



Physics Department

PHY-103

SCIENTIFIC MEASUREMENT

2011–2012

Contents:

- **Course information**
- **Laboratory instructions**
- **Lecture notes**

Student Name

This booklet contains:

0. A summary of information about the course.

1. Detailed instructions for the laboratory experiments. Bring this booklet, your laboratory notebook, and your report worksheets every time you come to the laboratory. Blank worksheets will be distributed in the laboratory.

2. Lecture notes. These cover the material considered essential for the course, but are not a substitute for your own record of the lectures.

PHY-103 SCIENTIFIC MEASUREMENT

Course Schedule 2010–2011

Week	Dates	Group				Marks
		A1 Monday	A2 Tuesday	B1 Thursday	B2 Friday	
1	Sept 26 – Sept 30	<i>Lectures on Tuesdays and Fridays 12 noon, in weeks 1–4</i> Complete experiments 1, 2, 3 (one per week, in weeks 2-4) according to the schedule in the laboratory				
2	Oct 3 – Oct 7					
3	Oct 10 – Oct 14					
4	Oct 17 – Oct 21					
		Monday and Tuesday		Thursday and Friday		
5	Oct 24 – Oct 28	Experiment 4 or experiment 5 <i>Lectures Tuesday and Friday</i>				25% or 15%
6	Oct 31 – Nov 4	Experiment 5 or experiment 4 <i>Lectures Tuesday and Friday</i>				15% or 25%
7	Nov 7 – Nov 11	<i>Reading week: write up experiment 4 Report</i>				
8	Nov 14 – Nov 18	Formative Assessment of Experiment 4 Report <i>Lectures Tuesday and Friday</i>				
9	Nov 21 – Nov 25	Choose one of experiments 6–12 <i>Lectures Tuesday and Friday</i>				40%
10	Nov 28 – Dec 2	<i>Continue (one of three parts per week)</i>				
11	Dec 5 – Dec 9	<i>Continue (one of three parts per week)</i>				
12	Dec 12 – Dec 16	<i>Write up Report 6, 7, 8, 9,10,11, or 12</i>				
		2 Homework exercises (due in weeks 4 and 6)				20%

Note: Updated information will be posted on the Scientific Measurement **website**, which should be consulted regularly at: <http://www.ph.qmul.ac.uk/~phy103/>

LABORATORY ETIQUETTE

The undergraduate laboratory is a place of work, and you should always conduct yourself accordingly. Remember that this is a scheduled class and you should use your time wisely. In doing so you will also be respecting your fellow classmates and the staff. In particular you should adhere to the following at all times:

- No food or drink should be taken into or consumed in the laboratory.
- Your mobile phone should be switched off while you are in the laboratory. Similarly you should not use MP3 players or other devices that would distract you from your surroundings.
- Aisles should remain clear at all times, as these are fire exit routes. Bags strewn across the aisles represent a trip hazard, and will be moved.

One final note: Everyone in the laboratory has a responsibility to use equipment in a safe and controlled way. Be aware of your surroundings at all times, and be aware of any hazards in your working environment. Safety is best obtained by you exercising care continually.

- **PHY-103**
SCIENTIFIC MEASUREMENT

Essential Information

Teaching staff

The two course organisers are:

Dr. Eram Rizvi (room 401, Physics) e.rizvi@qmul.ac.uk

Office hour: Wednesday's 11am-12pm

Dr. Jeanne R. Wilson (room 507, Physics) j.r.wilson@qmul.ac.uk

Office hour: Tuesday's 10-11am

There will be one academic demonstrator in the laboratory. There will also be six graduate students helping out during the term. In addition this course relies on the invaluable and enthusiastic assistance of our lab technicians, **Peter Crew** and **Saquib Qureshi** (technicians office in the 2nd floor laboratory, Physics). The lectures will be given by **Dr. Rizvi**.

Website

Basic information and up-to-date news about the course are given on its website, which is at:

<http://www.ph.qmul.ac.uk/~phy103/>

During the semester, homework solutions and comments on marking will be added, and when there is any new information, changes to normal arrangements, or general news they will be put on the course's web home page. **Be sure to consult this home page regularly.**

Summary of basic information

This course has two objectives: to teach techniques and skills that you will use in later courses, and to train you to think critically about experimental data. Teaching is mainly by practical work in the laboratory, supplemented by lectures and homework problems.

Assessment is entirely by coursework and continuous assessment — there is no written examination. **Attendance** at both the lectures and the laboratory sessions is **mandatory**. Attendance **will be checked**, and **warnings** will be given to those who are missing without a valid reason.

For each of the first four weeks you attend two 1-hour lectures and from the second week one 3-hour laboratory class. Lab classes start the second Monday of the term. The lectures deal with measurement, the treatment of errors, statistical distributions, etc. Two homework problem sheets will supplement the lecture material, due in weeks 4 and 6. Lectures take place on Tuesdays and Fridays at 12 noon (see the departmental timetable for the room location). The laboratory exercises cover electrical, optical, nucleonic and general techniques, in no set sequence — you work through each in turn. You will also learn to use the PCs in the laboratory for making graphs of laboratory data.

From the fifth week you typically attend one lecture, although this depends on the lecturer, and two 3-hour laboratory classes each week. The laboratory topics are:

- Weeks 5 & 6: short projects using a digital thermometer, and the oscilloscope and its uses.
- Week 8: formative assessment of the first lab report.
- Weeks 9, 10 and 11: a short project, which you choose from vibrations and waves, astronomy, computer control, mechanics, subatomic physics, electronics, spectroscopy or thermal efficiency.
- Weeks 7 ('reading week') and 12 are for writing up reports.

Laboratory classes are held in the Physics 2nd floor laboratory from **2.00–5.00 pm¹** on Mondays, Tuesdays, Thursdays and Fridays; demonstrators are only available during these periods. The laboratory can open at other times only by special arrangement with the Senior Technician, Mr Peter Crew. You will be assigned to one of two groups, A or B. Group A will attend lab classes on Mondays and Tuesdays and Group B on Thursdays and Fridays. For the first four weeks, the groups will be divided into subgroups A1, A2, B1 and B2 (see table). You always work in **pairs**. After you have completed laboratory exercises 1–5, you select one of exercises 6–12 and complete the three parts in three weeks.

If you finish experimental work before the end of the afternoon, it is a *mistake* simply to leave the laboratory. Use the time to start writing up your results, since the demonstrators are available there to answer questions and help you.

Before turning up to the laboratory

During this course you should make sure that you have read through the description of the laboratory experiment assigned to you (each session) prior to turning up to the department. This will help you identify the tasks that need to be completed, any questions you might have for the staff on hand, and to help you to use your time in the laboratory as efficiently as possible.

Laboratory notebooks

Laboratory notebooks (lab books for short) are scientific logs of your work. They are functional documents that you will use in order to collate and interpret results from source data that you take in the laboratory. There is a lot of good advice on the appropriate use of lab books in chapter 10 of the book by Squires (see below). The golden rule is to say what you are doing: write a sentence or two when you make measurements, put labels on graphs, captions on sketches, headings (with units!) on tables, etc. On reading your lab book you (and anyone else) should be able to reconstruct what happened that day in the lab. Only by writing things down systematically can you hope to do this — do not trust your memory. The following are some points to keep in mind while using your lab book:

1. You will need to bring your lab book to the laboratory whenever you are doing an experiment. This book should have your name clearly written on the front and inside cover in order for it to be easily identified.
2. Entries in the lab book do not need to be extremely neatly written, but **should be logically ordered and clear enough for you, and someone else to read**. Care should be taken to neither rush entering information into your book, nor painstakingly producing a work of art. The lab book is a working document: not a final draft.

¹ The laboratory technicians will be in the laboratory from 1.30pm for those students who wish to make an early start on their experiment.

3. Entries in the book should be clearly dated, with an appropriate title. For example:

28th September 2010 The Oscilloscope (Experiment Number 5)

4. Think about the aims of the experiment and what measurements you will need to make before you start to do anything else. This information can be obtained by reading through the description of the experiment from start to finish. If you have questions, **ask a demonstrator**, otherwise start to make a few notes of what you need to do, and then run through that list. For example, Experiment 1 may be summarised with

11/2/2009 14:05
STARTING LABEX. #1
MEASUREMENT OF g
Aim: Use a pendulum to test
 $T = 2\pi\sqrt{\frac{L}{g}} \dots$

Having identified the aim, you will then want to make a list of things you will need to do (and refer back to that list as the experiment progresses so that you make sure you don't forget to take a particular measurement. This may seem tedious, but it will help you to organize your time [= not waste time] in the laboratory! The list doesn't need to be too detailed, and may be as brief as:

CALCULATE θ GIVEN $\frac{\theta}{\sin\theta} = 1.02$
 SET l TO 30cm; COMPUTE $(\pm 5\%)$
- FIGURE OUT DISPLACEMENT CORRESPONDING TO θ ABOVE
- MEASURE T USING ...

where you might consider using check boxes that are ticked off as you progress through the experiment.

5. When preparing to take data, take a moment to think about what information you will need to record. Having done this you will be in a position to set up a logical order to log the data in your lab book.

6. All data (and calculations) that are required in order to write up an experiment should be written in your lab book. DO NOT USE, OR INSERT, SCRAPS OF PAPER when working out a calculation, as these can be easily lost. Any ancillary material added to your lab book (i.e. graphs printed from a computer or graph paper) should be firmly secured with glue or tape.

7. When you start to work on a new item on your list, note the time that you start working. Similarly if something goes wrong – make a note of what happened and when!

15:00 START TO MEASURE T WITH $L=60cm$
15:05 → PENDULUM STRING BREAKS
ASKED FOR A NEW ONE
[WILL RESTART MEAS.
FOR THIS L]

Laboratory report worksheets

As you work through the laboratory course you should take comprehensive notes in your lab book (see above). Once you have completed taking measurements, **all** of your measurements, calculations, graphical work and conclusions for laboratory experiments 1–3 & 5 **must** be entered directly onto the separate **lab report worksheets**. These worksheets have been designed to accommodate all the graphs, calculations and answers that you are expected to produce for these exercises.

The worksheets should not take too long to complete — two or three hours at most. **The deadline for handing them in is**

Experiment 1-3 and 5: **Before the start of your lab session the following week.**

They will be marked and returned as quickly as possible so you can learn from any feedback given on your worksheet or verbally by the demonstrators.

Rough working should have been completed in your lab book prior to you starting to write on the worksheet. We expect your worksheets to be clear, logically ordered and legible. You should record data and reproduce calculations neatly enough to read and marked by someone who is not familiar with your handwriting. If your work is illegible, then you will be required to prepare a legible copy from your original. Additional copies of the worksheet will be available on request.

Note that for **experiment 4** you *must* hand in *only* the **formal report**.

Formal laboratory reports

The course also requires you to produce **two formal reports**. The first is for **exercise 4**, and the second is for **one of exercises 6–12** towards the end of the semester. These reports must be typed using a word processor. If you are familiar with L^AT_EX, then you may use this to produce your report (this software is installed on the student network). As a general rule they should contain the following **sections**:

- Title
- Abstract
- Introduction
- Theory
- Experimental Details
- Results and Discussion
- Conclusions
- References

There is a good example of such a report in exercise 4. Use this as a guide for your own report (but note that yours will be much shorter). Writing reports will be covered in the lectures. It is also useful to read Squires, chapter 13 ('Writing a Paper') or Silyn-Roberts.

IMPORTANT: You should submit each formal report with the following specification: (i) single sided A4 (ii) use 11 or 12 point font using only one of Times, Times New Roman or Arial fonts (iii) bound loosely in **plastic binders** available from the technicians at cost. Do not staple, or otherwise fix the pages of your report together.

Teamwork vs. plagiarism

Worksheets and formal reports that you hand in *must* be your own work. Do not copy from other students. You will work in pairs for all experiments, and this means that you do the measurements as a team. However, you must write *your own* laboratory worksheets and reports. Submission of reports that are very similar, or that have parts that look as if they were copied from someone else, will be treated as possible *plagiarism* and may result in *serious disciplinary action*. See also the regulations for writing essays and reports in the Student Handbook.

Recommended books

The books listed here are available in the library, but can also be borrowed from the laboratory technicians. You are not required to buy any of these books, although you should consult them frequently:

An Introduction to Error Analysis by J.R. Taylor (University Science Books, 2nd edition 1997) is a very good statistics book, and you should go through some of its chapters in parallel with the lectures.

Statistics by R. Barlow (Wiley, 1989) is an excellent reference book on statistics that goes far beyond what is covered in this course.

Practical Physics by G.L. Squires (Cambridge Univ. Press, 4th edition 2001) is recommended as a guide to good laboratory practice.

Writing for Science by H. Silyn-Roberts (Longman, 1996) is about writing scientific documents, with much more computer-oriented information than Squires.

Use of computers

PCs running under Microsoft Windows are available for your use in this and other courses. There are additional PCs elsewhere running the same system. A wide variety of software, both commercial and locally-written, is available. You should feel free to use these computers at any time. There are laser printers available for making paper copies. These PCs normally run Windows on starting up.

Plotting programs: graphs and histograms

At various times you will need to use the computers to plot data as graphs or histograms, and to do simple line or curve fitting. There are a number of application programs available to you. You can use the following Windows programs:

- **PhysPlot** was written in the Physics Department with this course in mind. It produces suitable graphs and histograms and provides line and curve fitting. It is straightforward to use but has a few limitations. PhysPlot can be found in the Physics Applications program group. *You can download a copy of PhysPlot for use on your own computer from:* <http://hepwww.ph.qmul.ac.uk/PhysPlot>
- **Microsoft Excel** is a very powerful spreadsheet program which has extensive graphing and analysis capability. It has much more flexibility but is more complicated, and is also less oriented towards scientific graphs than PhysPlot. Note in particular that if you use Excel you **MUST** be able to plot uncertainties on the graphs and data points, otherwise use PhysPlot. Excel is found in the Microsoft Office program group.

You should try out these programs, or any others, and use whichever you feel happy with.

Word processing

An essential part of the course is to write two formal reports. These **must** be prepared using a **word processor**. The recommended application for word processing is **Microsoft Word**. You will find it in the Microsoft Office program group, in the Start menu. Like Excel it is fairly easy to use at a simple level, but there are many sophisticated and subtle features, some of which can save you a great deal of time and effort and are definitely worth learning about if you use it a lot. However, you are free to use any word processing application that you are happy with. If you are familiar with L^AT_EX programme for writing documents you may use this instead of Microsoft Word. You will find L^AT_EX on the student network (with an advice page on how to process your document on the departmental intranet).

E-mail and the web

We have already mentioned the course **website**, which contains much useful information and will be used for transmitting news and for posting homework solutions. In addition, individual contact between students and the course organisers is often most efficiently done by **e-mail**². You will be notified if you are in danger through not attending the course or not handing in work on time, and you can use e-mail to inform us of what might be causing such problems or to ask questions. We are very aware that e-mail is not a subtle medium and is often no substitute for face-to-face discussion, but it is an effective way to convey simple facts and queries without having to find someone (who may be halfway around the world at the time) in person.

General comments on computing

Do not be afraid to experiment — the worst that can happen is that you lose the data you have typed in (though saving it frequently will help to prevent that).

Always protect against **losing files** by **backing them up** to a memory stick. Loss of files will not be considered a valid excuse for not handing in work.

Finally, it is worth noting that some programs used for the **lecture demonstrations** on statistics are available in the Physics Courses program group.

² All communication to students will be via your college e-mail account. You will need to check this regularly.

Late and failed Submissions

The penalties for late work will be strictly enforced for both homework and laboratory scripts. If you have a valid reason (illness, bereavement etc) for late submission you must contact the course organisers as soon as possible, preferably before the deadline and submit an extenuating circumstances form to the departmental office.

Mark penalties are as follows:

Length of time after submission deadline	Mark penalty
<24 hours	-20%
1-3 days	-50%
>3 days	-100%

Experiments 1-3 do not contribute to the final course mark but it is compulsory to hand these in. This gives you a chance to act on the feedback you receive for this early work and apply what you have learnt to the later assignments that carry more marks.

It is a requirement that all work is submitted for this course, including the laboratory scripts that do not contribute to the final course mark (experiments 1-3). Any student with more than one piece of outstanding work will fail the course.

Homework

Homework problem solutions should be placed in the pigeon holes on level 1 by **2pm** on the following dates:

Homework 1: Thursday 20th October

Homework 2: Thursday 3rd November

Solutions will be available on the course web page 3 days after the hand-in deadline and late submission penalties detailed above will be strictly enforced.

Lab reports

As soon as you have completed a **laboratory exercise** you should hand in your laboratory report worksheet to one of the technicians in the laboratory. It will be marked and returned to you. **You may only start the next exercise once you have handed in the previous one.** This means that 'short' experiments must be handed in by the time of your laboratory session the following week.

You should have finished doing lab reports 1–3 by **Friday of week 4**. Your two **formal laboratory reports** (one on experiment 4, the other on one of experiments 6–12) should be handed in **to a laboratory technician** and a **receipt** obtained. The first long report and the script for experiment 5 should both be handed in by **Monday 14th November at 12:00h** and the second long report by **Friday 13th January at 12:00h**. The late submission penalties detailed above will be strictly enforced.

In order to pass the course, you must complete and hand in ALL experiments, homework exercises and reports.

Summary of deadlines

Thursday of week 4	First homework exercise
Friday of week 4	Laboratory experiments 1 to 3 should have been completed, with work handed in according to the due date for each Experiment.
Thursday of week 6	Second homework exercise
Monday of week 8	Due date for first formal report (experiment 4).
Monday of week 8	Due date for experiment 5.
Friday of week 1, Semester B	Due date for second formal report (one of experiments 6 to 12).

LABORATORY EXPERIMENTS

- 1 Electrical Measurements
- 2 Linear and Non-linear Behaviour
- 3 Nucleonic Measurements
- 4 Digital Thermometry
 - A Making the digital thermometer
 - B The cooling of coffee
- 5 The Oscilloscope
 - A Frequency measurement and the XY display
 - B Time constants and RC circuits

Choose ONE of the following:

- 6 Measurements of Wave Velocity (*do any two parts*)
 - A Sound waves in air
 - B Optical Measurements
 - C The vibrations of a copper rod
- 7 Measurements in Astronomy
 - A The angular resolution of telescopes
 - B Line spectra, chromatic resolution, and Doppler shifts
 - C The expansion of the universe
- 8 Light and Other Electromagnetic Waves
 - A Refraction of light
 - B The velocity of light
 - C X-ray diffraction
- 9 Fundamental and Subatomic Physics
 - A The charge-to-mass ratio of the electron, e/m
 - B Kinematics on a linear air track
 - C The decay of the π -meson
- 10 Digital and Analogue Electronics
 - A A counter, decoder and LED display
 - B Operational amplifiers
 - C A simple digital-to-analogue converter
- 11 Computing and Computer Control
 - A Visual BASIC
 - B Interfacing to devices
 - C Computer control of temperature
- 12 Thermal Efficiency
 - A Heat Engine Efficiency
 - B Heat Pump coefficient of performance
 - C Load for optimum performance

Laboratory Exercise 1 – ELECTRICAL MEASUREMENTS

Introduction

Electronic techniques and instrumentation are important in almost every branch of science. In this exercise you will meet some of the principles which govern the design and operation of such instruments. We restrict ourselves to circuits carrying steady, unchanging currents provided by constant voltage sources such as batteries (so-called **direct current** or **DC** circuits); in later exercises we meet time-varying currents in what are called, rather deceptively, **alternating current** or **AC** circuits.

Many substances can be loosely classified as **conductors**, which allow electric currents to pass relatively easily, or **insulators**, which offer much more resistance to current flow. Metals, with loosely bound electrons, are good conductors. So are various forms of carbon, such as graphite. An electronic component, manufactured to have a specific value of resistance, is called a **resistor**. When a current of I amps¹ passes through a resistor of value R ohms² there is a voltage drop or **potential difference** of V volts³ across the resistor. The familiar relation known as **Ohm's Law**, $V = IR$, is satisfied by most conductors including all metals. (However, it must be stressed that Ohm's Law does not apply to all substances. Modern electronics is possible only because of the development in the last half-century of a wide range of materials and devices for which the 'law' does not hold.)

A part of a circuit whose resistance is infinite (a perfect insulator, or a badly made 'dry' solder joint) is called an **open circuit**; one whose resistance is negligible is termed a **short circuit**.

Resistors commonly have values between a few ohms and several tens of megohms (1 megohm = 10^6 ohms). Some precision resistors (such as those in the variable resistance box you will use) are made entirely of metal, but most of those found in low-power electronic circuits are made of carbon, or an insulator thinly coated with a film of metal. These almost always have their resistance values marked on them by a simple and universally recognised colour code. There are three coloured rings near one end (figure 1); from the end these signify the first digit, the second digit, and the number of zeros to follow. A fourth ring at the other end shows the accuracy, useful in precision circuit work. The code is:

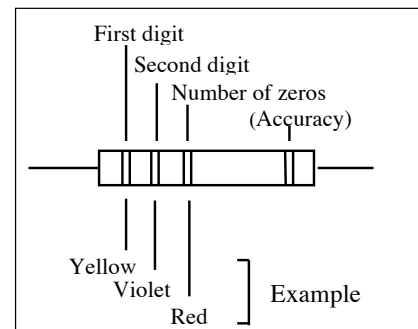


Figure 1 Resistance colour code

**Black = 0, Brown = 1, Red = 2, Orange = 3, Yellow = 4,
Green = 5, Blue = 6, Violet = 7, Grey = 8, White = 9**

A colour code chart is posted on the wall in the laboratory. In addition, there are several mnemonics for remembering this code.

In practice only a few resistance values are commonly used, selected so that each is about 20% larger than its predecessor. The basic ratios are 10, 12, 15, 18, 22, 27, 33, 39, 47, 56, 68 and 82. The resistor shown in figure 1 has a resistance of 4700 ohms = 4.7 kilo-ohms, usually read as '4.7 k', and often written as '4k7' to make the position of the decimal point more obvious on a crowded and perhaps poorly printed circuit diagram. (The 'k' is lower-case to distinguish it from 'K', the temperature unit degrees Kelvin.)

¹ The unit of current, the amp, is usually abbreviated to the symbol A.

² The unit of resistance, the ohm, is usually abbreviated to the symbol Ω .

³ The unit of voltage, the volt, is usually abbreviated to the symbol V.

Finally, we remind you that when resistors are joined in **series** (figure 2) the combined resistance is the sum of individual resistances:

$$R = R_1 + R_2 + \dots + R_n$$

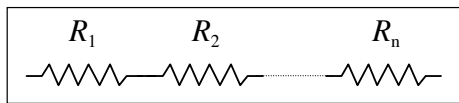


Figure 2 Resistors in series

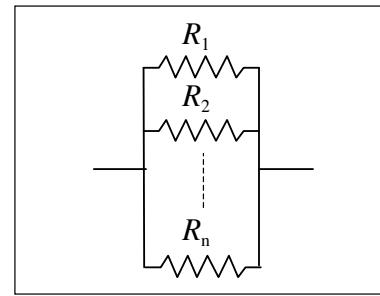


Figure 3 Resistors in parallel

but when they are connected in **parallel** (figure 3) the result is:

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} + \dots + \frac{1}{R_n}$$

Making simple circuits

In this exercise you will connect resistors in various ways. A power supply provides the voltage, and you will measure currents and potential differences with a digital multimeter (DMM, sometimes called a DVM when used to measure voltage).

Circuits are constructed on a so-called **breadboard**, an insulated board containing a grid of interconnected small sockets into which you can push the leads of electronic components such as resistors. Figure 4 shows the layout of a breadboard. The top and bottom lines of sockets are usually connected to the positive and negative terminals of the voltage supply. By convention voltage levels in the circuit are usually measured with respect to the negative line; several circuits or instruments can use the same negative potential (when it is called the **common** potential or, if the connection is made through the ground, the **ground** or **earth** potential). The first exercise illustrates the use of the breadboard.

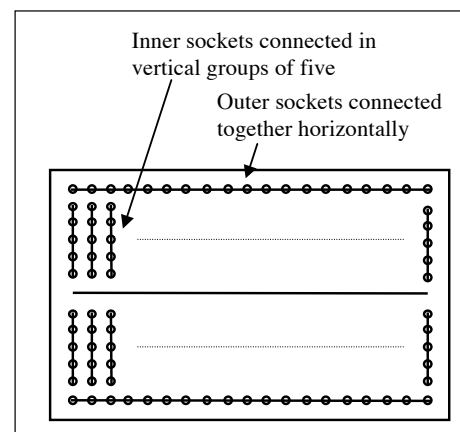


Figure 4 Breadboard connections

The potential divider

The simple circuit of figure 5 uses resistors in series to obtain a voltage lower than that directly available from the battery or other source. By Ohm's Law the current flowing is $I = V/(R_1 + R_2)$, so the potential difference between A and B is $V_{AB} = IR_1 = VR_1/(R_1 + R_2)$ while that between B and C is $V_{BC} = VR_2/(R_1 + R_2)$. Any voltage between ground potential and V can be obtained by an appropriate choice of R_1 and R_2 .

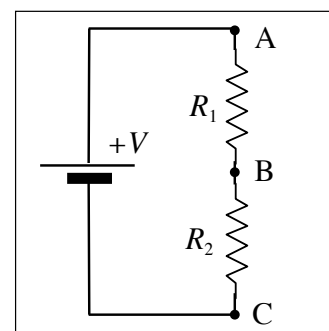


Figure 5 Potential divider

- Connect the circuit of figure 5 using the voltage supply provided. For R_1 and R_2 use resistors of 2.2 k Ω and 4.7 k Ω , respectively.
- First use the DMM to measure the actual resistances of R_1 and R_2 . Then use the DMM to measure the potential differences V_{AB} and V_{BC} , and compare with calculated values.
- Replace R_1 by the resistor labelled X , measure V_{AB} , and hence deduce R_X .

The Wheatstone bridge

This simple circuit (figure 6) uses the principle that if two voltage dividers R_1R_2 and R_3R_4 side-by-side have the same resistance ratios, that is if R_1/R_2 is the same as R_3/R_4 , then the voltages at points A and B must be equal. If three of the resistances are *known* then the fourth can be *calculated*.

This is an example of a **bridge** circuit. R_1R_2 and R_3R_4 are the two ‘arms’ of the bridge; when the bridge is ‘balanced’ there is no voltage between A and B so we don’t need a highly accurate voltmeter. (The principle of measuring a zero value is used in many so-called **null** methods of measurement.) This particular bridge circuit was devised by Sir Charles Wheatstone for use by telephone engineers.

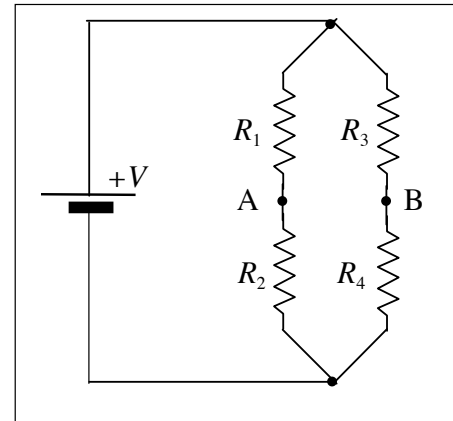


Figure 6 Wheatstone bridge

- Connect the circuit using equal resistors of 6.8 k Ω for R_1 and R_2 , the resistor labelled **Y** for R_3 , and the 1000 Ω variable resistor box for R_4 .
- Vary R_4 until the potential difference between A and B is zero. Hence determine the resistance of $Y=R_3$.

Input resistance

When you make measurements, you must always be aware that the instruments you use can themselves have an effect that changes the readings. For example, some meters have an **internal resistance** which is comparable with the resistances you are using, so when you measure a voltage with it you are in effect placing another resistance in parallel. On the other hand, a modern DMM has a high internal resistance than the circuit you normally want to measure, so it usually has little effect when measuring voltages. Here we will measure the internal resistance of the DMM meter.

- Connect the circuit of figure 7 using four equal resistors of 10 M Ω each.
- Measure the potential differences AB, BC and CD, using the DMM. Since $1/R_2 + 1/R_3 = 1/R_{BC}$, the resistance of the parallel section R_{BC} is half that of the series sections R_{AB} and R_{CD} , so V_{BC} should be half of V_{AB} or V_{CD} .
- Measure the open circuit voltage of the battery, i.e. the voltage across the battery terminals when no circuit is connected.

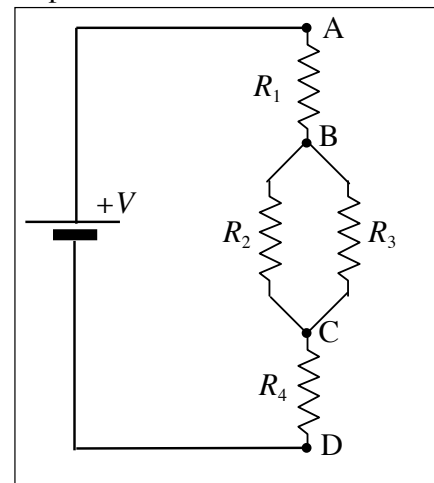


Figure 7 Input resistance

- From your results, and remembering that the total potential difference AD is constant, deduce the internal resistance of the DMM meter, and compare with the information given by the manufacturer on the back panel. Do this calculation twice, first using the value of V_{AB} , then V_{BC} . Your value of V_{CD} should be a useful check; is it?

The DMM has an internal resistance which is very high, so it has only a small effect on the voltage. An ideal voltmeter would have an infinite resistance and would then measure the **open circuit** voltage between two points. In practice, all instruments have some internal resistance

(usually called their **input resistance** or, when we are dealing with AC circuits, **input impedance**).

Output resistance

Any voltage source contains within itself some resistance to the flow of current, limiting the quantity of current that it can provide even when its output terminals are short-circuited. For many cells and batteries this **internal resistance** or **output resistance** is quite small, an ohm or less, as anyone who has accidentally dropped a spanner across the terminals of a car battery will appreciate.

- To measure the internal resistance R_{int} of the battery provided, connect it in series with the 1000 ohm variable resistance box and a switch (figure 8).
- With the switch open, measure the open circuit voltage.
- Then close the switch and vary the external resistance: (1) making a plot of R versus V , and (2) also finding the resistance, call it R , at which the measured voltage is *half* the open circuit value. Your circuit is then a potential divider in which half the voltage drop occurs across R_{int} and half across R , hence $R = R_{int}$.

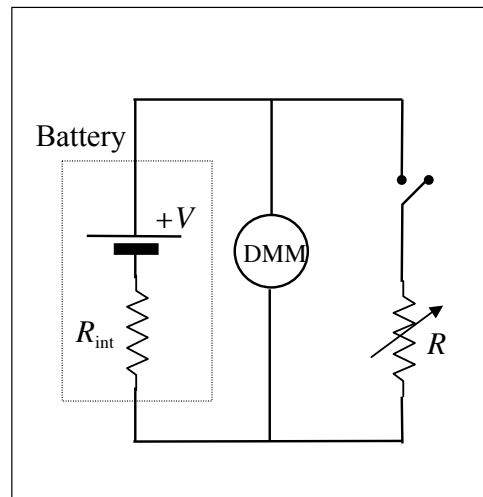


Figure 8 Output resistance

Caution: when the external resistance is low, large currents can flow. Do **not** leave the switch closed for more than a few seconds, otherwise you are likely to discharge the battery or burn out the external resistance.

Equivalent circuits

A theorem due to Thévenin states that a complex circuit of many voltage sources and resistors with two external terminals can be represented by a *single* output resistance R_T in series with a *single* voltage source V_T (provided it behaves linearly, i.e. obeys Ohm's Law). This very simple circuit is the (**Thévenin**) **equivalent circuit**. Another theorem states that the maximum power output of such a circuit occurs when the external, or **load**, resistance is equal to R_T . Given a two-terminal 'black box' containing an array of voltage sources and resistors, you are to measure V_T and R_T (that is, determine the equivalent circuit), and investigate its power output.

- Replace the battery of figure 8 by the 'black box'.
- With the switch open measure V_T , the open circuit voltage.
- Then with the switch closed reduce the load resistance R from 1000 Ω to a few Ω , recording both the potential difference V across R , and the value of V^2/R which is the power dissipated in the load resistance.

Plot the power output versus load resistance on two graphs (by hand – not computer):

- First, using **log-linear** graph paper, plot power (the **dependent variable** which you measure) along the linear axis versus resistance (the **independent variable** which you change) along the log axis to show the behaviour over a wide range of resistances. Use this graph to determine roughly the resistance range of the peak power.

- Second, using ordinary **linear** graph paper, plot an expanded (zoomed in) graph of power versus resistance showing only the region near the peak of the first curve. Take more data in the region of the peak to gain more accuracy. Use this second graph to find a value for the load resistance at maximum power output, and *compare this with R_T* , which as we have seen is the *load resistance when V is $0.5V_T$* .

Laboratory Exercise 2 – LINEAR and NON-LINEAR BEHAVIOUR

Introduction

Most students know something about the behaviour of elastic materials. They know that if a length of wire is pulled it stretches, and that for small forces (or loads, since the force is often applied by hanging a weight) the extension is proportional to the load; when the load is removed the wire reverts to its original length. If the load is too large, the wire becomes permanently stretched, and for sufficiently large loads it breaks.

The statement that extension x is proportional to applied force F , so that a graph of x versus F is a straight line, is Hooke's law. However, this so-called 'law' is nothing like Newton's laws of motion, which hold for any object; it is a description of what Hooke and others found to be the case for a few common substances, particularly metals. But for many materials the 'law' is a poor description of their behaviour under stress. Quite often it *appears* to be obeyed but careful measurements show that a graph of x versus F is not exactly a straight line. Observations of such **non-linear** behaviour can reveal a lot about the molecular structure of the material.

In this exercise you will study the stretching of a rubber band and find a better description of the relation between F and x than the Hooke's law, $F = Kx$. To do this, your measurements of extension have to be much more precise than the nearest millimetre, but to appreciate this fully you will start by measuring x as a function of F using a simple metre rule.

Extension measured with a metre rule

- The rubber band is hung from a thin rod and supports a weight hanger on which weights can be loaded — see figure 1. With a second stand clamp a metre rule so that the position of the bottom of the weight hanger can be read off on the millimetre scale.

- Record the scale readings as you add weights to the weight hanger, in steps of 20 grams up to 200 grams (larger loads have been found to result in permanent stretching of the band). Remember that the weight hanger itself has a mass of 20g. You should be able to estimate the reading to the nearest half-millimetre.

- Plot your results directly (see footnote¹) as mass in grams on the horizontal axis (**abscissa**) versus scale reading on vertical axis (**ordinate**).

- If you have been careful, your readings of the hanger position are probably accurate to within about a half-millimetre. To represent this draw vertical **error bars** on each side of your measured points to indicate that their uncertainty is ± 0.5 mm. The value of the mass is known comparatively accurately (unless you made a mistake in reading the numbers!) so no horizontal error bars are needed.

You will probably find that you can draw a straight line passing close to the points, through the majority of the error bars. You are unlikely to notice any obvious curvature or other systematic deviation from a straight line. In other words, *within the accuracy of your measurements* the behaviour of the rubber band is **linear** — it obeys Hooke's law.

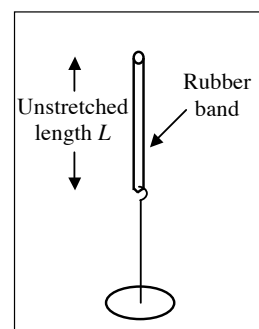


Figure 1 Set-up

¹**Footnote:** The mass m is of course directly proportional to the force mg . Your scale reading is not itself the *extension* since it is measured from some arbitrary position x_0 on the ruler, and it may decrease rather than increase, depending on which way up the ruler is clamped, but these factors will only change $F = Kx$ to $F = K(x_0 - x)$, which is still a linear (straight-line) relation.

Measurement Using a Travelling Microscope

To detect small departures from linear behaviour, you need to measure the extension more accurately. We shall use a **travelling microscope** focused on the weight hanger to follow the stretching of the rubber band, and a **digital** scale to measure the distance moved by the microscope.

- Try using the travelling microscope and learn how to operate it. It is capable of recording to an accuracy of 0.01 mm. Focus the microscope on the hanger so that the horizontal cross-wire is aligned with the bottom of the weight hanger. Check that the microscope can move downwards at least three centimetres, far enough to view the mark even when the band is fully extended. *Be careful not to confuse millimetres and centimetres.*
- Now repeat your measurements of extension versus load, adding mass in increments of 20 grams up to 200 grams. Measure and record the extension as accurately as you can. As you take the measurements, plot them as scale reading versus mass.
- Measure L , the unstretched length of the rubber band by placing the band on a flat surface and put a ruler or pen on top so that it lies flat. Then use the travelling microscope to measure this horizontal length of the band.

What do you find? Probably your data points lie on a gentle curve, not on a straight line at all — if this isn't obvious, hold the graph nearly level with your eye and look along the line of points. With the increased precision of measurement, the elastic behaviour is clearly non-linear.

Further investigation of the non-linear behaviour

If the expression $F = Kx$ is inappropriate, can we find a better description? Suppose the true relation is a power series:

$$F = A\left(\frac{x}{L}\right) + B\left(\frac{x}{L}\right)^2 + C\left(\frac{x}{L}\right)^3 + \dots$$

Note that we have divided the extension x by the original un-stretched length L to give the *fractional* extension, often called the **strain** — this ensures that every power term of x/L is dimensionless so the coefficients A, B, C, \dots all have the units of force. You have seen that the linear 'law', taking just the first term of the series, is quite a good first approximation, so let us investigate the effect of adding just the second (**quadratic**) term. Since x is much less than L this term will be much less than the first, even if the coefficients A and B are similar. Your results probably show that F increases slower than a straight line as the extension x increases, which we can achieve by making the coefficient B negative. If we also try putting A numerically equal to B we have the simplest possible non-linear expression:

$$F = A\left[\left(\frac{x}{L}\right) - \left(\frac{x}{L}\right)^2\right]$$

- Tabulate $(x/L) - (x/L)^2$ and plot this as the ordinate versus the mass m as abscissa. Draw your own conclusions.

Some final algebra

Even more precise investigations have shown that the expression:

$$F = K \left[\left(\frac{S}{L} \right) - \left(\frac{L}{S} \right)^2 \right]$$

is a good description of the elastic behaviour of rubber. Here S stands for the stretched length, $L + x$.

• By expanding this expression as a power series in x/L try to **derive** a series in which the simple quadratic expression you used above appears as the first two terms. You will need to use the binomial theorem in your expansion; it is as follows:

$$(1 - y)^n = 1 - ny + \frac{n(n-1)}{2!} y^2 - \frac{n(n-1)(n-2)}{3!} y^3 + \dots$$

where we must have $y^2 < 1$ and n can be positive or negative.

• How much more accurate would your measurements have to be if you wanted to show experimentally that the next (cubic) term in the expansion was needed? Hint: roughly estimate the size of this term for largest masses you have used.

Laboratory Exercise 3 – NUCLEONIC MEASUREMENTS

Introduction

Experimental techniques in nuclear and elementary particle physics can be extremely complex, requiring expensive apparatus and sophisticated treatment of data. However many of the principles can be illustrated by measurements that use radioactive sources, fairly familiar apparatus, and only simple arithmetic to derive results. This exercise uses a sample of the radioactive isotope cobalt-60 (^{60}Co) and a Geiger counter to demonstrate the absorption of γ -rays; it also illustrates the treatment of errors due to the variability of repeated measurements. Finally, the fact that radioactivity is an inescapable feature of our environment is illustrated by a short study of a natural radiation source that is commercially available.

The operation of a Geiger counter is described in many textbooks, though you do not need to know how it works in order to use it. There is a short description of how a Geiger counter works at the end of this lab script. You will use a standard commercial Geiger tube, the Mullard MX168, which has a thin (and delicate!) end-window made of mica that allows particles of low penetrating power to enter. Ionising particles such as α -, β - and γ -rays, and cosmic rays, will, if they cause sufficient ionisation, give rise to electrical impulses that can be recorded and counted by suitable circuits. The necessary electronics — a power supply of up to 500 volts, and a counter (called a **scaler**) are incorporated in a commercial instrument (either **Griffin** or **Philip-Harris**).

A note on radioactivity

Helium nuclei or electrons (α - or β -particles, respectively) are emitted when an unstable nucleus disintegrates, forming a **daughter** nucleus of a different element. Often the daughter nucleus is left with excess energy. The excess energy is lost via the emission of photons, just as an excited atom loses energy and gives off light. An important difference between the two cases is that the photons (γ -rays) from *nuclear* de-excitation are typically a million times more energetic than those from *atoms*. The isotope ^{60}Co is unstable, undergoing β -decay to yield the daughter nucleus nickel-60 (^{60}Ni) with a half-life of 5 1/4 years. This nucleus has excess energy[†] of 2.5 MeV which it then liberates by emitting γ -rays of energies 1.17 and 1.33 MeV in quick succession.

The source you will use is a tiny quantity of ^{60}Co securely sealed inside a small aluminium rod, which in turn is placed in a small hole in the base of a lead pot. The β -particles from ^{60}Co decay have low energy and are absorbed in the aluminium — only the γ -rays emerge from the top of the pot, and when this is covered by a lead brick even they are absorbed and there is no radiation hazard.

*All radioactive sources must be treated with respect and handled carefully. The ^{60}Co source must **not** be removed from its lead pot, nor should you gaze down at it. The lead brick should only be removed when you are ready to begin your measurements, and replaced when you have finished.*

The strength of a radioactive sample is measured by its **activity**, that is the number of disintegrations per second. This is measured in units called **becquerels** (1 Bq = 1 disintegration per second), although an older unit, the **curie** (1 Ci = 3.7×10^{10} disintegrations per second) is still in common use. Note the huge difference — 1 Bq is an incredibly weak source, while one

[†] **Footnote:** Energies in nuclear physics are conveniently measured in electron volts (eV) or the multiples kilo-eV (keV, 1000 eV) or mega-eV (MeV, 10^6 eV). 1 eV is the energy gained by an electron accelerated through a potential difference of one volt, and is equal to 1.6×10^{-19} joules.

curie is an extremely potent one. The sources used in this exercise had activities of 100 μCi ($100 \times 10^{-6} \text{ Ci}$) when they were produced as much as 15 and 25 years ago; their present activities, three to five half-lives later, have therefore declined to between $(1/2)^3$ and $(1/2)^5$ of their initial values, that is to between 12 and 3 μCi . So you will notice quite a wide variation in count rate, depending on which source is in your lead pot. As mentioned above, each disintegration is accompanied by the emission of two γ -rays. These are emitted in random directions so only a small number will pass out of the lead pot and into the Geiger counter, the rest passing into the lead walls where their energy is absorbed. The number of γ -rays recorded by a Geiger counter placed several centimetres above the pot will be of order ten per second. [Note the use of the phrase ‘of order ...’ or ‘of the order of ...’. It means ‘to the nearest power of ten’, so in this case it is telling you to expect a count rate *greater* than *one* per second but *less* than *one hundred* per second. In physics the phrase is always used in this technical sense.]

Measuring the count rate: the Poisson distribution

If random events such as radioactive decays are counted, the number N counted in a time t is not constant even if each interval is exactly t seconds long. We shall explore how N varies, and we shall see that once this variability is understood we can get, from only a *single* measurement, estimates of *both* the average value of N *and* its uncertainty.

- Before asking the demonstrator for your Cobalt-60 source take a background reading over 100s. Move the Geiger tube well away (at least a metre) from any sources in the laboratory and observe the number of **background** counts due to cosmic rays, natural radioactivity and electronic ‘noise’. Switch on the power supply and turn the **HT** (i.e. high voltage) control fully clockwise to 500 volts.

For the **Griffin** instrument the function switch should be set to **COUNT 10^3** and each measurement timed with a stop watch for 100 seconds. Lift the **RESET** switch on the instrument and release it to start. At the end of the time interval, press the switch down to **HOLD** the reading, and *keep it there* while you record it.

For the **Philip-Harris** instrument the time interval should be set to 100 seconds (10 seconds is too short, 1000 seconds is far too long!) and the slide switch set to **SINGLE READING**. Press **RESET** to start.

- Now read the source safety sheet and then ask a demonstrator for your Cobalt-60 source. Make a note of the source ID. Place the Geiger tube just above the mouth of the lead pot — see figure 1. We want to make repeated measurements of the number of counts in a fixed time, say one second — the procedure for doing this is slightly different for instruments from the two manufacturers:

Griffin: Set the function switch to **RATE 10^3 s^{-1}** .

Philip-Harris: Set **RANGE** to **1 sec.** and slide switch to **CONTINUOUS**.

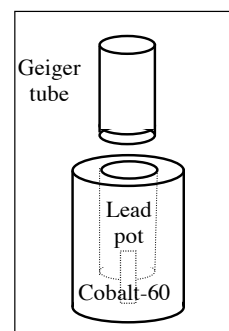


Figure 1 Set-up

- In each case the instrument will automatically record and display the number of counts in a one second interval, updating the display every 2 1/2 seconds. This gives you sufficient time to write down the number before it is updated. Record a large number of counts, say **50**.

