EXPERIMENTAL PHYSICS
Notes for Course PHYS2350

Jim Napolitano
Department of Physics
Rensselaer Polytechnic Institute

Spring 1999

Preface

These notes are meant to accompany course PHYS2350 *Experimental Physics*, for the Spring 1999 semester. They should make it much easier for you to follow the material and to be better prepared for the experiments. The course *will not* cover everything in these notes, but with some luck the notes will continue to be a useful reference for you.

The text is organized into two types of major sections, namely *Chapters* and *Experiments*, so that they follow in a more or less logical order. As much as possible, the Experiments only rely on material in preceding chapters. There is no index, but hopefully the table of contents will be good enough for the time being.

Thanks to helpful comments from many students and faculty, this has all gone through a number of revisions which I hope have made the material more useful and more clearly presented. In the latest version, I've reformatted everything into LaTeX2e, the new LaTeX standard. For the time being, I've removed the explicit distinction between "Experiments" and "Chapters", but the references should still be clear. (My apologies for any mistakes I've made which I didn't find in time!) This change allows me to use what I think are more a more clear postscript font.

Special thanks to Prof. Peter Persans for his comments, and for adding the Jarrell–Ash spectrometer to the laboratory for the *Atomic Spectroscopy* measurement. I've updated the "Procedure" section of that experiment to include a description of this instrument. Credit also goes to Peter for the expanded appendix giving a quick review of MATLAB commands.

Please give me any comments you might have on these notes, particularly if you see ways in which they may be improved.

Thanks for your help.

Jim Napolitano, January 3, 1999

Values of Physical Constants

The following table of fundamental constants is taken from the "Review of Particle Properties", published in Physical Review D I, v.50 (1994). The uncertainties in the values are very small and can be neglected for the experiments in this book.

Quanitity	Symbol	Value
Speed of light in vacuum	c	299792458 m/sec
Planck's constant	h	6.6260755×10^{-34} J sec
	$\hbar/2\pi$	6.5821220×10^{-22} MeV sec
Electron charge	e	1.60217733×10^{-19} Coul
	$\hbar c$	1.97327053×10^{-13} MeV m
Vacuum permittivity	ϵ_0	$8.854187817\times10^{-12}$ F/m
Vacuum permeability	μ_0	$4\pi\times10^{-7}$ N/A^2
Electron mass	m_e	0.51099906 MeV/c^2
Proton mass	m_p	938.27231 MeV/c^2
Deuteron mass	m_d	1875.61339 MeV/c^2
Atomic mass unit	u	931.49432 MeV/c^2
Rydberg energy	hcR_∞	13.6056981 eV
Bohr magneton	μ_B $= e\hbar/2m_e$	5.78838263×10^{-11} MeV/T
Nuclear magneton	μ_N $= e\hbar/2m_p$	3.15245166×10^{-14} MeV/T
Avogardro constant	N_A	6.0221367×10^{23} atoms/mole
Boltzmann constant	k	1.380658×10^{-23} J/K

Contents

Ch 1

Data Taking and Presentation

Progress is made in the physical sciences through a simple process. A model is developed, and the consequences of the model are calculated. These consequences are then compared to experimental data. If the consequences do not agree with the data, then the model is wrong, and it should be discarded. After enough successful comparisons with data, however, a model becomes widely accepted, and progress goes on from there.

Obviously, it is crucial that the data be "correct". Furthermore, the accuracy of the measurement must also be reported so that we know how strong a comparison we can make with the model. Finally, since it is likely that many people will want to compare their models to the data, the experimental results must be reported clearly and concisely so that others can read and understand it.

The purpose of this chapter is to give you some ideas on how to take data "correctly", and how to report it clearly. However, every experiment is different, so these guidelines can only serve as a broad basis. You will gain experience as you do more experiments, learning rules for yourself as you go along.

We will use some loose language, especially in this chapter. Experimental Physics is a subject that can only be truly learned from experience, and terms like "settings" and "uncertainties" will become much clearer when you've

done your time the laboratory. However, we attempt to at least roughly define terms as we go along. For starters, we take the term "quantity" to be the result of some measurement, like the number read off a meter stick or a voltmeter. Things that you can change by hand, which affect the "quantity" you want to measure, are called "settings".

