileo in 90 minutes, John and Mary Gribbin, Constable, 1997.

GCSE Electronics, Tom Duncan, John Murray, London, 1989.

Halley in 90 minutes, John and Mary Gribbin, Constable, 1997.

Henri Becquerel and Radioactivity, David Sang, Jill Sutcliffe and Mary Whitehouse, ASE, 1997.

Higher Physics, J. Jardine, Heinemann Educational, 1983.

How energy becomes matter… A first look at the world of particles, CERN, 1986.

How Everyday Things Work, Peter Goodwin, J. Weston Walch, 1992.

How Things Work, N. Ardley, Dorling Kindersley.

Intermediate Mechanics, D. Humphrey, Longman, 1971.

Leisure Physics (Science Now), Edward Fogden, Stanley Thornes, 1989.

Light, M. Sobel, University of Chicago Press, 1987.

Magnets: Advances in Permanent Magnet Technology, Mike Coey, TCD, CEAM (Commission of the European Communities).

Medical Physics 16-19 Project, M. Hollins, University of Bath/McMillan, 1989.

Medical Physics, Martin Hollins, Nelson, 1992.

Mendel in 90 minutes, John and Mary Gribbin, Constable, 1997.

Moons and Planets, W. K. Hartmann, Wordsworth, 1993.

My Name is Becquerel, Ernst Schwenk, Hoechst, 1994.

Newton in 90 minutes, John and Mary Gribbin, Constable, 1997.

Nuclear and Particle Physics, D. Sang, *University of Bath* • Science 16-19, Nelson, 1995.

Nuffield Science Teaching Project, J. Paull & D. Paull, 1982.

Of Magnets And Man, Z. Rahman, in ELEMENTS, University of Limerick, 1995.

One hundred years after the discovery of radioactivity, Special issue *Radiochimica Acta* **70/71**, Oldenbourg Verlag, 1996.

One hundred years of the electron, ASE, 1997.

Orbital Motion, A. E. Roy, Hilger , 1982.

Orbiting the Sun, F. L. Whipple, Harvard University Press, 1980.

Ordinary Level Physics, A. F. Abbott, Heinemann Educational Books, 1984.

Particle Physics Project, Institute of Physics, 1992.

Perceiving Shape from Shading, Vilayanur S. Ramachandra, *Scientific American*, August 1988.

Physics - A Textbook for advanced Level Students, Tom Duncan, John Murray, 1984.

Physics and Human Affairs, A. Hobson, Wiley, 1982.

Physics can be fun – A sourcebook of Practical Problems, Peter Goodwin, J. Weston Walch, K. Portland Maine, 1985.

Physics for All, J. J. Wellington, Stanley Thornes, Ltd, 1988.

Physics for the Health Sciences, C. and B. Nave, W. B. Saunders Company, 1985.

Physics Through Applications, J. Jardine (Ed.), Oxford University Press, 1989.

Physics Through Investigation, Glen Ireson, Hodder & Stoughton, 1998.

Physics, A. Alonso and E. J. Finn, Addison-Wesley, 1992.

Physics, D. C. Giancoli, Prentice Hall International, 1995.

Physics, D. Halliday and R. Resnick, John Wiley & Sons, 1978.

Physics, Physical Science Study, Massachusetts Institute, 1958.

Physics, R. Hutchings, *University of Bath • Science* 16-19, Nelson, 1991.

Physics: Longman Revision Guides, K. Palfreyman and C. Maunder, Longman, 1994.

Practical Electronics Handbook, Ian Sinclair, Heinemann, London, 1987.

Practical Physics at A-Level, T. Cross, Collins Educational, 1985.

Practical Physics Labs – A Resource Manual, Peter Goodwin, J. Weston Walch, 1990.

Principles with Applications, D. C. Giancoli, Prentice Hall, 1991.

Quarks - the Stuff of Matter, Harald Fritsch, Penguin Books.

Renewable Energy, Shell International Petroleum Co. 1994.