- Plot a **histogram** showing your results, grouping them into **bins** of equal width so that the largest bin contains perhaps ten counts, as shown in figure 2. We want to find $\langle N \rangle$, the **mean** of N for a sample of 50 measurements:

$$\langle N \rangle = \frac{\sum N}{50}$$

and also the quantity σ given by:

$$\sigma^2 = \frac{\sum (N - \langle N \rangle)^2}{49}$$

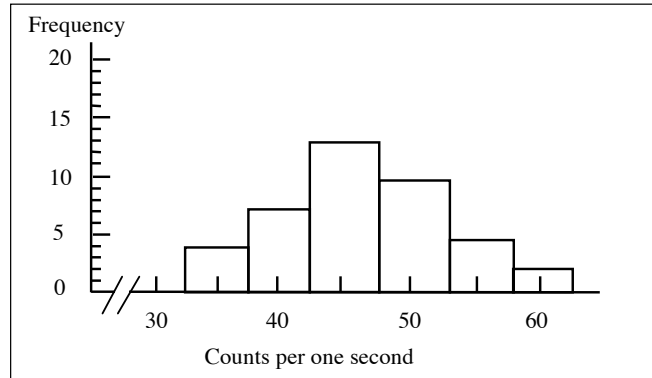


Figure 2 Histogram of number of counts per second

These calculations are tedious! Many electronic calculators are pre-programmed to perform them, and you should either use such a calculator, or even better use the computers in the teaching laboratory which will also prepare your histogram. (PhysPlot will work them out).

As you see, there is a lot of scatter in your results. It is a fact of life that no single measurement can ever be completely relied upon; the best we can do, if asked to measure radioactivity, is to quote $\langle N \rangle$, the **mean value**, and σ , the **standard deviation** which measures the uncertainty in

$$\sigma^2 = \langle N \rangle$$

$\langle N \rangle$. But does this mean we have to take lots of measurements in every case? Fortunately not — a simple result from statistical theory comes to our aid. This states that the distribution of counts shown by your histogram is a very special one called the **Poisson distribution**, which has the extremely useful property that the square of the standard deviation (called the **variance**) is equal to the mean value:

- Check this by comparing your own values of $\langle N \rangle$ and σ^2 . Since you've only taken 50 readings the agreement will not be exact but it should be reasonably good, probably within 20% (if not you have probably made an arithmetic blunder). If we make only one measurement, say N , then **N itself is the best estimate of the true mean value**, and the **square root of N is the best estimate of the true uncertainty**, or experimental error. There is more on means, standard deviations and distributions in the lectures, and in the recommended books.

Absorption of γ -rays in steel

It is found that the absorption of γ -rays in material is roughly exponential, that is the number N emerging from a thickness t is:

$$N = N_0 e^{-\mu t}$$

where N_0 is the number entering the material. The quantity μ , called the **absorption coefficient**, depends on the nature of the material and on the energy of the γ -rays. A knowledge of μ is essential in nuclear technology and medicine. We shall measure μ for steel, using the ^{60}Co γ -rays, which are close enough in energy to allow their average value, 1.25 MeV, to be taken as the energy at which μ is measured.

- Again with the Geiger tube above the lead pot, record the number of counts in equal time intervals, first with no steel interposed and then with successively more and more steel plates covering the mouth of the pot. Add the sheets four at a time until you reach 36 sheets, and count rates more than a factor of ten below where you started. The background count you measured earlier must be **subtracted** from each of your measurements to obtain the **true signal** due to the ^{60}Co γ -rays alone.

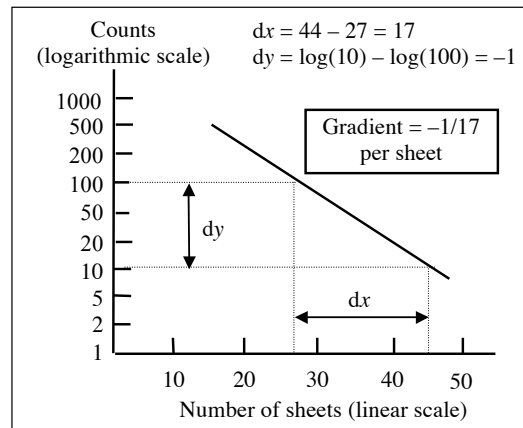


Figure 3 Finding slope of a log graph

By taking natural logarithms (i.e. log base e) of the expression above we obtain:

and a graph of $\ln N$ versus t should contain a set of points that lie in a straight line with a gradient $-\mu$. In numerical work it is more convenient to work with decimal number systems, so

$$\ln N = \ln N_0 - \mu t$$

converting to log base 10 gives:

$$\log N = \log N_0 - 0.4343 \mu t$$

where we have replaced log e by its value, 0.4343. A graph of $\log N$ versus t will have a gradient of -0.4343μ . To avoid the tedium of calculating a logarithm for every point plotted, we use printed graph paper which has one axis graduated with a logarithmic scale (see figure 3). The heavy lines at equal intervals correspond to the numbers 1, 10, 100, 1000, etc. whose logs are 0, 1, 2, 3, etc., and the intervening numbers are shown as lighter rulings more closely spaced as the number increases. You will soon find it very easy to use this **log-linear** (or **semi-log**) graph paper. Figure 3 shows how to calculate the gradient.

- Using log-linear graph paper, **plot** $\log N$ versus number of sheets of steel, **draw** a straight line, and find its **gradient**.
- Then **repeat**, this time using the **computer** to do the plotting, in order to obtain a printed graph and a more reliable value for the gradient.

Note: you may find that the point with *no* sheets is awkwardly high and seems to pull the graph up more steeply than the other points require. This is because some energetic electrons from the β -decay of ^{60}Co penetrate a small amount of steel. Disregard this point if necessary.

- Measure the thickness of several sheets using a **micrometer screw gauge**. Compute the average thickness, and use this to convert your gradient (units: per sheet) to a value for the absorption coefficient μ (units: per metre) at a γ -ray energy of 1.25 MeV.

Natural radioactivity

The laboratory technicians will give you a cotton ‘mantle’ for an incandescent gas lamp, such as those used by campers. The fabric is impregnated with a salt of the element cerium, which has the property of glowing brilliantly when heated. Cerium itself is not radioactive, but is so similar chemically to the radioactive element thorium that the two tend to be associated in nature. The same manufacturing process that concentrates cerium also selects thorium.

- Use the Geiger counter to observe the count rate from the gas mantle, comparing it with that from the ^{60}Co source. This is easiest if you measure both the gas mantle and the source at equal

distances from the Geiger counter. (If not, measure the distances from the Geiger tube to the ^{60}Co and to the mantle.)

- Make a rough estimate of the activity of the gas mantle, using knowledge of the activity of the ^{60}Co . (If the distances were not equal, take that into account using the inverse square law.) The laboratory technicians can tell you the original activity and age of each ^{60}Co source, and so you can work out your source's present activity. Note that for low-level sources it is important to take background into account.

This gas mantle, which was bought in a local High Street shop, is one of the most radioactive items in common use, though certainly not so active that its sale or possession is forbidden by relevant safety regulations. (The same is true of Brazil nuts, which are rich in uranium!)

Geiger counters

Geiger counters (or Geiger-Muller counters) are one of the oldest forms of radiation detectors that have been developed. They were invented in 1928, and modern Geiger counters are in widespread use today. This section gives a very brief overview of how a Geiger counter works. The aim of a Geiger counter is to detect a single particle of radiation. A single particle, such as an electron is difficult to detect with non-specialist modern electronics, the Geiger counter operates in a mode that results in a magnification, or amplification, of the signal to a level that is measurable. The device uses a high voltage signal to establish a high electric field between an anode wire and cathode (the walls of the Geiger tube). A gas fills the volume between the anode and cathode. Typically an inert gas is used to fill a Geiger tube (like Neon), with a small amount of a halogen quenching gas such as chlorine. A signal event occurs when radiation travelling through the gas ionizes one of the molecules. The ion-electron pair are accelerated in opposite directions in the electric field. As the field around the anode is large, the electron signal, which travels toward the anode, causes further ionization of the gas. This mechanism results in what is called an avalanche. Once an avalanche has been produced in a Geiger tube, it is possible that subsequent avalanches will occur until the device has been saturated. When the Geiger tube has been fully saturated it is not possible for any more charge to be released from the gas, and the tube will be insensitive to additional particles of radiation until it has recovered. This recovery period is called dead-time and this is typically of the order of 50-100 μs . The amplification factor achieved using this process is typically of the order of 10^9 to 10^{10} ion pairs. This corresponds to a signal pulse of the order of a few volts, which is straight forward to measure. Figure 4 illustrates three avalanches starting from an individual ionised electron in a Geiger tube.

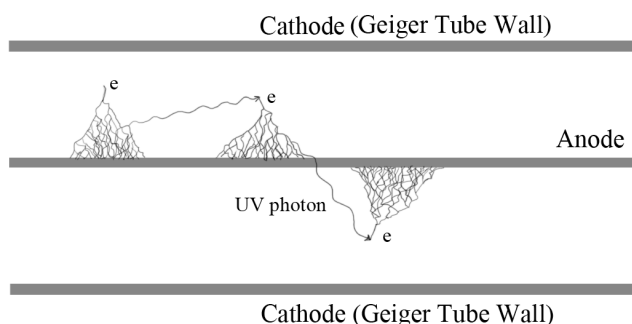


Figure 4: A Schematic of a Geiger tube with three avalanches initiated from a single ionised electron.

Laboratory Exercise 4 – DIGITAL THERMOMETRY

There are two parts to this exercise, which takes two 3-hour laboratory sessions. In part A you make and calibrate a direct-reading digital thermometer usable over the range 0–100 °C. In part B you use this thermometer to investigate the rate at which various liquids cool. There are few instructions provided for part B — instead you are given a copy of a published study of this topic, and asked to repeat the procedures and check the conclusions for yourself.

You must write a **short formal report** on this exercise, using the published paper as a model. **Submit** only the **report**. There is more on what must be included in the report on page 4–4.

Because the calibration of the thermometer is very sensitive to any changes in its electrical circuit, *you should do parts A and B on successive days, Monday/Tuesday or Thursday/Friday*. Your circuit will be left undisturbed on the bench between the two sessions.

Making the thermometer requires you to use and understand some of the electrical circuits covered in experiment 1, which you must have completed before starting this exercise.

Part A: Making the digital thermometer

Introduction

The temperature sensor is a semiconductor diode whose resistance varies with temperature. The diode is used as one arm of a Wheatstone bridge circuit, which will therefore only balance at one temperature. The off-balance voltage is measured with a digital multimeter (DMM) and adjusted with a potential divider to give a direct digital reading, in mV, of the temperature in degrees Celsius. There are three separate tasks: (i) measure the properties of the diode related to its temperature dependence; (ii) set up and adjust the Wheatstone bridge circuit; (iii) use a potential divider to adjust, calibrate and check the performance of your thermometer. You should allocate at most an hour to each of these tasks (including tabulating and plotting data), so as to complete part A in the first afternoon.

Characteristics of the diode

Diodes are electrical devices which allow current to pass in only one direction, the **forward** direction. In the **reverse** direction they have a high resistance, and so can act as one-way switches. The symbol for a diode is \rightarrow with the arrowhead indicating the forward direction, so a voltage applied thus $+\rightarrow-$ causes current to flow; the diode is then said to be **forward biased**. To make a diode, a semiconducting material (silicon in this case) is **doped** with an impurity in order to give a deficit of electrons, hence excess **positive charge**, in one region, and another impurity to give an excess of electrons, hence excess **negative charge**, in an adjacent region. The boundary between the regions is the **junction**. So the diodes you use here are called **silicon p–n junction diodes**. The diodes themselves are small, the size of a match head; the one you use has been encased in insulating mastic with only its two electrical leads left exposed.

The diode does not obey Ohm's law since the current I is *not* proportional to the voltage V . The relation between I and V is called the **characteristic**, and has the theoretical form:

$$I = I_s (e^{eV/kT} - 1)$$

shown in figure 1. Here e is the electronic charge, k is Boltzmann's constant, T is the temperature in degrees Kelvin (i.e. absolute) and I_s is the very small current which flows when the diode is reverse biased. The characteristic is not only strikingly non-linear, much more so

than the very mild non-linearity studied in exercise 2, but it also has an explicit dependence on temperature. If the current is kept constant then a change in temperature will be accompanied by a compensating change in voltage. Thus as T changes the resistance of the diode, that is the ratio $R = V/I$, also changes. You are to measure the characteristic of the diode and find its resistance at a suitable operating current.

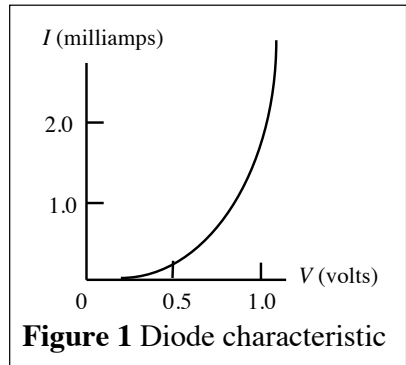


Figure 1 Diode characteristic

- Construct the circuit of figure 2 on the breadboard. The 10 kΩ variable resistor forms a potential divider, allowing you to vary the voltage across the diode.

- Measure and tabulate V and I (up to a few mA) and plot the forward characteristic of the diode.

- When used *later* as a temperature sensor the current should not exceed about 1 mA, in order to avoid excessive heating of the diode. **What voltage, approximately, does this correspond to?**

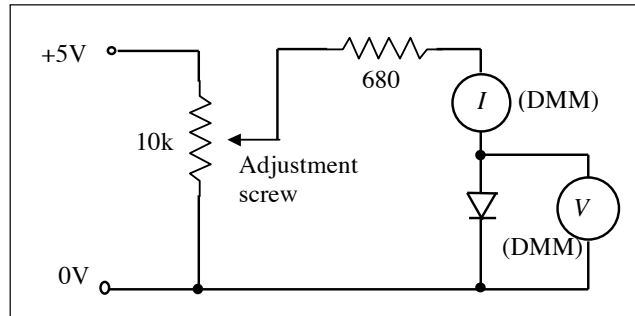


Figure 2 Circuit for measuring diode characteristic

- What is the resistance of the diode at this current and voltage?

- Use the characteristic equation given to **calculate** what voltage change will compensate for a temperature change of 1 °C at room temperature.

The Wheatstone bridge circuit

You may like to review the Wheatstone bridge circuit by referring to the lab script for Experiment 2 before continuing.

- Construct the bridge circuit of figure 3. First check that the 5V supply is stable using the DMM. The balance point of the Wheatstone Bridge will be affected if the power supply drifts. The leads on the small blue multi-turn variable resistor (helipot) are easily broken — do not stretch them!

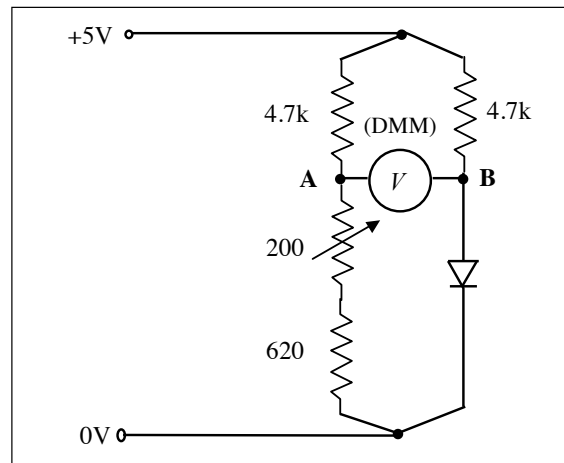


Figure 3 Wheatstone bridge circuit

We can easily place an upper bound on the current that will flow through the two arms of the bridge circuit by assuming that this circuit can be approximated by the two 4.7kΩ resistors. In other words by assuming that the other resistors and diode have zero resistance. The two 4.7 kΩ resistors in parallel present a combined resistance of 2.35 kΩ, allowing a current of no more than $5 \text{ V} / 2350 \Omega$. This corresponds to a current of about 2 mA to flow through the two arms of the bridge. At balance this will be divided equally between the arms, giving the desired 1 mA current through the diode. You should have calculated a diode resistance of several hundred ohms at this current; the 200 Ω helipot is adjusted to this value to balance the bridge.

- We wish to balance the bridge at 0 °C (so as to get a reading of zero at this temperature), so place the diode in a beaker of melting ice. Set the DMM to the 200 mV range, connect it across the outputs **AB** of the bridge, and adjust the 200 Ω helipot until the DMM reads zero. This is a

tricky adjustment, very sensitive to small movements of the helipot. At balance the DMM may still be fluctuating a few tenths of a mV on either side of zero.

- Take the diode out of the ice bath and see whether the DMM voltage increases or decreases as the diode warms up. If it decreases, reverse the meter connections to **A** and **B** so that the digital reading will be positive for temperatures above 0 °C.

Calibration of the thermometer

Your calculation of the voltage change accompanying a 1 °C temperature rise should suggest that at 100 °C the DMM will register well over 100 mV. To get a direct reading of temperature we need to reduce this using a potential divider, preferably one with a high input resistance so that it does not overly disturb the currents flowing in the bridge circuit. A resistance of 10,000 Ω should be sufficient.

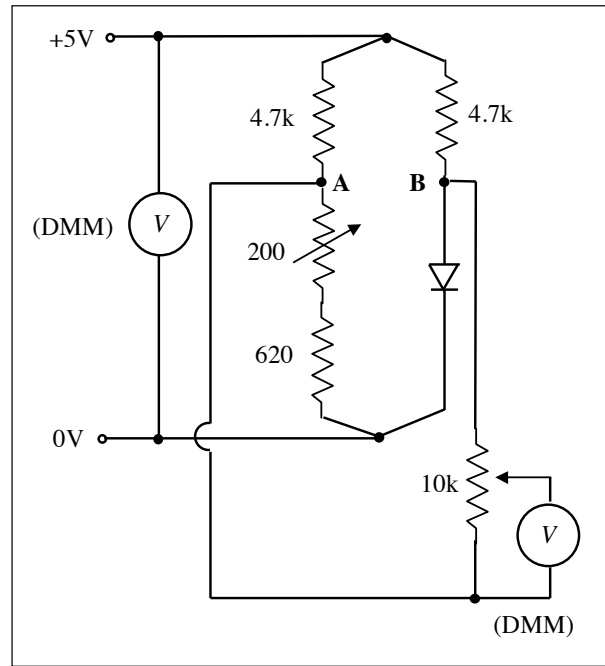


Figure 4 Final circuit for digital thermometer

- Replace the DMM across the output **AB** by the 10 kΩ helipot, and connect the DMM itself across the centre and an outside terminal of the helipot — this is shown in figure 4.
- Place the diode in boiling water and adjust this helipot until the DMM registers 100 mV. Then check that the reading is still zero when the diode is in melting ice, making small adjustments to the 200 Ω helipot if necessary. Repeat the sequence until readings of zero and 100 mV are obtained at 0 °C and 100 °C. The hot water cools quickly so you may only reach a temperature of 80-95°C. In this case you should adjust the helipot such that the DMM registers in mV the temperature of the water in degrees Centigrade. You now have a direct reading digital thermometer!
- Finally, check and correct the calibration against the alcohol-in-glass thermometer. Adjust your digital thermometer when the two thermometers are placed side by side in melting ice and boiling water. Then record the readings when they are both placed in water at some intermediate temperatures, say about 70 °C, 50 °C and 20 °C, using water from the cold tap for the latter. Try to estimate the glass thermometer reading to one-fifth of a degree, and comment in your report on the agreement between the two temperature scales.

Part B: The cooling of coffee

A paper by Rees and Viney on this subject is attached. **Read it before you start this part of the exercise.** Rees and Viney found that black and white coffee cooled at different rates, and sought to explain this. We are not asking you to attempt an explanation, but to repeat some of Rees and Viney's measurements and comment on whether your results agree with theirs and if not, what differences you find. You should be able to make the measurements and draw the graphs in one laboratory period — you may also have time to make some of the additional checks mentioned by Rees and Viney. We suggest that, after checking that your thermometer still records the ice and boiling points correctly, you measure the cooling curves of:

- 200 ml of plain water, brought to the boil and poured into the china mug.
 - Black coffee made by adding 200 ml of water to a level teaspoon of granules in the mug.
 - White coffee made by adding 20 ml of cold milk to a freshly-made mug of black coffee (200 ml again), stirring, and pouring out 20 ml to leave 200 ml. Do not re-boil the coffee.
 - Black coffee made by adding 20 ml of cold water to a freshly-made mug of black coffee (200 ml again), stirring, and pouring out 20 ml to leave 200 ml. Do not re-boil the coffee.
- Can you say **why** the fourth procedure might be informative?
- Use the digital thermometer to record the ambient room temperature.

Your **report** should describe what you did and the conclusions you draw, but it should not be as long as Rees and Viney's — about half the length is sufficient. It **must** be written using a **word processor**, a handwritten version is **not acceptable**. The report **must include**:

- A title.
- Author's name and affiliation.
- An abstract, which is a brief summary of a few lines *including results* but not too detailed.
- A short introduction — what are you describing, and why did you do it?
- A **brief** summary of the theory (you can refer to other publications for this, e.g. 'It is shown by Rees and Viney⁽¹⁾ that ...').¹
- A brief description of your digital thermometer.
- A brief description of what you did and measured.
- A summary of the results, together with calculated quantities. Show raw data as graphs (much more informative than tables), while derived quantities (such as time constants) can conveniently be tabulated. Use *log* rather than linear plots where appropriate.
- A discussion of the significance of the results, including any uncertainties due to measurement precision (errors) and whether or not differences found are meaningful.
- A short conclusion.
- A list of references with name(s) of author(s), journal name and volume number or book title and publisher, page number(s), and date.

¹ Note that it is unacceptable to copy verbatim from the attached paper. You should present your results in your own words. When you refer to the paper by Rees and Viney you should cite that reference appropriately. If you are unsure about this, please ask a demonstrator.

Marks will be deducted for poor spelling (use a spell checker), lack of clarity (proof read and be scientifically objective), and poor presentation. We will go over these points and answer questions about the reports in the lectures.

Remember to *proof read* and *spell check* your report. Look at it with print preview to check that it looks o.k. before you print it. Remember to save your work frequently, and keep a separate (*backup*) copy on your own USB stick. If you need help with word processing ask a demonstrator or laboratory technician. Print your final copy ***well before*** the hand-in deadline, since the printers may be very busy (or down!) just before the deadline!

On cooling tea and coffee

W. G. Rees

Research Fellow, Christ's College, Cambridge CB2 3BU, United Kingdom

C. Viney^{a)}

Research Fellow, Darwin College, Cambridge CB3 9EU, United Kingdom

(Received 22 December 1986; accepted for publication 5 September 1987)

Factors influencing the rate of cooling of hot coffee and tea have been investigated theoretically and studied experimentally using deliberately "domestic" apparatus. It is demonstrated that black coffee cools faster than white coffee under the same conditions. Under most (but not all) circumstances, if coffee is required to be as hot as possible several minutes after its preparation, any milk or cream should be added immediately, rather than just before drinking.

I. INTRODUCTION

You have just made a cup of coffee (or tea), which you intend to drink white. However, you are called away and so prevented from drinking it for several minutes. Assuming that you wish the coffee to be as hot as possible when you return, when should you add the milk—as soon as the coffee is prepared (scheme 1) or just before drinking it (scheme 2)? This question is hardly new, but we present here a simple realistic model and some actual measurements relevant to the problem, and discuss their practical significance.

Two opposing thermodynamic effects operate. When the coffee is hottest it loses heat more rapidly (as predicted by Newton's law of cooling). If this effect operated alone, we should of course add the (cold) milk as soon as possible in order to reduce the total heat loss during the cooling period. However, the addition of the milk cools the coffee directly, and the temperature reduction given by a simple law of mixtures is clearly greater the greater the initial temperature of the coffee. This effect operating in isolation would dictate deferring the addition of the milk to the last possible moment. It is necessary to make a mathematical analysis to determine which effect is more important.

II. THEORETICAL ANALYSIS

Newton's law of cooling states that, at least for small values of the excess temperature of a body relative to its surroundings (ΔT), the rate of cooling of a hot body is proportional to ΔT . The work of Dulong and Petit, and of Langmuir, led to the Lorentz cooling law, in which the rate of cooling is proportional to $(\Delta T)^{5/4}$. There is no contradiction, however, because the cooling law depends on the environmental conditions. The Lorentz law corresponds to the case of natural convection, whereas the Newtonian law corresponds to forced convection. The Newtonian law is

thus more appropriate under normal circumstances¹; this is fortunate, because it is rather easier to integrate than the Lorentzian formula.

Since the heat content of a body (in the absence of changes of phase) is linearly dependent on its temperature, we may express Newton's law as

$$\frac{d(\Delta T)}{dt} = -\frac{\Delta T}{\tau} \quad (1)$$

The solution to this differential equation is

$$\Delta T = \Delta T^{(0)} \exp(-t/\tau), \quad (2)$$

i.e., the temperature relaxes exponentially toward ambient with a characteristic time constant (τ) that depends upon the heat capacity of the body, among other factors.

Suppose that black coffee cools with a time constant τ_B and white coffee with a time constant τ_W . We shall assume these figures to refer to equal volumes of coffee. While in practice the volume of white coffee will frequently be somewhat greater, the effect of this difference is small, as discussed below. We shall further suppose that black coffee and milk have the same volume heat capacity, which will be close to that of water. Again, the effect of this approximation upon our conclusions will be small. With this simplification, the heat content of a volume of either fluid is proportional to its volume multiplied by its temperature (plus a constant).

Let the temperatures of black coffee and milk exceed ambient by ΔT_c and ΔT_m , respectively. If ν volumes of milk are added to 1 volume of black coffee, the resulting temperature excess (ΔT_w) of the white coffee is

$$\Delta T_w = (\Delta T_c + \nu \Delta T_m) / (1 + \nu). \quad (3)$$

According to the first scheme proposed, the milk is added immediately, and the resulting white coffee is allowed to cool for a time t . If the initial temperature excess of the

(black) coffee is $\Delta T_c^{(0)}$, this will be reduced, on adding the milk, to

$$(\Delta T_c^{(0)} + \nu \Delta T_m) / (1 + \nu). \quad (4)$$

After cooling for a time t , this temperature excess will have fallen to

$$\Delta T^{(1)} = [(\Delta T_c^{(0)} + \nu \Delta T_m) \exp(-t/\tau_w)] / (1 + \nu). \quad (5)$$

In the second scheme, the black coffee is first allowed to cool for the time t . Its temperature excess will fall during this period from $\Delta T_c^{(0)}$ to $\Delta T_c^{(0)} \exp(-t/\tau_B)$. Upon adding the milk, which we shall assume has been maintained at the excess temperature ΔT_m , the temperature of the mixture will fall to

$$\Delta T^{(2)} = [\Delta T_c^{(0)} \exp(-t/\tau_B) + \nu \Delta T_m] / (1 + \nu). \quad (6)$$

The problem thus resolves itself into that of finding which of $\Delta T^{(1)}$ and $\Delta T^{(2)}$ is greater. The usual solution involves assuming that $\tau_B = \tau_w = \tau_0$. In this case, it may readily be shown that

$$\Delta T^{(1)} / \Delta T^{(2)} = (1 + \nu \Delta T_m / \Delta T_c^{(0)}) / [1 + (\nu \Delta T_m / \Delta T_c^{(0)}) \exp(t/\tau_0)]. \quad (7)$$

If $\Delta T_m > 0$ (i.e., the milk is above ambient temperature), this expression will be less than unity and so scheme (2) *must* result in hotter coffee. If, on the other hand, $\Delta T_m < 0$ (as it will be if the milk is kept in a refrigerator), the expression will be greater than unity and scheme (1) results in hotter coffee. Indeed, this seems intuitive—if the milk is warm, add it later, and if it is cold add it straight away. However, if we now consider the case of $\tau_B < \tau_w$, which as we shall see in Sec. III is the case in practice, the situation becomes more complicated. The ratio $\Delta T^{(1)} / \Delta T^{(2)}$ is now

$$\Delta T^{(1)} / \Delta T^{(2)} = [(1 + \nu \Delta T_m / \Delta T_c) \exp(t/\tau_B - t/\tau_w)] / [1 + (\nu \Delta T_m / \Delta T_c) \exp(t/\tau_B)]. \quad (8)$$

If $\Delta T_m < 0$, this expression will always be greater than unity, and scheme (1) again wins. On the other hand, if $\Delta T_m > 0$, the better scheme will depend on the value of $(\Delta T_c / \nu \Delta T_m)$, on the time t , and on the two time constants.

When $\Delta T^{(1)} = \Delta T^{(2)}$, the above expression can be rearranged to give

$$\begin{aligned} \Delta T_c / \nu \Delta T_m &= [1 - \exp(-t/\tau_w)] / \\ &[\exp(-t/\tau_w) - \exp(-t/\tau_B)] \\ &= f(t). \end{aligned} \quad (9)$$

It therefore follows that scheme (1) produces hotter coffee so long as $(\Delta T_c / \nu \Delta T_m) > f(t)$, otherwise scheme (2) is better. Thus, in the realistic case of $\tau_B < \tau_w$, if the milk is cold it is still best to put it in straight away but, if it is warmer than the ambient temperature, the best time to put it in depends on the time for which the coffee is to be left to cool.

III. EXPERIMENTAL OBSERVATIONS

In order to investigate the aptness of the foregoing analysis, and to try to gain a general understanding of the significant influences on coffee cooling, we have made various measurements of the cooling of black coffee, white coffee, and a number of other liquids. The experimental hardware

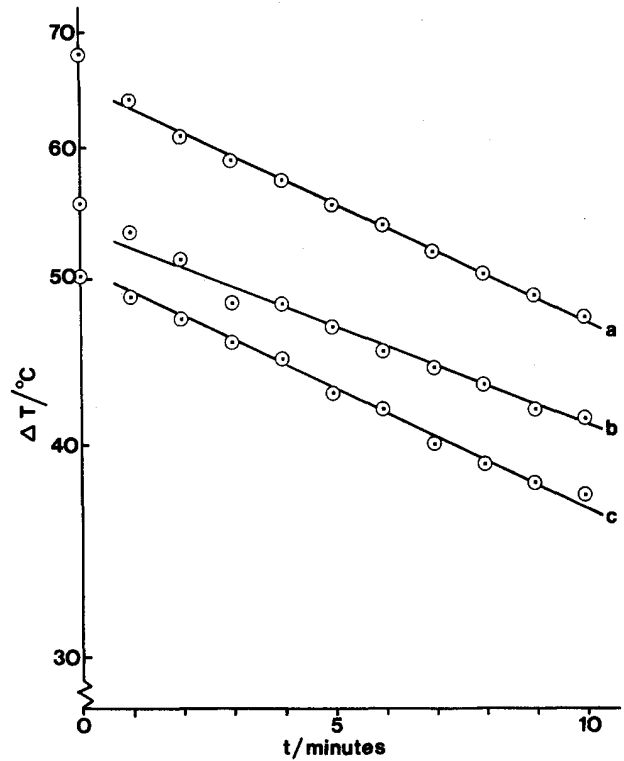


Fig. 1. Typical cooling curves, plotted as $(\log \Delta T)$ vs time: (a) black coffee ($\tau_B = 32$ min); (b) white coffee ($\tau_w = 38$ min); (c) hot water, under different insulation conditions ($\tau = 33$ min). The straight lines were drawn by least-squares regression, ignoring the first data point in each case.

consisted of a cylindrical glazed china mug, predominantly white in color, resting on sufficient insulation that heat loss through the base could be neglected. The mug had an inside diameter of 70 mm and a capacity of 320 ml, although it was normally filled with only 250 ml of liquid. Temperatures were measured using a standard -10 to $+110$ °C mercury-in-glass thermometer, which we estimated we could read to an accuracy of 0.3 °C. The thermometer rested with its bulb on the bottom of the mug during the cooling process, although immediately before making each measurement (once a minute) it was raised to the middle of the mug, used for stirring briefly and gently, read, and replaced. Our observations suggested that the precise method by which the temperature measurements were made did not significantly affect the deduced value of the cooling time.

The black coffee used in this work was prepared using a leading brand of instant coffee granules. The volume ratio ν of milk (ordinary pasteurized and homogenized cow's milk) added was 0.1; i.e., the white coffee consisted of 227 ml black coffee, with 23 ml milk stirred into it. Observations were made of the temperature excess ΔT , as a function of the time t , for a period of 20 min. Figures 1(a) and (b) show the results for black and white coffee during the first 10 min, plotted on a log-linear graph for identical experimental conditions. It can be seen that Newton's law of cooling is obeyed; indeed, within the experimental accuracy of the observations, we found no deviation from the predicted exponential decrease throughout the whole 20-min period, apart from a slightly accelerated cooling rate during the first minute. The cooling times τ were calculated,

using a least-squares regression analysis and ignoring the first minute's data in each case, to be 31.7 ± 0.3 min for black coffee and 38.3 ± 1.5 min for white coffee. This observed difference of 7 ± 1 min—i.e., nearly 20%—was found to be reproducible when changes were made to the thermal insulation of the mug.

We attempted further investigation to try to explain this difference in the cooling rate. The hypotheses that we attempted to test were that the difference might be caused by slight inaccuracies in measuring out the quantity of liquid; by differences in the blackbody radiative efficiencies; and by a difference in viscosity (which would influence the internal convection rate).

To test the theory that the cooling rates were influenced by the total volume of liquid, we performed similar experiments with hot water, first filling a mug (250 ml) and then half filling it. The cooling times were found to be 40 ± 1.5 min and 28.0 ± 0.7 min, respectively. This difference is not surprising in view of the much larger surface-to-volume ratio in the latter case, but it suggests that a variation of a couple of percent in the volume of liquid contained in the mug would cause no greater variation in the cooling time. This is evidently inadequate to explain the measured difference of about 20% between black and white coffee.

To investigate the possibility that the differences in cooling rates might be caused by differences in blackbody radiative efficiency, we repeated the original experiment with hemispherical domes of black paper or of aluminum foil surrounding the mugs. These domes had radii of about 150 mm, and had small holes at the top to allow the thermometer to pass through. They were shaped around a large mixing bowl, to ensure a consistent size. If the difference in cooling rate is attributable to greater blackbody radiation from black coffee, the dome of aluminum foil should have entirely removed the difference (since aluminum reflects practically all of the radiation incident upon it near the wavelength of maximum radiation, about $8 \mu\text{m}$ for a body at 70°C), whereas the dome of black paper (an absorber) should have accentuated it. In fact, we found cooling times of 50.1 ± 0.7 min for black coffee and 56.3 ± 0.8 min for white coffee with the aluminum dome, and 41.2 ± 0.5 and 48.6 ± 0.6 min, respectively, for black and white coffee with the paper dome. These figures suggest that the *difference* in cooling time is virtually independent of the radiation loss, although clearly radiation is a significant influence since the cooling times were longer under the aluminum dome than under the paper dome. The fact that all of these cooling times were longer than in the absence of any dome may be attributed to the retention of warm air by the domes.

To assess the likely effect of viscosity, we measured the cooling time for "Golden Syrup" (molasses). This is highly viscous, even at 90°C . Although the measured cooling time, 55 ± 3 min, was indeed longer than observed for other liquids, the change did not appear to be great enough for the difference between black and white coffee to be ascribed to (small) viscosity differences.

Finally, although this has no direct bearing on the *difference* between the cooling rates of black and white coffee, we investigated the effect of a draught of air over the surface of the cooling liquid. The draughts were provided by small electric fans and, even though precautions were taken to ensure that the mug and the surface of the liquid itself were protected from the flow of air, the effect was found to be dramatic. For coffee (black or white) or water, the cooling

time was reduced by about 40% for a draught of about 3 m/s, and by about 60% for a draught of about 10 m/s.

IV. DISCUSSION

The principal observation made in this work is that black coffee cools significantly faster than white coffee, by about 20% under normal conditions. Various influences on the cooling rate have been investigated.

A. Factors that affect black and white coffee equally

Not surprisingly, the cooling time was found to be approximately proportional to the ratio of volume to total surface area of the liquid, other things being equal. Thus, for a cylindrical container of (constant) radius r filled to a depth h ,

$$\tau \propto h / (h + r). \quad (10)$$

If $h = 2r$ (approximately true in practice),

$$\Delta\tau/\tau \sim \Delta h / 3h. \quad (11)$$

Thus the fact that a cup of white coffee is typically about 10% more full than the cup of black coffee from which it is prepared—although the two volumes were practically equal in our experiments—suggests that the white coffee should cool only about 3% slower. This predicted effect is small compared to the 20% difference in cooling times that we found for equal volumes of black and white coffee. Even if the suggested variation τ is proportional to volume-to-surface ratio is not strictly correct, the result (11) above should not be significantly in error.

The effect of varying the insulation below the mug (in the form of poorly conducting mats, piles of paper, and so on) was to change cooling times by no more than 5%. We took care always to compare cooling rates under the same insulation conditions; in any case, it is clear that differences in insulation would be unlikely to mask the intrinsic difference in rates.

Finally, we note that, of all effects likely to act equally on either liquid, that of a draught of air is greatest. A draught of only a few meters per second was sufficient to reduce the cooling time by 40%, and this is clearly sufficient to mask the intrinsic difference between black and white coffee. For this reason, all of our comparative experiments were performed in still air.

B. Factors that might affect black and white coffee differently

Our experimental investigation of the effect of blackbody radiation suggests that there is virtually no difference in radiative efficiency between the two liquids. Indeed, this is consistent with the fact that black and white pigments used in paints have very similar (and high) thermal emissivities. In other words, the color, which is, of course, determined by emissivity at visible wavelengths, bears no simple relationship to the total emissivity.² We might expect a similar relationship to hold for coffee.³

It also seems unlikely that a difference in viscosity (which influences the internal convection) can account for the difference in cooling times, since the viscosities of black and white coffee are not markedly different; also syrup (molasses), which has a viscosity several orders of magnitude greater, cools only twice as slowly.

The mechanism that we conjecture to be responsible for

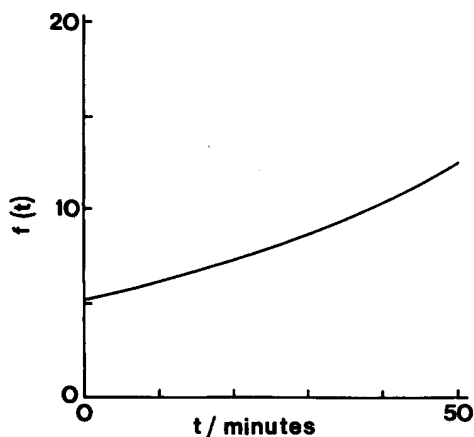


Fig. 2. Plot of $f(t)$, assuming $\tau_B = 32$ min and $\tau_W = 38$ min.

the difference in cooling times is a reduction in evaporation rate when milk is present in the coffee, but we were unable to perform a suitable experiment, with our deliberately “domestic” approach, to confirm this.

C. When to add the milk

Using the theory developed in Sec. II, and our measured values of $\tau_B = 32$ min and $\tau_W = 38$ min, we may evaluate $f(t)$. The result is shown in Fig. 2. If we assume that $\Delta T_c = 70^\circ\text{C}$, $\Delta T_m = 30^\circ\text{C}$, and $\nu = 0.2$, then $(\Delta T_c / \nu \Delta T_m) = 11.7$, for which value of $f(t)$ the time required is about 46.5 min. Thus the resultant coffee will be hotter if the milk is put in last only if it is left to cool for more than 46.5 min. However, the coffee will then be rather cool, at only 18°C above ambient.

If we assume that the longest time for which one might wish to leave the coffee is 10 min, then the critical value of $\Delta T_c / \nu \Delta T_m$ is about 6 (from Fig. 2). Thus if $\Delta T_c = 70^\circ\text{C}$, it can be seen that the effect of the difference between the two time constants will only be important if ν is large. For example, if $\nu = 0.5$, milk, which is less than 23°C above ambient, should be put in immediately whereas warmer milk should be put in as late as possible. This is illustrated graphically in Fig. 3. If the excess temperature of the milk is 35°C , the difference in drinking temperature amounts to nearly 1°C .

D. Factors omitted from our analysis

The analysis presented here omits a number of points. Walker⁴ has drawn attention to the cooling produced by the dissolution of sugar (for those who take it), and by vigorous stirring, which brings cool liquid to the surface faster than convection alone would do. He also points to the effect of leaving a metal spoon in the mug—which will absorb, radiate, and conduct heat from the coffee—and of the color of the mug itself. He has also investigated the effect of stirring cream into hot water, and of floating (whipped) cream on the surface.⁵ He plots temperature as a function of time for these liquids but, as he does not quote the ambient temperature, it is not possible to make a precise estimate of the cooling times. However, if the ambient temperature is estimated as 20°C , his data give cooling times of

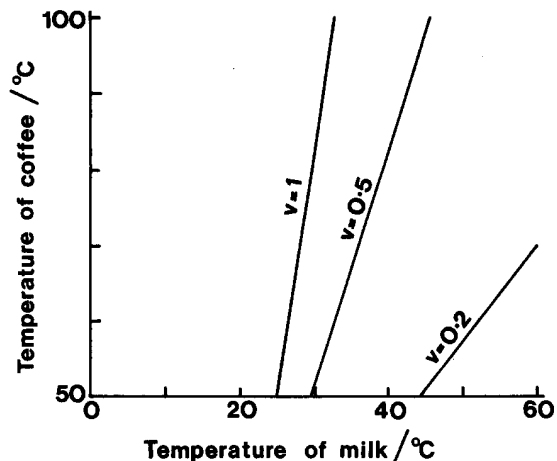


Fig. 3. Figure illustrating the best time to add the milk to the coffee, using the curve of $f(t)$ in Fig. 2 and assuming an ambient temperature of 20°C and that the coffee is left to cool for 10 min. Here, ν is the relative volume of milk added to 1 volume of black coffee. If the point representing the temperatures of the coffee and milk lies to the left of the relevant line, the milk should be put in immediately; if to the right, at the end of the 10-min cooling period.

30 min for hot water, 25 min for hot water with cream stirred in, and 43 min for hot water with the cream floating on top. The difference between the last two figures is presumably accounted for by the insulating effect of the layer of cream, as well as the prevention of evaporation. Smith¹ has analyzed the behavior of such a ‘Gaelic coffee’ system and demonstrated that it indeed invariably results in significantly hotter coffee than is produced by stirring in the cream.

V. CONCLUSIONS

The most significant conclusion to follow from this work is that black coffee cools faster than an equal volume of white coffee, by about 20%. The cooling in both cases obeys Newton’s law. This difference is not contributed to significantly by a difference in blackbody radiative efficiency and we suspect that the dominant factor is that of milk reducing the rate of evaporation.

If black coffee is prepared in a mug and is to be drunk white and as hot as possible some minutes later, the optimum time at which the milk should be added depends on its temperature. If the initial milk temperature is below ambient, it should be added straight away. If it is initially above ambient, the time depends on the volume fraction of milk to be added, on its excess temperature, and on the total time interval between making the coffee and drinking it.

This work has also illustrated cooling rates that result from a partially filled mug, a draught, and the absence of good thermal insulation below the mug.

^{a1} Present address: Department of Materials Science and Engineering, University of Washington, Seattle, Washington 98195.

¹A. C. Smith, *J. Naval Sci.* **6**, 184 (1980).

²V. M. Faires, *Thermodynamics* (Macmillan, New York, 1962).

³C. Viney, *Snippets No. 9* (Institute of Physics, London, 1985), p. 8.

⁴J. Walker, *The Flying Circus of Physics* (Wiley, New York, 1977).

⁵J. Walker, *Sci. Am.* **237**, 152 (1977).

Laboratory Exercise 5 – THE OSCILLOSCOPE

Introduction

The aim of this exercise is to introduce you to the oscilloscope (often just called a 'scope), the most versatile and ubiquitous laboratory measuring instrument. The oscilloscope is used to display and analyse electrical signals, either repetitive waveforms or transient pulses, which are changing too fast to be recorded by simple analogue meters like the AVO or by digital instruments such as the DMM. For many years oscilloscopes used *cathode-ray tubes* and *analogue* circuitry that swept a narrow beam of electrons across a fluorescent screen, as in a TV. These cathode-ray oscilloscopes, or CROs, use the signal to be analysed to control vertical movement of the electron beam. This produces a display that is effectively a graph of voltage (vertically) versus time (horizontally).

Analogue 'scopes need repetitious signals to produce the display. However, like so many electronic devices they are increasingly being replaced by *digital* 'scopes that can capture even a single pulse and display it from a memory. The use of computerised digital electronics also makes it possible for the display to be much more versatile and informative, with sophisticated mathematical treatment of the digital data available if needed. The use of thin liquid-crystal displays rather than bulky cathode-ray tubes makes 'scopes much more compact and lighter, and also allows inexpensive use of colour for the displays.

The horizontal axis of the display usually represents time, but sometimes it is useful for it to be controlled instead by a second time-varying voltage. We will not investigate this so-called *XY* mode. Note that instructions for the oscilloscope can be found at the end of this section.

Main features

Most 'scopes have similar basic features. They differ mainly in the speed of signals that they can handle, the number of parameters you can adjust, and the additional facilities that are offered. Because the data in modern 'scopes is digital, it can be stored and transferred to PCs by various methods for printing or adding to documents; indeed many of the more expensive digital 'scopes actually *are* PCs 'underneath', and this for example allows access to their data via the internet.

Digital 'scopes such as those in our laboratory can display signals from a few millivolts to a hundred volts, at frequencies up to about 60 MHz. More expensive models can work at up to a few GHz (1 GHz = 10^9 Hz). You do not need any detailed knowledge of the internal circuitry of an oscilloscope in order to use it effectively, but the main functions need explanation. We have provided separate notes describing how to do the things required for this experiment. However, our 'scopes also have clear instruction manuals, as well as built-in help (in a wide variety of languages!) on-screen via the **help** button.