I will often resort to saying something like "... and your intuition will get better after some experience." I apologize, but it is very hard to *tell* someone how to be a good experimenter. You have to learn it by being shown how, and then working on your own. There is at least one book, however, which contains many good ideas about carrying out experiments:

- *Practical Physics*, G. L. Squires, Third Edition
  Cambridge University Press (1991)

1.1 Your Log Book

Keep a log book. Use it to record your all your activities in the lab, such as diagrams of the apparatus, various settings, tables of measurements, and anything you may notice or realize as you go through your experiment. This log book will be an invaluable reference when you return to your data at any later time, and you want to make sense of what you did in the lab. The book itself should have a hard binding with pages that won't get ripped out easily. If you make plots on graph paper or a computer, they can be attached directly into the log book with tape or staples. However, a good log book has pages with both vertical and horizontal rulings, so that you can make hand drawn graphs directly on the page. *Never write data down on scratch paper so that you can do work with it before putting it your log book.*

Your log book should be kept neat, but not too neat. What's important is that you record things so that *you* can go back to them at a later time and remember details of what you did. Record your activities with the date and time, especially when you've returned to recording things after a delay. When you are setting up your experiment, don't worry about writing everything down as you go along, but wait until things make some sense to you. That

way, whatever you write down will make better sense when you go back to read it later.

Some scientists keep a log book as a daily diary, recording not only their measurements, but lecture and seminar notes and other similar things. A good tip is to leave the first few pages blank, and fill them in with a table of contents as you go along. How you organize your lab book(s) is up to you, but it is probably a good idea to keep a lab book specifically for your lab course.

1.2 Common Sense

For virtually any experiment, there are some good rules to keep in mind while you are taking data. It is a good idea to step back once in a while, during your experiment, and ask yourself if you are following these rules.

In later chapters we will be more precise with language regarding "experimental uncertainty" or "measurement error". For the time being, however, just take these terms at face value. They are supposed to indicate just how precisely you have measured the desired quantity.

1.2.1 Use Redundancy

If you measure something with the various settings at certain values, you should in principle get the same value again at some later time if all the settings are at their original values. This should be true whether you changed the settings in between, or if you just went out for a cup of coffee and left the apparatus alone. It is always a good idea to be redundant in your data taking. That is, check to make sure you can reproduce your results.

In practice, of course, you will not get the same result when you come back to the same settings. This is because any one of a number of things which you did not record (like the room temperature, the proximity of your lab partner, the phase of the moon,...) will have changed and at least some

of them are likely to affect your measurement in some subtle way. With some experience, you will be able to estimate what is or is not an acceptible level of reproducibility. In any case, the degree to which you can reproduce your results will serve as a measure of your experimental uncertainty for that quantity.

Be aware of any trends in your measurements as you take data. You can be redundant also by specifically taking data with settings that test any trends that you notice. If you expect data to follow a trend based on some specific model, then take more data than is necessary to determine the parameters of the model. For example, suppose you are testing the notion that temperature T is a linear function of pressure P, i.e.

$$T = a + bP$$

Then, the parameters a and b can in principle be determined with only two measurements of temperature at specific settings of the pressure. This is not redundant, however, and you should take data at more pressure settings to confirm that the linear relation is correct. If it is not, then that tells you something important about either your experimental setup or your model, or both.

It is natural to take data by changing the setting(s) monotonically. That is, to increase or decerease a setting over the range you are interested. It is a good idea to at least go back and take a couple of points over again, just to make sure things have not "drifted" while you took your data. A more radical alternative is to take your data at more or less random values for the settings.

Don't drive yourself crazy by changing more than one setting at a time while you are making measurements. Unless you are testing some trend you may have noticed, you will certainly want to go back to find out each of the settings affected the measured quantity.