Satellite Communications Systems, B. G. Evans, Institute of Electrical Engineers, 1991.

Science for Girls, A. Kelly (Ed.), Open University Press, 1987.

Science for Motor Vehicle Technicians, S. J. Zammit, Longman, 1973.

Seven Ideas that Shook the Universe, N. Spielberg & B. D. Anderson, John Wiley and Sons, 1995.

Shaping the Future 1 Making Physics Connect, Peter Campbell (ed.), Institute of Physics Publishing, 1999.

Shaping the Future 2 Physics in Mathematical Mood, Simon Carson (ed.), Institute of Physics Publishing, 1999.

Simple Mechanics, Ladybird Junior Series.

Study Topics in Physics (Series), W. Bolton, Butterworths, London, 1980.

Success in Electronics, Tom Duncan, John Murray, London, 1987.

Teaching Ethical Aspects of Science, Patrick Fullick and Mary Ratcliffe (eds.), The Bassett Press, 1996.

The Book of Inventions and Discoveries, Ian Marshall and Janet Ravenscroft (Eds), McDonald Queen Anne Press, 1991.

The Cartoon Guide to Physics, Larry Gonick and Art Huffman, Harper Perennial, 1991.

The Cosmic Onion, Frank Close, Heinemann Educational Books, 1983.

The discovery of radioactivity, *Physics Today*, February 1996, pp. 21–26.

The Effective Teaching of Secondary Science, John Parkinson, Longman Group Plc, 1994.

The Flying Circus of Physics, J. Walker, John Wiley & Sons (New York), 1977.

The IT in Science of Datalogging and Control., Roger Frost, IT in Science, 1995.

The IT in Secondary Science Book, Roger Frost, IT in Science, 1994.

The Particle Explosion, Frank Close, Michael Martin and Christine Sutton, Oxford University Press.

The Physical World, Dobson, Nelson, 1991.

The Physical World, Keighley and McKim, Pergamon Wheaton, 1975.

The Planetary System, D. Morrison & T. Owen, Addison Wesley, 1987.

The Realm of the Terrestrial Planets, Z. Kopal, Institute of Physics, 1979.

The Way Things Work, D. McCauley with N. Ardley, Dorling Kindersley, London, 1988.

The World of Particles, Georges Boixader & Brian Southworth, CERN, 1991.

The World of Physics, John Avison, Nelson, 1989.

Thermal Physics, M. Sprackling, Macmillan, 1991.

Thinking Physics is Gedanken Physics, L. Epstein, Insight Press, 1985.

Understanding Physics, M. Mansfield and C. O'Sullivan, Wiley, 1998.

University Physics, F. Sears and M. Zemansky, Addison-Wesley, 1971.

INDEX

Web Sites

Organisation	**Address**
American Physics Teachers' Association	http://www.psrc-online.org
Association for Science Education	http://www.ase.org.uk/
CERN	http://www.cern.ch
Department of Education and Science	http://www.irlgov./educ/
Institute of Chemistry of Ireland	http://www.iol.ie/~instchem/
Institute of Electrical Engineers	http://www.iee.org.uk/Schools/
Institute of Physics	http://www.iop.org/
Internet for Girls: World Wide Resources List	http://www.sdsc.edu/~woodka/resources.html
Irish Research Scientists' Association	http://www.irsa.ie
Irish Science Teachers' Association	http://indigo.ie/~istasec/
National Centre for Technology in Education	http://www.ncte.ie
New Scientist	http://www.newscientist.com
Particle Physics and Astronomy Research Council	http://www.pparc.ac.uk
Radiological Protection Institute of Ireland	http://www.rpii.ie
Royal Society of Chemistry	http://www.rsc.org
Schools Information Centre for the Irish Chemical Industry	http://www.ul.ie/~childsp/Sicici/index.html
Scoilnet	http://www.scoilnet.ie
The Chemists' Net	http://www.chemists-net.demon.co.uk