The *horizontal* axis is controlled by a **timebase** circuit which drives the display in the horizontal (*X*) direction at a constant rate, adjustable by a front-panel knob at the right. Selection of the *XY* display mode (see below) turns the timebase off and allows the display to be driven horizontally by one of the input voltage signals. Note, oscilloscopes cannot be used as ammeters directly.

Most oscilloscopes can display two or more signals, or **channels**. The signals are applied at front-panel sockets to internal **amplifiers** whose gain is adjustable by knobs; these are used to give convenient *vertical* deflections on the screen. Thus the 'scope is also like voltmeter. The signal displays can be moved up and down, even off the screen. You can display either or both traces, and if there are two signals then they can also be added or subtracted algebraically if desired to produce the display.

The amplifiers must have a stable gain at low and high frequencies if the 'scope is to be used for measuring voltages accurately. Each input channel has three modes for connection: AC, DC, and GROUND. The GROUND selection connects the channel to earth, allowing you to see the position of zero volts input. The AC setting introduces a capacitor in series with the input and is used to eliminate any DC component, so that a small time-varying signal can be displayed in the presence of a large DC offset without driving the trace off-screen. The DC setting should be used for low frequency signals. For high frequencies, the 60 MHz limit of our models limits them to signals longer than about 20 ns.

The third important circuit handles **triggering**. Its task is to synchronise the timebase with the arrival of a repetitive signal, so that at every sweep this appears in the same position on the screen. This has many complex features but we shall use only the simplest ones.

The display is **triggered** whenever the trigger signal exceeds a voltage that you can set. The source of this trigger signal can be either the channel 1 or channel 2 input signal, or a separate EXTERNAL signal fed into an adjacent socket. Every 'scope has an AUTOMATIC trigger mode, displaying a timebase even when no trigger signal is present so that you can see what might be there. NORMAL mode requires a signal to trigger the timebase. Finally, you can choose the slope (increasing or decreasing voltage) and the magnitude of the triggering signal.

Part A: Getting familiar with the 'scope and function generator

Displaying simple waveforms

This is a short exercise to familiarise you with basic use of the 'scope. Sinusoidal, square, and triangular waves are available from the function generator; their frequencies can be varied from 0.03 Hz to 3 MHz and their amplitudes adjusted up to about 20 volts **peak-to-peak**, i.e. from maximum positive (+10 V) to maximum negative (-10 V) with respect to earth.

- Using a breadboard connect the output of the function generator to the channel-1 input of the 'scope, making sure that the signal (red) leads, and earth are connected. Select sine waves — the square wave button should be out. Adjust the 'scope controls to obtain a stable waveform on the screen.
- Now try adjusting the controls on the oscilloscope and function generator to see what they do¹. Vary the frequency by factors of 10 between 10 Hz and 1 MHz, and compare the indicated frequencies on the function generator dial with the exact frequencies measured automatically by the 'scope. Do the respective calibrations agree over the whole frequency range? With care you should be able to check this to an accuracy limited only by the frequency dial on the function generator. You should compare what you see on the 'scope to the digital meter on the function generator.
- Also, at one frequency learn how to use the 'scope's **cursor** features. Use the horizontal cursors to measure the period (figure 1) of the waveform, convert this to frequency, and compare with the 'scope's automatic measurement and the function generator's digital meter measurement. Then compare the amplitude as measured by the vertical cursors, and compare

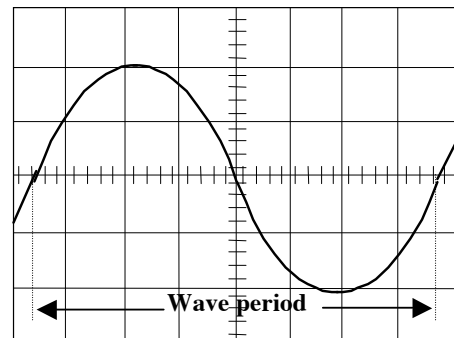


Figure 1 Sine wave

¹ Don't be afraid of *trying controls to see what they do* — this is really the only way to learn how to use a 'scope. Some of the buttons, knobs and menus will take a while to get the knack of, but once the general mode of operation is understood things will start to seem much easier.

with the automatic measurement. (The cursors are mainly meant for measuring the spacing of features that are *not* measured automatically by the 'scope.)

- Next, compare the calibration of the amplitude knob of the function generator with the amplitude measured by the 'scope (that of the oscillator is not intended to be very precise).
- Switch the trigger mode to NORMAL triggering and change the trigger amplitude. Study what happens when you change the trigger level or signal amplitude.

Comparing frequencies: direct display of waveforms

- An 'unknown' sine wave from an oscillator is available on the signal circuit terminals above your bench. Feed this signal (care with signal and earth connections again!) to channel 2. Trigger the timebase on this 'unknown' signal and, viewing both signals together on the screen, adjust your oscillator to match the 'unknown', getting as stationary a display as possible. Use the 'scope to measure the amplitude and the frequency of the 'unknown' waveform.

Part B: RC circuits

In exercise 2 you measured the input resistance of a DMM. In this exercise we show another way to do it which leads on to a demonstration, using the oscilloscope, of the effect of **capacitance** on electrical signals, and some more about the relative **phase** of wave trains.

A capacitor is a device that stores electric charge. Suppose a capacitor is charged, as in figure 2(a), by momentarily closing a switch connected to a battery. When the switch is opened again there is nowhere for the charge to go so it stays on the two plates of the capacitor, maintaining a voltage difference V between them. The capacitance C is the charge stored divided by the voltage: $C = Q/V$. But if a resistance R is placed across the capacitor, as shown in figure 2(b), the charge can leak through R from one plate to the other, discharging the capacitor. The larger R the slower the leak, and the larger C the more charge there is to transfer. So it is not surprising that both the charge on the capacitor and the voltage across it decrease with time t according to the exponential law:

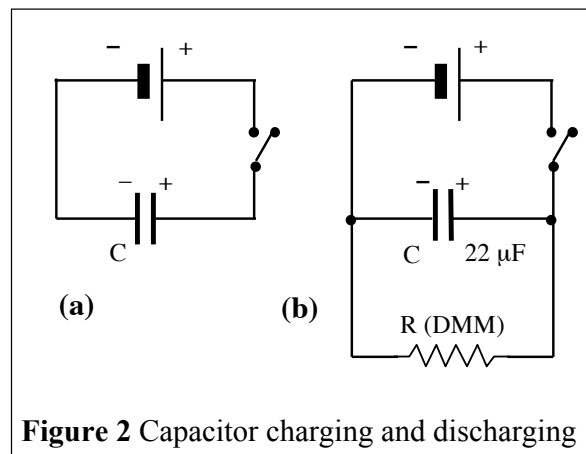


Figure 2 Capacitor charging and discharging

$$Q = Q_0 e^{-t/RC} \quad \text{and} \quad V = V_0 e^{-t/RC}$$

where the product RC , the **time constant**, is a measure of how rapidly charge and voltage decrease. We'll use this to measure the input resistance of a DMM.

Input resistance of a DMM

- On the breadboard make the simple circuit of figure 2(b), using a $22 \mu\text{F}$ capacitor and using the DMM as the resistor, set to measure voltage. Use a desk-top power supply connected to the breadboard with wires. The value of R is then the input resistance of the meter. The capacitor *must* be connected the right way round — the polarity is marked on it by a vertical line indicating +.
- Depress the switch and make contact for a few seconds, long enough to charge the capacitor up through the internal resistance of the battery. Release the switch and immediately start to

record the voltage as a function of time, every 15 seconds initially and then every minute after two or three minutes, until the voltage has fallen to less than a tenth of its initial value.

- Plot $\log V$ versus t and deduce a value for R from the slope of your graph (see exercise 4 to remind yourself how to do this). Comment on any unusual features of your graph.

Using the oscilloscope to measure time constants

The oscilloscope can be used to measure time constants as short as microseconds or less.

- Connect up the RC circuit of figure 3, and apply a square-wave of frequency about 50 kHz. Use the 'scope to observe, on channel 1, the voltage applied to the capacitor and, on channel 2, the current that flows into or out of the capacitor. Remember, the 'scope does not measure current directly, so instead measure the voltage across a 1 k Ω resistor. Just as previously, the positive-going and negative-going edges of the waves charge up the capacitor, which then discharges through the resistor R (the parallel input resistance of the 'scope is much too large to have any effect).

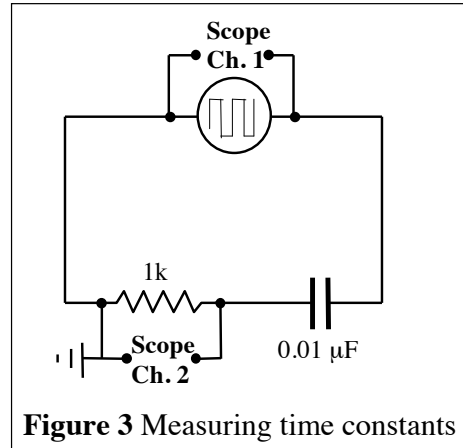


Figure 3 Measuring time constants

- Measure the time $t_{1/2}$ for this current, measured as the voltage across R , to fall to half its peak value. By taking logs of $V_0/2 = V_0 \exp(-t_{1/2}/RC)$ we see that $t_{1/2} = \log_2(RC)$. Does the value of RC obtained in this way agree with simply multiplying the known values of $R \times C$?

The RC circuit as a frequency filter

- High Pass Filter

- Raise the oscillator frequency to about 1 Mhz and notice that there is now too little time for the capacitor to discharge appreciably, and so the very short-period waves pass across unhindered. Switch to sine waves, and the input and output wave trains should still look the same.

- Now reduce the frequency; as you do so two effects will begin to appear: (a) The *amplitude of the output decreases*, becoming zero at very low frequencies since the capacitor remains completely discharged, with virtually no current flow, when the applied voltage is changing sufficiently slowly. For this reason the RC circuit is called a **high pass filter**, letting through only high frequencies. The frequency at which the output amplitude falls to $1/\sqrt{2}$ of its high frequency value is by convention called the **cut-off frequency**. It can be shown that the cut-off frequency is $1/(2\pi RC)$. (b) The *output signal begins to lag behind the input*. The capacitor causes a phase shift between voltage and current. Since this is a series circuit, the current must necessarily be the same everywhere in the circuit. Therefore the voltage across the capacitor will lag that current by at most 90° at very low

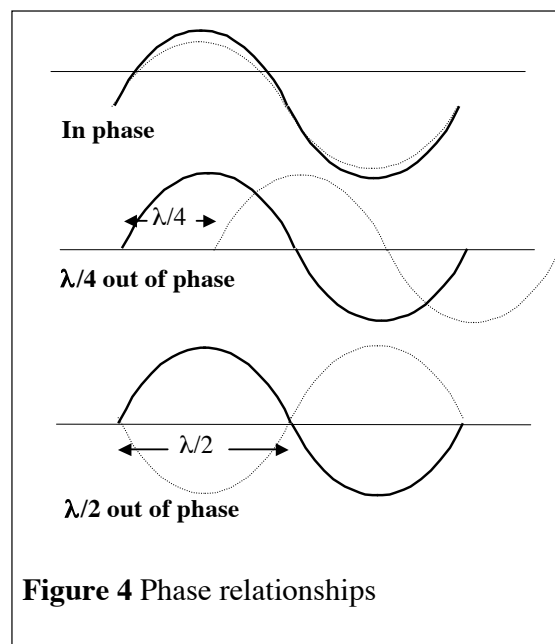


Figure 4 Phase relationships

frequencies, while at the same time the voltage across the resistor will be in phase with the current. This lag is called a **phase shift** (figure 4).

- Make a plot of the voltage across Ch. 2, the resistor, as a function of the logarithm of the frequency f : $\log f$.
- Measure the cut-off frequency and compare it with $1/(2\pi RC)$. Measure the phase shift at this frequency, converting your result to degrees.

- Low Pass Filter

Similarly to the high pass filter, you will now study the **low pass filter**, which lets through only low frequencies. Swap the resistor with the capacitor in your circuit.

- Make a plot of the voltage across Ch. 2, the capacitor, as a function of the logarithm of the frequency f : $\log f$.
- Measure the cut-off frequency and compare it with $1/(2\pi RC)$. Measure the phase shift at this frequency, converting your result to degrees.

Part C: RLC circuits

An **RLC circuit** consists of a resistor (R), inductor (L) and a capacitor (C) connected in series or in parallel.

An inductor is a circuit element consisting of a coil of wire on a core material made of ferrous or non-ferrous material. An inductor resists changes in the flow of electric current through it, because it generates a magnetic field that acts to oppose the flow of current through it, which means that the current cannot change instantaneously in the inductor. This property makes inductors very useful for filtering out residual ripple in a power supply, or for use in signal shaping filters. They are frequency-dependent devices, which means that their inductive reactance, or "effective resistance" to AC decreases as the frequency gets lower, and increases as the frequency gets higher. This property makes them useful in tone controls and other filters. Note that this is opposite to what happens with the capacitors.

RLC circuits can be used to select a certain narrow range of frequencies from the total spectrum of waves. There are two fundamental parameters that describe the behaviour of *RLC circuits*: the **resonant frequency** and the **damping factor**. We define ω as the angular frequency ($\omega = 2\pi f$), where f is the frequency. The resonant angular frequency ω_0 is given by $\omega_0 = 1/\sqrt{LC}$. The damping factor $\Delta\omega$ is given by $\Delta\omega = R/2L$ for the circuit considered in the following exercise.

Connect up a circuit as in figure 5 and apply sine waves of frequency f .

The self resistance of the inductor should be used in series with the resistor. However, we will assume that the resistance of the inductor is so small with respect to the resistance in the circuit to be neglected.

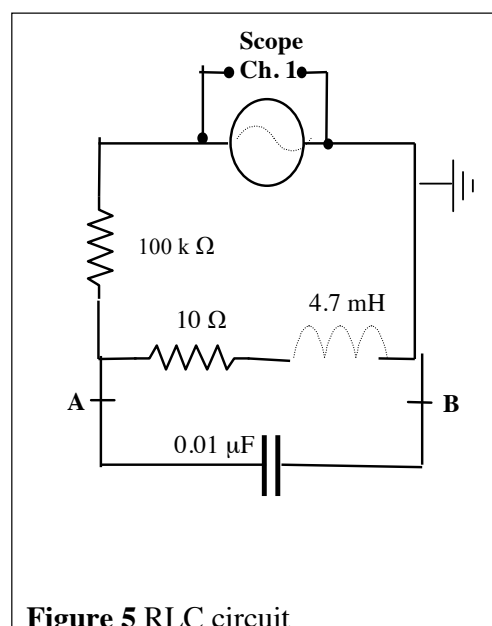


Figure 5 RLC circuit

- Make a plot of the variation of the voltage across A and B versus the angular frequency ω .

- Measure the maximum ω_0 of the distribution and compare it to $1/\sqrt{LC}$.
- For $\omega_0 \pm \Delta\omega$ the measured voltage is $1/\sqrt{2}$ of the output voltage. Measure $\Delta\omega$ and compare it to $R/2L$.

Change the sine waves to square waves of angular frequency ω_0 . You should still observe a sine wave across A and B. Vary the frequency and note that you still observe a *good* sine wave for frequencies in the range $\omega_0 \pm \Delta\omega$ (**narrow pass filter**).

- Store and print the voltage across A and B for $\omega=\omega_0$, $\omega=\omega_0+\Delta\omega$ and $\omega=\omega_0-\Delta\omega$. This experiment shows that a square wave can be represented as a sum of sine and cosine waves (Fourier transforms).

OSCILLOSCOPE QUICK-START INSTRUCTIONS

Introduction

Our digital oscilloscopes have good **instruction manuals**, and how to do something can usually be found by using the index. They also have, at the press of a button, **on-screen help** that is ‘context-sensitive’, i.e. related to the mode the ‘scope is in. Furthermore, on-screen messages and labels, as well as the help information, are available in a wide variety of languages. However, in order to get you going quickly, we have written here some **simple instructions** that follow the order of operations you need for laboratory exercise 6.

Displaying a single channel

Turn the ‘scope **on** using the button on top at the left. **Connect a signal** to channel 1 using a cable with a so-called BNC connector (plug it in and twist to lock) at one end, and banana plugs at the other end to connect to the output of the signal generator. Turn the **signal generator on** and select a sine wave signal of about 1 kHz. Press the **AUTOSET** button on the ‘scope and you should see the sine wave signal displayed — the ‘scope tries to set sensible time and voltage scales automatically.

You can adjust the voltage scale by using the channel-1 (CH 1) **VOLTS/DIV** knob, and the time scale by using the horizontal **SEC/DIV** knob. This allows you to optimise the display, and to see the waveform when you change the parameters of the input signals. You can also move the display up and down or left and right with the **POSITION** knobs.

The screen displays the voltage and time settings, as well as providing a measurement of the frequency of the trigger signal. There is also information about how the scope is being triggered, which will be discussed later.

Other settings affecting the operation of the display can be viewed and altered by pressing the buttons **CH 1 MENU** or **HORIZ MENU**. You can then use the unlabelled buttons along the right-hand edge of the screen to select the settings you want.

Setting the trigger

In order to display a periodic waveform, the ‘scope has to be triggered at a fixed point in its cycle. This is most often done by setting a particular voltage level, and specifying whether the signal should be rising or falling. The display can be triggered either by one of the signals being viewed (you can choose channel 1 or channel 2), or by a separate signal used only for triggering. The voltage level for triggering is set by the **TRIGGER LEVEL** knob. The trigger source and polarity can be selected by using the trigger menu, which you get by pressing the **TRIG MENU** button.

The trigger level is indicated on the screen by a horizontal arrow on the right hand edge of the display, as well as a numerical value at the bottom right of the display. A symbol indicates whether the trigger is on a rising or falling signal, as well as which input is being used. An arrow at the top edge of the screen indicates where in time the trigger is occurring.

It is well worth looking at pages 28–30 of the Tektronix instruction manual for a key to all the information that is normally available on the screen. We have photocopied these pages.

Making standard measurements

To make detailed measurements using old-fashioned analogue ‘scopes required counting grid boxes on the display. Digital ‘scopes automate measurements and do a far more precise job by offering two facilities: a standard set of quantities such as frequency and amplitude that are

calculated and can be displayed, and for other measurements a pair of on-screen cursors that you can move around to tell the 'scope where to measure.

To select measurements, press the **MEASURE** button in the middle of the top row of the control panel. You can select a mixture of measurements for one or both input channels, up to a maximum of five simultaneous quantities. There is a huge range available, but the most obvious ones are frequency, period, and amplitude (abbreviated to **Pk-Pk**, i.e. peak-to-peak). These will then be displayed along the right-hand edge of the screen as you alter the input signals.

Using the cursors

You can measure the horizontal distance (i.e. time) between two points on a signal, or you can measure the vertical distance (i.e. voltage). If you want both you must do them one after the other, you cannot do both at once.

Press the **CURSOR** button, below the **MEASURE** button. You then select whether to measure voltage or time, and choose channel 1 or channel 2. Two lines appear on the screen; their position is controlled by the channel-1 and channel-2 position knobs (which warn you of this by having an LED below them illuminated). The cursor measurement information appears on the screen. Delta is the distance between the two cursors, and Cursor 1 and Cursor 2 give the absolute positions of the cursors: time is referenced to the trigger position (arrow at top of screen), and voltage is with respect to 0 V.

Displaying two channels

The displays of channel 1 and channel 2 can be turned off or on independently, and other parameters set up, by pressing the **CH 1 MENU** or **CH 2 MENU** buttons. The display is helpfully colour-coded, as well as labelling each of the waveforms at its left-hand edge.

Setting up XY mode

Press the **DISPLAY** button (second row) and at the third item, Format, choose **XY**. In this mode the channel-1 voltage is on the horizontal (x) axis and the channel-2 voltage on the vertical (y) axis. Note that you cannot use the cursors in this mode.

Saving, printing, and using data on a PC

Screen images, as well as full numerical details of all the data points they contain, can be saved onto CompactFlash cards plugged into the 'scopes. These hold a large amount of information on small cards that need no special software, use no batteries or external power, and can be used like floppy or Zip disks.

Plug the CompactFlash card into the slot on the top rear right-hand side of the 'scope. When you have a display that you wish to save, press the **SAVE/RECALL** button (top row at left). Set the **Action** option to **Save Image**, and set the **File Format** of the graphics file to your choice; **TIFF** is probably best. Then select **Save**.

(In addition to saving a graphical picture of the screen, you can save the full numerical details of every dot on the waveforms in the form of a .csv (comma-separated) file that can be read into Microsoft Excel or PhysPlot and analysed or displayed using standard Excel facilities. To do this set the **Action** to **Save Waveform** and then **Save** to **File**.)

(The first time you use a particular CompactFlash card it has to be formatted. We have already done this with the lab's cards. To format or re-format a card, insert it into the 'scope, push the **UTILITY** button, select **File Utilities** from the menu, select **More** to show more of the menu, and select **Format**. Note that this *erases all existing data* on the card!)

When you are ready to transfer your files to a computer, remove the CompactFlash card by pressing the small Eject button next to it and lift it out. Take it to one of the lab's own computers (in the centre of the room, not against the walls). These are equipped with CompactFlash card-reader devices connected to the computer via a USB port at the rear. Plug the CompactFlash card into the reader.

Log in and double-click on **My Computer** and then on **Removable Disk (F:)**. You will see some folders and files. Your screen image(s) will have names like **FxxxxTEK**, where **xxxx** is a 4-digit sequential number. Double-click on the file icon to open the file in a simple image-viewer. This allows you to print it out, resize the image, and do several other simple manipulations. You can also copy the files to the computer's hard disk.

To insert the image into a Microsoft Word document, go down the **Insert** menu to **Picture** and select **From File ...**. Navigate to the file you want and select it — Word will paste it in.

Mathematical manipulations and Fast Fourier Transforms

The **MATH MENU** button allows you to select some useful simple operations such as adding or subtracting the signals in channel 1 and channel 2, which can be extremely useful. (For example, a lot of high-speed electronics transmits signals differentially, i.e. the signal is the difference between the voltage on two wires.)

More adventurously, there is a facility to do a Fast Fourier Transform (FFT) on a signal. To do this:

- Display a normal (voltage vs. time) waveform.
- Centre it vertically and make sure the top and bottom of the waveform are visible, not off-scale.
- If the waveform is not regular, make sure the 'interesting' part is in the centre eight horizontal divisions of the screen. If possible, display many signal cycles.
- Push the **MATH MENU** button, set the **Operation** to **FFT**, and select the channel.
- The display now shows frequency horizontally, and the vertical amplitude represents the contribution of that frequency to the waveform.
- You can change the frequency scale by using the **SEC/DIV** knob.

Eric Eisenhandler, 29/9/04

Laboratory Exercise 6 – MEASUREMENTS OF WAVE VELOCITY

Introduction

The theme of these experiments is to measure the speed of sound in different media: in air (6A), and in a copper rod (6C) and to investigate uses of wave phenomena in measurement techniques (6B). These are actually three separate exercises, each of which illustrates the use of the oscilloscope to display rapidly changing electrical signals.

*Each part can be completed in a single laboratory afternoon if you are fairly efficient, using the built-in measurement capabilities of the 'scopes. These are precision measurements and are capable of considerable accuracy, so in each case we want you to *compare* your results with those found by experienced investigators and given in textbooks.*

General principles

No detailed understanding of wave motion is needed to carry out this exercise. We simply use the relation **wave velocity = frequency × wavelength**, $v = f\lambda$, since in each part we measure the frequency and wavelength, and calculate the velocity. This is the usual way of finding wave velocities since it is extremely inconvenient to measure the time a wave takes to travel between a source and a receiver a long distance apart.

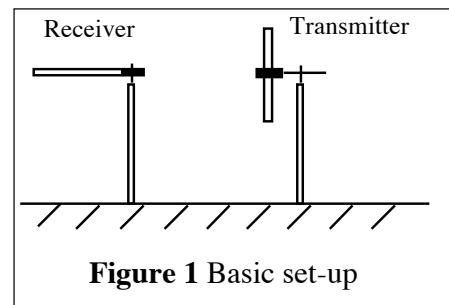
We also mention a relation that you will have to take on trust, namely $v_s = \sqrt{K/\rho}$; here v_s is the velocity of a sound wave (a compression or **longitudinal** wave, with to-and-fro motion of the molecules), ρ is the density of the solid, liquid or gas, and K is an **elastic modulus**, a quantity that measures the pressure needed to compress or deform the material by a given amount. There are different moduli depending on the material and how it deforms — you have probably met Young's modulus which measures the pressure needed to squeeze a solid in one direction. The relation enables the elastic modulus to be found from a measurement of sound velocity, and vice versa. Other types of waves (**transverse** waves involving side-to-side motion, as in a violin string) can also travel through solids, and for these there are similar but slightly more complicated relations between velocity and elastic modulus, which are mentioned and used in exercise 6C.

Exercise 6A: Sound waves in air

In this experiment you use several different methods to measure the wavelength of high frequency waves travelling through air. The sound waves are produced in a small transmitter driven by electrical signals from a sine wave oscillator, and are detected by a receiver which is a small microphone, similar to the transmitter, whose electrical output is displayed on the oscilloscope. Transmitter and receiver are placed facing each other on a graduated slide. Their response is sharpest near a frequency of 40 kHz, so your measurements relate to this frequency.

Frequency response

- Set the transmitter and receiver facing each other about 30 cm apart (figure 1). Vary the oscillator frequency and observe the response on the oscilloscope; there is a narrow band of frequencies for efficient reception. Set the oscillator to maximise the amplitude, measure the frequency with a laboratory frequency meter, and check it during the course of the experiment.



Direct measurement of wavelength

- Move the transmitter slowly towards the receiver. As you do so the relative phase of the transmitted and received signals changes, the wave trains shifting by one whole wavelength when you have moved the transmitter exactly this amount. There are millimetre scales on the shoeplates; measure the distance d that corresponds to a large number, N , of wavelengths, deduce the wavelength λ at this frequency, and so calculate the sound velocity v_s .
- Now recall the expression $v_s = \sqrt{K/\rho}$. In a gas the appropriate modulus K is the pressure, and in an ideal gas pressure divided by density is proportional to the absolute temperature T . So (show this for yourself) $v_s(T\text{ }^\circ\text{C}) = v_s(0\text{ }^\circ\text{C}) \sqrt{1 + (T\text{ }^\circ\text{C})/273}$. Find the temperature of the air in the lab, and so reduce your value for the speed of sound to the value at 0 °C. Compare this with tabulated information (e.g. Kaye and Laby). Estimate the errors in your measurement.

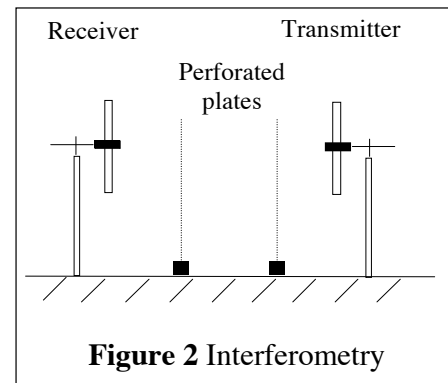
Standing waves

- Sound waves can bounce back and forth between the transmitter and receiver clamps. If the separation between transmitter and receiver is a whole number of half wavelengths, the there-and-back distance for one reflection is twice this, which is an exact whole number of wavelengths, and a **standing wave pattern** will be formed with **nodes** where the opposing waves cancel and **antinodes** where they reinforce. As you move the transmitter you can see the received signal increase in amplitude every half-wavelength as the standing wave pattern is formed. (In between there is a rather confusing variation of response because the travelling waves partially interfere). If this is difficult to observe try putting the aluminium discs on the receiver and transmitter, clamping them carefully so that the discs are flush with the fronts of the transmitter and receiver, and parallel to one another.
- Measure the change in distance corresponding to passing through a large number of nodes, deduce the wavelength, and compare with your earlier value.
- The change from one standing wave pattern to the next is accompanied by a change in the phase of the received and transmitted signals which you can see on the two-beam display of the 'scope. This phase change becomes more noticeable if you switch to XY display. In normal display mode the **horizontal** axis is controlled by a **timebase** circuit which drives the display in the horizontal (X) direction at a constant rate, adjustable by a front-panel knob at the right.

Selection of the *XY* display mode turns the timebase off and allows the display to be driven horizontally by one of the input voltage signals. Do this, and describe what you see.

Interferometry

- Place the perforated metal plates on the slide at right angles to the sound wave (figure 2). They let some of the wave through while reflecting part of it. Between these semi-reflecting plates the sound wave bounces back and forth, some of its energy escaping towards the receiver. There will therefore be a standing wave pattern between the plates when their separation is an integer number of half wavelengths, as we have seen. At these spacings the amplitude of the signal displayed on the 'scope will increase to a maximum, with weak minima at intermediate spacings.

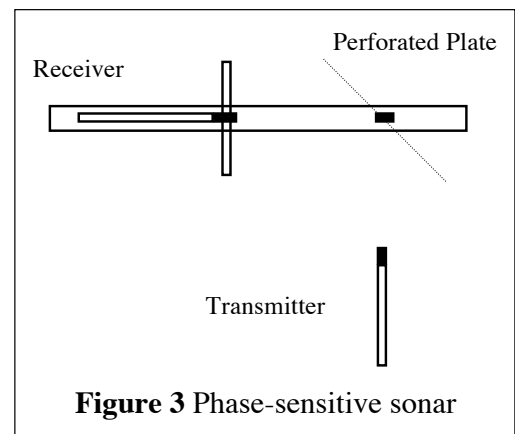


- Put the scope back into normal (not *XY*) two-beam operation, and move the perforated plates so as to measure the spacings of these maxima. This gives yet a third way of measuring the wavelength, so deduce it.

The technique of using semi-reflecting plates to form interference patterns of standing waves was developed by Fabry and Perot for use in optics, where it is widely used in spectrometers to measure, or select, different wavelengths. What you have here is a sonic analogue of the Fabry–Perot interferometer.

Phase-sensitive sonar

- Finally, take the transmitter off the slide and place it to one side with its ultrasonic waves directed at a perforated plate set at 45° (figure 3) so as to reflect the sound to the receiver. Viewing both traces on the 'scope, you will see that their relative phase changes as the plate is moved slightly towards or away from the receiver. A movement of one wavelength causes a complete phase shift of the same amount. The effect is seen more clearly if you switch to *XY* display, where the pattern is sensitive to quite small movements of the perforated plate.



- To increase the sensitivity still further, switch the oscillator from sine wave to square wave operation. You will see, on dual-trace display, that neither transmitter nor receiver can respond to the rapid changes of the 'square' wave so a sine wave is still displayed. Switch back to *XY* and observe the effect of *combining* the square wave with the sine wave in *X* vs. *Y*. You have an oblong display whose height (or width — it depends which way round your input channels are connected) is *extremely* sensitive to movement of the plate. Blow gently on it and you will agree!

- Set your pattern to fill most of the screen, estimate how large a change in the oblong shape you can detect (perhaps a millimetre or so?), and deduce what movement of the plate this corresponds to. You have built a sensitive device for remote measurement of small displacements — a phase-sensitive sonar.

Exercise 6B: OPTICAL MEASUREMENTS

Introduction

Light is a form of wave motion, the colour being determined by the wavelength — about 400 nanometres for blue light and 700 nanometres for red (a nanometre, nm, is 10^{-9} metre). Suppose two light waves travelling in the same direction are brought together. If the crests and troughs of one coincide with the crests and troughs of the other the waves reinforce each other, whereas if the crests of one fall exactly on the troughs of the other the waves cancel one another out. The first case, leading to enhanced brightness, is called **constructive interference**, the second, **destructive interference**. If the two waves have exactly the same wavelength (therefore colour) and the same amplitude, the cancellation will be perfect yielding complete darkness. If white light is used cancellation can be perfect only for one wavelength.

A common way of producing interference effects is to take a single light wave and split it in two, say by reflecting part of it from each of two surfaces a small distance apart. If the waves are recombined they will interfere — this is the origin of the colours of a soap bubble, as shown in figure 4. (You may have noticed that these colours, being due to the *absence* of light near a certain frequency, are different from the usual spectral colours of the rainbow which are due to the *presence* of particular frequencies.) The colour of the bubble can be used to calculate the thickness of the water film which, being typically much less than 10^{-6} m, is otherwise very difficult to measure. The measurement is easier if light of a single wavelength (that is, colour, hence **monochromatic**) is used because complete destructive interference (total darkness) can then occur.

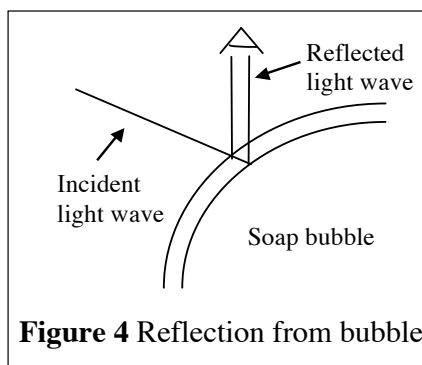


Figure 4 Reflection from bubble

In this exercise you will observe optical interference of (nearly) monochromatic light, and use it to measure the thickness of an air gap between two glass surfaces, one of which, a simple convex optical lens, is curved. You can then calculate the curvature of the lens surface. Comparison with similar measurements made with a mechanical curvature gauge, or **spherometer**, should convince you that optical methods are *much* more precise than mechanical ones for measuring small distances. Many analytical and process control instruments utilise interference techniques; a simple example is given at the end of this exercise.

A useful mathematical result: the sagitta rule

Figure 5 shows an arc of a circle of radius R , its ends connected by a chord of length $2a$. If we look at either of the two right-angled triangles,

$$R^2 - (R - h)^2 = a^2 \quad R^2 - 2Rh + h^2 = a^2$$

Since h is much less than R the term on the right is approximately $2Rh$, leading to the approximation known as the **sagitta rule**:

$$a^2 = 2Rh$$

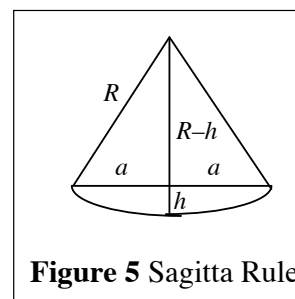


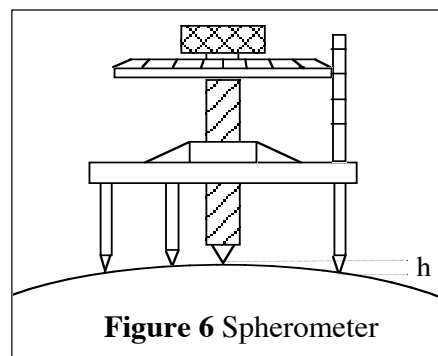
Figure 5 Sagitta Rule

(so-called from a resemblance between the diagram and a bow and arrow, the Latin *sagitta* meaning arrow). It applies also to a sphere, with h being the distance between a tangent plane and a spherical surface such as that of a lens, and R being the radius of curvature of the surface. We shall use the sagitta rule in both the mechanical and the optical measurements of R .

Measurements: mechanical

The spherometer (figure 6) is an instrument for measuring the curvature of a spherical surface such as that of the lens you are given. Its tripod feet are set around a circle of radius a , corresponding to the chord length in the sagitta rule. In use, its centre point is raised (or lowered) a distance h above or below the plane of the feet, allowing the radius of a convex (or concave) surface to be calculated using the rule. One turn of the screw corresponds to a fixed change in h , which for your instrument is 1.0 mm. The edge of the attached disc is divided into a scale of 100 equal graduations, each corresponding to 0.01 mm. To obtain the correct measurement you will need to subtract your scale reading from 1.0

- To use the spherometer, balance it on the convex lens with the point raised, screw the point down until the tripod feet are just free to move, and note the scale reading against the vertical bar. This may be positive or negative depending on the instrument you have, but we are interested in measuring the **distance travelled** obtained by subtracting this measured value from the measurement of a flat surface...

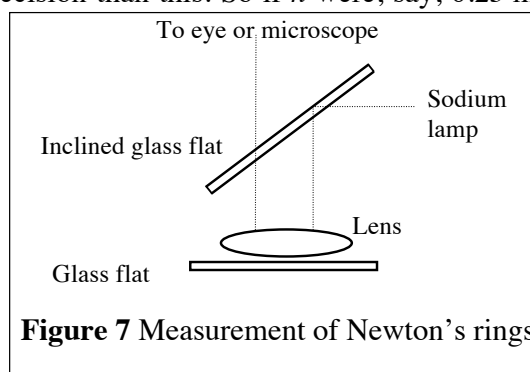


- Now stand the spherometer on the accurately flat *larger* sheet of glass, lower the point until it is again just touching the surface, and note the scale reading again. The difference between the two readings is distance h . **Repeat** the measurement of h several times.

- Find a by measuring the separation of the tripod feet which, by trigonometry, is equal to $a\sqrt{3}$. Check a by directly measuring the distance between any leg and the centre. Which of these measurements is more accurate, and why?

- Use your values of a and h to deduce R .

Your measurements of h will probably not agree to better than one scale division (0.01 mm), and anyway it is difficult to read the scale to better precision than this. So if h were, say, 0.25 mm, then your measurement would only be accurate to 1 part in 25, that is 4%. Also, there is an uncertainty on your measurement of a . Use the propagation of errors formula to calculate the uncertainty on R given your estimates of the uncertainty of a and of h . Remember, the **experimental error** or the **experimental uncertainty** in R , σ_R , can be expressed as either a percentage, $100 \times \sigma_R / R$, or as an actual value, $\pm \sigma_R$, using the symbol \pm to indicate our uncertainty: $R \pm \sigma_R$.



Measurements: optical

- Place the lens on the *smaller* flat sheet of glass underneath the inclined glass sheet in its holder, and put this on the stage of the travelling microscope. Turn on the yellow sodium lamp and shine the beam horizontally. The inclined glass sheet reflects some of the light down through the lens where part is reflected back from its bottom surface while the rest continues to the glass flat, is reflected, and retraces its path. The two returning light waves continue together back through the lens and up through the inclined sheet (figure 7).

- Look directly downwards. Your eye brings the two light waves to a focus on your retina where they interfere. If your vision is reasonably good you should be able to see a number of tiny concentric dark circles centred on the point where the lens touches the glass flat; these circles

$$a_0^2 = 2Rh_0$$

$$a_1^2 = 2Rh_1 = 2R(h_0 + \lambda/2) = 2Rh_0 + R\lambda$$

$$\vdots$$

$$\vdots$$

$$a_n^2 = 2Rh_0 + nR\lambda$$

are called **Newton's rings** (although they were actually studied first by Hooke). At a radius a from the centre there will be a dark ring if at that radius the distance h between the lens and the flat is exactly right for destructive interference between the reflected waves. At larger and smaller radii there are bright rings where constructive interference occurs. The sagitta rule relates a and h to the radius of curvature R of the lens surface. We move from one dark ring to the next whenever h increases by half a wavelength of light, since the extra distance travelled by one wave is $2h$, a whole wavelength. So starting with a dark ring at radius a_0 and separation h_0 , the sagitta rule gives for this and successive rings:

In these expressions λ is the wavelength of sodium light, which you can take to be 589 nm. A plot of a_n^2 versus n should yield a straight-line graph whose slope is $R\lambda$. Thus R can be found.

- Now for the measurements. The rings are much more easily seen through the microscope, and you can use the graduated scale on the horizontal bar to measure the position of the cross-wires as you move them from ring to ring. This scale has a **vernier** attachment to increase the reading accuracy (if you are not familiar with vernier scales ask a demonstrator). Practice moving the microscope horizontally by unclamping it, moving it to about the right position, clamping it firmly to the screw-drive, and then using the screw to move the microscope in a slow and controlled way.
- Set the cross-wire tangent to, say, the fifth dark ring to the left of the centre, and the same ring ($n = 5$) to the right, noting the scale readings at each position. Make sure that in going from left to right parts of the rings you pass through the *centre* of the rings. The difference between your readings is the diameter of the fifth ring. In a similar way measure the diameters of the 10th, 15th, 20th, ..., rings. You should be able to measure out to about $n = 50$, though it is difficult to avoid losing count. The best procedure is to take all the readings to one side first, then repeat on the other side of the centre. *Be careful not to confuse millimetres and centimetres.*
- Use the propagation of errors formula to calculate the error on a^2 from your measurement of a .
- Plot your measurements as you go along, check that a straight line is a reasonable fit to your data, and find its gradient. **Do not forget to plot the error bars!** Does your line pass through the origin? Then use the computer to print a neat graph and to calculate the gradient, which should be close to your hand calculation. Deduce a value for R . The computer also calculates the error in the gradient. Use this to find the experimental error $\pm\sigma_R$ in your value of R . How does this compare with the error from the mechanical measurement?

A practical application

Suppose the gap between the lens and the flat plate were filled with something other than air — a liquid, say, with refractive index μ . The refractive index measures how much more slowly light travels in a medium than in a vacuum (or in air, whose refractive index can for most practical purposes be taken equal to unity). As the refractive index increases so the velocity, and the wavelength, of the light decreases in inverse proportion:

$$\lambda_{\text{medium}} = \lambda_{\text{vacuum}} / \mu$$

So the condition for destructive interference becomes:

$$a_n^2 = 2Rh_0 + nR \lambda / \mu$$

and if you repeat the experiment with a liquid instead of an air gap the slope of your graph will be μ times smaller.

- Try it. Place a small drop of distilled water from the squeeze bottle on the flat plate, lay the lens on top, repeat your measurements and, using the value of R you have already found deduce the refractive index of water. Plot your data on the same graph as before.

This is the basis of some commercial instruments for measuring refractive index. The reason for using distilled water, incidentally, is to keep mineral deposits off the optical glass — the refractive index, even of London tap water, is scarcely different!

A question of physics

Notice that the centre of the pattern is dark. You might expect that where the two reflecting surfaces are so close together that the difference in distance travelled by the two waves (the **optical path length**) is negligibly small, they would interfere constructively leading to enhanced, not diminished, brightness. On the other hand, you might argue that since at the centre there is no gap at all, but a continuous light path in glass, there can be no reflection from either surface, hence no reflected light, hence darkness! Which, if either, of these two conflicting arguments is right? You might care to ask your lab demonstrator.

Exercise 6C: The vibrations of a copper rod

Gases and liquids respond to pressure in a simple way. Their change of density depends only on the magnitude of the compressive force, not on its direction. That's why bubbles in water are spherical, not flattened. The response is described by a single elastic modulus, the **bulk modulus** K , and there is only one sound velocity, $v_s = \sqrt{(K/\rho)}$.

Solids are more complicated. Pull a rubber band and it becomes thinner as well as longer. The molecules respond to the direction as well as the magnitude of the force. A solid requires at least **two** elastic moduli, K which measures the stress needed to change its volume, and the **shear modulus** (or **rigidity modulus**) G which measures the stress that changes its shape. The word 'shear' tells us that molecules are sliding past each other, a motion strongly resisted in solids but not in fluids (which fill any shape of container). As it is difficult to compress a solid object without changing its shape, a more useful modulus than K is **Young's modulus** E which measures the stress that compresses a solid in one direction while allowing it to change shape as it pleases in other directions. K , G and E are not independent — the relation between them is:

$$K = \frac{EG}{9G - 3E}$$

This extra complexity allows solids to transmit several types of waves. For example, **seismic waves** in the Earth travel at different speeds depending on whether compression or shear is the dominant motion. In this exercise you will use thin copper rods to study the vibrations which they support. These vibrations would travel as waves along very long bars, but in these short bars the waves reflect back-and-forth from each end, setting up **standing wave** patterns when their wavelength (and hence frequency) is just right for reinforcement to occur. An electrical method is used to make the bar vibrate in the desired way, and the movement of the bar is also measured electrically. Even large forces produce rather small deformations of the rod, and the vibration amplitude will be small *unless* the frequency of the driving force is exactly equal to one of the natural vibrational frequencies of the rod. This is an example of **resonance**, the familiar situation in which the stimulus is applied at the same time in every cycle, as when you push a child higher and higher on a swing.

In this exercise you set up both **torsional** (twisting) and **bending** vibrations, as sketched in figure 8. Mathematical analysis shows that the speed at which torsional waves travel is determined solely by the shear modulus: $v^{\text{torsion}} = \sqrt{(G/\rho)}$. Standing waves are set up when a whole number, n , of half-wavelengths fit into the length l of the bar (the same criterion as for the fundamental frequency and harmonics of an organ pipe). This requirement and the relation $v = f\lambda$ lead to the following expression for the resonant frequencies of torsional vibration:

$$f_n^{\text{torsion}} = \frac{n}{2l} \sqrt{\frac{G}{\rho}}$$

The analysis for bending waves is rather more complex. Their velocity is determined by Young's modulus, and also depends on their wavelength. The final result is:

$$f_n^{\text{bending}} = \frac{\pi a}{16l^2} (2n + 1)^2 \sqrt{\frac{E}{\rho}}$$

where a is the radius of the cylindrical rod. To be quite exact for $n = 1$ the factor $2n + 1 (= 3)$ must be replaced by 3.011.