1.2.2 Be Precise, But Don't Go Overboard

It is of course important to strive for as much precision as possible in your measurements. However, do not waste your time measuring one particular

quantity very precisely if the result your are ultimately interested in, depends on some other quantity which is known much more poorly. For example, suppose you want to ultimately determine the velocity v for some object moving in a straight line. You do this by measuring the distance L that it travels in a period of time t, i.e.

$$v = L/t$$

The relative precision of L and t contribute equally to the relative precision of t. (We will return to this in a later chapter when we discuss experimental uncertainties.) That is, if both L and t are both known to around 10%, they will both contribute to the uncertainty in v. However, there is no point in trying to figure out a way to measure t, say, to 1% if you cannot measure L to comparable precision. Your good idea for measuring t, although it may be useful and satisfying for other reasons, will not help you determine v much more precisely.

It is important to keep in mind how precisely you are measuring the various quantities that go into your final result. With experience, you will develop a good insight for knowing when enough is enough.

1.2.3 Measure Ratios

Whenever you can, use your apparatus to determine ratios of quantities measured at different settings. This is a very useful technique, since common factors cancel when you take a ratio, and the uncertainty in these factors cannot affect the ratio. Hence, a measurement of a ratio will be inherently more precise than a measurement of an absolute quantity. Some of the quantites we will measure in the experiments are ratios, and they typically are determined with relatively high precision.

Even if the ultimate goal of the experiment cannot be expressed as a ratio, try to find ratios among your data that you can use to test the model. For example, suppose you want to determine the resistivity ρ of some metal sample from the decay lifetime τ of some transient voltage signal. Your model says that τ depends on ρ through the relation

$$\tau = R^2/\rho$$

where R is the radius of the sample. Even though you cannot determine τ directly from a measurement of a ratio, you can measure τ for two different samples of the same metal, but with different radii R. The ratio of the lifetimes should be the same as the ratio of the squares of the sample radii, and this is a good check on your procedure. This and other examples will be pointed out along the way as we discuss the various experiments.

The determination of the lifetime of the free neutron is a good historical example of the triumph of ratios over absolute measurements. Free neutrons decay with a half life of around 10 minutes. Furthermore, up until quite recently, "samples" of free neutrons were only available in fast moving streams from nuclear reactors. Through the 1970's, the neutron lifetime was determined through two *absolute* measurements, one of the decay rate as the stream passed through some detectors and the other of the flux of neutrons in the stream itself. The various measurements of different groups *did not agree* with each other, and the resulting large uncertainty in the lifetime had serious consequences in astrophysics and particle physics. Then in the 1980's, using a result based on the accepted model for neutron decay, a different group measured a *single ratio* which agreed with previous, but less precise, measurements of this ratio and finally pointed the way to the correct value of the neutron lifetime. A fine account of these measurements is given in "How Long Do Neutrons Live?", by S.J. Freedman, in Comments in Nuclear and Particle Physics, **19**(1990)209.

1.2.4 Avoid Personal Bias

Nobody starts working on an experiment without at least a rough idea of what he or she is supposed to measure. It is impossible, therefore, to have no notion of what to expect from the measurement of some quantity for some range of settings. Sometimes, though, the result of a measurement is quite surprising and may be an important clue to how nature works! You must walk a pretty fine line between what you expect and what you are trying to learn. This will again become easier and more natural with experience.

Never fudge your data to give you the answer you expect! You will not learn anything from this, and you may miss something very important. There

are several great examples in the history of science where highly regarded researchers end up with egg on their faces for not keeping this in mind. One example of this is documented in a very readable paper, "How the First Neutral Current Experiments Ended", by Peter Galison, in Review of Modern Physics, **55**(1983)477.

1.3 Tables and Plots

A picture is worth a thousand words. It is always best to display your data using either a table or a plot, or both. Tables are particularly useful if you want someone else to be able to take your numbers and test a different model with them. Plots are best if you want to show features in your data that may be particularly important, such as "peaks" or "valleys" that demonstrate some phenomenon happening at a particular setting, or "trends" like linear or exponential behavior which may or may not support some specific model.

It is a real art to know just how much information to include on a table or plot. Too little data can leave the reader without enough to figure out what can be concluded from the experiment. On the other hand, if you put too much on the page it is very frustrating to know exactly what the important point is. As with most things in Experimental Physics, experience will be the most important teacher.