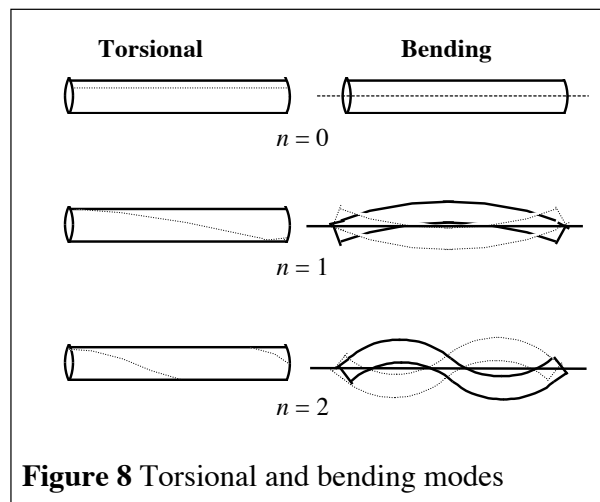
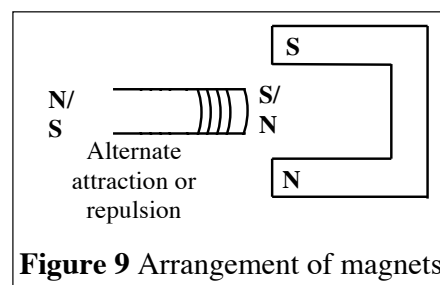


Figure 8 Torsional and bending modes

1. Bending vibrations

- Two copper rods are provided, one for bending and the other for torsional vibrations. Use the rod with circular coils wound near its end for the bending modes. Suspend it horizontally so that it can vibrate freely and connect one of the coils to the function generator via the black amplifier box. This box is required because of the poor impedance matching between the 50Ω output of the function generator and the 3Ω impedance of the coil. Arrange a permanent magnet to give a field which, by interaction with the current in the coil, will force the rod into bending vibrations (figure 9). Detect the vibrations by arranging a second magnet in a similar way near the other coil so as to induce a small voltage signal in that coil which, after amplification by the grey amplifier box, can be registered by the oscilloscope.



- Check that the apparatus is set up correctly by switching off the oscillator and striking the bar gently in the middle, when the fundamental vibration (i.e. $n = 1$, figure 8) should be excited and you should see its oscillation on the 'scope. From the trace, roughly estimate the frequency.

- Turn on the oscillator and look for the $n = 1$ resonance in this region, varying the oscillator frequency until a sharp maximum is obtained. Display both input and output signals on the 'scope, making the final adjustment of frequency so as to make the output as large as possible. Estimate the phase relationship at resonance. The frequency may not be indicated reliably by the oscillator dial, so use the oscilloscopes.

- Plot a suitable graph relating harmonic number n to frequency, measure the length and radius of the rod, and deduce a value for Young's modulus E of copper. The density may be found in Kaye and Laby, which also lists currently-accepted values for elastic moduli.

2. Torsional vibrations

- Use the other rod to set up torsional vibrations. The rectangular coil wound at the end of this rod produces a magnetic field in a different direction from the previous circular coil, so by arranging the permanent magnets appropriately (perpendicular to the coil, with the two ends different by 90°) you can both induce and detect twisting motion.

- Once again using the frequency generator set to generate a sine wave scan through and determine the resonant frequencies, and so deduce a value for the shear modulus G of copper. From your measurements of E and G , deduce a value for the bulk modulus K .

Resonant vibrations in quartz

Longitudinal sound waves, that is waves of compression running to-and-fro along the rod with velocity $v_s = \sqrt{E/\rho}$, are easy to produce mechanically. But there are some crystalline materials, for example quartz, in which these vibrations can be produced electrically. An applied voltage produces a small change in the crystal's length (the **piezo-electric** effect), so an AC voltage causes a wave of expansion and contraction to run through the crystal. The resonant frequency depends on the size of the crystal; a thin slice of quartz, for example, has a resonant frequency of hundreds of kilohertz. Only a tiny input of electric power is needed to sustain these resonant oscillations, whose frequency is determined by the size of the crystal and is thus extremely stable. These crystal oscillators are at the heart of every digital watch and computer clock.

- The speed of sound waves along such a quartz crystal is 5440 m/s and the density of quartz is 2600 kg/m^3 . Calculate the value of Young's modulus, and find the thickness of a slice of quartz

whose fundamental resonant frequency is 1 MHz. Note that the requirement for resonance is the same as that for torsional vibrations.

Laboratory Exercise 7 – MEASUREMENTS IN ASTRONOMY

Part A: The angular resolution of telescopes

Introduction

For astronomical observations, reflecting telescopes have replaced the refracting type of instrument. Large mirrors with great light-gathering power are lighter, more rigid and easier to control than equivalent lenses; lenses are also plagued by chromatic aberration which brings different colours to a focus at different points. One of the largest refractors made, the 28" at the Old Greenwich Observatory, is a beautiful instrument but suffers badly from chromatic aberration as you will find if you get an opportunity to use it. Both reflectors and refractors, however, have an intrinsic limit to their resolution, that is to their ability to distinguish two objects which appear very close together in the sky with a very small angular separation. This quality of a telescope, called its **angular limit of resolution**, or just its **resolution**, is determined by its aperture — the larger the aperture, the greater the resolution (and hence the smaller the angular separation that can be resolved). The object of this part of the exercise is to set up a reflecting telescope, to measure its resolution, and to compare this both with that of the unaided eye and also with an estimate from theory. An example of an angular measurement in astronomy completes this part.

Adjusting the telescope

The primary mirror of a reflecting telescope is a paraboloid, which brings parallel light from a distant object to a focus in front of it. It is therefore necessary to move the image out of the incident beam so that it can be observed without obscuring the incoming radiation. In the Newtonian arrangement this is achieved by reflecting the light off a small flat mirror set at 45° to the telescope axis and just in front of the focal plane (see figure 1). The image is then observed by an eyepiece which acts essentially as a microscope.

- The components of the telescope can be clamped to the triangular section of rigid steel, the 'optical bench'. Measure the **focal length** of the primary as accurately as you can (the error should not be greater than 1 cm) by using a distant light as a source and finding the distance of its image from the vertex of the mirror. View this image using a paper screen moving along the triangular beam.

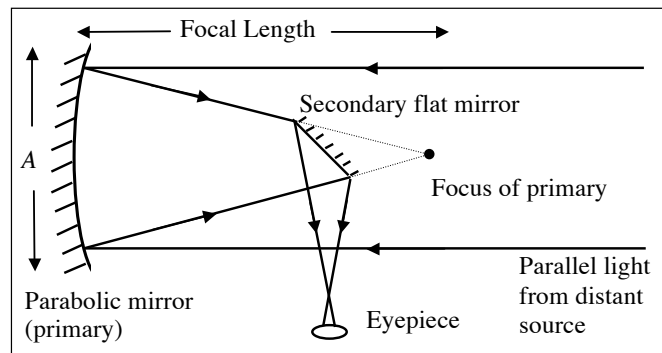


Figure 1 Reflecting telescope

- Measure the diameter of the primary mirror (its **aperture**, A) and also the smaller diameter of the secondary mirror.

- Place the secondary at a position such that when it is at an angle of 45° the image is seen clearly in the eyepiece. (It may help to look first without the eyepiece and barrel and adjust so that your eye is seen reflected.) Ideally, the image of light reflected from the primary should completely fill the secondary.

- Make a scale drawing like figure 1 to see whether this is the case. Make any final adjustments while viewing the distant light source through the eyepiece.

Measurements

- Use the resolution test card to measure the resolving power of both the telescope and your eye. Place the card in the laboratory as far as convenient from the telescope (at least 10 m if possible). View the sets of lines through the telescope and determine the set of smallest line spacing which can be resolved into individual lines. Use the ability to determine the direction of the lines correctly as evidence for resolution. Use the travelling microscope provided to measure the distance d between consecutive line centres in the set which can just be resolved. Measure also the distance D between the test card and the primary mirror. Hence calculate the small angle $\theta = d/D$ which is the experimental angular limit of resolution.
- By similar measurements, find the angular limit of resolution of your eye — you may wish to try this for left and right eyes separately as well as together. Do this wearing spectacles or contact lenses if you normally do so.

Note that this expression for θ gives the angle in units of **radians** rather than in degrees; recall that $\theta(\text{radians}) = (\text{arc length})/\text{radius}$ (see figure 2). Thus 2π radians corresponds to a full circumference, or 360° , and so a radian is approximately 57.3° . A measure commonly used in astronomy in the **arcsecond**, or $1/3600$ th of a degree. Therefore, one radian is approximately 2.06×10^5 arcseconds.

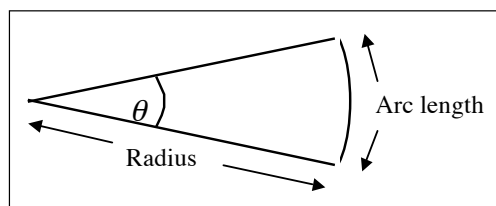


Figure 2 Definition of radian

Theoretical estimates of angular resolution

Light, being a wave motion, must be treated by a theory which includes the effects of **diffraction**, that is the bending of wave trains as they go around obstacles or as they pass through apertures (as you see on a large scale when sea waves spread out on entering a harbour mouth). The primary mirror is an obstacle to the incoming light waves so they are diffracted to some extent, thus smearing out and confusing the images of two sources having a small angular separation. The smallest angular separation, θ_{\min} , which can just be distinguished (that is, the resolution) is determined by the ratio between λ , the wavelength of the light, and A , the diameter of the primary mirror. The ability to resolve overlapping images is a rather personal thing and there are several prescriptions (criteria) for the exact formula to use. One of them, the **Rayleigh criterion**, is based on the mathematical form of the diffraction patterns from circular apertures: it states that $\theta_{\min} = 1.22 \lambda/A$. Other prescriptions such as the **Abbé criterion** (which omits the factor 1.22) are sometimes used.

- Taking λ to be 550 nm, near the peak of the spectral curve for white light, use these criteria to calculate the theoretical angular resolution of the telescope and of the eye (for which the aperture is the diameter of the pupil, typically 3.0 mm). Compare the results with your experimental values.

Angular measure and the distance of astronomical objects

Astronomers measure angular positions in the sky. As the Earth circles the Sun the position of a nearby star shifts relative to that of more distant stars since its direction in the sky changes slightly. Figure 3 shows that in six months the position angle changes by an amount equal to the diameter of the Earth's orbit divided by the star's distance. Therefore, this distance equals the radius of the Earth's orbit divided by the deviation of the position angle from its average value (a quantity called the **parallax** of the star). The distance will be in kilometres if the Earth's orbital radius is also given in kilometres and the parallax is in radians. Astronomers find it more convenient not to bother with the exact value of the orbital radius but simply to call it an

Astronomical Unit (AU), and also to measure angles in arcseconds (the parallax of most stars is less than 1 arcsecond). Using these units the distance of a star is measured in **parsecs (pc)** — the distance in parsecs is the reciprocal of the parallax in arcseconds. Thus $1 \text{ pc} = 2.06 \times 10^5 \text{ AU}$, or approximately $3 \times 10^{13} \text{ km}$.

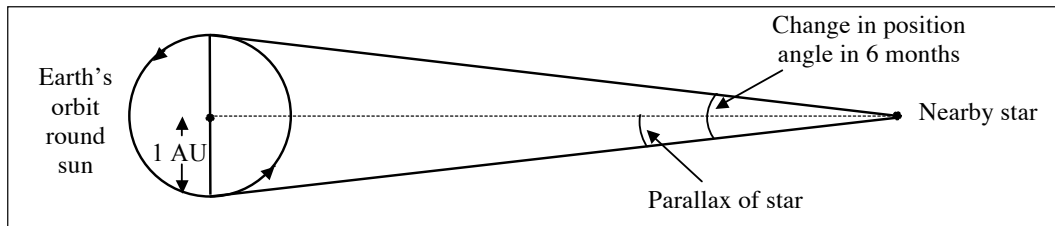


Figure 3 Parallax

- Besides their annual parallax, some stars show a real movement (**proper motion**) relative to more distant stars. Figure 4 shows two photographs of the same area of the sky, taken 10 years apart. One star, whose parallax is known to be 0.55 arcseconds, has moved appreciably during this time. Find it, measure the distance it has moved (estimate to a quarter of a millimetre using a good graduated rule), convert this distance to an angle using the fact that the photographs measure 40.5 arcminutes (1 arcminute = 60 arcseconds) in the horizontal direction, and calculate the velocity of the star across the line of sight.

Part B: Line spectra, chromatic resolution and doppler shifts

Introduction

The spectrum of light from stars contains many sharp features, light or dark bands called **spectral lines**, that tell astronomers about the chemical composition and physical conditions on its surface. This is done by comparison with simple spectra produced in the laboratory. In this part you will use a diffraction grating to produce the line spectra of a number of elements, and will investigate how close in wavelength two lines can be before they become indistinguishable. This property is called the **chromatic resolution** of the instrument or sometimes, when there is no possibility of confusion with angular measurements (see part A) just the **resolution**. The better the resolution, the more closely one can investigate small wavelength changes caused, for example, by the Doppler effect. Finally, you will measure wavelength shifts in the spectrum of a star system that are actually due to relative motion of the stars.

The grating spectrometer

In this part you use a diffraction grating mounted on a precision rotating table. A parallel beam of light of wavelength λ striking one side of the grating will be **diffracted**, that is, deviated in all directions, as it passes through the apertures between the opaque lines of the grating. The transmitted light will interfere constructively (see exercise 1) at angles determined by λ and by the spacing d between the lines. In these directions (the **principal maxima**) the transmitted light intensity is greatest and a telescope set at these angles will see images of the source, a narrow slit parallel to the grating lines, in the colour of the wavelength selected. The expression for the angles of these principal maxima is:

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

where the number $n = 1, 2, 3, \dots$ is called the **order** of the principal maximum. The grating you will use has a small value of d so the right-hand side is large; since $\sin \theta_n$ must be less than one, only the first few orders will be visible.

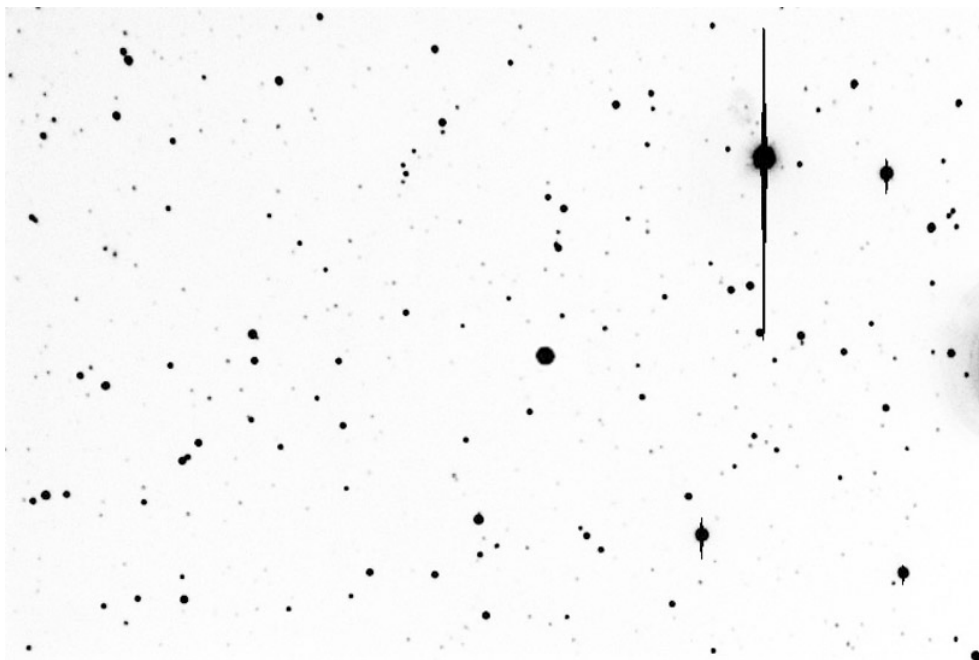
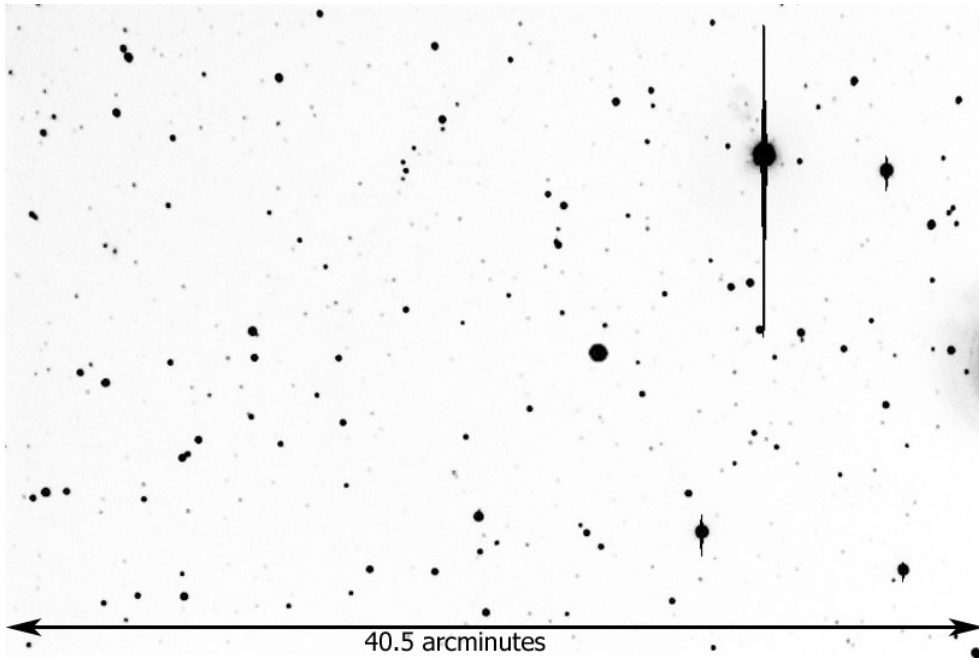


Figure 4 Proper motion of a star

The spectrometer is shown in figure 5. The collimator, which produces a beam of parallel light coming from the adjustable slit source, and the telescope can be rotated around the centre of the table. Vernier scales give accurate measurements of the angular position.

- Place the grating over the centre of the table and perpendicular to the collimator axis. Clamp the table and the collimator firmly, ensuring that the telescope can swing freely on both sides of the straight-through position.

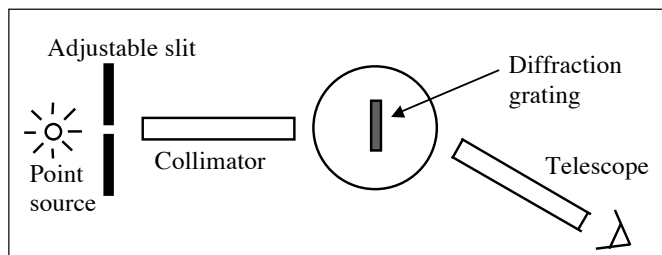


Figure 5 Grating spectrometer

- In this position observe the direct image of the slit, and adjust the slit, collimator and telescope as needed. You want the slit to be as **narrow as possible** while still appearing uniformly bright, and to be **vertical** (the lines of the grating are vertical). The focusing should be as sharp as you can get it. A useful **hint** is to take the spectrometer carefully into the main lab and adjust the telescope by focusing on a distant object on the horizon. Then adjust the collimator to focus the image of the slit. Record the angle of the telescope in the straight-through position, and subsequently always record the angles of a diffraction maximum on *both left and right sides*. Not only does this give you *two* independent measurements of the angle, but also since the average of left and right readings should be the initial straight-through reading, it also gives you a valuable *check* against blunders *and* an estimate of your measurement accuracy.

- At this point you should check that the grating is accurately perpendicular to the light falling on it from the collimator. Think of a way to do this.

Observation of spectral lines

- In order to calculate λ from measurements of the diffraction angle θ , we first need to know the grating line-spacing d . Calculate this from the number of lines per mm (or inch) which is engraved on the grating.

- Using the gas discharge tubes provided, measure a selection of prominent spectral lines of sodium, mercury, cadmium and hydrogen. Observe first, second and (where possible) third orders of principal maxima. Use the following list of prominent lines to *identify* them, and then *calculate their wavelengths*.

Element	Colour	Intensity
Sodium	green/yellow	weak
	green/yellow	weak
	yellow	very strong
	yellow	very strong
Mercury	violet	strong
	turquoise	weak
	green	very strong
	yellow	quite strong
Cadmium	yellow	quite strong
	blue	quite strong
	blue	strong
	green	strong
Hydrogen	red	quite strong
	blue	weak
	deep red	weak

- The two yellow lines of the sodium doublet near 589 nm, the famous ‘D-lines’ of sodium, should be well resolved if you have set up the instrument carefully. Be sure you identify the lines correctly, using the diffraction equation to decide whether the longer or the shorter wavelength is diffracted through the greater angle.
- Finally, look up the *true values* of the wavelengths of all the lines, and *compare* your results with these.

Chromatic resolution [*This section is optional; do it at the end if you have time*]

Diffraction theory shows that the angular separation of two nearby wavelengths is increased by (i) going to higher orders, as we have seen above, and (ii) increasing the *total* number of lines on the grating which contribute to the diffraction. If the Rayleigh criterion for angular resolution (see part A) is used, then it can be shown that the minimum wavelength difference that can be resolved is

$$\delta\lambda = \lambda/Nn$$

where N is the total number of lines and n the order. It is convenient to define the **resolving power** as the ratio $\lambda/\delta\lambda$ between the wavelength itself and the smallest difference that can be measured at that wavelength. The larger the resolving power the better the chromatic resolution. From the expression above, the resolving power equals Nn .

- The sodium D-lines should be well resolved when you use the full width of the grating. A simple way to vary the illuminated width is to clamp vernier callipers just in front of the grating and adjust the opening of the jaws. Find the narrowest opening that still allows you to resolve the D-lines with confidence. Evaluate the quantity Nn and compare with the value of $\lambda/\delta\lambda$. Repeat this several times for both orders. Do the experimental and the theoretical estimates of resolving power agree? Remember that the Rayleigh criterion is somewhat arbitrary.

[*End of optional section*]

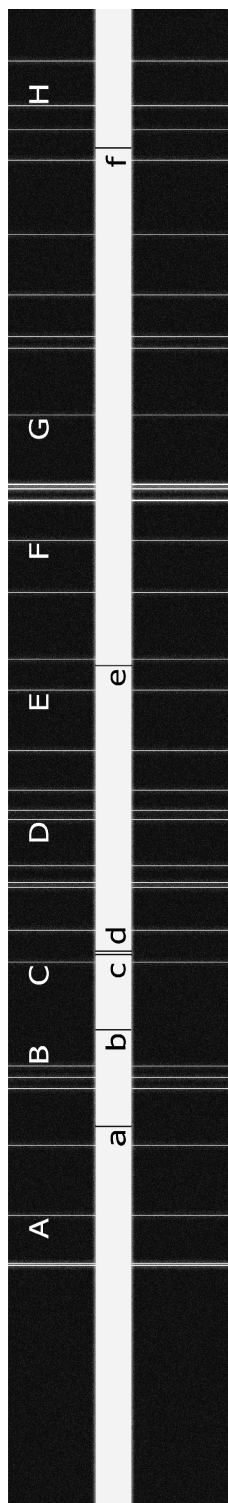
Doppler shift of spectral lines

A change in wavelength occurs when a source of light moves towards or away from an observer. The change, $\Delta\lambda = \lambda' - \lambda$, where λ is the normal wavelength and λ' is the changed wavelength, is proportional to the relative velocity v :

$$\frac{\Delta\lambda}{\lambda} = \frac{v}{c}$$

where c is the velocity of light. This is the Doppler shift. When source and observer are moving apart, v is positive, λ' is larger than λ and the light becomes redder. If they are approaching, v is negative and the light becomes bluer. The effect is very small but can be detected in the light from some stars and galaxies.

• Figure 6 is a photograph of part of the spectrum of iron, taken by focusing diffracted light from iron vapour onto a long strip of film. The principle maxima appear as bright lines, some of which are labelled A–H. Their wavelengths are well known, and are tabulated below. Superimposed across the centre of the spectrum of iron is that of a star, taken by directing the light from a telescope onto a spectrometer fitted with a camera. Some faint dark bands labelled a–f can be discerned; these are known to be lines of the elements hydrogen and calcium, whose wavelengths are also well known and tabulated below. However, because the star is moving relative to the Earth its light is Doppler shifted, so these stellar spectral lines do not appear at exactly the wavelengths measured in the laboratory. From the information below, and careful measurements of the positions of the stellar lines relative to the iron lines, find the apparent wavelengths of the lines a–f, deduce whether the star is moving towards or away from us, and estimate the relative velocity.



Stellar lines:

Line	λ (nm)
a'	388.90 hydrogen
b'	393.38 calcium
c'	396.86 calcium
d'	397.01 hydrogen
e'	410.17 hydrogen
f'	434.05 hydrogen

Iron lines:

Source	Line	λ (nm)
A		388.71
B		395.67
C		400.52
D		407.17
E		413.21
F		420.20
G		426.05
H		440.48

The lines a'–f' given in the table above are unshifted and correspond to the shifted lines a–f in figure 6.

Figure 6 Doppler-shifted spectrum

Part C: The expansion of the Universe

Introduction

In this part you will use the ideas and methods of parts A and B to measure the distances and the velocities of five distant galaxies, from photographs and spectra. A plot of distance versus velocity should then show that the two are related, a finding first made by the American astronomer Hubble in 1929. Hubble found that the spectra of many distant galaxies are shifted towards the red; he interpreted this as a Doppler shift and showed that the apparent velocity of recession increased linearly with the distance of the galaxy. Although there continue to be arguments about the value of the constant of proportionality (**Hubble's constant**) in this linear law, and even about the interpretation of the red shift as due to motion, there is no doubt about the observations. The quality of the material you will have to work with is considerably better than that available to Hubble.

The distance of the galaxies

Finding astronomical distances is difficult. This is the method we will use for these distant galaxies. Galaxies tend to form large clusters containing perhaps hundreds of members, some big and some small. It is found that in nearby clusters whose distances can be measured reasonably accurately by other means, the *brightest* elliptical galaxy is usually about 30,000 parsecs in diameter. (Galaxies are broadly classified as spiral or elliptical — the latter are often sufficiently close to spherical that it makes sense to refer to a single value for the diameter.) If we assume that this is true for all galactic clusters, then we can estimate the distance of a far-off cluster by measuring the apparent angular diameter θ of its brightest elliptical member and inferring the distance D from the expression (figure 7): $D = 30,000/\theta$ parsecs, where θ must be in radians.

- The scale of the photographs is given by the barred line, which is 150 arcseconds long. Measure the diameter of each galaxy's image, taking the mean of several measurements in different directions, and deduce the angular diameter of the galaxy in the sky. Convert this from arcseconds to radians, and deduce the distance D to the cluster.

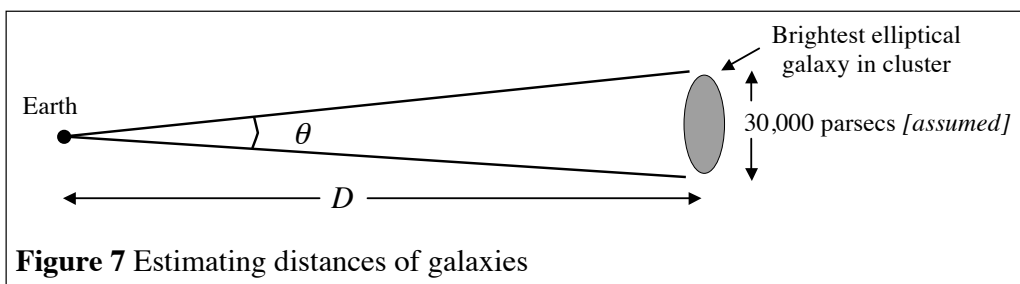


Figure 7 Estimating distances of galaxies

The velocity of the galaxies

The spectrum of each of the five galaxies shows two strong dark lines, prominent in elliptical galaxies and due to calcium. These, usually called the calcium H and K lines, are the lines b and c in the stellar spectrum you measured in part B, where their wavelengths are given. The arrows on the photographs show the extent to which these lines have been shifted towards the red. The comparison spectra above and below the galactic spectra are of helium. The lines marked $a-g$ in these spectra have the following wavelengths:

- By measurements similar to those in part B, find the **redshift** $\Delta\lambda/\lambda$ of the calcium lines. Hence calculate the recessional velocity of the galaxy.

Line λ (nm)

<i>a</i>	388.87
<i>b</i>	396.47
<i>c</i>	402.62
<i>d</i>	414.38
<i>e</i>	447.15
<i>f</i>	471.31
<i>g</i>	501.57

Hubble's law

- Plot a graph of recessional velocity v versus distance D for the five galaxies, and hence determine a value for H in the expression for Hubble's law:

$$v = HD$$

Give your value for H , which is **Hubble's constant**, in its usual units of $\text{km s}^{-1} \text{Mpc}^{-1}$.

- Use your value of H to estimate the age of the Universe.

Laboratory Exercise 8 – LIGHT AND OTHER ELECTROMAGNETIC WAVES

In the three parts of this exercise you will study some of the properties of electromagnetic waves. Whatever their wavelength, all e.m. waves travel at the same speed in a vacuum, can be reflected, refracted and scattered, and show interference effects. All these and more will be dealt with in detail in future physics courses; here we demonstrate some of them using two very different wavelengths, from 10^{-10} m (X-rays) through 10^{-6} m (visible light). The techniques used are different but the phenomena you will see are the same, although on quite different scales.

Part A: Refraction of light

Introduction

Newton found that a glass prism separated white light into its spectral colours. In this part you will carry out a refined version of Newton's experiments using a **prism spectrometer** to measure the deviation of light of different wavelengths when it passes through a prism. The property of bending light is called **refraction**, and is measured by the **refractive index**, μ . (The refractive index is actually the ratio of the speed of light in a vacuum to the speed in the transparent medium.) When μ , and hence the angle of bend, varies with the wavelength λ the effect is called **dispersion**. Every transparent material exhibits dispersion. In some optical instruments this can be a nuisance, producing coloured effects when white light is used, but in the prism spectrometer a large dispersion is useful because it improves the instrument's **resolution**, or ability to separate two close wavelengths.

The refractive index of colourless transparent materials decreases as the wavelength increases. This behaviour is called **normal dispersion**, although we now know that the opposite behaviour, naturally called anomalous dispersion, is just as common. In 1836 the mathematician Cauchy suggested that normal dispersion was well described by the expression:

$$\mu = A + \frac{B}{\lambda^2}$$

Here A and B are constants for the material. In fact Cauchy added a third term on the right, C/λ^4 , but this is very small and usually ignored.

In this part you will check Cauchy's formula using a number of spectral lines whose wavelengths are accurately known, and you will do it for glass prisms of both low and high dispersion. Although there are some exceptions, denser glass has higher refraction, and greater dispersion, than less dense glass and the commercial names for different glasses (Light Crown, Extra Dense Flint and so on) reflect this. The usual way to describe the optical properties of spectrometer glass is to list not the Cauchy constants A and B , but some combination of the refractive indices at certain specified wavelengths. By measuring these you should be able to use a table of glass types to find out what type of glass your prisms are made of.

The refractive index is measured as follows. Light passing through the prism is deviated by an angle D . As the prism is rotated some position is found at which D has a minimum value, D_{\min} , the **angle of minimum deviation**. Textbooks on optics show that

$$\mu = \frac{\sin \frac{1}{2}(A + D_{\min})}{\sin \frac{1}{2}A}$$

where A is the angle of the refracting edge of the prism — see figure 1. Light passes symmetrically through the prism when it is refracted through the angle of minimum deviation.

Setting up the spectrometer

This instrument is capable of considerable precision when properly adjusted, so it is worth spending a little time doing it carefully. The aim is to make light from the slit parallel as it passes through the **collimator** and to bring this parallel light to a focus on the cross-wires of the **telescope**'s eyepiece. The prism on its table must refract this parallel light in a plane perpendicular to the common rotation axis of table, collimator and telescope. Follow each step of the following simplified procedure in sequence — refer to figures 1, 2 and 3.

- Looking through the telescope eyepiece against a bright white background, bring the cross-wires into sharp focus.
- Carefully take the instrument into the main lab. Adjust the objective lens of the telescope relative to the eyepiece/cross-wire combination so that the image of a distant object on the horizon is sharply focused on the cross-wires. The telescope is now adjusted.

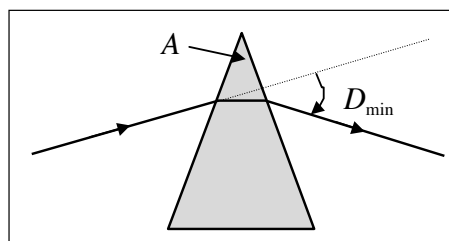


Figure 1 Definitions of A and D_{\min}

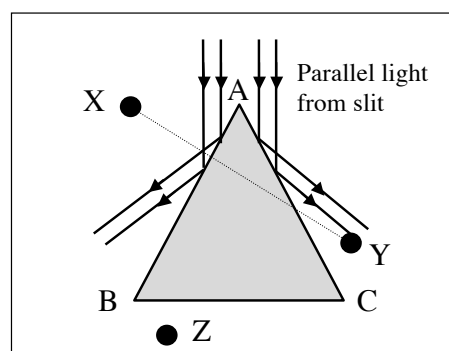


Figure 2 Setting up the spectrometer

- Illuminate the slit with a sodium lamp and view it through the telescope in the straight-through position, without a prism. Adjust the collimator to give a sharp image of the slit, and make this as narrow as possible while still passing light along its whole length. Make sure the slit is precisely vertical. The collimator is now adjusted.
- Place the prism on the table with its refracting edge pointing to the collimator and one of its refracting faces AB perpendicular to the line XY joining two of the levelling screws, as in figure 2. View the slit by reflection in AB, and by adjusting X and Y centre the image on the telescope's cross-wires. Repeat this for reflections in the face AC but this time adjust *only* screw Z. Check the reflection in AB and make small adjustments to X and Y, going through the sequence again if necessary until both faces reflect the light centrally down the telescope. The prism on its table is now adjusted.

NOTE: The prisms are precision optical pieces. Handle them only by the top and bottom triangular surfaces. Do not touch the refracting faces. If they get finger-marked, ask for an optical wipe and clean them carefully.

Hint: In what follows you need to view an image of the slit after the light has reflected off, or passed through, the prism. **Always first** push the telescope out of the way and use your naked eye to see the image, **then** when you know what to look for and roughly where it is, look through the telescope. *The commonest cause of frustration in optical measurements is squinting through an instrument when it's pointing in the wrong direction.*

Measuring the refracting angle A

- With the prism as shown in figure 2 and the prism table securely clamped, set the image of the slit, reflected in face AB, as accurately as you can across the intersection of the cross-wires, clamping the telescope and using the fine movement screw for the last delicate adjustment. Record *both* vernier scales which register the telescope's angular position. Unclamp the telescope and set it equally carefully to view the slit reflected in face AC, and again record its angular position on both scales. The *difference* between the two settings is $2A$. Take the average of your two determinations.

Measuring the angle of minimum deviation

- Rotate the prism table to a position like position 1 in figure 3, where you judge the light to be passing roughly symmetrically through it. Use your naked eye to see the image of the slit — the yellow spectral lines of sodium will be prominent but you will also see fainter lines of other colours. Using the telescope, view the yellow lines (you may be able to resolve the two lines — if so choose the one of shorter wavelength, which is deviated more than the other)

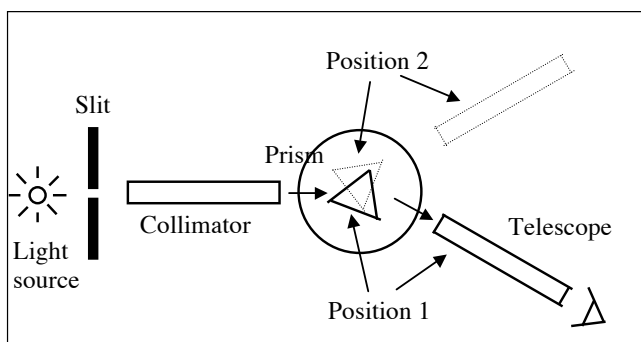


Figure 3 Measuring angle of minimum deviation

while you rotate the prism table back and forth; the image will move towards and then away from the straight-through position. It is at its most forward position when the light undergoes minimum deviation.

- Clamp the table and use the fine adjustment screw to set the precise position of minimum deviation. Set the cross-wires exactly on the image and check again that the deviation is a minimum. Record the vernier readings of both telescope scales.

- Now rotate the prism table to position 2, where the deviation is in the opposite direction, and repeat the measurements on this side.

- The angle turned through by the telescope between positions 1 and 2 is $2D_{\min}$. Take the average of your determinations and, with your knowledge of A , find the refractive index for the yellow sodium lines.

- Replace the sodium lamp with mercury, cadmium and hydrogen lamps, and repeat your measurements of D_{\min} for the brighter spectral lines from each. The wavelengths can be obtained from the laboratory technicians. Plot μ versus λ *as you take your measurements*. [Students who do this later and find that a point lies well off a smooth curve because they have misidentified a spectral line have no excuse. *Take longer in the lab* to calculate and plot graphs.]

- By means of another suitable graph, attempt to verify Cauchy's relationship.

Dispersion in different glass

- Replace the first prism by the second, made of a different, denser, glass. Repeat the measurements you have just made so as to obtain values for the refractive index for the following spectral lines: sodium at 589.0 nm (the D line), hydrogen at 486.1 nm (the F line), hydrogen at 656.3 nm (the C line). Hence calculate the **reciprocal dispersive power** V :

$$V = \frac{\mu_D - 1}{\mu_F - \mu_C}$$

for both prisms. With this information on the optical properties of the two types of glass, attempt to find a closely similar type in the table in Kaye and Laby.

Part B: The velocity of light

Introduction

In this exercise you will measure the velocity of a beam of light of wavelength λ about 500 nm, both in a vacuum and when travelling through transparent plastic. The methods used in exercise 7 are no use here. As you know, the speed of light, c , is about 3×10^8 m/s, so substitution in $c = f\lambda$ yields a frequency f of about 6×10^{14} Hz, which is too high to be measured by an electronic frequency meter even if the very short wavelength could be measured accurately. The trick here is to vary the intensity of the light source at a frequency very much less than f . The change in intensity is carried along by the light wave, travelling at the same speed c and appearing as a sinusoidal variation of amplitude with a very much longer wavelength than that of the **carrier wave** itself. In other words, the carrier wave is **modulated** by the lower frequency signal — see figure 4. [This is the way that AM (amplitude modulation) radio works. The signal to be transmitted is impressed on a constant *frequency* carrier wave as a change in *amplitude*. In FM (frequency modulation) radio the *amplitude* of the carrier wave stays the same but its *frequency* is modulated.]

The modulation frequency used here is 50 MHz or 60 MHz, depending on which apparatus you are using — if in doubt, check with a demonstrator. **Deduce** the corresponding modulation wavelength λ_m . The light is received by a photodiode detector whose response follows the modulation at 50 or 60 MHz. Even this reduced frequency is too high to be displayed on the oscilloscopes we use, so another electronic technique is used to reduce the frequency still further. A separate oscillator unit is tuned to a frequency slightly different from 50 or 60 MHz, e.g. 59.9 MHz. When this signal is mixed with a 60 MHz signal from the light source or photodiode the two ‘beat’ together producing a combined signal whose frequency is the *difference* of the two — this is called the **heterodyne** technique. The mixed signal has a frequency of ~ 100 kHz and can easily be displayed. Figure 4 shows that a 50 or 60 MHz signal which has travelled only a few metres to the receiver will be considerably out of phase with the signal from the source. A feature of the heterodyne technique is that it preserves this phase relation, so a measurement of the phase difference between the mixed low-frequency signal from the source and the mixed low-frequency signal from the receiver is a direct measure of the phase lag in the 50 or 60 MHz signal. Converting this phase difference to a fraction of the modulation period gives the time for the wave to travel a known distance to the receiver, and hence the wave velocity.

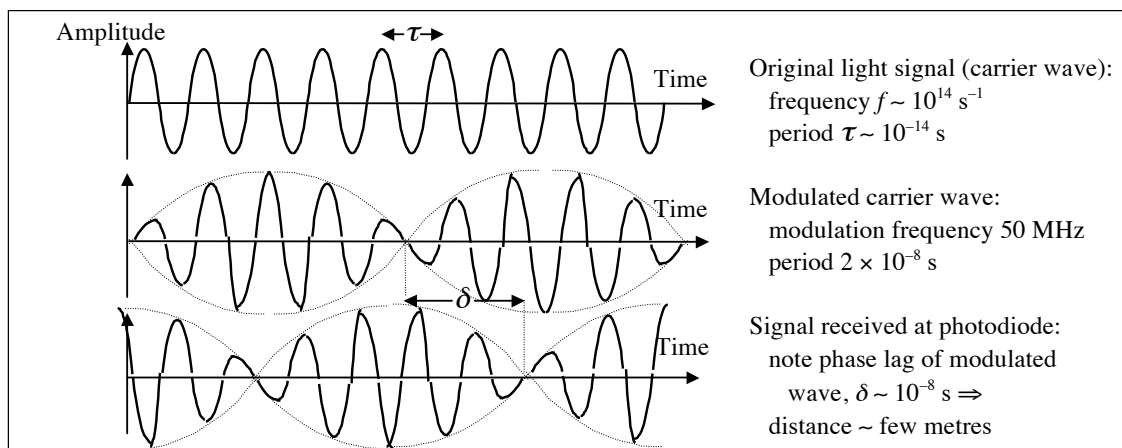


Figure 4 Amplitude modulation of carrier wave

Apparatus

This consists of a box containing the light source, modulator, receiver, and mixer, and a there-and-back path of a few metres for the light. The source is a light-emitting semiconductor diode (LED) which emits red light; a 50 or 60 MHz oscillator modulates the voltage across the LED. The light passes through a lens L1 (see figure 5) whose function is to produce a wide parallel beam which is sent down one arm of an optical bench and back along the other arm after reflection in two 45° mirrors M1 and M2. Another lens L2 focuses the parallel returning light onto a light-sensitive photodiode whose output signal oscillates at 50 or 60 MHz in phase with the returning light. Correct optical alignment is essential in this experiment, and it is worth spending some time getting it right.

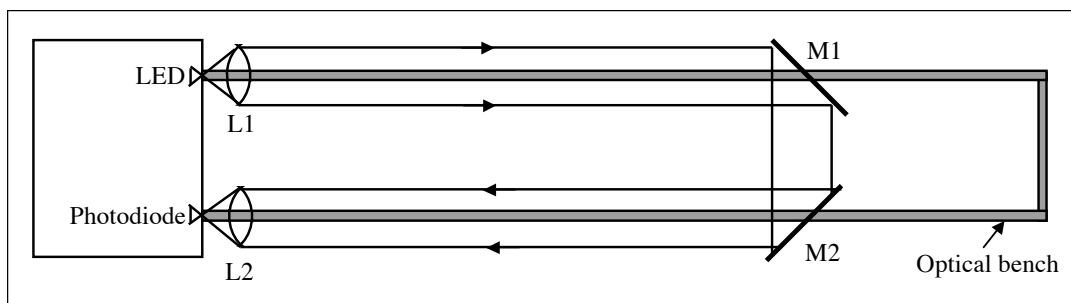


Figure 5 Apparatus

- Place L1 with its centre accurately level with the LED. Let the light that has passed through L1 fall on the screen carrying a circle of the same diameter as L1 and L2. This screen is used to trace the path of the light to the mirrors and back to L2. A truly parallel and aligned beam will exactly fill the circle and remain at the same height all the way along the path. This needs to be done in a darkened lab. Move the screen along the rail to M1 and adjust L1 as necessary to keep the beam parallel and on axis. Repeat with the screen on the return rail, adjusting only M1 and M2 by the screws on their back face (L1 should not need further adjustment if you have been careful) to bring the beam centrally onto L2, which should then be adjusted so as to focus the beam onto the photodiode. Final adjustments can be made by displaying the signals from the heterodyne mixer on the oscilloscope and maximising the response of the receiver relative to the transmitter. Ideally, the received signal should be the same size for all distances of the mirrors.

Measurements

- Use the XY mode of the 'scope to display the Lissajous figures formed by the outgoing and the incoming signal. Each mirror distance corresponds to a different phase lag between the two waves, so the figures in general are elliptical. If the sine waves have the same amplitude and are exactly in phase the ellipse becomes a straight line at 45° to the X axis; if they are 180° out-of-phase the line slopes the other way. A circle is produced at 90° phase difference. By adjustment of the channel gains you should be able to get 'scope traces of roughly the same amplitude in both X and Y. There is a phase control knob on the supply unit which allows you to select to some extent the light path which gives no apparent difference in phase (clearly this is necessary since the phase difference is actually zero only when the light path is zero, which is experimentally most inconvenient!). Check that the phase can be changed from 0° to 180° as the mirrors are moved along the length of the bench.
- When you are satisfied, make careful measurements of the spacing between the two mirror positions corresponding to 0° and to 180° phase difference. The extra distance travelled by the light between these two positions is a half-wavelength of the 50 or 60 MHz modulation wave, so the distance the mirrors move is one-half of this, that is $\lambda_m/4$.

- Repeat this measurement with a number of different settings of the phase control knob so as to get a good idea of the variability of the readings and hence their statistical error. Deduce the velocity of light in air. The uncertainty in your value will include a contribution from the statistical error just mentioned, and also a systematic error due to uncertainty about the precise frequency of the 50 or 60 MHz oscillator.

Speed of light in acrylic plastic

Two transparent acrylic rods (Perspex, Lucite and Plexiglas are trade names for this plastic) are to be placed on the outward and return rails close to the lenses. The mirror unit is brought close behind them and the phase control knob adjusted for a straight line display, whether 0° or 180° phase difference is immaterial. The rods are removed and the mirrors moved away until the *same* straight-line Lissajous figure is obtained. The distance between the two mirror positions represents the extra time taken for the light to travel the combined length L of the perspex rods in comparison with the same length of air.

To the light wave, the rods appear to be longer than the same actual length of air. This apparent length due to the slower light speed is called the **optical path**. It is equal to the actual length multiplied by the ratio of light speeds in air (almost the same as vacuum) and plastic, a ratio that as you know is the **refractive index** μ . So the movement S of the mirrors introduces an extra distance $2S$ equal to the difference between the optical path in plastic, μL , and the optical path in the same length of air which is the actual length L :

$$2S = (\mu - 1)L$$

- Make several measurements of the mirror movement needed, using slightly different initial settings, and deduce a value for μ . With enough measurements, ten or more, you will be able to assign a standard error to μ with some confidence. Hence find the velocity of light in the plastic.

The invariance of c

Suppose there are two sets of apparatus like this in the lab, at right angles to one another. Careful measurements of the speed of light in a vacuum were carried out by Michaelson and Morley using such a geometrical set-up (but a quite different technique). They had expected to find a difference because of the motion of the Earth, just as the measured sound speed on a windy day depends on the direction of the wind. The motion of the Earth through the ‘æther’ in which the light waves were thought to be travelling should have produced a similar effect. Michaelson and Morley found no evidence at all for this ‘æther wind’ effect; your measurements will not be precise enough for you to make such a claim. Independently, Einstein had concluded that the ‘æther’ is unreal, and that the speed of light is a universal constant independent of the motion of source or observer — a conclusion that led directly to his theory of relativity.

Part C: X-ray diffraction

Introduction

In this part we study what happens when electromagnetic waves pass through a regular three-dimensional array of small scatterers. The effects were first studied by the Braggs, father and son, who shone X-rays on crystals and found strong reflections in some directions but nothing elsewhere. They interpreted this as the effect of scattering not just from individual atoms but from whole sheets of atoms lying in parallel planes a distance d apart (see figure 6). When the glancing angle θ is such that the extra distance travelled by the X-rays between successive sheets of atoms, $2d \sin \theta$, is $n\lambda$, a whole number of wavelengths, all the scattered X-rays are in phase

and interfere constructively. Thus a large signal is seen in that direction as if the sheets of atoms had, like mirrors, reflected the X-rays. But this is not specular reflection — an enhanced signal is seen *only* in those special directions for which the so-called **Bragg condition**

$$2d \sin \theta = n\lambda$$

is satisfied. Knowing λ , the atomic spacing d can be found from these **X-ray diffraction** studies. This is the basis of much modern crystallography and molecular biology.

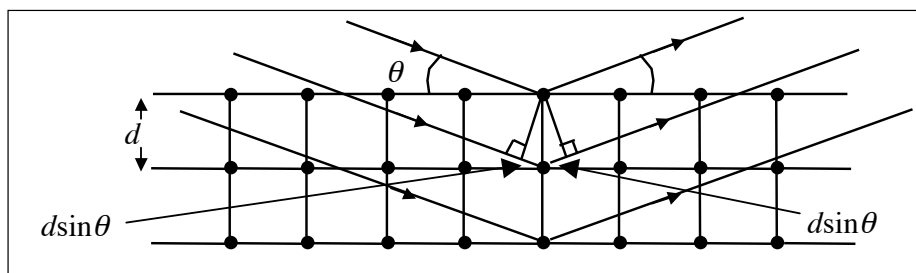


Figure 6 Scattering by a crystal

In this part you are given an X-ray diffraction pattern that has been recorded photographically. The pattern is produced by shining a narrow and tightly parallel (within 0.1°) X-ray beam on a thin wire of (in this case) tungsten, comprised of millions of tiny crystals lying in completely jumbled and random orientations (figure 7). The majority of these crystals are not in any special position and the beam passes through them unscattered. By chance, though, a few will be within about 0.1° of one of the Bragg angles, θ , relative to the direction of the beam, and these will give Bragg reflections at an angle of 2θ to the beam. So these reflected X-rays will travel outwards along a cone of apex angle 4θ , and will be recorded where they strike a strip of film wrapped around a cylindrical tube. The X-ray beam enters and leaves this tube through holes, and the whole arrangement is an **X-ray camera for powder diffraction** (the word ‘powder’ simply meaning that the target is not a single crystal but is made up of many tiny crystals). You can see on your film that the images are actually curved, since they are sections of a conical surface.

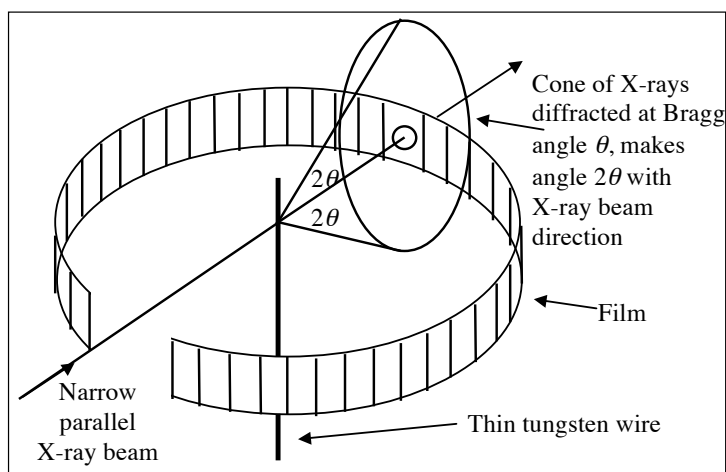


Figure 7 Powder diffraction

angles, θ , relative to the direction of the beam, and these will give Bragg reflections at an angle of 2θ to the beam. So these reflected X-rays will travel outwards along a cone of apex angle 4θ , and will be recorded where they strike a strip of film wrapped around a cylindrical tube. The X-ray beam enters and leaves this tube through holes, and the whole arrangement is an **X-ray camera for powder diffraction** (the word ‘powder’ simply meaning that the target is not a single crystal but is made up of many tiny crystals). You can see on your film that the images are actually curved, since they are sections of a conical surface.

Measurements

The diameter of the camera used to take these photographs was 57.30 ± 0.02 mm. A measurement of the spacing between corresponding images to left and right of the beam hole is a measurement of the arc length around the original cylinder which, divided by the camera’s radius, will give the apex angle 4θ in radians. Note that the diameter of the camera is carefully chosen so that exactly *half* of the left-right separation, measured in mm, is numerically equal to the angle θ in degrees (check that you understand this).

- **Do not remove the film** from its envelope. Tape it down onto a sheet of paper, fix an accurate ruler along the equatorial line of the film, and measure the coordinates of corresponding lines to the left and to the right. Try to estimate distances to one-fifth of a millimetre division. Tabulate your readings and evaluate the angle θ for each line. Note that the *sum* of the left and right

readings for each line should be constant — the extent to which it varies gives you an idea of the measurement inaccuracy and hence the error in the results.

Evaluation

You will have about eight values of θ , each corresponding to Bragg reflection. All these are first-order, i.e. $n = 1$. Each corresponds to a different value of d , the atomic spacing. To see this study figure 8, which shows a square crystal lattice of side a in just two dimensions. Besides the lines of atoms, a apart, running vertically and horizontally, there are other lines, less densely packed with atoms, running at angles as shown. A little geometry using Pythagoras' theorem will convince you that the spacings between these various sets of lines are all of the form

$$d = a/\sqrt{h^2 + k^2}$$

where h and k are small integers — they are the number of repetitions of the simple **unit cell** that defines the direction of the lines. In the examples shown:

$$h = 1 \text{ and } k = 0 \Rightarrow d_1 = a/1 = a$$

$$h = 1 \text{ and } k = 1 \Rightarrow d_2 = a/\sqrt{2}$$

$$h = 1 \text{ and } k = 2 \Rightarrow d_3 = a/\sqrt{5}$$

and so on.

A similar rule applies in three dimensions. The spacings between sheets of atoms responsible for each of the Bragg reflections are:

$$d = a/\sqrt{h^2 + k^2 + l^2}$$

where h, k and l are all small integers. Inserting this expression into the Bragg equation with $n = 1$, squaring and rearranging gives

$$\sin^2 \theta / (h^2 + k^2 + l^2) = \lambda^2 / 4a^2$$

That is, when divided by the appropriate small integer ($h^2 + k^2 + l^2$), each value of $\sin^2 \theta$ gives the same value, namely $\lambda^2 / 4a^2$. Hence knowing λ , the atomic spacing a can be found.

- You can do this by trial and error, finding what sequence of integers gives the same value, within errors, for each diffraction image. Or you can seek, again by inspection, the **highest common factor** of your $\sin^2 \theta$ values — make an informed guess at the number which will divide each of them an (almost) exact whole number of times. Having found the highest common factor roughly, divide each entry in your table by the *exact* whole number, which is the value of ($h^2 + k^2 + l^2$) for that line, and average the quotients. (For reasons which do not concern us here, not all such numbers are possible.)

- The X-rays used to take this photograph had wavelengths of 1.540562×10^{-10} m and 1.544390×10^{-10} m, one being rather more intense than the other. You may be able to distinguish separate Bragg reflections from the two wavelengths. Deduce the atomic spacing a in tungsten.

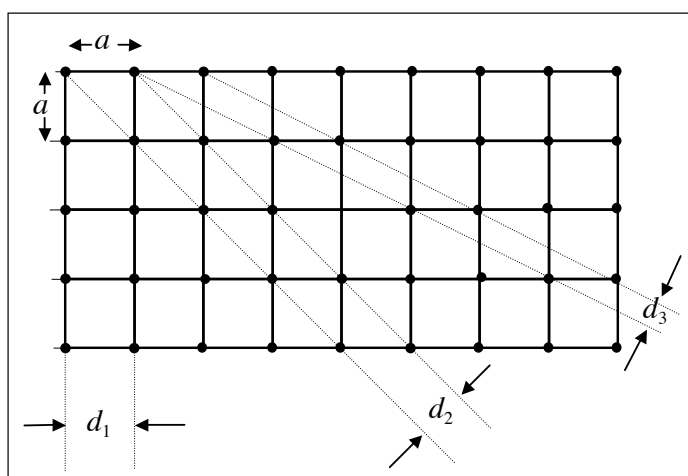


Figure 8 Unit cells

Laboratory Exercise 9 – FUNDAMENTAL and SUBATOMIC PHYSICS

The first and last parts of this exercise are concerned with two subatomic elementary particles — the **electron**, which is stable, and the **pion**, an unstable particle that lives fleetingly before it decays. Both measurements are based on the concept of momentum, so although the middle section on kinematics may seem incongruous, all three parts are linked by similar principles.

Part A: The charge-to-mass ratio of the electron, e/m

Introduction

A force acts on a charged particle moving in a magnetic field, tending to push the particle sideways. [The same force is responsible for the movement of a current-carrying conductor between the poles of a magnet, the basis of electric motors.] The force acting on an electron of charge e and mass m moving with velocity v in a field B is equal to Bev . Introducing the momentum $p = mv$, this becomes Bep/m . If the magnetic field is constant the electron undergoes a constant sideways acceleration. That corresponds to motion in a circle. Equating the centripetal force to the product of mass and inward acceleration v^2/r gives

$$\frac{Bep}{m} = \frac{mv^2}{r} = \frac{p^2}{mr}$$

so the momentum is

$$p = Bmr \left(\frac{e}{m} \right)$$

The kinetic energy of the electron comes from acceleration through a potential difference (voltage) V . Equating potential energy lost to kinetic energy gained gives

$$eV = \frac{1}{2}mv^2 = \frac{p^2}{2m}$$

Substituting for p we find

$$\left(\frac{e}{m} \right) = \frac{2V}{B^2 r^2}$$

so the charge-to-mass ratio can be found by measuring the radius of the circular path, and knowing both the magnetic field and the accelerating voltage.

The apparatus comprises a spherical bulb or ‘tube’, with ‘guns’ to produce and control a narrow beam of electrons, surrounded by a pair of **Helmholtz coils** for producing the magnetic field within the bulb. There are separate power supplies for tube and coils.

Electron beam tube

The electrical features, shown in figure 1, are: a **cathode** which produces electrons, an **anode** which accelerates them through voltage V , and an additional **deflector** electrode which can steer the beam slightly. Connections are made through the base and neck of the tube. The tube is almost completely evacuated, allowing the electrons to travel freely with few collisions with gas atoms. A small quantity of helium is present; its atoms emit greenish light if ionised by collision, so making the beam visible when the ambient light level is low. Part of the glass bulb is coated with

luminescent paint. There are two independent electron beams selected by a switch on the base of the tube, one directed across a diameter of the bulb, the other directed tangentially upwards. You will use the tangential beam, which can be bent into a complete circle by the magnetic field. Before switching anything on identify all the components, trace the circuit connections, and follow the instructions carefully.

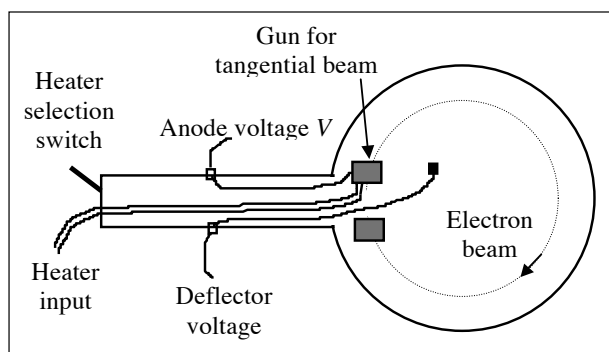


Figure 1 Electron beam tube

Each electron gun comprises an indirectly heated cathode and a conical anode with a small hole to allow the electrons to emerge; the deflector electrode nearby controls either beam. There are separate controls and meters for the anode (0–300 V) and the deflector electrode (0–50 V). When the power supply is switched on, current flows through the heater of whichever electron gun is selected. The heaters must *not* warm up while there is an accelerating voltage on the anode, so *before switching on make sure that the anode voltage is zero* by turning the control fully anticlockwise. **Failure to do this can cause serious damage.** Likewise, *reduce the anode and deflector voltages to zero before switching off.*

Helmholtz coils

A pair of coils arranged as in figure 2 and carrying equal currents I produce a substantially uniform magnetic field in the region between them, which is where the electron beam is produced. The essential feature of this **Helmholtz pair** geometry is that the separation of the coils is equal to their average radius R . If there are N turns of wire on each coil then the magnetic field intensity is

$$B = \frac{32\pi NI}{R\sqrt{125}} \times 10^{-7} \text{ teslas}$$

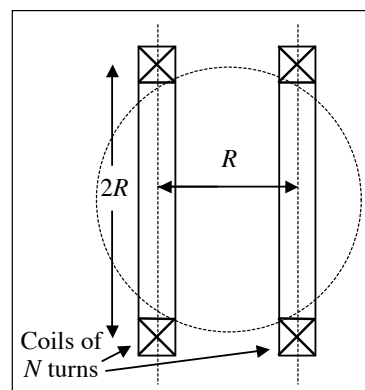


Figure 2 Helmholtz coils

These coils are wound with 320 turns of enamelled copper wire, and their mean diameter is 13.6 cm. Hence their radius is 6.8 cm.

Make sure, using convenient spacers or in other ways, that the coils are parallel and that the separation between the middle of their coil windings is also 6.8 cm. Currents up to 1 A are provided from a separate power supply.

Measurements

- With the Helmholtz coil supply off and the anode voltage zero, switch on the electron tube supply. **Wait one minute** before applying voltage to the anode. Raise the voltage slowly. At about 50 V you will see the first sign of the beam, by 70 V it should have enough energy to cross the tube, and by 100 V it is a bright filament. Switch on the Helmholtz coil supply and increase the current. If the beam is a spiral rather than a circular arc, the coils and the tube may be slightly misaligned. This is difficult to correct (with this apparatus) but you might try turning the whole assembly around — the extra deflection may be due to the effect of a local magnetic field. A small spiral deflection doesn't matter. A voltage can be applied to the deflector plates to make sure that the electron beam leaves the anode in the right direction, but you will probably not need to use this control much.

- Measuring the diameter of the circular beam is difficult. You can hold a transparent ruler in front of the tube, or try placing a mirror behind the tube and lining up a ruler with its reflection, or even try a travelling microscope if the electron beam is narrow.

When combined, the expressions for the value of e/m and for B yield

$$\frac{1}{d} = \left(\frac{16\pi N}{5 \times 10^7 R} \sqrt{\frac{e}{10mV}} \right) I$$

where d is the diameter of the circular beam. Hence a graph of $1/d$ versus coil current I , for constant accelerating voltage V , should yield a straight line from whose slope e/m can be found. (Note that I should be plotted on the x -axis and $1/d$ on the y -axis because we know I quite precisely, while $1/d$ has substantial measurement errors.)

- Use at least two values of V , but preferably three or four, starting near 100 V and going no higher than about 270 V. For each value of V , increase I and measure d for each of about ten diameters, aiming for about 5 mm intervals. The largest value should have the beam skimming the phosphorescent screen, the smallest can be about 40 mm. The beam may be very blurred, so its centre line is uncertain. Keep a close check on V and correct it if it tends to drift.
- If you are having problems obtaining data as above because d is difficult to measure, try keeping d constant while varying V and I to give a sufficient number of data points.

Analysis

- Plot $1/d$ versus I for each value of V . From the gradient of each of your graphs deduce a value of e/m . Average these, and estimate the uncertainty of your answer. Considering the difficulty of the d measurement, this experiment can give results remarkably close to the accepted value of the electron e/m , which is $1.76 \times 10^{11} \text{ C kg}^{-1}$. For example, from readings taken with anode voltages of 90 V and 180 V a Course Organiser obtained 1.78×10^{11} and $1.69 \times 10^{11} \text{ C kg}^{-1}$ for lines drawn through the origin.
- If all of the magnetic field comes from the Helmholtz coils, then your straight lines should pass through the origin. Consider whether or not this is the case. If it is not, then the electrons are being bent even when $I=0$, possibly by a small and constant additional magnetic field B_0 due to the Earth's field or to iron in the lab. If I is such that it produces a field B equal and opposite to B_0 then there will be no bending, so d is infinitely large and $1/d$ is zero. This corresponds to where the straight line crosses the x -axis. To deduce B_0 , use the value of this x -intercept: set B_0 equal to the value of B from the Helmholtz coils and so show that at this point

$$B_0 = -\frac{32\pi N}{\sqrt{125} \times 10^7 R} I$$

- Use this expression to deduce B_0 if your data appear to justify it.

Part B: Kinematics on a linear air track

Introduction

The linear air track has a number of holes through which air is blown to support metal riders or 'vehicles'. It provides an almost friction-free means of studying collisions and investigating the principles of classical mechanics as formulated by Newton. What we study here are the properties of objects in motion, that is **kinematics**. Kinematics is an extension of **statics** (bodies at rest) and is concerned with laws that apply whatever the forces that cause the motion. The study of the forces and their effects is **dynamics**.

This neat distinction works excellently in everyday life, but Einstein showed that it is an approximation only true in our sluggish low-speed world. In the relativity theory which has replaced Newtonian mechanics as an *exact* theory, statics does not exist because nothing is at rest

in every frame of reference. Einstein's kinematics is more complicated than Newton's, but if the dynamics is formulated carefully it contains the same great fundamental laws of classical mechanics — the laws of conservation of energy and momentum. The linear air track is a good place to study these.

Setting up

The manufacturer's booklet has a description of the apparatus. For measuring the speed of the vehicles we use a lamp and a light-sensitive detector, arranged so that a card on the vehicle breaks the light beam for a time which is measured digitally to an accuracy of 1 ms. There are two of these systems per track so that the speed of two vehicles can be measured independently.

- Horizontal levelling is critical. After a rough adjustment using a spirit level, the air supply should be turned on and the final levelling carried out with an unloaded vehicle as an indicator. It is not enough simply to bring a moving vehicle to rest by altering the levelling, as this will over-compensate. After each adjustment the vehicle should be brought to rest by hand and any tendency for the vehicle to move one way or the other further compensated. Appendix 2 of the manufacturer's booklet shows that a difference in level of no more than 0.25 mm between the ends can be tolerated.

- The speed of the vehicles is measured by a photoelectric timing technique. Set up a lamp with a focusing lens on one side of the track and the light sensitive detector (photodiode) on the other, 10 to 15 cm away. The electronic timers are configured so that they start to count when the beam is broken by the card carried by a vehicle, and stop counting when the card has passed. To do this connect the light source to the supply and connect the START terminals together with a shorting lead, connect the photodiode to the STOP terminals, set MODE to 1, and select TIME. Adjust the light beam if necessary so that the timer is switched off when light falls on the photodiode, and on when it doesn't. Then push the vehicle, with card mounted, into the beam and note the positions where counting just starts and where it just stops. The distance the vehicle has moved is the *effective* length of the card. Do this for each vehicle.

Friction as a cause of speed loss

A moving vehicle bouncing back and forth between two tightly stretched rubber bands will eventually come to rest. It loses its kinetic energy both through viscous drag (friction) from the air pad supporting it, and also through loss of energy in each collision with the rubber bands. The second is not likely to be too important, especially as in most measurements the rebound is not recorded, but the first will cause the vehicle to slow down at a steady rate so it will introduce a definite uncertainty into most measurements. We must therefore study the speed loss in detail.

- Appendix 2 of the booklet shows that the effect of friction is to reduce the vehicle's speed after travelling a distance s from its initial speed v_0 to a lower speed $v = v_0 - Ks$, where K is a constant term involving the viscosity of air and the dimensions and mass of the vehicle. The effect of speed loss at rebounds is not dealt with in the Appendix, but you should try to **show** that if the same fraction of the vehicle's kinetic energy is lost at each rebound then the velocity after n rebounds will be $v = v_0 e^{-Cn}$, where C is another constant.

- Position a photodiode assembly about halfway along the track. Release a vehicle from one end by catapulting it at a moderate speed (less than 1 m/s) and measure its speed on each traverse. Plot speed versus n , the number of rebounds. If *friction* were the only cause of energy loss we would expect a linear relation between v and n (which is proportional to s). If the *rebounds* were the only cause of energy loss we would expect an exponential relation, in which case a graph of $\log v$ versus n would be a straight line. Probably both effects occur.

- Inspect your graph to decide whether it makes sense to treat the high-speed data as being dominated by rebound energy loss and the low-speed data by frictional energy loss. If so (experience suggests it does!) draw a straight line through the low-speed data, extrapolate it back to the origin, and find K from the gradient. The *difference* Δv between the measured and the extrapolated speed at each of these points represents the effect of rebound loss, which drops to virtually nothing after a few rebounds.
- Plot $\log(\Delta v)$ versus n to confirm this, and find C , the fraction of velocity lost in each rebound.
- The constant K is the loss in speed per metre of track, due to friction. You can use it either to correct your results, which might be quite difficult, or to estimate the error in speed measurements made over a given length of track.

Conversion of potential to kinetic energy

Your earlier work in physics, as well as exercise 2, has taught you that energy is needed to stretch a rubber band, that the energy stored is proportional to the area under the force/extension graph, and that if this graph is linear the area is proportional to the square of the total extension. On the air track this energy can be converted to kinetic energy $mv^2/2$, where m is the mass, allowing a check of these kinematic relations.

Here is an analysis of what to expect. Let x_0 be the unstretched length of rubber band, x its length when stretched between supports, and D be the difference between x and x_0 . From Hooke's law the potential energy stored is proportional to D^2 . When you pull back the band you extend it by an additional amount d , say, giving it a total extension of $(D + d)$. The potential energy stored is now proportional to $(D + d)^2 = (D^2 + 2Dd + d^2)$. So the difference should be proportional to the kinetic energy $mv^2/2$ transferred to the vehicle when it is released.

- If you substitute typical values for x and x_0 , and use Pythagoras' theorem to find a typical value for d , you may find that the third term in this bracket, d^2 , is small compared to the first two terms (show this!). This is convenient because if we drop the third term we find

$$\frac{1}{2}mv^2 \propto 2Dd$$

suggesting that a graph of v^2 against d should be a straight line. If on the other hand d^2 is not small enough, then it will be necessary to plot v^2 against $d^2 + 2Dd$. (This might be easiest using Excel.)

- Measure the velocity with which the vehicle is launched for various extensions of the rubber band, using the adjustable screw to position the band precisely. Take several readings at each extension to see how reproducible they are. Catapulting off-centre will introduce wobble; ignore any obviously wobbly launches. The extension of the band is calculated from the geometry of the triangle formed by the unextended band and the two halves of the extended band. Plot v^2 against d (or $d^2 + 2Dd$, see above) and comment.
- Couple two vehicles together, and confirm that for equal catapult extensions the kinetic energies $mv^2/2$ of all combinations are equal, in other words that v is proportional to $1/\sqrt{m}$.
- These measurements show, at best, that the kinetic energy acquired by the vehicle is *proportional* to the potential energy of the stretched band; they do *not* show equality. The value of the kinetic energy for a particular launch is easily calculated, given the mass of the vehicle. The potential energy can be measured by hanging a scale pan on a rubber band and finding the weight needed to produce the corresponding extension. Make this comparison for several extensions, and comment on the results.

Elastic collisions

This phrase is used to describe collisions with **no loss of energy**. Two vehicles fitted with repelling magnetic buffers will collide almost elastically as long as their relative speed is low enough — if not there will be an audible ‘clink’ as they collide.

- Place one large vehicle at rest in the centre of the track, launch the other towards it, and observe the sequence of collisions. If these, and the rebounds at the ends, were perfectly elastic the sequence would continue for ever. As it is, more energy will probably be lost in a rebound at one end than at the other, and the vehicles will eventually come to rest at that end of the track.

Your observations should clearly suggest that in a perfectly elastic collision between two equal masses, one of which is at rest, *all* the energy of the moving mass is transferred. The moving mass comes to a complete halt and the other rebounds at the speed of the mass which struck it. This is an example of the law of **conservation of momentum**: the initial momentum is mv and is concentrated in the moving mass, the other having no momentum. The conservation law states that the final momentum must also be mv . The law doesn’t indicate how this momentum is shared between the two vehicles, but since their masses are equal they could share the speed v in any proportion as long as their velocities add up to v . *But* the total energy must still be $mv^2/2$, and this too can be shared in any proportion as long as the squares of the velocities add up to v^2 . These two requirements, one from conservation of momentum and one from conservation of energy, are impossible to fulfil together unless one of the final velocities is zero. (Try it — convince yourself that if $a + b = c$ and $a^2 + b^2 = c^2$ then either a or b must be zero.)

- No actual measurements of speed were needed yet. Now replace the launch vehicle with the lighter one and catapult it towards the heavy vehicle at rest. Measure the speed v of the launch vehicle before collision and the speed V of the struck vehicle after collision. Use the laws of conservation of energy and momentum, as above, to calculate what the velocity V should be in terms of v and the masses of the two vehicles. Check the validity of this calculation for a range of initial velocities, making corrections for loss of speed due to friction if you think it necessary.
- Repeat, with the roles of the heavy and light vehicles reversed. These measurements are best done by two students working together.

Inelastic collisions

These involve some loss of kinetic energy, usually by heat and sound. Any deformation of the colliding vehicles will heat them up and cause loss of kinetic energy, which is why the phrase **rigid body** is often used in mechanics; it is an ideal non-deformable object that always collides perfectly elastically [hardened steel ball-bearings are close to perfectly elastic objects]. The easiest way to make the air-track vehicles collide inelastically is to fill the hollow end of one with plasticine and use a buffer with a pin on the other so that the two vehicles stick together when they collide. The squashy plasticine absorbs a lot of the kinetic energy. Add plasticine to each vehicle as necessary to keep it balanced when it slides, and so that the two heavy vehicles still have the same mass.

- Repeat the sequence of measurements under **elastic collisions**. Your data should show that although kinetic energy is not conserved in the collisions (we’ve made sure of that), *momentum is still conserved*. This is a very important result. Whereas energy can take many forms, is sometimes difficult to account for, and is eventually dissipated away as random heat (as you see even on the air track), momentum is recognisably the same quantity in all branches of physics and is not dissipated away. You might ask where the momentum of the catapulted vehicle came from. Can you provide a satisfactory answer?

Explosions

- The opposite of collisions, objects coming together, are explosions, objects flying apart when kinetic energy is released. Momentum is conserved in explosions, too. Hold two vehicles fitted with magnetic buffers close together so that they repel one another. With practice it is possible to release the vehicles from rest so that they fly apart without wobbling. When you can do this, measure their respective speeds for various ratios of mass. The ratios 1:1, 3:2, 2:1 and 4:1 are all possible. Check the conservation of momentum in each case.

Part C: The decay of the π -meson

Introduction

The π -meson is an unstable particle which exists for a few tens of nanoseconds before breaking up into two elementary particles. If the π -meson is electrically charged these particles are a **neutrino** and a **muon**, the neutrino being nearly massless and neutral (hence its name) and the muon having the same charge as the π -meson, either $+e$ or $-e$. In such transmutations of elementary particles the total energy is conserved, as always, but in calculations one has to be careful to include not only the kinetic energy but also the energy associated with the mass of the particles according to the Einstein relation $E = mc^2$. As we have seen in part B, momentum also is always conserved. In this exercise you will study a number of examples of break-up or **decay** of charged π -mesons; by applying the laws of conservation of energy and momentum the **mass** of the π -meson can be found.

The photographs

You are given some photographs of happenings in a bubble chamber, which is in a magnetic field so that charged particles move in spirals. The pictures show the gently curving tracks of a beam of positively charged π -mesons entering at the left, some of which stop in the chamber and then decay to positive muons (which are visible) and neutrinos (which are not). After travelling a short distance (due to receiving some kinetic energy from the decay of the π -mesons, which are heavier) the muons come to rest and also decay, after about 2 microseconds. The muons decay into three other elementary particles: two neutrinos, which are nearly massless, and a positron (the anti-particle of the electron) — of these only the positrons are visible in the chamber.

- Identify the decay sequences π -meson \rightarrow muon \rightarrow positron, which form unmistakable ‘hooked’ patterns in which the muons leave short tracks between the ends of the π -meson tracks and the start of the distinctive sharply-curving spiral tracks of the positrons. These curl up to smaller radius as they lose energy, and hence momentum, in the liquid filling of the chamber.

Applying the conservation laws

We now show that all that is needed to calculate the mass M of the π -meson is the kinetic energy T of the muon, together with a knowledge of the muon’s mass m . (In the same way only the velocity and mass of one of the air track vehicles was needed in order to calculate the complete kinematics of the ‘explosions’ in the last section of part B.)

First, however, we must modify the laws of classical mechanics so they work in the relativistic world of elementary particles travelling at nearly the speed of light. If a positron were travelling at a low speed v it would have kinetic energy $T = mv^2/2$ and momentum $p = mv$, and the relation between the two would be $T = p^2/2m$. But the positrons are travelling almost as fast as light, so their rest-mass energy must also be considered. The appropriate expression in relativistic mechanics for the total energy E of the positron is

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

which comes from applying Pythagoras' theorem to a triangle whose sides are the rest-mass energy mc^2 and the momentum times velocity of light, pc — see figure 3(a). [This expression is in fact *always* true, but in our sluggish non-relativistic world we usually don't realise it. Consider figure 3(b), where an arc corresponding to rest-mass energy mc^2 has been marked on the hypotenuse, the kinetic energy T making up the rest of the total energy. Then

$$\begin{aligned} T &= E - mc^2 \\ &= \sqrt{m^2c^4 + p^2c^2} - mc^2 \\ &= mc^2 \sqrt{1 + p^2/m^2c^2} - mc^2 \\ &= mc^2 \left(1 + p^2/2m^2c^2 + \dots \right) - mc^2 \end{aligned}$$

by binomial expansion; if the momentum p is very small only the first term of the expansion is needed, and the expression simplifies to $T = p^2/2m$, which is the familiar result already quoted.]

The relation $E^2 = m^2c^4 + p^2c^2$ gives the *total* energy E , including the rest-mass energy, in terms of the momentum p of a particle, and its mass; c is the speed of light. But the total energy E of the muon is also equal to the sum of its kinetic and rest-mass energies, $T + mc^2$. Since the neutrino has nearly zero mass its total energy can be taken to equal pc , where p is its momentum. Equating the rest-mass energy of the π -meson to the sum of the muon's and neutrino's energies (conservation of energy) gives

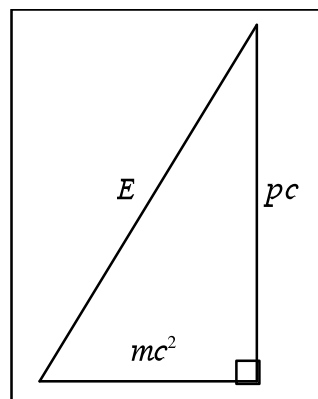


Figure 3(a)

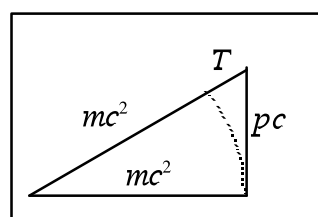


Figure 3(b)

$$Mc^2 = T + mc^2 + pc$$

The π -meson at rest has no momentum, so (conservation of momentum) the momentum of the neutrino must be equal (and opposite) to the momentum of the muon:

$$p = \frac{\sqrt{E^2 - m^2c^4}}{c}$$

After substituting $E = T + mc^2$ and eliminating p between the equations for conservation of energy and momentum we find

$$Mc^2 = mc^2 + T \left(1 + \sqrt{1 + \frac{2mc^2}{T}} \right)$$

showing as was stated that only T and m are needed to find M . The muon mass m is known from measurements to be 1.88×10^{-28} kg.

- You must find T from the photographs.

Finding the kinetic energy of the muon

A charged particle travelling in a bubble chamber loses energy by ionising the atoms it passes — the tracks you see are the trails of bubbles that form round each of these clusters of ionisation. The greater the rate of ionisation the faster the initial kinetic energy is lost and the shorter the track. Experiments have shown that the length L of the track of a particle with charge $\pm e$ and mass m is related to its initial kinetic energy T by the expression

$$T = 6 \times 10^{-12} \sqrt{L} \text{ joules}$$

Thus a measurement of the muon track length L , expressed in metres, gives T .

All the muons start off with the same kinetic energy and there is little variation in their rate of energy loss, so it should be clear that all the muon tracks should be the same length. But on your photographs they are obviously *not* all the same length. Each photograph shows only the projection of the track onto the plane of the film. The tracks of muons travelling towards or away from the camera will be foreshortened, so you have the task of reconstructing the actual track length from many projected tracks of different lengths. The actual track length is clearly longer than most of the tracks you measure. The Appendix shows that this problem has a simple mathematical solution. You measure a large number of track lengths L , find their average $\langle L \rangle$, and multiply by 1.273. The result is your best estimate of the actual track length.

Measurements

- Study about 15 photographs, which should yield about a hundred clear muon tracks. Use the scales printed on the clear acetate sheet to measure the lengths of as many as possible, taking care not to omit some very short tracks which are clearly extremely foreshortened; to do so would introduce a bias towards long tracks.
- Deduce a value for M , the mass of the π -meson. The accepted value is 2.49×10^{-28} kg.

APPENDIX

Consider a number of line segments each of length l , oriented at random in space as the muon tracks are. Suppose one of them makes an angle θ with the perpendicular to the plane of the film. Its projected length is $l \sin \theta$, which is the length L that you measured. There is an equal chance of the line segments pointing into any small element $d\Omega$ of solid angle, so the mean value of $L = l \sin \theta$ is found by integrating it over all solid angles and dividing by the total solid angle:

$$\langle L \rangle = \langle l \sin \theta \rangle = \frac{\int l \sin \theta d\Omega}{\int d\Omega} = \frac{\pi}{4} l$$

Derive this result for yourself. Hence $\langle l \rangle = (4/\pi)\langle L \rangle = 1.273\langle L \rangle$, as stated above.

Laboratory Exercise 10 – DIGITAL and ANALOGUE ELECTRONICS

In this exercise you will use electronic ‘chips’ to build some useful circuits. The first part illustrates the use of digital integrated circuits (ICs) with a counter and display, and the second extends this to making a rudimentary timer. The third part is a study of analogue operational amplifiers (‘op-amp’) ICs, while the last part unites the analogue and digital worlds by using an op-amp to build a simple digital-to-analogue converter.

Part A: Counter, decoder, and LED display

Introduction

In this section you will build a circuit for counting and displaying signals, using digital integrated circuits (ICs, or ‘chips’). ‘Digital’ means that the signals have only two discrete levels, which may be called true or false, high or low, on or off. For consistency, we use binary digits 0 and 1 to indicate low and high, and represent them by voltages 0 V (‘close to zero’) and +5 V (‘about five volts’). These levels are separated by a clear margin, so they don’t have to be generated very precisely. They are called TTL signals (from Transistor–Transistor Logic). We will use so-called CMOS chips (Complementary Metal-Oxide-Silicon), which are made up of many transistors used as switches. These chips are common and inexpensive. Some are very simple, but others are extremely complex. Their operation is summarised in the *CMOS Data Book*, of which we have copies in the laboratory.

The family of chips we will mainly use is the 4000 series, with numbers of the form 4XXX. They always need to be connected both to a power source, typically labelled V_{DD} and between +3 to +15 V (we always use +5 V), and to earth (0 V), labelled V_{SS} .

CMOS devices are easily damaged by electrostatic charge, due to their very high input resistance. When not plugged into a circuit board they are normally stored on aluminium foil, conducting plastic or conducting foam. Do not rub them on man-made fabrics, and it is often a good idea to wear an earthing strap to avoid static discharge when handling them.

So-called ‘**pinout**’ diagrams show where power, earth, and all logic signals must be connected. Note that the pin numbering, as viewed from *above*, always has pin 1 at bottom left and goes around the chip *anti-clockwise*. You can tell which end has pin 1 from a small dot or notch on the chip package. (Packages like this with two rows of pins are called dual in-line, or DIL.) The pinout diagram should indicate which pins are for input signals and which for outputs. If an **input** to a chip is *not used*, you *must* nearly always connect it to an appropriate level, either low or high, to be sure that the logic will work as planned — beware of any ‘floating’ connections.

The equipment for this exercise consists of a test box with a **breadboard** on top and **power supplies** inside. It also has **switches** to give voltages corresponding to **logic 1 when up** and **logic 0 when down**, and **light-emitting diodes** which **light up when a logic 1 (high) is applied**.

The top row of holes on the breadboard are all connected together, and so is the bottom row (see figure 4 in exercise 2). Normally, you should connect the top row of holes to the +5 V supply connector on the rear of the unit, and the bottom row of contacts to the earth connector. This allows fairly short, neat connections to any part of the breadboard.

Although very simple logic can be used to construct complex circuits, it is better to use ready-made chips that integrate entire tasks. This makes for easier design and construction of practical, reliable equipment. In this part we will count and display electronic pulses using counter, decoder, and display chips. We will start at the display end and work backwards.

Binary-coded decimal (BCD) representation

Digital logic is closely tied to the use of binary, or base 2, numbers. This is fine inside computers, but humans prefer decimal notation. One common way to do this is by using binary-coded decimal (BCD), which uses four binary bits for each decimal digit. In principle, four bits can count from 0 to 15 (i.e. 16 combinations) but in BCD the four bits, representing 1, 2, 4, and 8, only count from 0 to 9 — the combinations adding up to more than 9 are simply not used. (*If you are unfamiliar with binary notation or do not understand BCD then ask a demonstrator.*)

7-segment LED display

Light-emitting diodes (LEDs) behave like ordinary $p-n$ diodes. The direction of current flow is indicated by the diode symbol. The forward current through the diode must be limited by a series resistor. The current allowed varies with the size of the diode, but a typical value is $\sim 10\text{--}15$ mA. Thus for a +5 V supply a typical resistor would be $\sim 330\text{--}500$ Ω . (For rough calculations like this you can neglect the small resistance of the conducting diode.) LEDs are commonly available in red, yellow, green and infrared (used for remote controls and fibre optics). Blue LEDs are rarer and dearer.

Seven-segment displays for displaying numbers are very common. Each digit is made up of an array of seven LEDs in the form of strips (see figure 1), with each strip denoted by a letter between a and g . This arrangement also allows a few letters to be displayed, but this has problems since it confuses characters such as b and 6 , or D and 0 and O .)

Each of the LEDs must be supplied from a voltage source via a series resistor. The number of connections needed is minimised by having a common cathode connection to earth for all the diodes. Another simplification is to use a ‘resistor pack’ containing eight resistors in a DIL package similar to an IC. The resistors are connected between corresponding pins on opposite sides of the package.

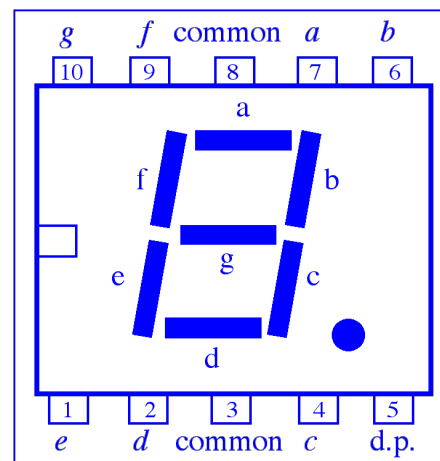


Figure 1 Segments and pinout of LED display

➤ Start to **wire up** the circuit shown in figure 2. Plug chips into the breadboard with their *pins on either side of the central gap*. Put the **display** at the *right-hand end* of the breadboard, and the **resistor pack** (150 Ω) next to it. Leave room at the left-hand end for the decoder/driver and counter chips, which each have 16 pins. Connect the *common cathode* to earth — two pins are provided but only *one* of them has to be connected. Ignore the decimal point (*d.p.*).

4511 Decoder/driver

This is a fairly complex chip; its pinout is given in figure 3. First, it **latches** (i.e. stores) binary-coded decimal (BCD) input data, so that the display remains lit even while the inputs (labelled A , B , C , D) are changing. Next, the BCD (1, 2, 4, 8) data are **decoded** into separate signals for each of the seven segments of the display. Finally, the signals are **amplified** by drivers. These can sustain ≤ 25 mA each, which is ok since the LED display segments must be limited to 20 mA. This would seem to imply resistors of at least 250 Ω , but in fact there is some resistance in the 4511 output driver, so that resistors of ≥ 150 Ω are acceptable. (A full specification for the 4511 is available in the *CMOS Data Book*.)

➤ **Connect the 4511** to the display via the resistor pack. **Take care** that the output connections are not shorted to earth or each other. Connect V_{DD} , \overline{BL} and \overline{LT} to +5 V, and V_{SS} and LE to earth.

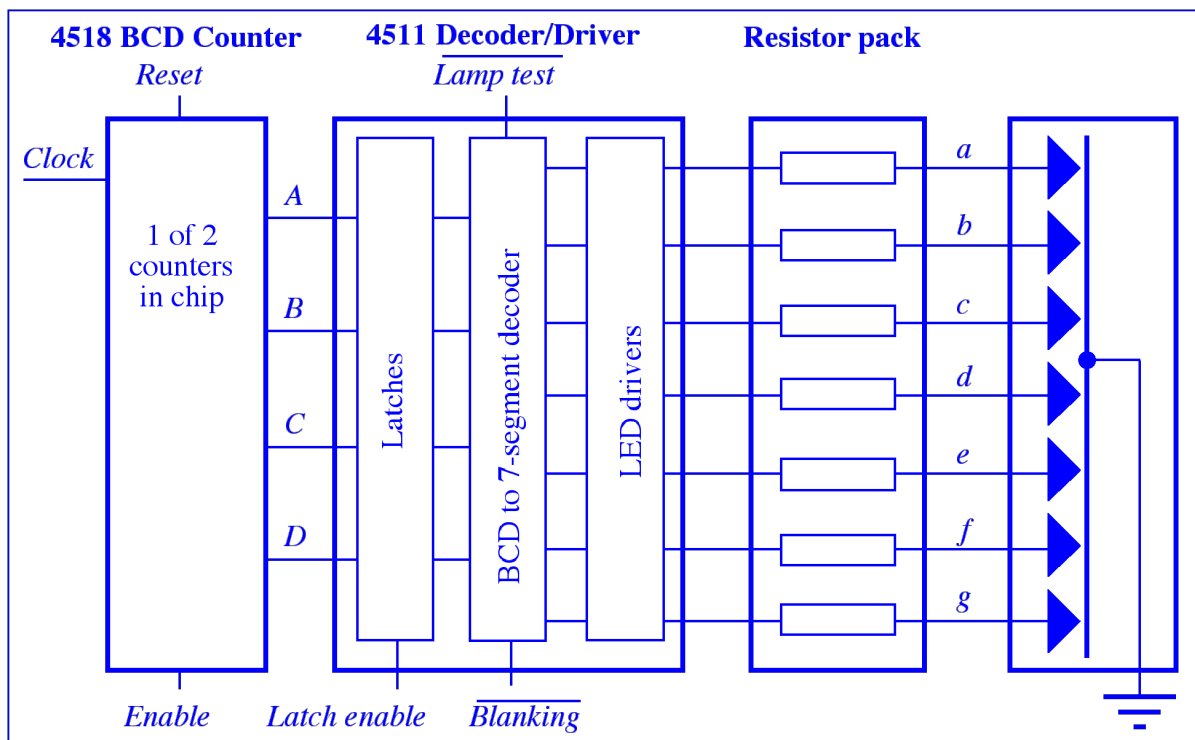


Figure 2 Overall layout of counter, decoder, and display circuit

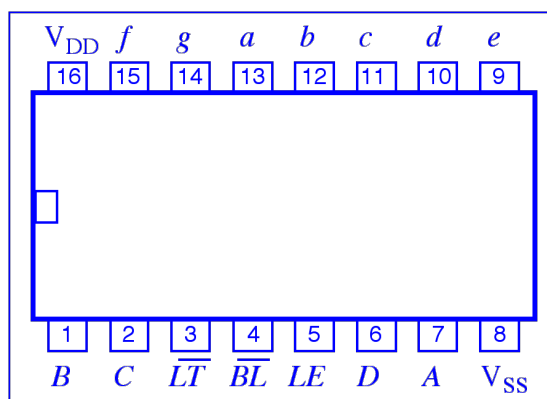


Figure 3 Pinout of 4511 decoder/driver chip

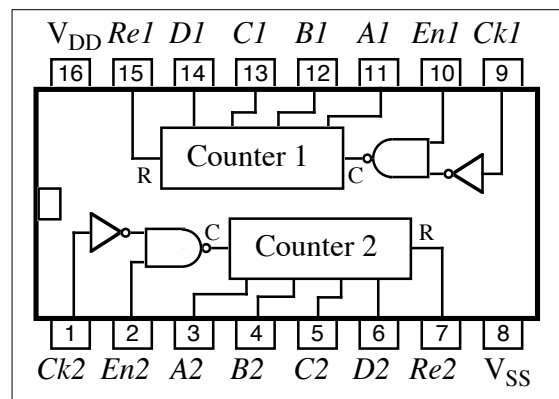


Figure 4 Pinout of 4518 dual decimal counter chip

➤ **Examine the behaviour of the 4511** using the test-box switches to supply the inputs A , B , C , D , with the least significant binary digit (A) on the right. **Check the truth table (table 1)** given below. (X means ‘don’t care’, in other words the signal can be either 0 or 1 without having any effect.) Since BCD goes only from 0 to 9, find out what happens if ‘invalid’ codes 10, 11, 12, 13, 14, 15 are applied.