1.3.1 Tables of Data and Results

In the old days, data was recorded directly from the instruments into the log books. In modern times, however, we usually use some sort of computerized interface to gather the data. In either case, it is a good idea to keep the "raw data", as we call it, in the log book. Of course, be judicious in what you call "raw data". If you read line positions from a spectrum with a thousand data points in it, just record the line positions and not all the data points!

It is smart to always record data points exactly as you read them from the instrument, instead of doing any conversions in your head or (heaven

forbid!) on scratch paper. Record all conversion factors or offset values in your log book.

Always put labels at the head of columns or rows. The labels should be terse but descriptive of the setting value or of the measured quantity, and you should keep using that notation as you do calculations and analysis in your log book. Always include the units along with the label, and try to stick with standard conventions. When recording numbers, make sure you keep enough significant figures (depending on the precision you expect to be important), but not too many.

Let's do an example. Suppose you are measuring the time period ΔT of some oscillating signal using an oscilloscope as a function of some relative pressure setting P_{REL} measured with a vacuum gauge. (We will discuss various laboratory instruments in a later chapter.) You make a table in your log book that looks like the following:

P_{REL} (in. Hg)	ΔT (div.)
27.5	5.3†
25	5.0
20	4.5
14.5	4.1
5	7.0‡
0	6.6
†0.1 ms per division	
‡50 μs per division	

Notice that in the middle of taking the data, you found it was better to switch the time base of the oscilloscope. You did so and noted the different conversion factors.

Now let's suppose that you want to do some calculations with this data so that you can test some model. If you leave room in the table, you can put the results of the calculations right there. (In this case, there is not so much data and we can do this without making the table too crowded.) The model is best described by its dependence of the frequency $\nu = 1/\Delta T$ on the

internal pressure P. The table might then be extended in the following way:

P_{REL} (in. Hg)	ΔT (div.)	P/P_{ATM}	ν (kHz)
27.5	5.3†	0.083	1.89
25	5.0	0.17	2.00
20	4.5	0.33	2.22
14.5	4.1	0.52	2.44
5	7.0‡	0.83	2.86
0	6.6	1	3.03

†0.1 ms per division
‡50 μs per division

This is a clear, concise description of the data you took, and the numbers are available to someone who may have some other idea of how to look at your data. If you want to examine how well a particular model might compare to this result, the first thing to do is make a plot.

1.3.2 Making Plots

It is handy to plot the results listed in a table. That makes it easy to refer back and forth between the table and the plot, picking off details visually on the plot and reading the relevant numbers from the table. For the data listed in the above table, we've plotted the analyzed quantites in Fig. 1.1. This picture could easily and quickly be made by hand, directly in the log book.

Some important things can immediately be learned from this plot. First, we've drawn a straight line through the data points and it is clear that our results show that to a good level of accuracy, the frequency depends linearly on pressure. Note that we have plotted the data with a "suppressed zero" on the vertical axis. This is a useful technique when the data covers only a limited range, but you should be careful to make it clear when an axis does not start at zero. The slope of the line can easily be read off the plot, and its value compared to the model prediction.

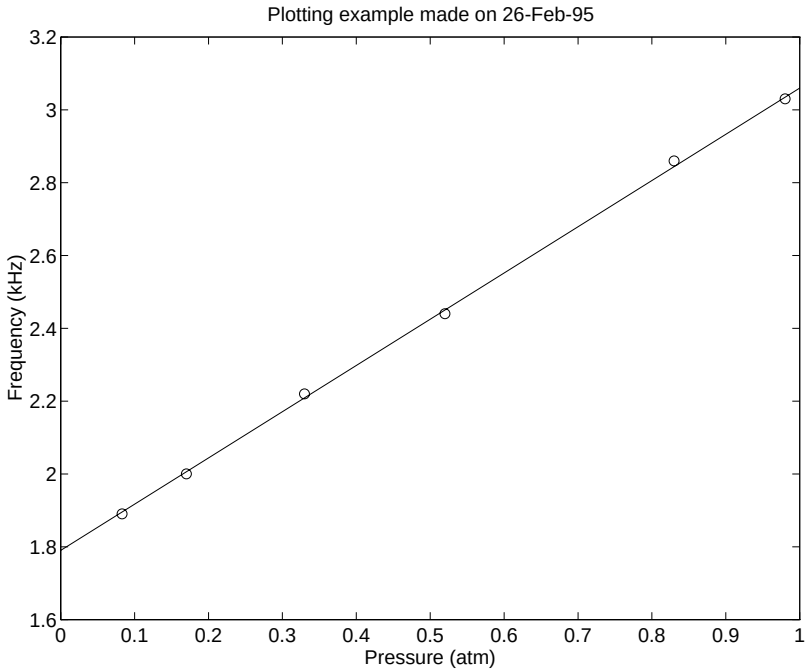


Figure 1.1: An Example of Plotting Data

It is a good idea to choose the axes of your plot so that you can compare the behavior to a particular model. You can do this easily by eye if the model predicts a straight line when you plot your data. If the model predicts a linear dependence (like in the above example) then a simple plot on linear axes will do. However, if it predicts some other kind of dependence, you have to resort to different ways of plotting the data.

For example, if the model predicts an exponential dependence, e.g. $N = N_0 e^{-t/\tau}$ as in radioactive decay, then it is best to plot $(\log N) = (\log N_0) - (\log e/\tau) \times t$ versus t in which case you again get an easy-to-see straight line dependence, where the slope determines the value of τ. If the expected dependence is a power law function, e.g. $g = g_0 V^n$ as in the gain of a photomulitplier as a function of voltage, then $\log g$ plotted against $\log V$ gives a straight line whose intercept determines g_0 and slope determines n.

Special graph paper (or axis scaling) called "semilog" or "log-log" allows you to plot the quantites directly without having to take the logarithms

Table 1.1: Linear Axis Scaling in Plots

Model	Best Scale	Slope	Intercept
$y = Ax + B$	Linear	A	B
$y = Ae^{Bx}$	Semilog	$B \log e$	$\log A$
$y = Ax^B$	Log-Log	B	$\log A$

yourself. These different choices are listed in Table 1.1.

In cases where the model is more complicated, you can still plot the data in a way that allows you to easily see a straight line dependence. For example, if $y = \sqrt{Ax^{2.3} + B}$, then plot y^2 as a function of $x^{2.3}$. In experiment 12 on *Compton Scattering*, you will learn that the scattered photon energy E' depends on the incident energy E and the scattering angle θ in the following way:

$$E' = \frac{E}{1 + (E/m)\,(1 - \cos\theta)}$$

In this case, you can plot the quantity E/E' as a function of $(1 - \cos\theta)$ and the result should be a straight line with slope E/m and intercept at 1.

The data plotted in Fig. 1.1 simply shows data points. We will later see how we determine an "uncertainty" with each of these data points, usually associated with the quantity plotted on the vertical axis. In this case, the data points are plotted with an "error bar", that is, a symbol with a vertical line drawn through it. The limits of the vertical line indicate the range of "uncertainty" associated with that point. We will see many examples of this when we describe the experiments.

1.4 Using Computers

Nothing can replace a hand-drawn plot in your log book, as you take your data, as a check that things are proceeding normally. I urge all of you to follow this practice when you are actually running your experiment. However, a neater presentation is of course possible using any of a number of computer

programs designed to tabulate and plot data. A different use of computers is to actually help you *analyze* data, not just plot it, and most programs allow you to do some of both.

Be aware that even though different programs may all claim to be flexible at some level, they are all written with specific priorities and audiences in mind. The program you like to use will likely come down to your personal taste. Following are some thumbnail sketches of programs that run either on PC's or Mac's, or ones that run on the Rensselaer Computer Services (RCS) Unix system. For most examples in this book, however, I will use the program MATLAB which is available on PC's, Mac's, and on RCS.

1.4.1 Programs for the PC

There are zillions of PC plotting and analysis programs out there. Some are very inexpensive and some are very pricey. What they are capable of is pretty much correlated with their cost, but that doesn't mean that *you* will find more expensive programs more useful. Following are some of the programs either on the PC's in the Physics Department, or available at the ITS Product Center. A more complete list and descriptions were published in the Spring 1994 Physics Courseware Communicator.[1]

- The student edition of MATLAB. I recommend this program. More on this below.
- GRAPH III (CRICKET GRAPH). This is a simple, easy-to-use, plotting program with data entered on a spread sheet. The plots are high quality and have plenty of useful options. The program allows for some very basic analysis options and curve fitting, but not advanced enough for many of things you will need for this course.
- F(G) SCHOLAR. This is a good program for scientific data plotting, at a reasonable price. Besides producing fine plots, it has very sophisticated tools for data analysis including curve fitting.