➤ Be sure also to **check the blanking (\overline{BL}), lamp test (\overline{LT}), and latch enable (LE)** features of the chip to understand what they do. Note that \overline{BL} and \overline{LT} are ‘active low’ (i.e. they cause things to happen when they are at 0 V, not +5 V), while LE is ‘active high’.







4518 BCD counter

We will count so-called ‘clock’ pulses using half of a 4518 dual decimal counter. (We ignore the other counter for now.) This works by counting in binary, but some extra logic causes it to reset when it has counted to 10 so that we get BCD. Its pinout is given in figure 4, and full details are given in the *CMOS Data Book*. (In the book, A is called $Q1$, B is $Q2$, etc.) The counter increments either on positive-going edges of the *Clock* (Ck) input or negative-going edges of the *Enable* (En) input, and is reset by the *Reset* (Re) input. This is summarised in **table 2**.

Table 1 Truth table for 4511 Decoder/Driver

Inputs							Outputs							
<i>LE</i>	\overline{BL}	\overline{LT}	<i>D</i>	<i>C</i>	<i>B</i>	<i>A</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	Display
X	X	0	X	X	X	X	1	1	1	1	1	1	1	8
X	0	1	X	X	X	X	0	0	0	0	0	0	0	
0	1	1	0	0	0	0	1	1	1	1	1	1	0	0
0	1	1	0	0	0	1	0	1	1	0	0	0	0	1
0	1	1	0	0	1	0	1	1	0	1	1	0	1	2
0	1	1	0	0	1	1	1	1	1	1	0	0	1	3
0	1	1	0	1	0	0	0	1	1	0	0	1	1	4
0	1	1	0	1	0	1	1	0	1	1	0	1	1	5
0	1	1	0	1	1	0	0	0	1	1	1	1	1	6
0	1	1	0	1	1	1	1	1	1	0	0	0	0	7
0	1	1	1	0	0	0	1	1	1	1	1	1	1	8
0	1	1	1	0	0	1	1	1	1	0	0	1	1	9
0	1	1	1	0	1	0	0	0	0	0	0	0	0	
0	1	1	1	0	1	1	0	0	0	0	0	0	0	
0	1	1	1	1	0	0	0	0	0	0	0	0	0	
0	1	1	1	1	0	1	0	0	0	0	0	0	0	
0	1	1	1	1	1	0	0	0	0	0	0	0	0	
0	1	1	1	1	1	1	0	0	0	0	0	0	0	
1	1	1	X	X	X	X								

Table 2 4518 BCD counter actions

Clock	Enable	Reset	Action
	1	0	Increment counter
0		0	Increment counter
	X	0	No change
X		0	No change
	0	0	No change
1		0	No change
X	X	1	$A = B = C = D = 0$

➤ Connect earth, then connect the *A1*, *B1*, *C1*, *D1* outputs of the 4518 to the corresponding inputs of the 4511, and then connect +5 V. Connect *En1* to one of the test-box switches, set to 1, and *Re1* to another switch, set to 0.

➤ **Test** the counter by using a breadboard switch, connected to *Ck1*, as a source of input pulses. Check the operation of *Enable* and *Reset*. Then **speed it up** by changing the input to use an oscillator at a rate of less than 10 Hz. (If your oscillator will not go to a low enough frequency, you can use the other half of the dual counter, Counter 2, to divide the pulse rate by 10. Connect the output *D2* of Counter 2 to *En1* of Counter 1, and earth the *Ck1* input. The output *D2* goes from 1 to 0 whenever 10 pulses are counted, and this increments the second counter.)

DO NOT DISMANTLE YOUR CIRCUIT – IT IS NEEDED FOR PART B

Part B: A seconds timer

Introduction

In this section you will study the 4018 divide-by- N counter chip, which has facilities for dividing the number of input pulses by a programmable factor ranging between two and ten. You will then use it, together with a circuit to rectify a 50 Hz sine-wave voltage derived from mains electricity, to turn your counter into one that counts seconds.

4018 Divide-by- N counter

This is a 5-stage counter with facilities for dividing the input by any factor between two and ten. It is made more complex by having a 'preset' facility, to begin counting at any 5-bit (i.e. ≤ 32) binary number as well as zero. A pinout diagram and information on its division options are in figure 5, while a diagram showing how the logic works is given in the *CMOS Data Book*.

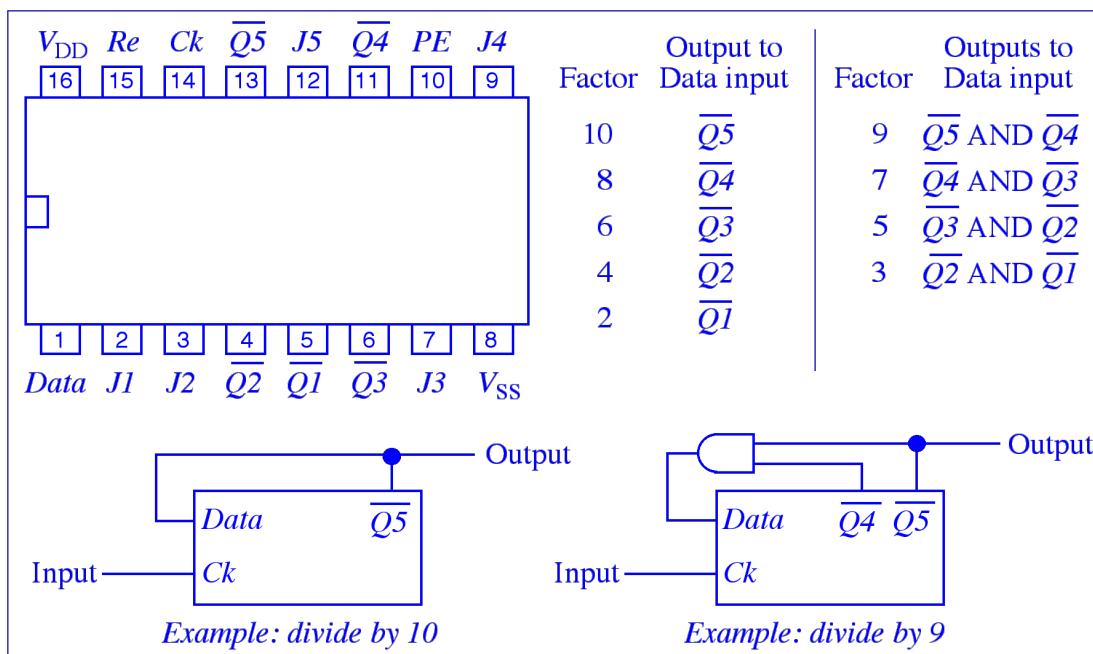


Figure 5 Pinout and division options for 4018 divide-by- N chip

The various divide options are selected by connecting different \overline{Q} outputs back to the Data input. For example, $\overline{Q5}$ selects division by 10. In each case, the divided-down output appears on the \overline{Q} pin that has been fed back to the input. For example, to see the input pulse train divided by 10 you use the $\overline{Q5}$ output. Note that division by even factors simply requires a connection, while for odd factors two outputs must go through an additional AND gate, e.g. 4081. (For odd-factor output, look at the higher-numbered \overline{Q} pin.) Examples are shown in figure 5.

For normal operation the *Preset Enable* (PE) is held low. To start counting at a non-zero value, *Preset Enable* is set high and the desired binary number is put onto the five *Jam* (J) inputs. We will *not* test this facility.

➤ Connect up the 4018 to the 4518 counter and LED display from part A. On the 4018, *Preset Enable* should be earthed, and the *Jam* inputs can be left unconnected. The TTL output of the signal generator should be connected to the Clock input. Reset (Re) can be connected to a switch, normally low, to allow the chip to be cleared. The output being used should be connected to the 4518 counter. However, because the outputs are inverted we'd like the 4518 to

increment on negative-going pulses. To do this, connect the 4018's output to the Enable input of the 4518, and connect the Clock input of the 4518 to earth (see table 2).

➤ Test the chip dividing by an **even** factor. To do this, display the pulses going in and coming out on a scope, and also count the outputs using the counter and LED display. Note that the outputs are inverted. When you have this working, divide by an **odd** factor as well.

Rectification of AC and conversion to TTL pulses

A half-wave rectifier using the 9 V AC supply is shown at the left of figure 6. *Be sure to use the AC outputs of the supply and not the earth output.* This rectifier only lets through the positive half of the sine-wave mains cycle, since the diode chops off the negative half. The symbol shown in the middle is a Zener diode, a device that does not allow the voltage across its terminals to exceed a given value. We use a Zener diode here to limit the voltage to a little under +5 V, 4.7 V to be precise. This should give us square-ish 50 Hz pulses that can drive TTL logic, since the 'low' should be close enough to 0 V. The circuit must also include a resistor of $\sim 100\ \Omega$ to limit the current through the Zener diode when it starts to conduct. We finally add a simple TTL circuit to clean up the pulses; almost anything would do but a very simple choice is a 7404 inverter since it only needs one input to produce an output. Use pin 1 as input, pin 2 as output, connect +5 V to pin 14, and earth pin 7. The TTL input and output need load resistors of $\sim 1\ \text{k}\Omega$.

➤ **Build this circuit** and **examine** the waveform produced on the scope. Sketch both the original and the cleaned-up pulses. These 50 Hz pulses will form the basis of the seconds timer in the next section.

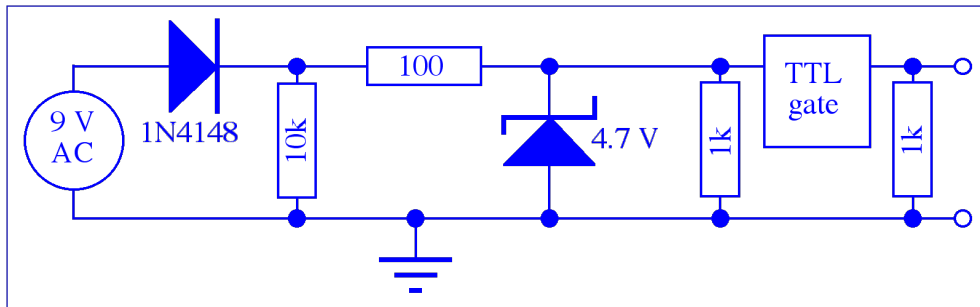


Figure 6 AC to TTL converter, including 'clean-up' circuit

Seconds timer

The source of the timing information is 50 Hz AC, converted to TTL as above. The pulse rate must then be reduced by a factor of 50, using two 4018s to divide first by 10 and then by 5 to give 1 Hz. This can then be fed to half of the 4518 dual decade counter, which will drive the seconds display. A block diagram of the entire system is shown in figure 7. For a ten-second timer you do not need any of the circuits along the right-hand side of the diagram; these are used for the *optional* 100-second timer extension that follows.

➤ **Build** and **test** the ten-second timer circuit.

If you are feeling ambitious and have some time, this can be extended into a two-digit 100-second timer. This requires the use of two LED displays and 4511 decoder/drivers, as in figure 7. Production of the 1 Hz signal is just as before. It is fed to half of the 4518 dual decade counter, which drives the seconds digit. In order to get the tens-of-seconds digit you can also use this half of the 4518 to divide the rate by 10, as described at the end of part A. Use the other half of the 4518 to drive the tens-of-seconds display. Use a common reset signal for both digits.

➤ **Optional (for extra credit): Build** and **test** this circuit.

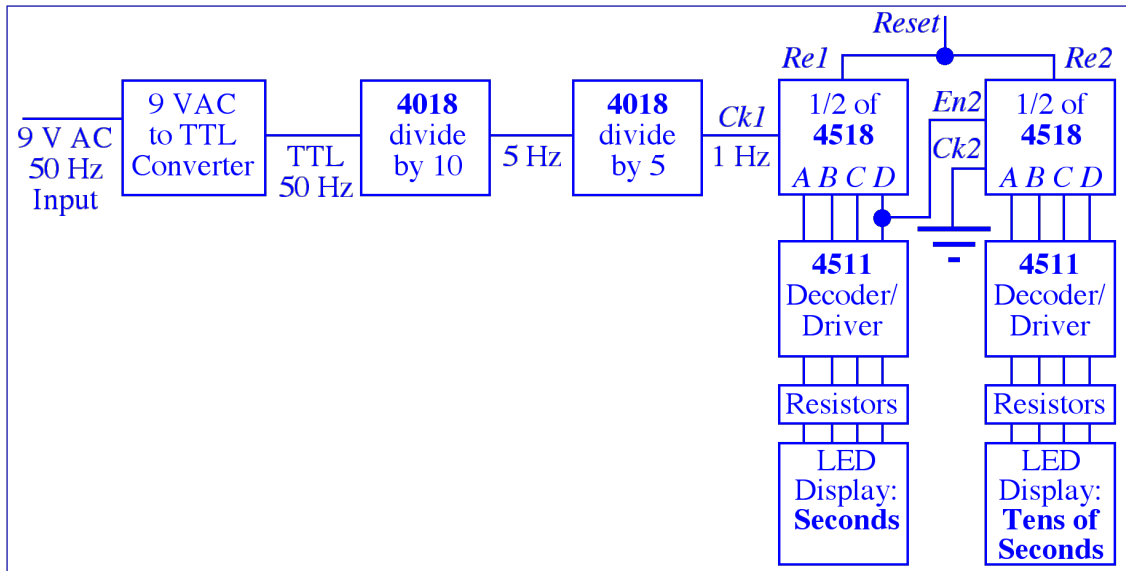


Figure 7 Block diagram of seconds timer. The chips along the right-hand side are only needed for the optional 100-second timer — omit them for a ten-second timer.

Part C: Operational amplifiers

Introduction

The ideal operational amplifier, or op-amp, is a high-gain differential-input amplifier denoted by the triangular symbol shown in figure 8. It has two inputs, called inverting (−) and non-inverting (+), and requires both positive and negative voltage supplies but no explicit earth. The op-amp amplifies the voltage *difference* between the two inputs, so that the output voltage is

$$v_{\text{out}} = A_o(v_+ - v_-)$$

where A_o is called the open-loop, or open-circuit, gain. For a good op-amp A_o is typically of the order of a million at low frequencies, so that even a voltage difference of a few microvolts will give an appreciable output. (We use lower-case letters like v and i for voltages and currents that can change with time, in contrast to steady DC ones which would be called V and I .)

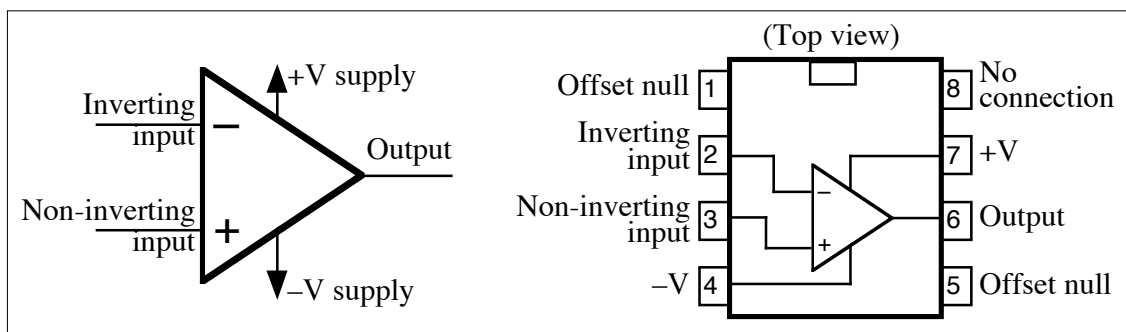


Figure 8 Operational amplifier, and pinout of 741 and 081

In reality it is impractical to use the amplifier in this way, since most of the time the output will simply saturate to nearly the level of the positive or negative power supply because the input signal is too large. It is almost always necessary to use *negative feedback*, which means that the output is connected back to the inverting input. This reduces the gain to a reasonable level.

In addition, an ideal op-amp has a very high input resistance (ideally infinite, typically at least of the order of megohms) and a low output resistance (ideally zero, typically of the order of a hundred ohms). Typical op-amp limitations are that the gain decreases at high frequencies, and the response to very rapidly-changing inputs is not instantaneous.

When using an op-amp with feedback, we can get quite far in understanding what happens simply by applying two ‘golden rules’:

- The output attempts to do whatever is necessary to make the voltage difference between the inputs zero.
- The inputs draw essentially no current.

In this part we will use two common and inexpensive op-amps, the 741 and the 081. Their pinout is shown in figure 8. Pins 1 and 5, *offset null*, are used to fine-tune the output to zero when the two inputs are equal. We can ignore them.

Inverting amplifier

This simple configuration is shown in figure 9. It uses an input resistor R_i and a feedback resistor R_f . **Build the circuit** step-by-step as follows.

➤ First, place a 741 across the gap on the breadboard, and connect pin 7 to the +12 V and pin 4 to the -12 V power terminals on the back of the test box. Connect the bottom horizontal row of breadboard connectors to the earth terminal of the power supply.

➤ Switch power on and use a multimeter to **check** the supply voltages, connecting one meter lead to earth. Also **measure** the open-circuit output voltage.

➤ Switch power off and connect the two resistors. For R_f use 100 k Ω connected between pins 6 and 2, and for R_i use 10 k Ω connected between pin 2 and a free breadboard connection.

➤ Connect one of the oscillator outputs to R_i and the other to earth. Connect channel 1 of the scope to the oscillator output, and connect channel 2 to the op-amp output (pin 6). This will allow a direct comparison of the op-amp’s input and output signals. At least one scope **earth** lead should be connected to the earth of the breadboard. (Note that even though the common earth is not shown in many op-amp diagrams, it will be present since the power supply is connected to it internally.) Using the scope, **set up the oscillator** for sine waves of about 1 kHz with an amplitude of roughly 100 mV.

➤ Switch power on and **measure the voltage gain** obtained with your amplifier by comparing the amplitudes of the input and output signals on the scope. **Compare** the voltage gain with the predicted value, which is:

$$\text{Voltage gain } A \equiv \frac{v_{\text{out}}}{v_{\text{in}}} = -\frac{R_f}{R_i}$$

This result, which depends *only on the resistor values and not at all on the op-amp chip*, follows from the ‘golden rules’. The first rule says that the inverting input is at ~ 0 V (it’s called a ‘virtual earth’), so the voltage across R_f is v_{out} and the voltage across R_i is v_{in} . The second rule says no current flows into the op-amp, so the currents in the two resistors must be equal. By Ohm’s Law we have:

$$I = v_{\text{out}}/R_f = -v_{\text{in}}/R_i$$

which can be rearranged to get the equation above. The minus sign indicates the inversion of the signal, which you should see on the scope. This equation is valid over a wide range of resistor values and voltage levels. The *resistor* values must lie below the input resistance (~ 2 M Ω for the 741 and $\sim 10^{12}$ Ω for the 081) and above the output resistance (~ 75 Ω). The *voltages* will be above the op-amp’s ‘offset’ (\sim mV) but cannot exceed the voltage supply levels.

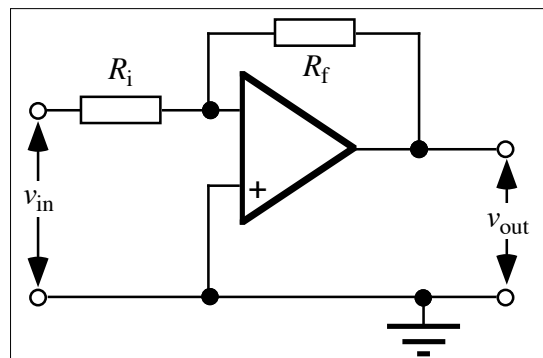


Figure 9 Inverting amplifier

The other restriction on the validity of the simple gain equation is frequency. At DC and low frequencies there is no problem, but at high frequencies the gain decreases.

➤ **Measure the frequency response** by increasing the sine-wave frequency of the oscillator, and measuring and plotting the gain at a number of points. The 741 should not show any decrease below about 10 kHz, so few points are needed below this value; concentrate your measurements where things start to change. **Plot** the results — a *log* scale is needed for the horizontal axis because of the wide frequency range, i.e. an axis labelled 1, 10, 100, 1000 etc.

➤ **Replace** the 741 with an 081, and **repeat** the measurements and gain vs. frequency **plot**. Notice the improved high-frequency response. If the oscillator will not go to sufficiently high frequencies to see the fall-off, try increasing R_f or decreasing R_i , since (as discussed below) the fall-off occurs at lower frequencies when the gain is higher.

There is always a trade-off between high-frequency response and gain. This is expressed by the **gain–bandwidth product**. The **bandwidth** is defined as the high-frequency cut-off, i.e. the frequency where the voltage gain has fallen to $1/\sqrt{2}$ of its maximum (usually mid-frequency) value. ($1/\sqrt{2}$ in voltage is equivalent to half the output power). For a 741, the gain–bandwidth product is $\sim 10^6$, so for example if the gain is 100 then we'd expect a bandwidth of $\sim 10^4$ Hz.

The **decibel (dB)** is often used in connection with amplifier gain because it is a logarithmic unit, and so can deal with widely-ranging values. Decibels are defined in terms of *power* ratios; the relative power gain in decibels is

$$\text{dB} = 10 \log_{10} \frac{P_{\text{out}}}{P_{\text{in}}}$$

In other words, a power ratio of 10 is +10 dB. An increase in power by a factor of 2 is $10 \log_{10} (2) = 3.01$ dB, which is normally called 3 dB, while a decrease by a factor 2 is –3 dB. If we wish to use voltage rather than power, then

$$P = \frac{V^2}{R} \quad \text{so dB} = 10 \log_{10} \frac{P_{\text{out}}}{P_{\text{in}}} = 10 \log_{10} \frac{V_{\text{out}}^2/R}{V_{\text{in}}^2/R} = 10 \log_{10} \frac{V_{\text{out}}^2}{V_{\text{in}}^2} = 20 \log_{10} \frac{V_{\text{out}}}{V_{\text{in}}}$$

Thus a *voltage* ratio of 2 means $20 \log_{10} (2) = 6.02$ dB, approximately +6 dB, while a ratio of 0.5 means –6 dB. A ratio of 10 is +20 dB, while 0.1 is –20 dB.

Hence, a gain decrease of –3 dB represents a decrease in power of a factor 2, and a decrease in voltage of a factor $1/\sqrt{2}$.

If we plot an amplifier's gain in decibels vs. logarithm of the frequency, we have a log–log plot on which the drop-off in gain as a function of frequency should be roughly a straight line. The usual units for the slope of this line are dB per decade, i.e. the drop in gain (in decibels) for each factor of 10 increase in frequency.

➤ **Re-plot** the frequency responses of the 741 and 081 in this way so that you can **estimate** the **slopes** at which they fall off. Estimate the **bandwidth** of each chip, and hence find their **gain–bandwidth products**.

Non-inverting amplifier

This circuit is shown in figure 10. The input signal now goes to the non-inverting (+) input, with the feedback still connected to the inverting (–) input.

➤ **Wire up this circuit**, again using 100 k Ω for R_f and 10 k Ω for R_i .

➤ **Measure** the **gain**, and **compare** with the predicted voltage gain, which is:

$$A = \frac{v_{\text{out}}}{v_{\text{in}}} = 1 + \frac{R_f}{R_i}$$

This can again be deduced from the ‘golden rules’. Here, $v_+ = v_- = v_{\text{in}}$, and v_- comes from what is effectively a voltage divider:

$$v_- = v_{\text{out}} \frac{R_i}{R_i + R_f}$$

Simple substitution gives the gain of the amplifier. As with the inverting amplifier, we see that if the op-amp behaves ideally then the gain depends *only* on the *external resistor values* and not on the properties of the op-amp chip itself. The positive value means that there is no inversion, and the absolute value of the gain is higher than that of the inverting amplifier by 1.

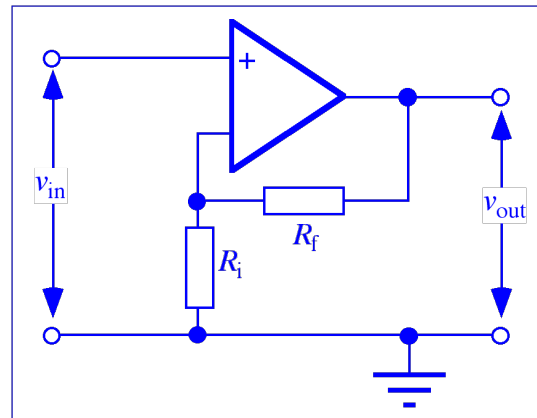


Figure 10 Non-inverting amplifier

Part D: A simple digital-to-analogue converter

Introduction

The world around us is full of analogue signals. The availability of sophisticated but cheap digital logic and microprocessors means that it is also essential to have converters, both to digitise analogue signals for processing, and to turn digital results back into analogue. In this part you will first use an op-amp to sum input signals, and then study an electronically controlled switch. Finally, these will be put together to build a simple 4-bit digital-to-analogue converter (DAC) in order to understand some of the basic principles.

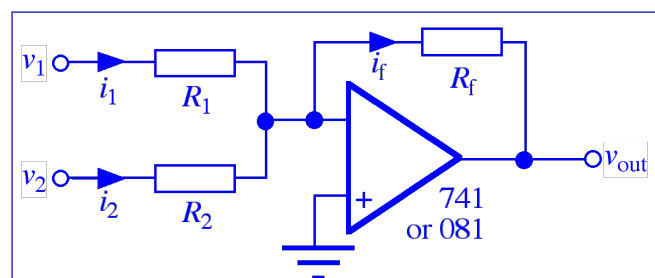
The DAC constructed in this exercise works by summing currents, each of whose magnitude is proportional to the binary bit that it represents. The binary number is formed by using an electronically-controlled switch for each current, so that each binary bit can be turned on or off. To show this circuit in action, the binary number is taken from a counter so that it varies in a simple and easily understood manner.

However, this DAC has very limited performance. For real applications demanding higher precision and speed it is far better to buy ready-made integrated converter chips than to try to design your own circuits. Very high-performance conversion is not easy, and prices go up rapidly as the number of bits and/or the speed increase. For example, the 16-bit DACs used to play CDs were initially state-of-the-art but have now become both more sophisticated *and* very cheap.

Summing amplifier

This circuit is a variation on the inverting amplifier of part B, and is shown in figure 11. Since no current flows into the op-amp input (golden rule), the currents in the two input resistors R_1 and R_2 are summed in the feedback loop resistor R_f : $i_1 + i_2 = i_f$ or in terms of voltage:

$$\frac{v_1}{R_1} + \frac{v_2}{R_2} = -\frac{v_{\text{out}}}{R_f}$$



If $R_1 = R_2 = R_f$, then $v_{\text{out}} = -(v_1 + v_2)$. **Figure 11** Summing amplifier

This means that, apart from the sign, the output depends on the *sum* of the inputs.

➤ **Build the summing circuit**, using ± 12 V supplies for the op-amp. The choice of resistance is not critical — 10 k Ω is suitable for all three resistors.

➤ **Test** the circuit by putting a sine wave signal into one of the inputs, and +5 V DC into the other input. Observe the results on the scope, noting the DC level as well as the sine wave.

DO NOT DISMANTLE YOUR CIRCUIT — IT IS NEEDED LATER

4066 CMOS switch

An ideal switch has no resistance when closed and an infinite resistance when open. Mechanical switches come close, while electronic ones are further from ideal. However, the electronic versions allow much higher switching speeds. We will use a 14-pin CMOS chip which has four single-pole switches and four control inputs. The switches have a resistance of about 90 Ω when closed, and can switch 5 V signals at speeds up to 5 MHz. The pinout is given in figure 12. Note that there are two modes, *analogue* and *digital*. We will use **digital** mode. The rules for use are:

- A switch is off when its control voltage equals the pin 7 voltage (i.e. 0 V).
- A switch is on when its control voltage equals the pin 14 voltage (i.e. +5 V).
- Signals passing through the switches must never go below the pin 7 voltage, nor above the pin 14 voltage.
- The switches are bi-directional, and there is no difference between input and output terminals. In other words, the switches behave just like a 90 Ω resistor.

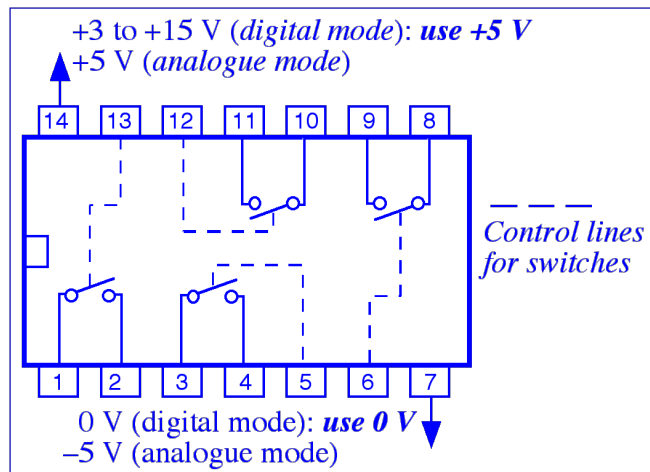


Figure 12 Pinout of 4066 switch

➤ **Check** how the 4066 works by using a multimeter to measure the resistance of one of the switches (e.g. pins 1 and 2) when it is open and closed. **Control** the switch by using one of the breadboard's switches to set the control voltage (e.g. on pin 13). **Record** the resistance with control settings of 0 V and +5 V.

➤ **Use** the 4066 by putting two of its switches in series with R_1 and R_2 to control the inputs of the op-amp summing amplifier previously built. **Connect** +5 V to one side of one of the switches (e.g. pin 1), and R_1 to the other side of the same switch (e.g. pin 2). **Connect** the TTL output of the signal generator to one side of another switch (e.g. pin 3), and R_2 to the other side of the same switch (e.g. pin 4). **Control** the two switches by using two of the breadboard switches (connected to e.g. pins 13 and 5, respectively). Use the scope to see what happens: make **drawings** of the waveforms with one switch on, the other switch on, and both switches on. (Pay attention to DC levels as well as waveform.) Is this what you would expect?

Digital-to-analogue converter

The aim now is to build a 4-bit DAC, using the circuit shown in figure 13. An op-amp is used to sum four inputs, each representing one of the binary bits. The bits will be generated by +5 V signals, with the resulting currents weighted by using resistors in the ratio $8R : 4R : 2R : R$. Each of the bits will be turned off or on by one of the CMOS switches of a 4066. This means that the output voltage of the op-amp will be:

$$v_{\text{out}} = -v_R \left(\frac{1 \text{ or } 0}{1} + \frac{1 \text{ or } 0}{2} + \frac{1 \text{ or } 0}{4} + \frac{1 \text{ or } 0}{8} \right)$$

Note that the output is inverted. This could be corrected by adding an inverting amplifier with unity gain after the summing amplifier. The most-significant bit, or MSB (i.e. the left-hand one in binary) comes from the smallest resistor, and the least-significant bit (i.e. the right-hand one in binary) from the largest resistor.

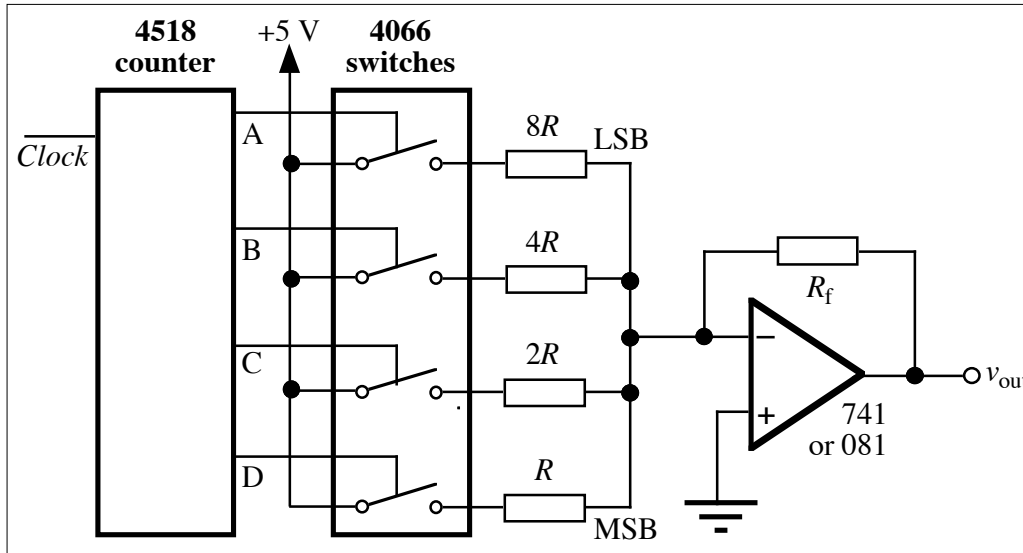


Figure 13 Digital-to-analogue converter

In order to test the DAC, the circuit includes a 4518 counter, to count TTL pulses from the TTL output of the signal generator. This gives binary outputs that ramp up from zero to maximum value and then reset.

Suitable resistor values are 3.3 k Ω , 6.8 k Ω , 15 k Ω and 27 k Ω , with a feedback resistor of 3.3 k Ω . Note that the switches also add 90 Ω each, and that due to the imprecise resistor values ($\pm 5\%$) the binary ratios will only be approximate. It is possible to do much better by using higher-precision resistors.

➤ **Build this circuit.** Note that the op-amp uses ± 12 V supplies, while all the other chips use +5 V.

➤ **Examine** and **sketch** the waveforms at the input and output of the circuit, using the scope, with the TTL input to the counter going at a rate such that everything can be seen clearly. If all is well, you should see a pattern like a staircase. Do you understand the number of steps?

➤ Finally, **measure** the output voltage corresponding to each of the 10 possible binary (BCD) numbers. This is easier if you replace the TTL input to the counter by one of the breadboard switches. This allows you to clock up the counter one count at a time. Values can be read off the scope, with care, or you can use a DMM. Make a **graph** of the results, with counter values on the x -axis and the measured voltage on the y -axis. Is the plot **linear**?

Practical DACs need more than four bits, and using this ‘binary ladder’ method would demand resistors with a huge range of values. (An 8-bit DAC would require the largest resistor value to be 256 times the smallest one.) Therefore, a slightly more complicated method called the ‘ R - $2R$ ladder’ is used; as the name implies, only two resistor values are needed.

Laboratory Exercise 11 – COMPUTING AND COMPUTER CONTROL

Part A: Basic Visual Basic

The Visual Basic environment

Basic is an example of a high-level computer language. This means that it uses words and mathematical symbols rather than machine code, which is a list of binary numbers suitable for direct action by the computer's microprocessor chip. Like other high-level languages, Basic is written as a series of commands that are either interpreted into machine code as execution proceeds, or compiled (i.e. pre-interpreted) beforehand. Basic has a very long history, and there are numerous versions and 'dialects'. Microsoft Visual Basic is relatively recent, and also incorporates pre-prepared GUI (Graphical User Interface, pronounced 'gooey') elements such as menus, windows, scroll bars and buttons. This greatly improves the user-friendliness of the programs, and makes the problems of input and output much simpler for the programmer. In addition, Visual Basic uses the modern paradigms of being 'object oriented' (the GUI elements are examples of 'objects'), and of being 'event driven' (the program waits for external stimuli, such as mouse clicks, to initiate certain activities).

The Visual Basic application provides a modern 'integrated development environment' (IDE) for preparing, editing, compiling, debugging, and running programs before they become standalone applications. It can be started using a shortcut that we have put onto the desktop (see below), or via the Windows 'Start' pop-up menu, and has controls similar to many other Windows applications. For this part of the exercise, we have written a simple Visual Basic project called **VBtrainer** for you to use. This has a place to put in various snippets of Basic code in order for you to become familiar with the language. You will work from your own copy of VBtrainer so that you can alter it and try things out independent of other students.

- Ask the lab technicians to give you an **account** on one of the PCs used for this project. Unless there are problems, you must always use the same PC and account. When you have logged on to your PC, click on the start button (bottom left of screen) then select My Computer from the menu. Under **Network Drives** you will see two drives. One is **Software3 on Teaching Samba Server**. Double click on this drive to view the folders stored on it. You will not be able to save any files onto this drive, or make any changes to the files on there, but you can make copies of them. This drive contains the folders **Trainer** and **Heater**.

- VB trainer is found inside the folder **Trainer**. Copy the Trainer folder onto the other network drive called **xyz on Teaching Samba Server** where **xyz** is your user name. This is the drive into which you must save all your work. If you do not know how to do this then ask a demonstrator. You will be able to use, alter and save any changes you make to this copy. Later on you can save differing versions (some perhaps with altered controls) under different names. At the end of each session on the PC, you must be sure to back up our copy of the project you are working on (**VB trainer** or later **Heater**) to a memory stick. Note that the standard filename extension for Visual Basic projects is **.vpb**, so double-clicking on any file with that extension will bring up the Visual Basic immediately.

- At any stage you may use the blue-grey Play button (right-facing triangle) on the toolbar to test-run the program. As on a tape or CD player, there are also buttons for Pause (||) and Stop (square). You can also start by using the Go! button of the program, and stop by using the Quit button. (Equivalent controls are also available from the menu bar, where they are called Start, Break and End.) Stopping brings you back to the editing mode. Try out these controls now.

You can learn about other facilities, especially the easy-to-use debugging system and how to add more graphical controls to the program, from the online help system or by asking the demonstrators as the exercise progresses.

As a useful tip for when you come to write your report, remember that in Windows you may, at any time, do a CTRL-‘Print Screen’ combination keystroke. This puts an image of the screen onto the clipboard, from which it may be pasted into e.g. a Word document and later trimmed or edited. Additionally, you may wish to copy (CTRL-C) the text of your programs so that you can then paste (CTRL-V) it into any other file. In this way, you can show the reader of your report what would be seen on-screen at any time, or make a copy of your program code so that it can be included in your report.

Before starting any new task or sub-experiment, copy all VBtrainer (or later Heater) files into a suitably named folder in the folder with your username, and be sure to work from that.

Basic programs

We shall now try some simple Basic instructions. The complete list, and assistance on any topic, may be reached via the Help menu, but the object of this part of the exercise is to learn and familiarise yourself first by some simple examples.

- View the VBtrainer code and look through it to find the part with the heading:

```
Private Sub cmdGo_click()
```

The program comes to this subroutine whenever the Go! command button is clicked. Under the heading, enter the following line *exactly* as shown:

```
picOutput.Print "Hello World!"
```

Then execute the program by pressing the Go! button. In the line you typed, `picOutput` is an ‘object’ which happens to be the big picture window on the screen, and it has an associated ‘method’ or action called `.Print` which puts text into the window. Later we will see how to get output printed on paper by referring to `Printer.Print`, in which a physical object called `Printer` has a similar method.

- Press the Quit button. Now edit the program so the command reads:

```
picOutput.Print 2+3
```

Run the program again. Note the difference between `"2+3"`, which merely prints whatever is inside the quotation marks, and `2+3`, which evaluates the expression and thus gives us a kind of calculator. To make sure you understand this point, try changing the line to:

```
picOutput.Print "2+3=",2+3
```

- Try other combinations and operations, including `*` and `/` for multiplication and division, and `^` to raise to a power; `^0.5` gives the square root. Basic also contains most mathematical and trigonometric functions, so try `Log(10)`, `Cos(3.1)`, `Exp(1.0)` and similar examples. Note that `Log` is actually the natural log (base e), and that angles are expressed in radians not degrees. A nice feature is that if you make an error in the code it will be highlighted.

Now we can generalise the process, by introducing variables `a`, `b`, and `c`. As soon as we use a variable the program allocates it a memory location to store its current value in.

- Put the following sequence into the program and run it:

```
a = 2
b = 3
c = a^b + 2*(b-a)
picOutput.Print "The value of c is ",c
c = c+1
picOutput.Print "The value of c is ",c
```

This shows that more complicated calculations can still be performed quite simply. But this example also illustrates a very important point about most high-level computing languages. In normal algebra $a^b + 2*(b-a) = c$ is perfectly valid and means exactly the same thing as $c = a^b + 2*(b-a)$, but in Basic the first of these is *wrong*. The reason is that the = sign has a very different meaning than in algebra — it is an *instruction to take the value* of whatever is on the *right-hand side* and *transfer* it into the single *variable* that is named on the *left-hand side*. So in this case the value of $a^b + 2*(b-a)$ is calculated, and then c is given that value. A more drastic illustration is the line $c = c+1$, which is nonsense in algebra — in Basic it says to add one to the value of c and then set c equal to that new value. You will get used to this different way of doing things. You might also like to think about the order in which operations are carried out. For instance, in the example above a is raised to the power b , and *not* the power $b + 2*(b-a)$.

Next we learn how to input information to the program. In the program above, changing the values of a and b requires you to change the program itself, which is not usually good practice. VBtrainer has been set up with two text boxes, so that when you run the program you can enter numbers in them before clicking Go!. Within the program these text boxes have the names `txtIP1` and `txtIP2`, so that you can write code that looks like `a = txtIP1.Text`. Note that `.Text` is now a ‘property’ of an object rather than a method, i.e. something it *has* rather than something it *does*. However, you should not necessarily count on Basic converting from the text to the numeric value automatically. So you should properly write `a = Val(txtIP1.Text)`.

- Alter the program so that the values of a and b are read from the text boxes rather than entered directly into the program code.

A very powerful technique in computing is to be able to do different things, depending on intermediate results. We do this using an `IF` statement. Hopefully you can guess what is going on in the following (where ... means that some additional program statements would normally be inserted):

```
IF a>b THEN
...
ElseIF a<b THEN
...
Else
...
END IF
```

- Try to write your own program using this sort of `IF` block, reading in a and b from the text boxes. If a is bigger than b then output the result of $a-b$, if b is bigger than a then output the result of a/b , and if a equals b then output the square root of a (or b). Be sure to *record* the program code so that you can display it in your report, in order to allow it to be marked easily.

By now you will probably have made some mistakes, and seen that Visual Basic tries to correct you if you type anything that does not make sense. (If you have not yet made any mistakes, try typing something wrong on purpose just to see what happens!)

Visual Basic is set up to *ignore* any line whose first character is an apostrophe ('). In that case the whole line is typed in green as a **comment**. You should *use comments frequently* from now on to let others know, in English, what is going on in your programs. Equally important, comments also remind *you* (especially when you have not looked at the code for a while!) why you wrote what you did.

Another crucial programming concept is to do something repeatedly. This is achieved using a **loop**, as illustrated in the following example.

- Try the following in order to test an idea for generating prime numbers:

```
For i=1 to 10
c = 2^i - 1
```

```
picOutput.Print "Is ",c," a prime number?"
Next i
```

This goes round a set of instructions, starting with $i=1$ and increasing i by 1 each time round until it has the value 10, when it will pass the `Next` instruction and carry on with what follows. Normally the variable changes by +1 each time the loop is performed, but we can use a `Step` instruction (e.g. `For i=1 to 10 Step 2`) to go forward or backward by a given number of integers each time.

- See if you can write a program to produce the factorial ($n!$) of a number n . These numbers get huge very quickly, so be careful.

In statistics the logarithm of a factorial turns out to be extremely useful, so much so that Stirling gave as an approximation $\ln(n!) = (n + 1/2) \ln(n) - (n - 1)$.

- Test this out by calculating both $\ln(n!)$ and the approximation, and printing the ratio as n gets larger.

Again, *record* the code for your programs so that you can display it in your report.

We have only scratched the surface of Visual Basic. There are many more Basic statements and GUI elements. To discover them, get used to using the online help system.

Part B: Interfacing with the outside world

Our main aim in this part of the exercise is to see how to program the PC to communicate with apparatus in a simple way, and to do this we have to look at doing input and output with the world outside the computer's mouse, keyboard and screen. This is a very important aspect of the power of computers, but to understand what is going on we have to be familiar with the binary (base 2) numbers that are used.

Binary numbers

Computers communicate with other electronic devices using **binary** (i.e. base 2). For example, transistors are turned on or off, which corresponds to 'true' or 'false', or logic '1' or '0', using **binary digits (bits for short)**. Silicon 'chips' are made up of a large number of semiconductor devices which manipulate very large numbers of these bits very quickly, for example:

- storing them, i.e. **memory**;
- adding, subtracting, or carrying out logical operations such as AND, OR;
- moving them about.

Binary digits are written in ascending powers of 2, starting at the right:

$$\begin{array}{r} 2^4 \ 2^3 \ 2^2 \ 2^1 \ 2^0 \\ 16 \ 8 \ 4 \ 2 \ 1 \end{array}$$

For example, binary number 10111 = 1×16
 0×8
 1×4
 1×2
 1×1
 so in decimal it is 23

In order to avoid writing out or displaying the long strings of bits that are a feature of binary, it is customary with computers to express the numbers in the much more compact hexadecimal notation ('hex'), which uses base 16 rather than the conventional decimal base 10 or the binary

base 2. The conversion to and from binary is easy since $2^4 = 16$, so one hex digit represents four binary digits and so decimal numbers up to 15. (Too bad we don't have 16 fingers, though.)

The conversion is:

Hex	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
Decimal	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Binary	0	1	10	11	100	101	110	111	1000	1001	1010	1011	1100	1101	1110	1111

Two hex digits (i.e. 8 binary bits) can represent a number from 0 up to 255 (decimal; $2^8 - 1$) = FF (hexadecimal) = 1111 1111 (binary). Four hex digits (16 binary bits) can represent from 0 up to 65,535 (decimal; $2^{16} - 1$) = FFFF (hexadecimal) = 1111 1111 1111 1111 (binary).

- Example: 2B (hex) = $(2 \times 16) + 11 = 43$ (decimal). What is the binary form of this number?

An 8-bit computer is one that can handle and store 8 bits at once (8 bits is usually called a **byte**), that is, each memory location can store 8 bits (or 2 hex digits), and it can move these 8 bits on 8 parallel wires called the **data bus**. The data bus carries information within the computer and also to and from the outside world via connectors called **ports**. Later computers were 16-bit, storing and transporting two bytes at once, and nowadays they are 32-bit or 64-bit.

Input and output

It is common for computers and other electronic devices to talk to each other by transferring 8-bit bytes of information to an address or port that in practice is a socket on the outside of the computer. However, in the outside world most electrical signals are *analogue*, i.e. they can take any of a continuous range of values, while computers require *digital* information that can be represented by discrete binary integers and so is not continuous but quantised.

On the computers that are used for this exercise, we have installed electronic boards called DAQ (**Data Acquisition**) cards. These include analogue-to-digital converters (ADCs) to change an analogue signal from outside (e.g. a voltage from an external thermometer) into a byte of binary information for use in calculations within the computer. That allows *input* to the computer. For *output* from the computer we can either use digital information directly (e.g. one binary bit indicating 0 or 1 to decide whether a switch is turned off or on, possibly for a heater), or in addition the DAQ cards provide a digital-to-analogue converter (DAC) which will take a floating-point number in the range -10 to $+10$ from the computer and convert it to a corresponding voltage on an external wire. We will use all of these facilities.

Light and switch interface box

CAUTION: Before starting this section, make sure that the *heater*, used later in part C, is **NOT plugged in to the mains**. The reason will become obvious later!

Although we can use our outside connections to communicate immediately with various pieces of lab equipment, it is instructive first to use the interface boxes to become familiar with how things work. These have eight toggle switches and eight LEDs, allowing you to set the binary value of a byte, read it into the PC, and then display all of its bits. The software to do this is usually specific to a given model of DAQ card, and in VBtrainer we have hidden this complication from you in subroutines that you can call. (You are of course welcome to look at them and make copies!) To *receive* a byte of data into the PC the necessary code in your cmdGo_click subroutine is `Call DigIn(arg1)`, where `arg1` represents a variable containing a byte of input data read in from an external device, for example the eight switches. To *transmit* data out to an external device the code is `Call DigOut(arg2)`, where `arg2` is a variable set by your program, for example to display on the LEDs. Both `arg1` and `arg2` are positive 8-bit binary

integers, i.e. their values are between 0 and 255. Note that `arg1` and `arg2` represent variables, and you can actually give them any names you like.

- Write a program that simply reads in the binary setting of the switches and displays them on the lights. Print the value on the screen as well (this is effectively a binary-to-decimal converter). *Record* the program code so that you can display it in your report, in order to allow it to be marked easily.

- Write a program to loop through all 256 possible values (0–255) of an 8-bit byte, displaying them on the lights. (Switches not needed here.) A simple loop will run too fast for you to be able see it incrementing, so you will have to slow the program down. There are various ways to do this. One is simply to insert a **delay loop** in the code to ‘waste’ time. For example:

```
For i=1 to 1000
Next i
```

Vary the number of passes through this loop to alter the delay. How many passes are needed to slow things down enough?

- Modify the program so that if a particular number which you specify via the statement `a = txtIP1.text` is selected on the switches, something special happens. For example, arrange the program to put up a special message when that number is input (in binary) on the switches.

Again you are reminded to *record* all code that you write for your report.

Digital-to-analogue converter

A digital-to-analogue converter (DAC) is an integrated circuit which outputs a voltage proportional to the binary number sent to it. VBtrainer includes a subroutine to do this — to generate a voltage of `arg3` volts at the interface box via the DAC all you have to do is insert in your code the statement:

```
Call AnaOut(arg3)
```

- Arrange your program so that you can type a voltage value in *volts* into an input text box and then output it using the DAC. Check that the DAC generates the correct value using a DMM connected to the green and red pair of terminals on the interface box. What range of voltages can you get?

- By changing the values of the voltage and reading the variations in the DMM, try to get some idea of the resolution of the DAC and *comment* on it.

- Next write another program that generates a pattern of output voltages one after another in a loop, and displays the output on an oscilloscope. Remember to ensure that your program has a way of ending! Start with a sine wave; this could include something that might look like:

```
pi = 3.14159
steps = 500
c = 2*pi/steps
For n = 0 to steps
a = 10*sin(n*c)
Call AnaOut(a)
Next n
```

- Use a similar technique to plot the following:
 - a) A squarewave.
 - b) A sawtooth, or a ramp and a triangle.
 - c) Another function of your choosing.

Part C: Temperature control by computer

The aim of this part of the experiment is to use information coming into the computer to tell it what actions to perform on signals that it puts out. Specifically, you will accurately control the temperature of a water-bath using computer software.

To allow you to do this, we have provided one further subroutine, `Call AnaIn(arg4)`. This gives the voltage coming from a thermocouple and thus measures temperature. In addition, `Call DigOut(arg2)` now takes on the specific job of digitally switching a heater on and off.

Equipment and program

- Familiarise yourself with the wiring of the equipment. As before, the DAQ card in the PC is connected to the interface box. There is a control wire from this box via a 5-pin plug to a small grey box, which also has a mains input, and an output to the heater itself. This box has a red safety light which is lit whenever mains voltage is on at the heating element. When your experiment gets under way you should see the LED flashing, with a frequency dependent on how much heat you want to deliver. The software you use is ideally 'fail-safe', but things can always go wrong so look at the LED from time to time in case the heater is accidentally left on. ***If there is a problem, simply unplug the grey box.*** In addition, be sure to switch off and unplug all equipment at the end of your activities for the day.

CAUTION: *Ensure that the heating element is always immersed up to a level between the two indented marks, otherwise it will be permanently damaged.*

To measure the temperature, a chromel–alumel thermocouple is used which has a temperature sensitivity of approximately $40 \mu\text{V}/^\circ\text{C}$, and is therefore amplified before the ADC chip can be used. The thermocouple is plugged into the side of the interface box. The end result is an analogue input to the computer which is roughly 10 mV per $^\circ\text{C}$. You will have to calibrate this more accurately later as a first step in the experiment.

- Most of the controlling program has already been written for you. It incorporates updating graphics to enable you to see what is happening continuously. ***Always*** work from ***your own copy*** of this Visual Basic project, called **Heater**, so ***copy it to the folder with your username.***
- Before inserting your own code, run the program as it is. The green Start button begins a data-taking run, and immediately turns into a red Stop button to allow you to terminate the run. Starting a new run will clear all information from the last run and set everything to default values. The Print button prints just the graphical output, and the Save button enables you to write out the data to a `.dat` file which may then be input into PhysPlot to allow more control over the graph before printing it. (Be careful with file naming in order to avoid overwriting any previous files that you still want to keep!)

It is important to understand the general mode of operation of this program. It spends 99.9% of its time within the subroutine `USER` (the only one that you are allowed to modify). Once the call is complete, the subroutine is called again and again. Every once in a while (`Dt` seconds, as set up on the controls), an interrupt is generated and a fast monitoring call is made. The counter `ncount` is updated, the variable `T1` is filled with the latest temperature, and the graph is updated with this value. Then control is returned to `USER`. Within this routine you should be able to discover that such an update has been made by checking whether `ncount` has increased, using your own internal storage variable.

If `Dt` is set low it is easier to control the temperature. Start with `Dt = 5` (seconds) when you are getting the feel of apparatus and the task. However, your final aim is to use `Dt = 30`, and show your skill in using more sophisticated control algorithms to overcome the lack of knowledge

from less frequent readings. We do this because in the real world we might, for example, be using one PC to control thousands of temperatures in this way.

The aim of the experiment is to switch the heater on and off repeatedly from the controlling program so that the temperature of the water goes to and then remains as close as possible to a target temperature. The target temperature that you are actually aiming for is called T_0 , and it is set by another control within `USER`. There are also outputs, including a timer, the last input voltage (`ADC`), and the temperature (T_1) resulting from calibration of the ADC. There are two calibration constants, `a` and `b`, for converting ADC voltage readings to temperature; linearity is assumed. You must enter these. Within the program there is a line $T_1 = a \cdot \text{ADC} + b$, where `ADC` is the voltage reading from the ADC, and you should find that `a` is roughly 100 and `b` is roughly 0.

Experiment

The aim of the first step is to find accurate values of `a` and `b`, so as to give accurate values of T_1 in $^{\circ}\text{C}$ agreeing with the alcohol-in-glass thermometer. Note that the thermocouple has a faster response time to temperature changes than the bulkier thermometer.

- Start by setting `a = 1` and `b = 0`. This means that T_1 is the actual thermocouple voltage (after amplification). By putting the thermocouple first in iced and then in very hot water, as well as two intermediate temperatures, and letting things settle down, you should be able to calibrate the ADC values against the corresponding thermometer readings. Make a graph of temperature vs. voltage using PhysPlot, and then fit a straight line — the gradient and intercept will give you `a` and `b`. Thereafter *they should be entered before each run. Be sure to record everything!*

You are now in a position to start altering the world! (Or at least the temperature of the water-bath.) As before, `Call DigOut(1)` turns on the first LED light on the interface box. However, when the grey heater control box is plugged into the interface box it *also* controls the heater. `Call DigOut(1)` turns the heater on, and `Call DigOut(0)` turns it off. Furthermore, the program always does a `Call DigOut(0)` when it closes down, to make it relatively fail-safe.

- Double-click on the heater project `Heater.vbp` to open the program coding. Insert your user code into the subroutine `Private Sub User()` near the end. Try `Call DigOut(1)` by putting it into a `FOR` loop which turns the heater on for half the time and off for the rest. You should see the red light on the interface box flashing in time with the safety light on the heater switch box, so this works as your own indicator.

For the rest of the experiment it is your task to set a target temperature (T_0) to which, hopefully, T_1 converges. You should test your software thoroughly by using both a *low* and a *high* target value for each version, say 40°C and 90°C . You should *not*, yourself, obtain a new value of T_1 at any time but rely upon the updates every `dt` seconds. In this way you are gradually forced to use more sophisticated algorithms on the limited information available, as with the problem in real control situations where the computer has to monitor very many devices.

- Write versions of `USER` based on the following algorithms, of increasing sophistication:
 - (a) With `dt = 5` seconds, turn on the heater if the last value of $T_1 < T_0$, and off otherwise. Naturally there will be overshoots and hence oscillations.
 - (b) If $T_1 < T_0$ turn on the heater for a fraction of time which is proportional to $|T_1 - T_0|$, i.e. the difference between T_1 and the required temperature. Start with `dt = 5` seconds, but when things are working well try increasing `dt` to 30 seconds.

You are eligible to get extra marks by inventing other ways of quickly converging on T_0 . For example, use the difference in the last *two* recorded temperatures (i.e. the linear trend), as well as $|T_1 - T_0|$, to decide how long to turn on the heater. Remember to state clearly, for all cases, what T_0 was when you write up your results. Your **report** must include a **diagram of the set-up**, **graphs**, and **printed listings of all programs** discussed.

Laboratory Exercise 12 – THERMAL EFFICIENCY

Heat Engine

A heat engine uses the temperature difference between a hot reservoir and a cold reservoir to do work. Usually the reservoirs are assumed to be very large in size so the temperature of the reservoir remains constant regardless of the amount of heat extracted or delivered to the reservoir. This is accomplished in the Thermal Efficiency Apparatus by supplying heat to the hot side using a heating resistor and by extracting heat from the cold side using ice water. In the case of the Thermal Efficiency Apparatus, the heat engine does work by running a current through a load resistor. The work is ultimately converted into heat, which is dissipated by the load resistor (Joule heating).

A heat engine can be represented by a diagram (Figure 1). The law of Conservation of Energy (First Law of Thermodynamics) leads to the conclusion that $Q_H = W + Q_C$, the heat input to the engine equals the work done by the heat engine on its surroundings plus the heat exhausted to the cold reservoir.

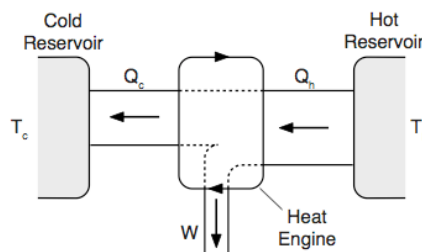


Figure 1: Heat Engine

The efficiency of the heat engine is defined to be the work done divided by the heat input

$$e = \frac{W}{Q_H}$$

So if all the heat input was converted to useful work, the engine would have an efficiency of one (100% efficient). Thus, the efficiency is always less than one.

NOTE: Since you will be measuring the rates at which energy is transferred or used by the Thermal Efficiency Apparatus all measurements will be power rather than energy. So

$P_H = dQ_H/dt$ and then the equation $Q_H = W + Q_C$ becomes $P_H = P_W + P_C$ and the efficiency becomes

$$e = \frac{P_W}{P_H}$$

Carnot showed that the maximum efficiency of a heat engine depends only on the temperatures between which the engine operates, not on the type of engine.

$$e_{Carnot} = \frac{T_H - T_C}{T_H}$$

where the temperatures must be in Kelvin. The only engines which can be 100% efficient are ones which operate between T_H and absolute zero. The Carnot efficiency is the best a heat engine can do for a given pair of temperatures, assuming there are no energy losses due to friction, heat conduction, heat radiation, and Joule heating of the internal resistance of the device.

Adjusted Efficiency

Using the Thermal Efficiency Apparatus, you can account for the energy losses and add them back into the powers P_W and P_H . This shows that, as all losses are accounted for, the resulting adjusted efficiency approaches the Carnot efficiency, showing that the maximum efficiency possible is not 100%.

Heat Pump (Refrigerator)

A heat pump is a heat engine run in reverse. Normally, when left alone, heat will flow from hot to cold. But a heat pump does work to pump heat from the cold reservoir to the hot reservoir, just as a refrigerator pumps heat out of its cold interior into the warmer room or a heat pump in a house in winter pumps heat from the cold outdoors into the warmer house.

In the case of the Thermal Efficiency Apparatus, heat is pumped from the cold reservoir to the hot reservoir by running a current into the Peltier device in the direction opposite to the direction in which the Peltier device will produce a current. A heat pump is represented in a diagram such as Figure 2

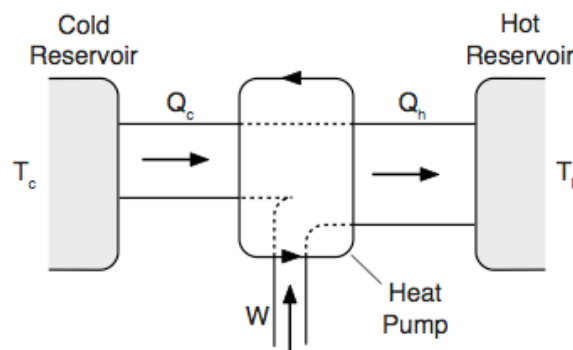


Figure 2: Heat Pump

NOTE: By conservation of energy, $Q_C + W = Q_H$, or in terms of power $P_C + P_W = P_H$.

Coefficient of Performance

Instead of defining an efficiency as is done for a heat engine, a coefficient of performance (COP) is defined for a heat pump. The COP is the heat pumped from the cold reservoir divided by the work required to pump it

$$\kappa = COP = \frac{P_C}{P_W}$$

This is similar to efficiency because it is the ratio of what is accomplished to how much energy was expended to do it. Notice that although the efficiency is always less than one, the COP is always greater than one. As with the maximum efficiency of a heat engine, the maximum COP of a heat pump is only dependent on the temperatures.

$$\kappa_{max} = \frac{T_C}{T_H - T_C}$$

where the temperatures are in Kelvin. If all losses due to friction, heat conduction, radiation, and Joule heating are accounted for, the actual COP can be adjusted so it approaches the maximum COP.

Measurements with the Thermal Apparatus

Three quantities may be directly measured with the Thermal Efficiency Apparatus: temperatures, the power delivered to the hot reservoir, and the power dissipated by the load resistors. The details of how these measurements are made follow.

Temperatures: The temperatures of the hot and cold reservoirs are determined by measuring the resistance of the thermistor imbedded in the hot or cold block. To do this, connect an ohmmeter to the terminals located as shown in Figure 4. The switch toggles between the hot side and the cold side. The thermistor reading can be converted to a temperature by using the chart located on the front of the Thermal Efficiency Apparatus. Notice that as the temperature increases, the thermistor resistance decreases (100 kΩ is a higher temperature than 200 kΩ). **NOTE:** To get the exact temperature reading you must interpolate between numbers on the chart.

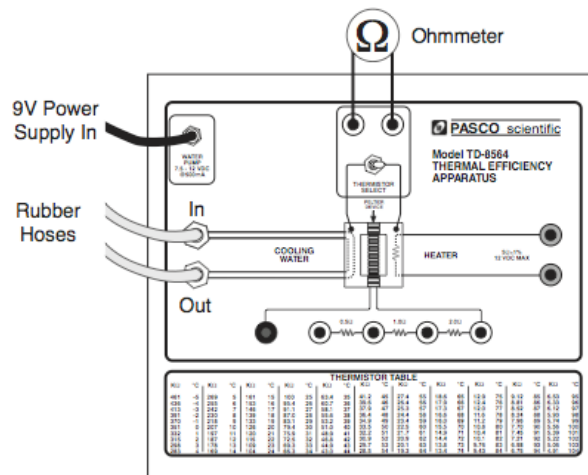


Figure 3: Thermal Efficiency Apparatus

Power Delivered to the Hot Reservoir (P_H): The hot reservoir is maintained at a constant temperature by running a current through a resistor. Since the resistance changes with temperature, it is necessary to measure the current and the voltage to obtain the power input. Then $P_H = I_H V_H$.

Power Dissipated by the Load Resistor (P_W)

The power dissipated by the load resistor is determined by measuring the voltage drop across the known load resistance and using the formula:

$$P_W = \frac{V^2}{R}$$

The load resistors have a tolerance of 1%. When the Thermal Efficiency Apparatus is operated as a heat pump rather than as a heat engine, the load resistors are not used so it is necessary to measure both the current and the voltage. So the current into the Peltier device is measured with an ammeter, and the voltage across the Peltier device is measured with a voltmeter and the power input is calculated with the formula $P_W = I_W V_W$.

Indirect Measurements: It will be necessary to know three additional quantities in the experiments: 1) The internal resistance of the Peltier device; 2) The amount of heat conducted through the device and the amount radiated away; 3) The amount of heat pumped from the cold reservoir. These quantities may be determined indirectly with the Thermal Efficiency Apparatus in the following ways.

Internal Resistance: Before the adjusted efficiency can be calculated, it is necessary to calculate the internal resistance. This is accomplished by measuring the voltage drop across the Peltier

device when an external load is applied. First run the Thermal Efficiency Apparatus with a load resistor (R) as in Figure 4.

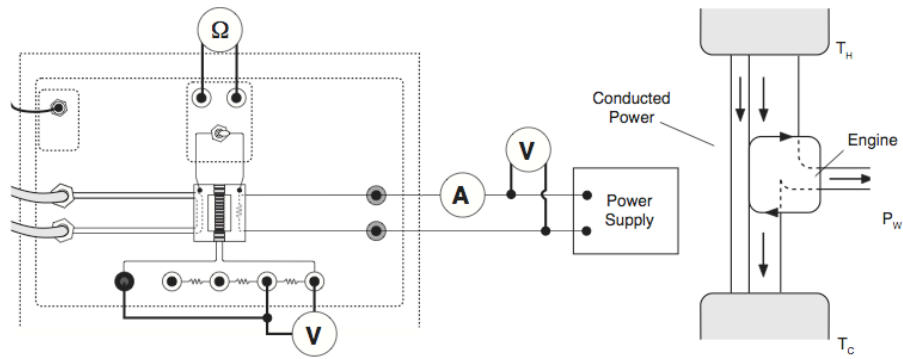


Figure 4: Heat Engine with a load

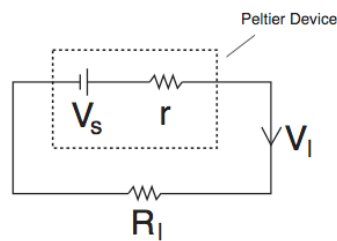


Figure 5: Procedure for finding internal resistance

The electrical equivalent of this setup is shown in Figure 5. Kirchoff's Loop Rule gives

$$V_S - Ir - IR = 0$$

Next, run the Thermal Efficiency Apparatus with no load, as in Figure 6.

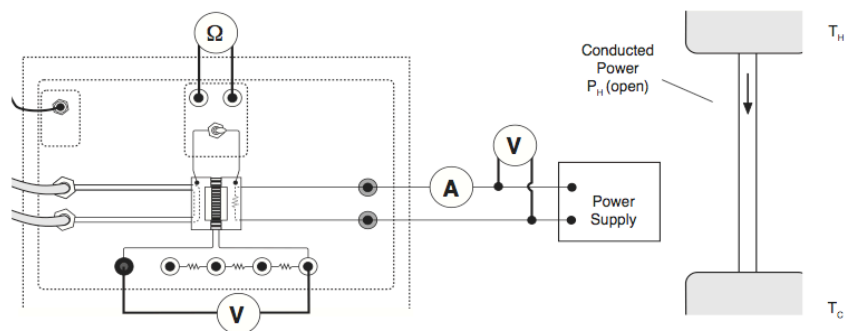


Figure 6: No load

Since there is no current flowing through the internal resistance of the Peltier Device, the voltage drop across the internal resistance is zero and the voltage measured will just be V_S . Since we have measured V_w rather than I in the heat engine mode, the equation above becomes

$$V_S - \left(\frac{V_w}{R}\right)r - V_w = 0$$

Solving this for the internal resistance gives us

$$r = \left(\frac{V_S - V_w}{V_w}\right)R$$

You may also find the resistance by measuring the currents for two different load resistors and then solving the resulting loop rule equations simultaneously.

Heat Conduction and Radiation: The heat that leaves the hot reservoir goes two places: part of it is actually available to be used by the heat engine to do work while the other part bypasses the engine either by being radiated away from the hot reservoir or by being conducted through the Peltier device to the cold side. The portion of the heat which bypasses the engine by radiation and conduction would be transferred in this same manner whether or not the device is connected to a load and the heat engine is doing work.

The Thermal Efficiency Apparatus is run with a load connected to measure P_H (Figure 4) and then the load is disconnected and the power input into the hot reservoir is adjusted to maintain the temperatures (less power is needed when there is no load since less heat is being drawn from the hot reservoir). See Figure 6. $P_{H(\text{open})}$ is the power input to the hot reservoir when no load is present. Since, while there is no load, the hot reservoir is maintained at an equilibrium temperature, the heat put into the hot reservoir by the heating resistor must equal the heat radiated and conducted away from the hot reservoir. So measuring the heat input when there is no load determines the heat loss due to radiation and conduction. It is assumed this loss is the same when there is a load and the heat engine is operating.

Heat Pumped from the Cold Reservoir: When the Thermal Efficiency Apparatus is operated as a heat pump, conservation of energy yields that the rate at which heat is pumped from the cold reservoir, P_C , is equal to the rate at which heat is delivered to the hot reservoir, P_H , minus the rate at which work is being done, P_w (Figure 2). The work can be measured directly but the heat delivered to the hot reservoir has to be measured indirectly. Notice that when the heat pump is operating, the temperature of the hot reservoir remains constant. Therefore, the hot reservoir must be in equilibrium and the heat delivered to it must equal the heat being conducted and radiated away. So a measurement of the heat conducted and radiated away at a given temperature difference will also be a measurement of the heat delivered to the hot reservoir. The heat conducted and radiated is measured by running the device with no load and measuring the heat input needed to maintain the temperature of the hot side (Figure 6).

Part A: Heat Engine Efficiency

In this experiment you will determine the actual efficiency and the Carnot efficiency of the heat engine as a function of the operating temperatures.

- Prepare the ice-water bath and immerse both rubber tubes from the Thermal Efficiency Apparatus into the bath.
- Plug the 9V transformer into the wall socket and into the pump on the Thermal Efficiency Apparatus. You should now hear the pump running and water should be coming out of the rubber hose marked “out”.
- Plug the ohmmeter into the thermistor terminals.
- Connect a DC power supply and a voltmeter and ammeter to the heater block terminals. Adjust the voltage to about 11V. **NOTE:** This is just a suggested value chosen to make the hot temperature nearly at the maximum allowed. Any voltage less than 12V is suitable. The Thermal Efficiency Apparatus should not be run for more than 5 minutes with the hot side above 80°C. A thermal switch will automatically shut off the current to the heater block if it exceeds 93°C to prevent damage to the device.
- Connect the 2Ω load resistor with a short patch cord as shown in Figure 4. Connect a voltmeter across the load resistor. The choice of the 2Ω load resistor is arbitrary. Any of the load resistances may be used.

Preliminary Procedure

Allow the system to come to equilibrium so that the hot and cold temperatures are constant. This may take 5 to 10 minutes, depending on the starting temperatures. To speed up the process, increase the voltage across the heating resistor momentarily and then return it to the original setting. If it is desired to cool the hot side, the voltage can be momentarily decreased. Remember that the thermistor resistance goes down as the temperature increases. Measure the temperature resistances of the hot side and the cold side by using the toggle switch to switch the ohmmeter to each side. Record the readings and convert the resistances to temperatures using the chart on the front of the device and record these temperatures. Record also the voltage (V_H) across the heating resistor, the current (I_H), and the voltage across the load resistor (V_W). Lower the voltage across the heating resistor by about 2 V and repeat the procedure until data for five different hot temperatures have been taken.

Analysis

For each of the data runs, calculate the power supplied to the hot reservoir, P_H , and the power used by the load resistor, P_W , and record these values in a Table. Calculate the temperature difference for each trial, the actual efficiencies from the powers, and the Carnot (maximum) efficiencies from the temperatures. Report all these values in a Table. To compare the actual efficiency to the Carnot efficiency, construct a graph. Plot the Carnot efficiency vs. ΔT and also plot the actual efficiency vs. ΔT . This may be done on the same graph.

Questions:

- The Carnot efficiency is the maximum efficiency possible for a given temperature difference. According to the graph, is the actual efficiency always less than the Carnot efficiency?
- Does the Carnot efficiency increase or decrease as the temperature difference increases?
- Does the actual efficiency increase or decrease as the temperature difference increases?
- The Carnot efficiency represents the best that a perfect heat engine can do. Since this heat engine is not perfect, the actual efficiency is a percentage of the Carnot efficiency. The overall (actual) efficiency of a real heat engine represents the combination of the engine's ability to use the available energy and the maximum energy available for use. From the data taken, what is the percentage of available energy used by this heat engine?
- The actual efficiency of this heat engine is very low and yet heat engines of this type are used extensively in remote areas to run things. How can such an inefficient device be of practical use?

Detailed Study:

You will now determine the actual efficiency and the Carnot efficiency of the heat engine and then compensate for the energy losses to show that the compensated actual efficiency approaches the Carnot efficiency.

- Prepare the ice-water bath and immerse both rubber tubes from the Thermal Efficiency Apparatus into the bath.
- Plug the 9V transformer into the wall socket and into the pump on the Thermal Efficiency Apparatus. You should now hear the pump running and water should be coming out of the rubber hose marked “out”.
- Plug the ohmmeter into the thermistor terminals.

To obtain all the necessary data for the heat engine it is necessary to run the Thermal Efficiency Apparatus in two different modes. The Heat Engine Mode determines the actual efficiency of the Peltier device. The Open Mode determines the losses due to conduction and radiation. Data from both modes is used to calculate internal resistance and the Carnot Efficiency.

Heat Engine Mode

Connect a DC power supply and a voltmeter and ammeter to the heater block terminals. Turn on the voltage to about 11 V. NOTE: This is just a suggested value chosen to make the hot temperature nearly at the maximum allowed. Any voltage less than 12V is suitable. The Thermal Efficiency Apparatus should not be run for more than 5 minutes with the hot side above 80°C. A thermal switch will automatically shut off the current to the heater block if it exceeds 93°C to prevent damage to the device.

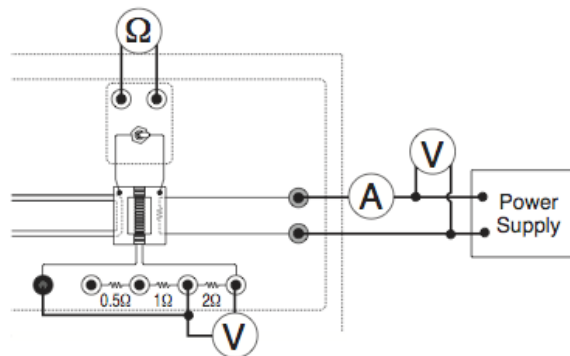


Figure 7

Connect the 2Ω load resistor with a short patch cord as shown in Figure 7. Connect a voltmeter across the load resistor. Allow the system to come to equilibrium so that the hot and cold temperatures are constant. This may take 5 to 10 minutes, depending on the starting temperatures. To speed up the process, increase the voltage across the heating resistor momentarily and then return it to 11V. If it is desired to cool the hot side, the voltage can be momentarily decreased. Remember that the thermistor resistance goes down as the temperature increases. Now measure the temperature resistances of the hot side and the cold side by using the toggle switch to switch the ohmmeter to each side. Record the readings in a Table. Convert the resistances to temperatures using the chart on the front of the device. Record the voltage (V_H) across the heating resistor, the current (I_H), and the voltage across the load resistor (V_W) in the Table.

Open Mode

- Disconnect the patch cord from the load resistor so no current is flowing through the load and thus no work is being done. Now all the power delivered to the heating resistor is either conducted to the cold side or radiated away. Leave the voltmeter attached so that the Seebeck voltage (V_s) can be measured. (see figure 6)
- Decrease the voltage applied to the hot side so that the system comes to equilibrium at the

same hot temperature as in the Heat Engine Mode. Since the temperature difference is the same as when the heat engine was doing work, the same amount of heat is now being conducted through the device when there is no load as when there is a load. (It may not be possible to exactly match the previous cold temperature.). Record the resistances in a Table and convert them to degrees. Also record V_H , I_H and V_s .

Calculations:

Actual Efficiency: Calculate the actual efficiency using

$$e = \frac{P_W}{P_H}$$

where $P_w = V_w^2/R$ and $P_H = I_H V_H$

Record the powers as in the example Tables shown below:

Internal Resistance = $r =$ _____

Mode	T_h (K)	T_c (K)	P_h	P_w	I_w
Engine (2Ω load)					
Open					

	Actual	Adjusted	Maximum (Carnot)	% Difference
Efficiency				

Maximum Efficiency: Convert the temperatures to Kelvin. Calculate the Carnot efficiency using the temperatures and record as shown above.

Adjusted Efficiency: The purpose of the following calculations is to account for all the energy losses and adjust the actual efficiency so that it matches the Carnot efficiency.

First, the work done in the actual efficiency calculation only includes V^2/R for the power dissipated by the load resistor R but, to account for total work done by the device, it should also include $I^2 r$ for the power dissipated by the internal resistance, r , of the device. This Joule heating of the Peltier device is not counted in the actual efficiency because it is not useful work. Thus, in the adjusted efficiency, the total work done in terms of power is

$$P'_W = P_W + I_W^2 r = \frac{V_W^2}{R} + I_W^2 r$$

where $I_w = V_w/R$. Calculate I_w for the 2Ω load and record it.

Second, the heat input must be adjusted. The heat that leaves the hot reservoir goes two places. Part of it is actually available to be used by the heat engine to do work while the other part bypasses the engine either by being radiated away from the hot reservoir or by being conducted through the Peltier device to the cold side. The portion of the heat which bypasses the engine by radiation and conduction would be transferred in this same manner whether or not the device is connected to a load and the heat engine is doing work. Therefore this heat can be considered to not be available to do work and should not be included in the heat input in the adjusted efficiency.

$$P'_H = \text{available heat} = P_H - P_{H(open)}$$

The Thermal Efficiency Apparatus is run with a load connected to measure P_H (Figure 4) and then the load is disconnected and the power input into the hot reservoir is adjusted to maintain the temperatures (less power is needed when there is no load since less heat is being drawn from the hot reservoir). See Figure 6. $P_H(\text{OPEN})$ is the power input to the hot reservoir when no load is present. Since, while there is no load, the hot reservoir is maintained at an equilibrium temperature, the heat put into the hot reservoir by the heating resistor must equal the heat radiated and conducted away from the hot reservoir. So measuring the heat input when there is no load determines the heat loss due to radiation and conduction. It is assumed this loss is the same when there is a load and the heat engine is operating.

Having accounted for the obvious energy losses, the adjusted efficiency should match the Carnot efficiency which assumes no energy loss. The adjusted efficiency is

$$e'_{adjusted} = \frac{P'_W}{P'_H} = \frac{P_W + I_W^2 r}{P_H - P_{H(open)}}$$

Calculate the internal resistance, r , using the equation

$$r = \left(\frac{V_p - V_W}{V_W} \right) R$$

Record this resistance, then calculate the adjusted efficiency and record the result. Calculate the percent difference between the adjusted efficiency and the Carnot (maximum) efficiency.

$$\%Difference = \frac{e_{max} - e_{adjusted}}{e_{max}} \times 100\%$$

Questions:

- If the difference between the temperature of the hot side and the cold side was decreased, would the maximum efficiency increase or decrease?
- The actual efficiency of this heat engine is very low and yet heat engines of this type are used extensively in remote areas to run things. How can such an inefficient device be of practical use?
- Calculate the rate of change in entropy for the system which includes the hot and cold reservoirs. Since the reservoirs are at constant temperature, the rate of change in entropy is

$$\frac{\Delta S}{\Delta t} = \frac{\Delta Q / \Delta t}{T} = \frac{P}{T}$$

for each reservoir. Is the total change in entropy positive or negative? Why?

Part B: Heat Pump Coefficient of Performance

Before doing this experiment, it is necessary to perform the HEAT ENGINE EFFICIENCY (Part A) experiment to get the data necessary to determine the internal resistance of the Peltier device. To complete the measurements for this experiment, use the following instructions to run the apparatus as a heat pump (pumping heat from the cold side to the hot side):

Prepare the ice-water bath and immerse both rubber tubes from the Thermal Efficiency Apparatus into the bath. Plug the 9V transformer into the wall socket and into the pump on the Thermal Efficiency Apparatus. You should now hear the pump running and water should be coming out of the rubber hose marked “out”. Disconnect the power supply to the hot side. Connect the power supply directly across the Peltier device with no load resistance. See Figure 8. Connect an ammeter and a voltmeter to the power supply.

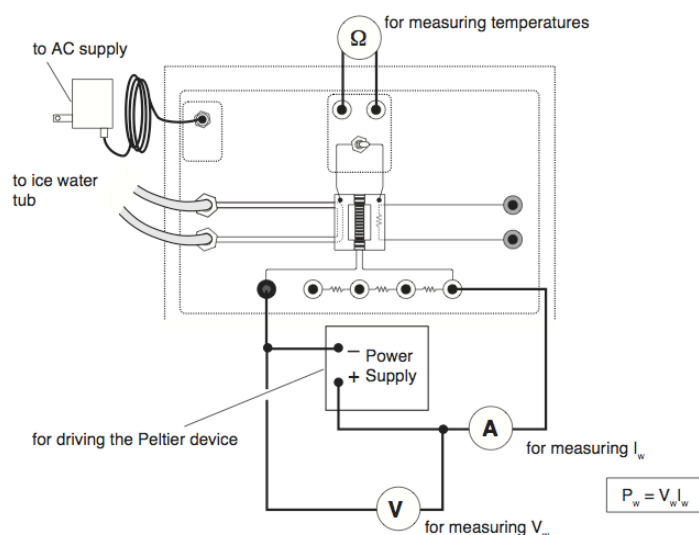


Figure 8: Heat pump mode

Procedure:

Increase the voltage until equilibrium is reached at the same hot temperature as in the previous experiment. The hot side is now being heated by heat pumped from the cold side rather than the heater resistor. Record the resistances and convert them to degrees. Also record the voltage (V_w) and the current (I_w) in a Table.

Analysis:

Actual Coefficient of Performance: Calculate the actual COP using the data taken in the Heat Engine experiment.

$$\kappa = \frac{P_C}{P_W} = \frac{P_H(OPEN) - P_W}{P_W}$$

Record this result in a Table. Maximum Coefficient of Performance: Calculate the maximum COP using

$$\kappa_{MAX} = \frac{T_C}{T_H - T_C}$$

Adjusted Coefficient of Performance: Part of the power being applied to the Peltier device is being dissipated in the Joule heating of the internal resistance of the device rather than being used to pump the heat from the cold reservoir. Therefore, to adjust for this, I^2r must be subtracted from the power input to the Peltier device. Then the COP becomes the heat pumped from the cold reservoir divided by work done to pump the heat, rather than dividing by the work done to pump the heat and heat the internal resistance. In terms of the power,

$$\kappa_{Adjusted} = \frac{P_{H(OPEN)} - P_W}{P_W - I_W^2 r}$$

Record this in the Table then calculate and record also the percent difference between the adjusted COP and maximum COP:

$$\% \text{ Difference} = \frac{\kappa_{MAX} - \kappa_{adjusted}}{\kappa_{max}} \times 100\%$$

Questions:

If the difference between the temperature of the hot side and the cold side was decreased, would the maximum COP increase or decrease?

Calculate the rate of change in entropy for the system which includes the hot and cold reservoirs. Since the reservoirs are at constant temperature, the rate of change in entropy is

$$\frac{\Delta S}{\Delta t} = \frac{\Delta Q / \Delta t}{T} = \frac{P}{T}$$

for each reservoir. Is the total change in entropy positive or negative? Why?

Part C: Load for optimum performance

This experiment finds the load resistor which maximizes the power output of the heat engine. The power delivered to the load resistor, R , is $P = I^2 R$. The amount of current that flows through the load resistor varies as the load is varied. From Figure 5, $V_S = I(r+R)$ where V_S is the Seebeck voltage and r is the internal resistance of the Peltier device.

So the power can be expressed in terms of the Seebeck voltage, the internal resistance, and the load resistance:

$$P = \left(\frac{V_S}{r + R} \right)^2 R$$

Assuming the Seebeck voltage remains constant if the temperatures of the hot and cold reservoirs are constant, the power can be maximized with respect to the load resistance by taking the derivative and setting it equal to zero:

$$\frac{dP}{dR} = \frac{V_S^2 (r - R)}{(r + R)^3} = 0$$

This shows that when the load resistance is equal to the internal resistance of the Peltier device, the power delivered to the load will be a maximum.

Procedure: Connect a DC power supply and a voltmeter and ammeter to the heater block terminals. Turn on the voltage to about 11V (Any voltage less than 12V is suitable. The Thermal Efficiency Apparatus should not be run for more than 5 minutes with the hot side above 80°C. A

thermal switch will automatically shut off the current to the heater block if it exceeds 93°C to prevent damage to the device.). Connect the 0.5 Ω load resistor with a short patch cord as shown in Figure 9. Connect a voltmeter across the load resistor.

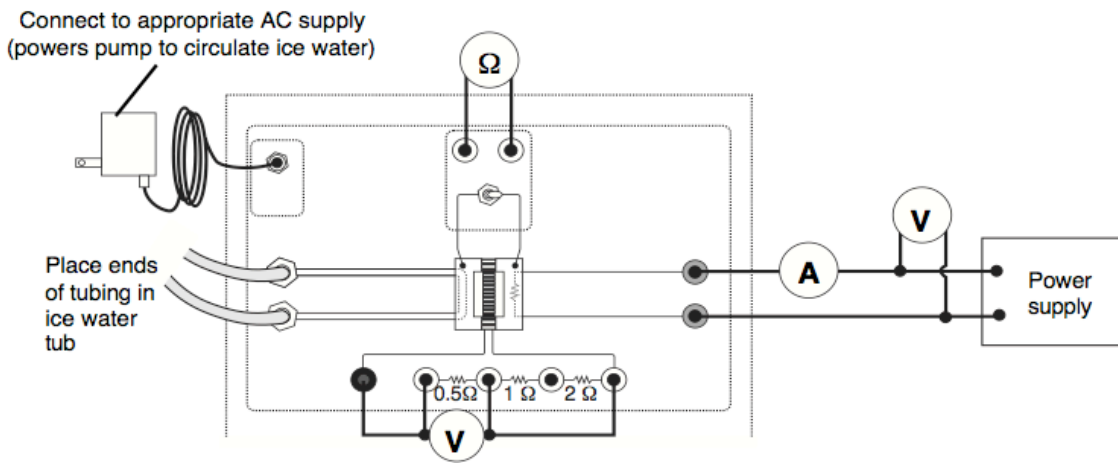


Figure 9: Connecting the 0.5Ω load resistor

NOTE: Alternatively, a variable power resistor (rheostat) may be used in place of the load resistors supplied with the Thermal Efficiency Apparatus. This has the advantage of being able to continuously vary the load resistance. However, it will be necessary to measure the resistance of the load.

Allow the system to come to equilibrium so that the hot and cold temperatures are constant. This may take 5 to 10 minutes, depending on the starting temperatures. To speed up the process, increase the voltage across the heating resistor momentarily and then return it to 11 V. If it is desired to cool the hot side, the voltage can be momentarily decreased. Remember that the thermistor resistance goes down as the temperature increases.

Measure the temperature resistances of the hot side and the cold side by using the toggle switch to switch the ohmmeter to each side. Record the readings in a Table. Convert the resistances to temperatures using the chart on the front of the device as explained earlier. Record the voltage (V_H) across the heating resistor, the current (I_H), and the voltage across the load resistor (V_W). Then calculate the power input to the hot side, $P_H = I_H V_H$, and the power dissipated by the load resistor, $P_L = V_W^2 / R$. Calculate the efficiency, $e = P_L / P_H$ and record all these values in a Table. Adjust the power input to the hot side to keep the temperature of the hot reservoir at the same temperature as it was for the 0.5 Ω resistor while repeating the steps above for the other possible load resistances: 1, 1.5, 2, 2.5, 3, and 3.5 ohms.

Questions

- For which load resistor is the efficiency a maximum?
- How does the load resistance for optimum efficiency compare with the internal resistance measured in Part A?

Lecture Notes

Contents	Page
Recommended reference books on statistics.....	2
Graphs	2
Straight-line graphs	3
Measurements and errors	4
Mean, standard deviation, and standard error of the mean	5
Histograms and distributions.....	7
Combining or ‘propagating’ errors	9
Probability	11
Binomial distribution	11
Poisson distribution	12
Gaussian distribution.....	13
Fitting data	15
Straight-line fits.....	15
The χ^2 test.....	16
Weighted means	16
Tables of Gaussian integrals	

LECTURE NOTES

Recommended reference books on statistics

For general reference in the course, we have recommended the book:

- *Practical Physics*, by G.L. Squires (Cambridge University Press, 4th ed. 2001)

This has material on experimental methods and preparation of reports, as well as some chapters on statistics. However, for a more thorough treatment of statistics there are better books. Two that we like are:

- *Statistics*, by R.J. Barlow (Wiley, 1989)
- *An Introduction to Error Analysis*, by J.R. Taylor (University Science Books, 2nd ed. 1997)

We do *not* require you to buy these. They will all be available for consultation in the Short Loan Collection of the Main Library, and can also be borrowed from the technicians in the laboratory. Either of the two statistics books would be a useful reference to have for future work.