[1]This quarterly publication reviews physics-related software, mainly for educational use. It is available from the Physics Courseware Evaluation Project (PCEP) at North Carolina State University. Their email address is PCEP@NCSU.EDU.

- DeltaGraph Pro. This is a higher level program than graph III (and costs around twice as much). The graphics are a bit more sophisticated, and there are some more options for data analysis, but the primary audience is not scientific.
- PSI-Plot. This is a relatively sophisticated package, aimed at scientists and engineers. A new release (version 3.0) contains many of the analysis options you are likely to encounter in this course. It is a bit expensive, but we do have a copy on one of the PC's in the student laboratory. You are welcome to try it out.
- excel. Many of you are familiar with this program, basically a spread sheet with graphics. However, it attempts to be very broad based, and is therefore hard to adapt directly to the sorts of things you will need to do.

1.4.2 Programs on RCS

The people at ITS maintain a bunch of programs that you can use These programs are generally more sophisticated that what you get on a PC, for two reasons. One is that RCS has lots more memory and disk space than what you get on a PC, and unless there is a lot of traffic, the computers you use are a lot faster. The second reason is that the University pays for the programs, and they can afford some very nice packages. If anything, you might want to buy some documentation for the program or programs you settle on, but in some cases, that documentation is free and available on RCS itself. You can use the Unix man pages to find out more about these programs, and where to go to get more documentation.

Note that you can use the SUN workstations and the mtools utilities to reac PC-compatible floppy disks on RCS. This is a fine way to transfer data from the lab to RCS. Another way is to use ftp with PC's that are connected to the campus network.

- matlab. I will use matlab for most of the examples in this course. More information is in the next sections.
- gnuplot. This is a pretty simple-to-use plotting program, but it has

almost no analysis capabilities. One very nice feature is that you can plot combinations of standard built-in functions on top of your data points. The program should be able to do all you need in this course, so far as plotting is concerned.

- XMGR. The subtitle for the manual calls XMGR "Graphics for exploratory data analysis", and that is pretty accurate. You can actually produce wonderful looking plots, and do rather sophisticated things with your data, including fitting and manipulations. The program also works with the X11 interface so that most of your control can be window-driven, although you don't have to do things that way. The biggest problem with the program is that the documentation is not easy to read, and it will take some practice to get good at it.

- MAPLE. Most of you are familiar with MAPLE from your math courses. Recall that this program is designed for *symbolic* manipulation, not data manipulation. That is, it works well with formulas, but can be hard to use when massaging data. It can be used this way, however, so if you're adept at MAPLE, you might want to use it for data analysis as well as plotting.

1.4.3 MATLAB

MATLAB is a *numerical* analysis package that is ideally suited for data analysis. It is easy to use, and has most of the features you will need already built-in. These include fitting, integration, differentiation, and the like. The name comes from "MATrix LABoratory", which reminds us that data is internally stored and manipulated as matrices.

We will refer to MATLAB throughout these notes, including specific examples for the various experiments. General information for data analysis can also be found in Sec. 6.2.3, and various sections of Chapter 9, in particular sections 9.2.1 and 9.5.

Making Plots with MATLAB

MATLAB also has sophisticated plotting capability. Note, however, that your emphasis should be on data analysis, not making beautiful plots. The plot in Fig. 1.1 was made with MATLAB using the following commands:

```
x=[0.083 0.17 0.33 0.52 0.83 0.98];
y=[1.89  2.00 2.22 2.44 2.86 3.03];
xl=[0    1   ];
yl=[1.79 3.06];
plot(x,y,'o',xl,yl)
xlabel('Pressure (atm)')
ylabel('Frequency (kHz)')
title(['Plotting example made on ',date])
print -dps plexm.ps
clear x y xl yl
```

In this example, data is entered line-by-line. The semicolon (";") after each data line is not necessary, but if it is not included, MATLAB echos the values of the newly created variable. The plot function has a number of arguments, and we specify the 'o' which means "plot the points as circles" for data points, and no option to just connect points with a straight line. The axes are labeled and a title is added as shown. Note that the arguments to these functions are in fact matrices of character strings, and that is why we enclose the argument to "title" in square brackets. The "print" command as used here generates a POSTSCRIPT file which can be stored or sent to your favorite printer. If no options are given to the "print" command, then the output automatically goes to the default printer. Finally, the data variables we defined in the beginning are cleared, freeing up the memory they required.

You can change lots of things on plots, like the character size for example, using the "handle graphics" capability in MATLAB. Refer to Appendix D to learn the basics.

Entering Data into MATLAB

For larger amounts of data, you can tell MATLAB to retrieve data from a separate file, instead of having to type all the numbers in by hand. This is done with the MATLAB command "load", which, for example, will read a two-column ascii data file with n lines into a $n \times 2$ matrix. The name of the matrix is the same as the name of the file with the extension stripped off. Individual vectors of data can be extracted from this matrix. For example, if the name of the file is "mine.dat", then the MATLAB commands

```
load mine.dat
x=mine(:,1);
y=mine(:,2);
```

create two vectors x and y, each of which contains the n elements of the two columns in the file. A different approach is to "read" the numbers in whatever format they were written using commands like

```
fid=fopen('sc1.lis');
a=fscanf(fid,'%f')';
fclose(fid);
```

which reads a column of numbers in the file "sc1.lis" into a vector `a`. (The format control `%f` should be familiar to C programmers. Note that the vector is created by transposing the list read with `fscanf`.) These techniques should be particularly useful to you when reading data transferred to RCS from a floppy disk or through an ethernet connection.

Keeping Track of Things

Anytime in the middle of a MATLAB session, you can type the command `whos` to get a list of the variables you've created and their type. The command `who` just gives you the list of names. These can be very useful if you get confused regarding what's been created in the course of entering commands.

Commands do not need to be entered to the command line for MATLAB. Instead, they can be created with some editor, and stored in a file with the extension ".m". Just entering the name of the file (without the ".m") to the MATLAB command line executes the command in this file.

Further Documentation on MATLAB

Remember that you should use MATLAB primarily for data *analysis*, not data *plotting*. We will refer to the relevant commands and show examples along the way in the rest of this book. There is a built-in help documentation for MATLAB that should help you find your way, once you get started. There are a number of other sources of MATLAB documentation:

- *The* MATLAB *Documentation Set*, The MathWorks,
  which contains several separate publications on various ways to use and modify MATLAB. The ones most useful to you at the beginning would be
  - *The* MATLAB *User's Guide*, a short description of what MATLAB can do and a tutorial introduction.
  - *The* MATLAB *Reference Guide*, which is a complete listing of all the standard MATLAB functions.
- *The Student Edition of* MATLAB, Prentice Hall (1994)
  which combines the User's Guide and Reference Guide from the standard documentaion set, and can be purchased separately from the program. It comes with the software package you can purchase from the ITS Product Center.
- *Numerical Methods for Physics*,
  Alejandro Garcia, Prentice-Hall (1994),
  a good book on numerical methods which uses MATLAB for most of the programming examples.

You might also browse the World Wide Web home page of The MathWorks, at http://www.mathworks.com/, which contains lots of useful information including a list of books which use and refer to MATLAB.

1.5 Formal Lab Reports

When you are finished with an experiment, or some part of it, you may have to write a formal report on what you've done. This is certainly the case if you want to publish your results in a scientific journal. Different people have different ideas about what these reports should look like, and papers for journals have well defined formats that have to be followed.

Here's a hint for writing up papers or lab reports. Before you start writing, think about how you might explain your work to someone. Better yet, find someone who will listen to you explain your experiment to them, but who is *not* familiar with it. You'll be surprised to see how clearly you can organize your thoughts this way.

The main sections of a formal report or paper are likely to include the following:

- **Title.** Give some thought to the title of the report. It must be terse, but still let the reader know what it is about. Don't forget that titles of papers are entered in data bases used for computerized literature searches, so try to include words that will make your paper show up in a typical search on the subject.