Graphs

Most experiments involve measurements to be plotted in one form or another. This is because people find it much easier to see both expected and unexpected behaviour pictorially than in a long table containing many numbers. A graph could, for example, help to pick out elementary mistakes in measurements, find a peak (maximum) or a trough (minimum), draw a calibration for obtaining values from later measurements, or verify a theoretical relationship and find its parameters. Consider, for example, using a prism spectrometer to measure the angle of refraction θ for each of a number of spectral lines of known wavelength λ . Although we might not know the exact relation between θ and λ , experience suggests that it will be smooth without any sudden kinks (we've all seen a rainbow with one colour blending smoothly into the next), so on a simple plot of θ versus λ any spectral line that was misidentified, or any angle that had been misread, would not fall on a smooth curve. This plot could also be used as a calibration graph for the spectrometer since we could use it to find the wavelength of an unknown line. Finally, suppose we want to check the relationship between refractive index n and wavelength:

$$n = A + B/\lambda^2$$

and find the constants A and B . We would calculate n from the refraction angle, and plot n versus $1/\lambda^2$. This will produce a straight line if the relation is valid, giving A and B as the intercept and slope of the line respectively.

Note on graph terminology: When we say 'plot n versus $1/\lambda^2$ ' it matters which we say first: what we mean is that n should be on the vertical (y) axis and $1/\lambda^2$ should be on the horizontal (x) axis. The x -axis (*abscissa*) is usually an *independent* variable (i.e. something that we *control*), while the y -axis (*ordinate*) is a *dependent* variable (i.e. something that we *measure*).

Other mathematical relations can also be 'rectified' as we did above, that is converted to a linear relation that will yield a straight line. For example, taking logs of the relation $y = AB^x$ gives:

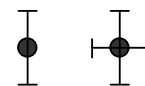
$$\log y = \log A + x \log B$$

We would plot $\log y$ versus x . to find intercept A and slope $\log B$. Similarly, taking logs of the relation $y = Ax^n$ gives:

$$\log y = \log A + n \log x$$

and now we would plot $\log y$ versus $\log x$ to find intercept A and slope n .

It is often helpful to plot several sets of data using the same set of axes, so they can more easily be compared. If you do this you *must* use different symbols (\times , \oplus , \bullet , etc.) for the different sets of data. **Error bars** should be included for either (or both) x and y variables. Show them like the ones at the right.



Error bars should ideally be standard statistical errors if you have a reliable way of calculating these (e.g. from repeated measurements), but failing that your own estimates or best guesses are better than nothing. Error bars may be very different in size for different parts of the graph.

The points on a graph need not be uniformly distributed. Straight lines only need few points, but peaks, troughs and sharp bends need more points to define them precisely. A quick measurement scan covering the entire range first can save a lot of time later since this gives us the limits of the variables for the axes, and tells us where points are most needed (experience counts here!).

Make a rough plot of the ‘raw data’ as you go, straight into your worksheet or notebook. Don’t worry about neat straight line axes drawn carefully in coloured pencil — one reason why we asked you to use the quadrille ruled notebooks from experiment 8 onwards is that rectangular boxes are already drawn and all you need add are tick marks showing the scales. Make these scales convenient to work with — 10 divisions should normally represent 1, 2 or 5 units. Sometimes 4 units is acceptable but certainly not 3! (Why not? Because it is virtually impossible to plot points correctly and quickly, and then to read off values along the axis.)

For plotting over a very wide range of values, graphs with logarithmic scales on one or both axes is available. This might be useful if, for example, you are measuring something at different frequencies, say 1 Hz, 10 Hz, 100 Hz, 1000 Hz, 10,000 Hz, etc. No ordinary graph could show all these points adequately, but the log scale shows each power of ten the same distance apart on an axis (note that all log graphs use logs to base 10, not natural logs). Care is needed when finding the gradient of a graph with a logarithmic axis — you must remember that the increment is $(\log y_2 - \log y_1)$, not $(y_2 - y_1)$, and that this is a dimensionless quantity, unlike y itself.

Straight-line graphs

These are often the most useful. A good idea of the line which best represents your measurements can be obtained by eye. (Try holding the paper nearly level with your eye and looking along the line of points.) A well-fitted line should leave roughly equal numbers of points on each side of the line. As for the errors in the slope and intercept, the method of deliberately drawing a line which is *not* the best (colloquially and quite wrongly called the ‘worst’ lines method — there are many worse lines than that!) is not satisfactory although it gives a rough idea of what the slope and intercept could not be.

A far better method is to use a least squares fit as, for example, is done in the computer program **PhysPlot**. This takes the sum of the squares of the deviations (distances) of the points from the line and finds the values of slope and intercept that minimise it. In other words, the expression:

$$\sum_i (y_i - mx_i - c)^2$$

is minimised. **PhysPlot** gives the slope m and the intercept c , as well as the errors in these, $\pm m$ and $\pm c$, together with a so-called χ^2 which is a number that indicates how well the points fit to a straight line — the lower the number the better the fit. There is more on this later.

We emphasise that before doing a full fit, the graph *must* be sketched first to see if it really is anything like a straight line, or at least whether some part of it is linear.

Measurements and errors

A measurement consists of three important ingredients. First there is a *number*, and then *units* of some kind. In addition we need some indication of how *reliable* the measurement is, or what confidence we have that a repeated measurement would produce a similar result. *Precision* and *accuracy* are other words that are used in this context, but most often we use the word *error*, which is meant to imply there is some correct value that we are trying to find. In this context, *error* does *not* mean that we have made a mistake! Like many words, its usage in technical English is not the same as in everyday English. In general, if we repeat measurements we will not get exactly the same result each time. The *error* is an estimate of how much *spread* or *dispersion* we would expect to see in such repeated measurements. These lectures will gradually add a more precise meaning to the concept of the error by introducing the subject of statistics.

However, there are three very important ideas that you should make sure you absorb as quickly as possible since, if nothing else, you will lose marks if you don't! They are:

- It is always better to make a very *rough guess* of the error on a final result than to quote *no* error at all. Even if it is a pure guess, you at least tell the reader your feeling of how accurate you think your measurement is.
- The number of significant figures (s.f.) that you use to quote a result already has the effect of implying some kind of accuracy. If you say $c = 2.98 \text{ ms}^{-1}$ there is an implication that you believe it is not likely to be either 2.97 or 2.99 ms^{-1} . Bear this in mind every time you write down a result from a calculator or a computer program! You will probably go years between occasions when it is justified to use 6, 7 or 8 significant figures. Normally you should use about 3 s.f. If you have the kind of calculator that allows you to pre-set the accuracy with which results are displayed, it is worth doing that to remind you. You will not lose calculation precision, and even if you did you probably could not justify needing that precision anyway.
- We shall find, and you should never forget, that any estimate you make of an error always carries a big error itself (the 'error of the error'). If you ever find that your calculated error on an error is bigger than that for the original quantity, you should realise that you are wasting your (and the marker's) time. A large part of these lectures will involve teaching and justifying various approximations and 'rule-of-thumb' techniques. Try to understand and then to utilise these as early as possible.

There are basically two distinct *types* of error, *systematic errors* and *statistical errors*. We will define them in turn.

Systematic errors

These are incorrect, or shifted, results that arise from the *system* or method of measurement. In general, if the measurement is repeated without changing the system you will get roughly the same incorrect measurement. Examples are using a tape measure that has been stretched, or a clock that runs fast, but often it's more subtle, e.g. the effect of temperature on your length measurement, the angle that you are forced to view your apparatus from (parallax), or even some theory that you are using to work out intermediate results not being justified in the circumstances you are using.

With some justification, the book by Squires suggests that the term systematic error is a euphemism for experimental mistake. It is next to impossible to make a good estimate of likely systematic errors, since you have to be aware of the possibilities in order to be able to investigate them. Once you do know about them, you can usually correct for their effect. Bear in mind that despite our image of performing measurements with an open mind, the truth is that whenever we make a measurement we have some expectation of what the likely result will be. This is not an entirely bad thing, as it helps us to be naturally suspicious of things going wrong.

At undergraduate level, the best we can hope to do is that once our suspicions are aroused we can check, for example by exploiting any natural symmetries the system may have. For example, a given length should be the same no matter which end we measure from, and it should also be the same whichever tape measure we use. If we only have two tape measures, and they give conflicting results, we do not know if one has stretched or the other has shrunk. It would be better to be able to resolve the issue, but under these circumstances we might fairly quote the difference as a possible systematic error of using tape measures in general.

We say that some systems of measurement are intrinsically ‘more accurate’, for example a micrometer screw gauge compared to a tape measure. In general, this means that we are saying that we can assign a smaller systematic error to that system.

Statistical (or random) errors

The remainder of this lecture course will concern the vast majority of measurements that we make which, because of many small random effects, are slightly different each time we repeat them. If we use a ruler to measure a length, fluctuations in where we place the end, the precise way we estimate the graduations, or even the lighting conditions or eyesight may well give slightly different answers each time. With a screw gauge the differences will be much smaller since these effects are not contributing; probably we are now sensitive to random minor imperfections on the bounding surface being measured. In either case, the random nature of the fluctuations make the results susceptible to statistical analysis. With repeated measurements, not only can we make a better estimate of what the ‘true’ value might be, we can also estimate to what degree of confidence we might believe that this is the case. In turn, eventually this will tell us how confident we may be that a result is consistent, or otherwise, with a given theory. If there are two competing theories, we may then be able to discard one as being too unlikely, and science has taken another forward step!

Mean, standard deviation, and standard error of the mean

When we make a set of repeated measurements, we generally notice the distinctive feature that they cluster around some value. This is called the ‘central tendency’. To estimate the *correct* value we could use at least three different procedures. They are:

- Simply put all of the measurements in numerical order, and choose the middle one. This is called the *median* of the set. If there is an even number of measurements this can be resolved by taking an average of the middle two.
- Choose the most frequent or popular measurement. This is called the *mode* of the set.
- However, it can be shown that the *best* estimate of the true value is to take an *arithmetic average* of all results. So for n measurements we add them all together and divide the result by n . This is called the *mean* of the set. Mathematically, for a quantity x (e.g. a length) this will be denoted by \bar{x} or $\langle x \rangle$. If there are n different measurements, each denoted by x_i , we define the mean as:

$$\langle x \rangle = \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

The mean is not the only important factor that describes a set of measurements. The next most important parameter is their *spread* or *distribution* about the mean. This tells us a lot about the quality of the measurements — to what accuracy can we believe the mean value, and hence how many significant figures should we write when we quote the mean value. In short it constitutes an *error* on the mean value. In general we would expect that a more accurate technique, like the micrometer screw gauge compared to a ruler, would give a much smaller spread of values.

How can we ‘design’ one number to specify the width of our distribution? Clearly the expression $d_i = (x_i - \bar{x})$ (which is called the **residual** of measurement i) gives us some measure of the distance, or deviation, of each point from the mean, so you might be tempted to think that the average value of all the residuals would be a good candidate. However, this is a *signed* quantity (either positive or negative), and indeed by definition this average is *zero*. To get around this, we average the *square* of this quantity to give us the **variance**. More usefully, if we now take the square root of the variance we get the **standard deviation** of the set of measurements. Mathematically, the variance is denoted by σ^2 , so that the standard deviation, σ , is:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

People familiar with electrical signals might recognise this as the **root-mean-square** (i.e. the square root of the average square) or **r.m.s.** value. It is important to note that σ has the same units as the original quantity being measured, x . Mathematically, it is possible to show that if we had an infinite number of measurements instead of just the n available to us, then the standard deviation would be slightly bigger and that a *better* estimate would be:

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

The replacement of $(n-1)$ for n is referred to as Bessel’s correction. It also ensures that we don’t try to measure a spread with only one measurement, but numerically it is seldom very important in practice.

We will show later that, in general, roughly two-thirds of all measurements should lie within $\pm\sigma$ of the mean, and the remaining one-third outside of that. If that is not the case, then you should suspect that either there is something funny about the data, or that the error has been calculated incorrectly.

The quantities \bar{x} and σ are, of course, characteristic of a particular set of n measurements that we make. If we were to *repeat* all the measurements we will have n slightly different values of these quantities. It can be shown (see for example Squires, chapter 3) that these repeated \bar{x} values themselves tend to cluster around the true value, with a standard deviation given by:

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

and so this is called the **standard error of the mean** or just the **standard error**. In your experiments or homework exercises it would be correct to think of σ as the error in a *single* measurement, and σ_m as the error in the *mean* of those measurements. Note that as n gets bigger the spread of results, characterised by σ , stays the *same*, but the standard error of the mean gets *smaller*, i.e. your result becomes more reliable because you have determined the mean more accurately. However, this only improves like the square root of n .

Calculating standard deviation

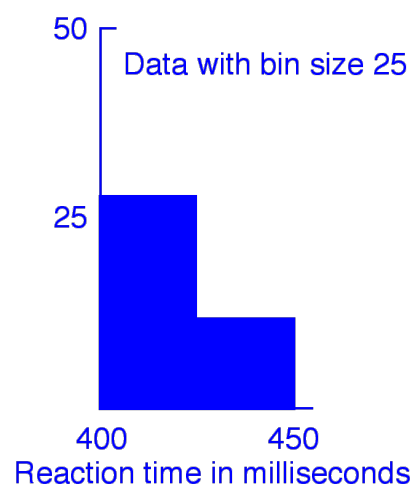
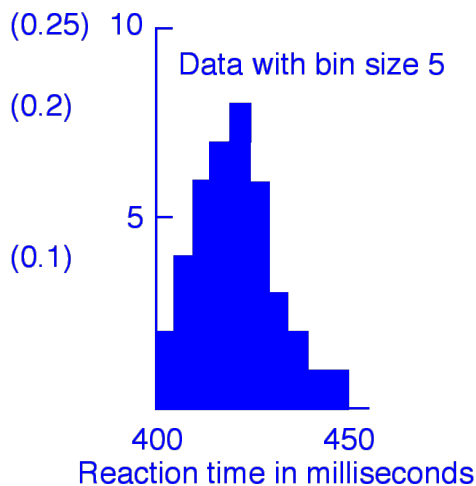
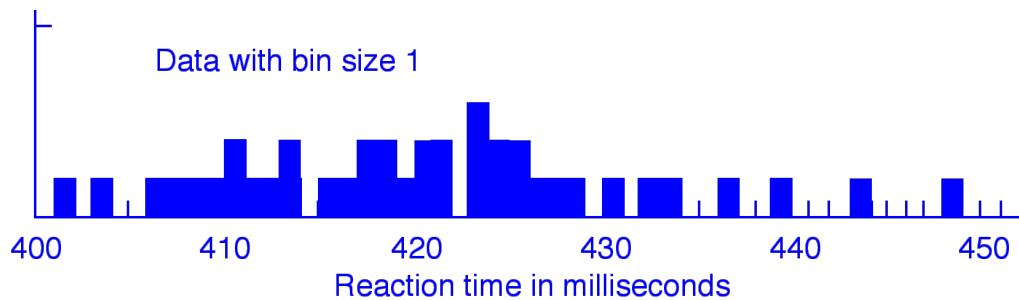
If we want to calculate the standard deviation as we keep on adding data measurements (e.g. using a computer), there is a difficulty in doing a running calculation since in the formulas given above we need the overall mean in order to work out the standard deviation, and every time we add a data point the mean will change slightly. We can get around this by doing some algebra:

$$\begin{aligned}
\sigma^2 &= \frac{1}{n} \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right) \\
&= \frac{1}{n} \left(\sum_{i=1}^n (x_i^2 + \bar{x}^2 - 2x_i\bar{x}) \right) \\
&= \frac{1}{n} \left(\sum_{i=1}^n x_i^2 + n(\bar{x})^2 - 2\bar{x} \sum_{i=1}^n x_i \right) \\
&= \frac{1}{n} \sum_{i=1}^n x_i^2 + \bar{x}^2 - 2\bar{x} \frac{1}{n} \sum_{i=1}^n x_i \\
&= \frac{1}{n} \sum_{i=1}^n x_i^2 + \bar{x}^2 - 2\bar{x} \cdot \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \\
\sigma^2 &= \frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i \right)^2 = \overline{x_i^2} - \bar{x}^2 = \langle x_i^2 \rangle - \langle x \rangle^2
\end{aligned}$$

This shows that if we work out both the sum of the measurements and the sum of the squares of the measurements as we go along, then we can calculate the standard deviation without having to find a new value of the mean each time.

Histograms and distributions

In order to represent a set of measurements pictorially we can show them in boxes or *bins*, for which the horizontal position shows the measurement values and the width of the bin represents the range of values that each box includes. The number of boxes in a given bin, or the height of that bin, is proportional to the number of measurements giving values within that range. Such a display is called a *histogram* and examples are given below.



A convention used for histograms is that if the left hand side of a bin is labelled, say, 4.0 and the bin width is 1.0 (so that the next bin is labelled 5.0), then that bin will contain all measurements from 4.0 up to 4.999... but not including 5.0.

The choice of the width of the bins always requires some thought. Our aim is to display best *how* the measurements are distributed about the mean value. All three histograms above display the same set of measurements. If the bin width is *too narrow*, as in the top histogram, there will only be very few measurements or none within each bin, so that the height within each bin will not represent anything useful. If the bins are *too wide*, as in the lower right histogram, then most of the measurements will be in only a few bins and we will not learn much about the shape — the extreme case is one bin containing all the measurements, which is of no value! The lower left histogram is a much better choice, since the shape is shown well and the number of measurements per bin varies fairly smoothly.

As we take more and more measurements under the same conditions, the general shape of the histogram is preserved but the bins can be made narrower. We can then learn more about the shape, and so obtain more detailed information. For very large numbers we eventually get a smooth-looking curve in most cases.

It is *much* easier to compare two histograms if the vertical axis has the same scale no matter how many measurements have been made. We can do this by dividing the number of measurements in each bin by the total number of measurements that have been made. This is called **normalising** the data. For the data shown in the figure there are 40 measurements, so that each measurement counts as 0.025. On the lower left histogram in the figure we have shown in parentheses how the vertical axis would be relabelled. The sum of all bins is then unity, i.e. 1.

If the measured quantity is x , the bin width is often denoted as Δx or δx as it gets smaller and dx as it eventually tends towards zero. If the y axis is made to represent not just the *number* of measurements but the *number per unit x* then we say that the y axis is a **distribution function**, $f(x)$. The number of measurements in a given bin must then be $f(x)\Delta x$, or $f(x)dx$ in the limit. If in addition $f(x)$ is normalised by dividing it by the total number of measurements, then $f(x)\Delta x$ becomes the **fraction** of measurements that occur in the narrow interval from x to $x + \Delta x$. This is then the **probability** of getting a measurement in this range. We can sum up several such bins to get the probability for a wider range, i.e.

$$\sum_{i=p}^q f(x_i)\Delta x$$

In the limit as Δx tends to zero this gives us the formula:

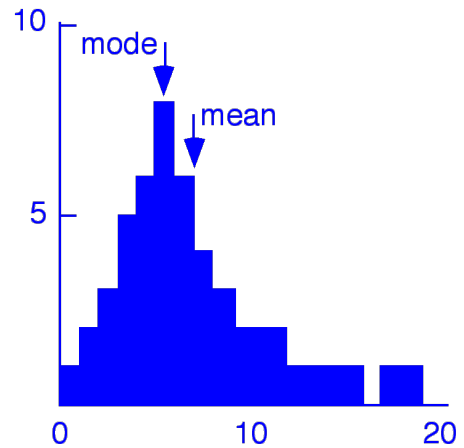
$$\int f(x)dx$$

which says that the probability to get a result in a given range of measurements is just the area underneath the curve in that range. Because we have a normalised distribution, the area under the *entire curve* must be *unity*.

The histograms drawn on the previous page are more or less symmetric, and when that happens the mean (420.3 for the example shown) is approximately equal to the mode (the most frequent measurement; 423 for the example) and is also approximately equal to the median (the value for which half the measurements are below and the other half above; 421 for the example). ‘Overall’ quantities like the mean, mode and median are much less useful if there is a complex structure, for instance several well-defined peaks. However, even when there *is* one peak it’s sometimes not symmetric, for example there might be a long tail on one side. A measure of how asymmetric, or skewed, a distribution is can be given by a quantity called the **skewness**:

$$\text{skewness} = \frac{\text{mean} - \text{mode}}{\sigma}$$

The skewness is zero if the distribution is symmetric, and can be positive or negative. For the example to the right, the mean is 6.48 and the mode is 5. The standard deviation is 4.07, so the skewness is +0.36.



Combining or 'propagating' errors

It is usually straightforward to estimate errors on the primary quantities that we measure directly. For example, half a division width on an instrument dial (or possibly a tenth of a division if they are widely spaced, and you have good eyesight!) is often a good enough estimate. For a more important variable you should try if possible to repeat some measurements, and so get a real estimate of the standard deviation.

The problem now is how to work out the error on a *derived* quantity which depends directly on the *measured* quantity via some known formula or expression. In other words, we want to know what will be the likely variation of some important variable caused by the lack of precision, or error, on the primary measured quantity.

The technique is usually to break up an expression into a series of smaller algebraic steps such as taking powers, multiplying and dividing, adding and subtracting (in that order, according to rules you have hopefully learnt already). The rules for combining the errors in these sub-cases are proved in the lectures, but it's only the results we are interested in so they can be summarised here. Note that to keep things relatively simple, we always assume that variables x and y are independent of each other, i.e. are *uncorrelated*.

Addition or Subtraction: e.g. $Q = x + y$ or $Q = x - y$. In *either* case, we **add the errors** on x and y *in quadrature*. This form of addition occurs frequently in physics and merely means that we take the sum of the *squares* of the errors on x and y (i.e. the variances) to get the *square* of the error on Q (i.e. its variance):

$$(\sigma_Q)^2 = (\sigma_x)^2 + (\sigma_y)^2$$

Note here a strange-seeming quirk in making measurements which involve a difference between two quantities. If $Q = A - B$ and A and B are about the same size, it is quite possible that the error on Q will end up being much bigger than the size of Q itself!

Multiplying by a constant: e.g. $Q = kx$. The *fractional error* on Q is the *same* as that on x :

$$\frac{\sigma_Q}{Q} = \frac{\sigma_x}{x} \quad \text{so since } Q = kx \quad \text{we get } \sigma_Q = k\sigma_x$$

Multiplying or Dividing two variables: eg $Q = xy$ or $Q = x/y$. In either case the *fractional errors* on x and y are **added in quadrature**:

$$\left(\frac{\sigma_Q}{Q}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2$$

Raising to a power: e.g. $Q = x^p$ In this case the *fractional error* on Q is **p times the fractional error** on x :

$$\left(\frac{\sigma_Q}{Q}\right) = p\left(\frac{\sigma_x}{x}\right)$$

Example: The volume, V , of a sphere may be found by measuring its diameter, D . The error on D , (ΔD), could be for example a known calibration error of a micrometer. Now,

$$V = \pi D^3 / 6$$

The fractional error on V is just the same as the fractional error on D^3 , as the $\pi/6$ is just a multiplicative constant. The power rule then gives us:

$$\left(\frac{\sigma_V}{V}\right) = 3\left(\frac{\sigma_D}{D}\right)$$

Calculus

There are more complicated cases than those involving the simple arithmetic combinations listed above. As might be expected since it is a theory of small differences, calculus can often come to our aid. For example, the expression for the refractive index of a glass prism of angle A is:

$$\mu = \frac{\sin \frac{1}{2}(A + D)}{\sin \frac{1}{2}A}$$

where D is the minimum angle that light is deviated by the prism; see lab. exercise 9. If we *know* the value of A (i.e. it has no error) we can differentiate this expression to get $d\mu/dD$ and then we can identify the error σ_μ with $d\mu$ and the error σ_D with dD , thus enabling us to calculate one in terms of the other. Remember if you are using calculus that dD , the differential, has a *sign*, but we identify it with an error that could go in either direction so σ_D will just be the absolute value.

General functions of more than one variable

For those of you familiar with partial derivatives, if Q is a function of more than one variable $f(x, y, \dots)$ a **general expression** for the variance of Q that can be used in all sorts of cases is:

$$(\sigma_Q)^2 = \left(\frac{\partial f}{\partial x}\right)^2 (\sigma_x)^2 + \left(\frac{\partial f}{\partial y}\right)^2 (\sigma_y)^2 + \dots$$

Some common sense

You can see that once we start to introduce calculus, it is very easy for an error calculation to get out of hand and you may find yourself spending more time and space calculating an error than the original quantity! This may *just* be worth the effort if you are doing a research project, but it is *not* justified in general, and you should have a very good reason if you find yourself doing too much of that in any undergraduate experiment. On the other hand, completely *ignoring* the effect of an error because of the difficulty of calculating it is an even bigger sin!

In the end, an error is an *estimate* of how far we might be away from the true value. In the lectures you will be given an argument why you cannot in any case expect the accuracy on any computed error to be much better than of the order of 10%. For this reason, it is worthwhile getting to know some of the (acceptable!) short cuts. In cases which involve calculus it is sometimes much quicker, in this age of calculators, merely to insert two close values of a variable, for example D in the above, and then look at the resulting change in μ . The fractional change in one can then be related to the fractional change in the other. We can discover approximately how the error is propagated quite simply, by brute force!

We have seen in error theory that we often add sources of error from two different quantities in quadrature. If one source of error (or fractional error if applicable) is three times bigger than another it contributes *nine* times as much to this sum, so after taking the square root the smaller source of error is negligible to within the accuracy we expect our final error to have anyway. This is an excellent justification for simply ignoring many sources of small errors — they are just not important. Rarely is it worth combining more than two errors. Rough estimates of comparative contributions of different errors pay dividends in not wasting time later.

Probability

Probability is a very important concept in physics, especially quantum physics. It is expressed as a *dimensionless* number P , the probability or likelihood for something to happen or for a certain measurement to occur. P is always in the range 0 to 1, where $P = 0$ means something *cannot* happen, and $P = 1$ means it *always* happens. For example, $P = 0.5$ is the estimate to get heads if you toss an unbiased coin.

If we take a frequency histogram $N_i/\Delta x$ and divide the contents of each bin by the total number of measurements N_{tot} to get $(N_i/N_{\text{tot}})/\Delta x$ in each bin, then we have a normalised histogram. N_i/N_{tot} is the fraction of measurements with values between x and $x + \Delta x$, and thus is the *probability* of getting a measurement in this range. Such a histogram is called a *probability histogram*.

Binomial distribution

As the bin size in such a histogram gets smaller and smaller, the measurements become a smooth curve $f(x)$. Can we find theoretical expressions for $f(x)$ which fit for various sets of physical results? In order to simplify the problem, we first consider a simple case where the outcome is an integer — the tossing of a coin. Getting all heads (or tails) appearing on all tosses can only occur in one way, while other combinations having a mixture of heads and tails can occur in more ways than this and hence are more probable in a given trial. Take six coin tosses:

0 heads (6 tails) can occur in only	1 way
6 heads (0 tails) can occur in only	1 way
1 head can occur in	6 ways
5 heads (1 tail) can occur in	6 ways
2 heads and 4 tails can <i>each</i> occur in	15 ways
3 heads (3 tails) can occur in	20 ways

There are thus a total of 64 distinctly different ways that the tossing of a coin six times can turn out. This can be plotted in a histogram. If we actually try, say, 64 trials of six tosses each we almost certainly won't get *exactly* this shape (see figure at right for an example), but if we increase to, say, 6400 trials and divide the contents of each bin by 6400 we should get a reasonably accurate *frequency distribution*. This is called the **binomial distribution**. There are only two possibilities of result, heads or tails, and the coefficients of the expansion below represent the frequency of occurrence. If p and q are the probabilities of getting heads and tails then $p + q = 1$ (heads or tails is a certainty), and $p = q = 1/2$ if the coins are unbiased. We do a *binomial expansion*:

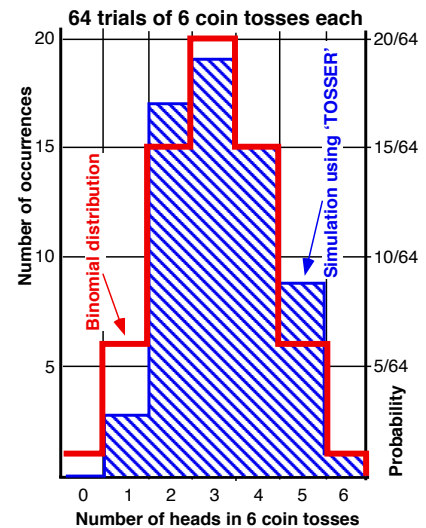
$$\left(\frac{1}{2} + \frac{1}{2}\right)^6 = \left(\frac{1}{2}\right)^6 + 6\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right) + \frac{6 \cdot 5}{2}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^2 + \frac{6 \cdot 5 \cdot 4}{3 \cdot 2}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^3 + \dots$$

In general, if n is the number of tosses (e.g. 6) then the value of the k th term, which gives the probability of getting k heads (or k tails), is given by:

$$f(k) = \frac{n!}{k!(n-k)!} p^k q^{n-k} = {}_n C_k p^k q^{n-k}$$

where ${}_n C_k$ is a symbol representing a combination of k things out of n things.

The binomial distribution has a mean value of $\lambda = np$. (For $p = 1/2$ it is symmetric, with median and mode both equal to the mean value). Its standard deviation is $\sigma = \sqrt{npq} = \sqrt{np(1-p)}$. In our coin-tossing example, the mean is therefore 3.0 and the standard deviation is $\sqrt{6/2} = 1.225$.



Two special cases which are very important in physics can be derived from the binomial distribution. The first is the Poisson distribution, and the other is the Gaussian or normal distribution.

Poisson distribution

Here p is very small and n is very large, in such a way that their product $\lambda = np$ (the mean) is finite. In that case, the binomial distribution becomes the **Poisson distribution**:

$$f(k, n, p) = \frac{(np)^k}{k!} e^{-np} \quad \text{or} \quad f(k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

The second expression is very useful because *it no longer contains n or p , but only λ* . There are various ways of deriving this, for example in the books by Barlow or Squires. In the lectures we use one version of Stirling's approximation for factorials of large numbers:

$$n! \approx e^{-n} n^n \sqrt{2\pi n}$$

and also that in the limit as $n \rightarrow \infty$, an expression of the form shown gives an exponential:

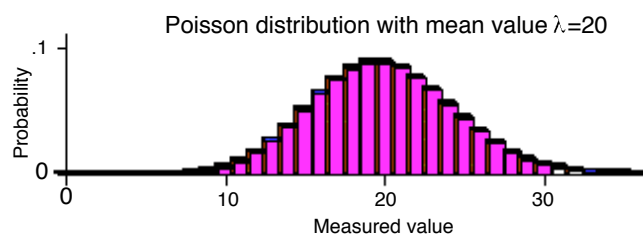
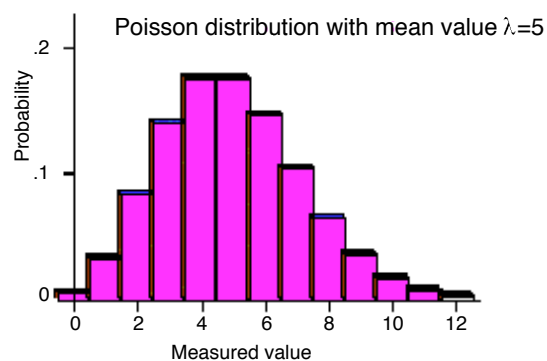
$$\left(1 - \frac{x}{n}\right)^n \rightarrow e^{-x}$$

One of the most important applications of the Poisson distribution in physics is to radioactive decay, where we have a very large number of atoms n (e.g. Avogadro's number $\sim 10^{23}$), each of which has a very small probability of decay p , such that their product λ is finite. For example, if we count particles emitted from a radioactive source for one hour and get say 1800 counts, how many counts do we get in repeated 10-second measurements? We expect the mean value of counts per 10 seconds to be $1800/360 = 90/18 = 5$. This mean value is $\lambda = np$. However, if we operate the counter for many 10-second intervals we certainly won't get $k = 5$ counts every time, but anything ranging from 0 to 10 or even more. For any value of k ,

$$\text{Probability of a given } k = \frac{5^k e^{-5}}{k!} \quad \text{where } k = 0, 1, 2, \dots$$

The predicted distribution is shown in the diagram. It is skew for small values of λ , but becomes symmetric for large numbers, as shown in the examples with means of 5 and 20.

Start with the binomial standard deviation $\sigma = \sqrt{np(1-p)}$. Since p is small, $1-p$ is approximately equal to one, so we get $\sigma = \sqrt{np} = \sqrt{\lambda}$. The number of counts k that we measure is our best (and only) estimate of λ , and we see that one measurement actually tells us **both** the mean **and** the standard deviation, since σ is given by the square root of the measured number of counts. In addition, the fractional error on λ is given by $\sqrt{k}/k = 1/\sqrt{k}$. This means that the more measurements we make the better we know λ , but this improves only as a square root, so we need to make, for example, four times as many measurements to halve the fractional error, and 100 times as many to reduce it by a factor of 10.



Gaussian distribution

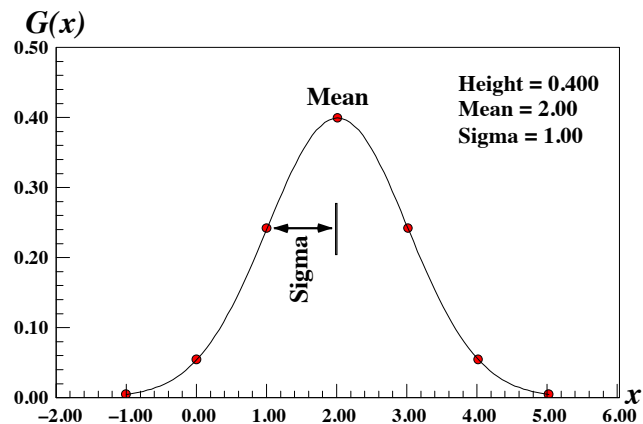
Binomial and Poisson distributions concern *discrete* occurrences. More often in physics we must deal with *continuous* distributions. These are described by a *probability density* $f(x)dx$. For reasons discussed below, the **Gaussian distribution** (sometimes called the *normal* distribution) is the most important one to know about, since it describes the situation of random errors due to a large number of independent disturbances in the system. If we start from a binomial distribution for a probability that is not close to zero or one, and let $n \rightarrow \infty$ (which means that $\Delta x \rightarrow 0$, i.e. that we are taking a very large number of measurements) then the distribution gets smoother, and in the limit we have a bell-shaped curve that can be written:

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

This is normalised, i.e. the integral of $G(x)$ is one, which is why it has a messy constant in front of the exponential. It has a mean value of μ and a standard deviation of σ , which can be checked from the definitions:

$$\mu = \int_{-\infty}^{\infty} x G(x) dx \quad \text{and} \quad \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 G(x) dx$$

The distribution is symmetric, so that the median and the mode are both equal to the mean μ . If we alter μ then the curve shifts position along the x -axis but does not change its shape. If we change σ then the centre of the distribution does not move but it gets taller and narrower or shorter and wider. At $x = \mu \pm \sigma$, $G(x)$ is 0.61 of its value at the peak. A Gaussian with $\mu = 2.0$ and $\sigma = 1.0$ is shown in the figure. If we work in terms of a new variable $z = (x - \mu)/\sigma$, so that distances from the mean are measured in units of the standard deviation, then the Gaussian can be written in a much simpler way as



$$G(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

In addition to thinking of the Gaussian as the limit of a binomial distribution for large n , we can also remember that a Poisson distribution is what happens to a binomial for large n and small p such that the mean $\lambda = np$ is finite, and that a Gaussian is very similar in shape to a Poisson distribution provided that λ is large enough, ≥ 10 for practical purposes.

Why Gaussians are so important — the Central Limit Theorem

Suppose you have made n measurements (x_i) of some quantity x , and obtained a mean value of your measurements $\langle x \rangle$. Whatever the distribution of the x 's, *even if they are not Gaussian*, if you repeat the entire experiment many times then the different *mean values* $\langle x \rangle_j$ that you find will follow a Gaussian distribution, with a mean value (of all the means!) μ and a standard deviation (i.e. standard error of the mean) of σ/\sqrt{n} . This is called the Central Limit Theorem (CLT).

The consequence of the CLT is that variables that vary randomly due to the action of a large number of small effects (even where the effects themselves are *not* necessarily Gaussian, provided only that they can either increase or decrease the variable being measured) end up with the net result that they follow a Gaussian distribution. That is why Gaussian distributions are so common in the 'real world', and therefore why it is very important to know about them.

How well do we know the standard deviation?

In addition to the standard error of the mean, it is also possible to say something about how *well* we know the standard deviation (i.e. the spread) of a set of measurements — what we might call the ‘error of the error’. Without proof, we state that the *fractional error on σ* is

$$\frac{\sigma_{\sigma}}{\sigma} = \frac{1}{\sqrt{2n}}$$

Using this expression, $n = 10$ for example would give a fractional error of 0.22 (22%), and while $n = 50$ is an improvement it is only 0.1 (10%). This shows that in most experiments we really don’t know the value of σ very precisely. That is why it is just not worth agonising *too* much, or quoting more than one or at most two significant figures for the error.

More on Gaussian probabilities

Since the Gaussian is a probability distribution, the total area under the curve is one. We can use the area under *regions* of the curve to work out the probability for, say, getting a measurement *within* a certain number of standard deviations of the mean (in either direction):

$$\int_{-z}^{+z} G(z) dz$$

Subtracting this integral from one gives the probability for getting a measurement *more* than that number of standard deviations from the mean.

In addition to this ‘two-tailed’ probability, we can also find the ‘one-tailed’ probability for getting a result *smaller* than a particular value of z :

$$\int_{-\infty}^z G(z) dz$$

By changing the limits of integration to run from z to $+\infty$ this becomes the probability for getting a result *bigger* than a particular value.

However, and unfortunately, $G(x)$ is not an easily integrated function, so it is still true that even with computers and calculators the quickest way to solve some problems is to use tables of numbers showing the results of these integrals for different numbers of standard deviations. Two such tables are given at the end of these notes: a ‘two-tailed’ table in terms of distance (in units of σ !) from the mean, and another table for ‘one-tailed’ probabilities, i.e. that z is above (or below) some value. The symmetry of the Gaussian distribution, and the fact that it is normalised to unity, means that these can be used to deduce other things, such as the probability that a value is within a specified number of standard deviations *above* the mean.

These tables show that for a Gaussian there is a 0.6827 (68%) probability of a measurement being within one standard deviation of the mean, 0.9545 (95%) within 2σ of the mean, and 0.9973 (99.7%) within 3σ . This is why we said in the first lecture that we would expect about one-third of measurements to lie *more* than one standard deviation from the mean or true value. We also expect about 5% to lie more than two standard deviations away. We can also say that 90% of measurements are expected to lie within 1.645σ of the mean, 95% within 1.960σ , and 99% within 2.576σ .

Fitting data

Suppose we have some data measurements of the form $(x_i, y_i \pm \sigma_i)$ and we want to know whether they agree or disagree with a particular theoretical distribution or shape, of the form $y = f(x)$. In addition, the theoretical distribution might have some parameters that can be adjusted to try to fit the data — for example for a straight line the slope and intercept. We need a method to find the values of the parameters that give the *best fit* to the data. We then plot both the data and the theory on a graph to compare them. Although our eyes can tell us a lot, particularly if data and theory disagree for some or all of the measured values of x , things are not always clear-cut, and we would like a method that can say in a more *quantitative* way how well the data agree or disagree with the theory. This will be done by looking at how far the measurements are from the theory, in terms of σ_i — just *one* precisely measured point can occasionally destroy a theory.

Straight-line fits

We will illustrate the fitting principle by showing how to fit straight lines to data. For simplicity, we will assume that the data points have the form shown above, i.e. there is no error in measuring x , and that all points have equal errors σ on y . The form of the straight line is $y = ax + b$, and what we have to do is to find the best possible values of the slope a and the y-intercept b . To do this, we take the residuals (i.e. distances) $d_i = [y_i - (ax_i + b)]$ of each data point from the line, and consider the sum of their squares:

$$E = \sum_{i=1}^n [y_i - (ax_i + b)]^2$$

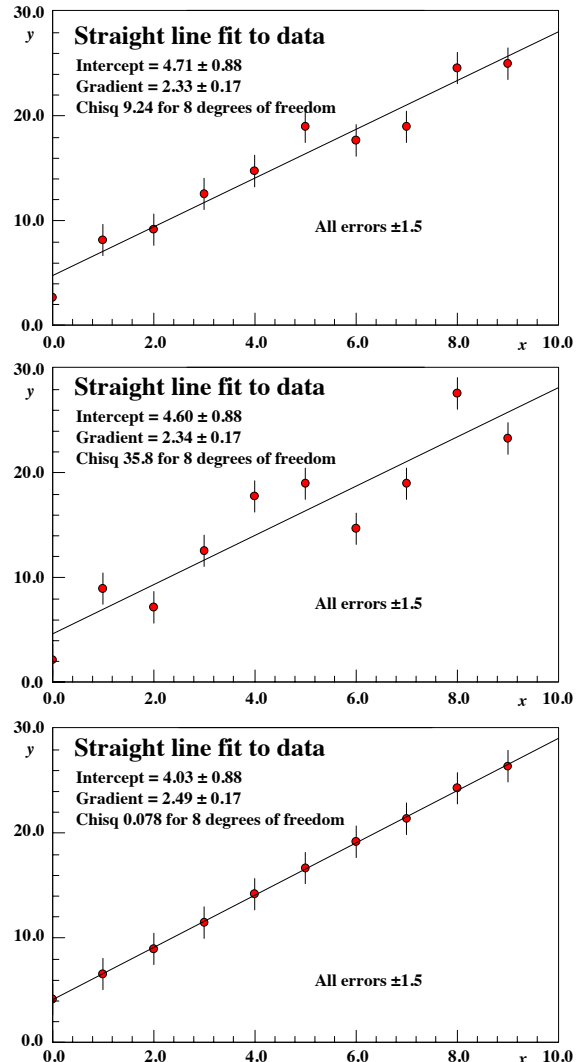
We take the *squares* because the residuals can be either positive or negative, and we only want to know how *far* the measurement is from the line, in either direction. We then minimise this sum E with respect to both a and b by setting the partial derivatives with respect to both variables equal to zero:

$$\partial E / \partial a = 0 \quad \text{and} \quad \partial E / \partial b = 0$$

Skipping the algebra, which can be found in the recommended statistics books, the results are:

$$a = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \quad \text{and} \quad b = \frac{\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i}{n}$$

These can be messy to work out by hand, but both PhysPlot and many calculators do it all for you. Since we minimise a sum of squares, this procedure is called the *method of least squares*. Formulas for the errors on a and b can also be found in the recommended books.



The figures show fits to three seemingly similar sets of data. The first is a good fit, the second is poor, the third is ‘too good’ — the line is *always* much less than 1σ from the points. More on this below.

The χ^2 test

You may already have noticed that when you do a fit using PhysPlot, the box showing its parameters also mentions a quantity called chi-squared (χ^2). Once more we look at the residuals of the measurements from the fit, $d_i = y_i - f(x_i)$. This time, however, we work in units of the error on each measurement, σ_i , and χ^2 is defined to be the sum of their squares:

$$\chi^2 = \sum_{i=1}^n \left(\frac{d_i}{\sigma_i} \right)^2 = \sum_{i=1}^n \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2$$

For a really good fit to the data, we expect that the residuals should depend only on the measurement errors σ_i , and so *on average* the residuals should be about 1σ . This would give a total χ^2 roughly equal to the number of measurements, n , and so we should have $\chi^2/n \approx 1$. This is the case for the first data set in the figure. If χ^2/n is much *bigger* than one we do not have a good fit, as shown in the second data set. This can happen if the errors are underestimated, or if the data really can’t be described by a straight line. If χ^2/n is much *less* than one something is also wrong, as in the third part of the figure — perhaps the errors are badly overestimated. Be *very* suspicious in situations like this where the fitted line is well inside almost all of the error bars!

If we fit a more complicated curve to the data, for example a polynomial, we should expect a better fit simply because there are more adjustable parameters; for example a cubic has four. In order to allow for this, i.e. to favour simpler functions, a quantity called the **number of degrees of freedom**, or ***n.d.f.***, is defined. This is the number of points, n , minus the number of variable parameters in the fitted function (e.g. two for a straight line). So what is calculated is $\chi^2/n.d.f.$ In detail, the probability of getting a given χ^2 for a specified *n.d.f.* is not simple and must be looked up in a table, but the general rule that a value close to one is a good fit is enough for this course.

Weighted means

So far, we have usually assumed that all the measurements have equal errors. Although we will not show what happens if this is not the case for all the results given in the lectures, it is particularly useful to be able to calculate the mean value of a set of measurements correctly, such that points with small errors count for more than points with large errors. What we need is called the **weighted mean**. We start with n measurements of the form $x_i \pm \sigma_i$, and we give each measurement x_i a **weight** $w_i = 1/\sigma_i^2$. It can be proved that the best value for the mean is:

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} = \frac{\sum_{i=1}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2}}$$

and that the error on this mean value is:

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{\sum_{i=1}^n w_i}}$$

It is easy to show that if all the errors are equal these formulas simplify to the previous ones.