- **Abstract.** This is a concise, self-contained summary of the experiment. It should report the method, conclusion, and an assessment of the accuracy and/or the precision of the result. *The abstract is a summary of the whole paper. It is not an introduction.*

- **Introduction and Theory.** Write what you expect to learn and a general description of the experiment. You should include relevant equations and formulas, and refer to previous work on the subject.

- **Experimental Setup.** Describe the apparatus including all relevant detail. Diagrams with symbols, standard where available, are a good idea. Refer back to these diagrams when writing the Procedure.

- **Procedure and Data.** Indicate how you proceeded to take data. Basic analyses used to process the data can be included here. Tables

summarizing the results are a good idea. Keep in mind aspects of the procedure that affect the accuracy and the precision of the data. What limits the precision?

- **Interpretation and Discussion.** This section should contain any detailed analysis on the data, particularly where it applies to testing a certain hypothesis. Discuss the result, and whether or not it makes sense. Derive whatever quanitites you can from the data, and interpret them.
- **Conclusion.** Summarize the experience of this measurement. You may want to include suggestions for further work, or for changes and/or improvements to the apparatus.
- **References.** List all cited references in a separate section.
- **Appendices.** If you want to include things like raw data, calculations, detailed equipment descriptions, and so on, you should put them in Appendices. They should be there if the reader wants to go into the work in more detail, but should not be necessary for understanding the motivation or interpretation of the measurement.

It is important to include citations to important literature relevant to your work. Tables and plots should be used wherever appropriate to make your point.

1.6 Exercises

1. The following table lists data points for the decay rate (in counts/sec) of a radioactive source:

Time (sec)	Rate (/sec)	Time (sec)	Rate (/sec)	Time (sec)	Rate (/sec)
0.6	18.4	2.0	3.02	3.6	1.72
0.8	10.6	2.4	2.61	4.0	1.61
1.2	8.04	2.8	2.08	4.2	1.57
1.6	6.10	3.0	1.50	4.3	1.85

a. Plot the data using an appropriate set of axes, and determine over what range of times the rate obeys the decay law $R = R_0 e^{-t/\tau}$.

b. Estimate the value of R_0 from the plot.

c. Estimate the value of τ from the plot.

d. Estimate the value of the rate you expect at $t = 6$ sec.

2. An experiment determines the gravitational acceleration g by measuring the period T of a pendulum. The pendulum has an adjustable length L. These quantities are related as

$$T = 2\pi\sqrt{\frac{L}{g}}$$

A researcher measures the following data points:

Data Point	L (prulp)	T (klotz)
1	0.6	1.4
2	1.5	1.9
3	2.0	2.6
4	2.6	2.9
5	3.5	3.4

One of these data points is obviously wrong. Which one?

Ch 2

Basic Electronic Circuits

Nearly every measurement made in a physics laboratory comes down to determining a voltage. It is therefore very important to have at least a basic understanding of electronic circuits, before you start making physical measurements. It is not important to be able to design circuits, or even to completely understand a circuit given to you, but you do need to know enough to get some idea of how the measuring apparatus affects your result.

This chapter introduces the basics of elementary, passive electronic circuits. You should be familiar with the concepts of electric voltage and current before you begin, but something on the level of an introductory physics course should be sufficient. It is helpful to have already learned something about resistors, capacitors, and inductors as well, but you should get what you need to know about such things out of this chapter, at least as far as this course is concerned. There is a very little bit at the end about diodes and transistors, but there is more on them in the experiments in which they are used.

This chapter is not a substitute for a course in electronics design. There are of course lots of books on the subject, and you should get one that you are comfortable with. Solid state electronics is an ever growing field, so don't get hooked on a very old book. An excellent, up-to-date text and reference book on electronics that most people in the business use, or at least have a copy of is:

- *The Art of Electronics*, by Paul Horowitz and Winfield Hill
  Second Edition, Cambridge University Press (1989)

A student manual for this book is also available. Another nice book which includes a few introductory chapters on solid state electronics, including the physics behind diodes and transistors, is

- *Experimental Physics: Modern Methods*, by R. A. Dunlap,
  Oxford University Press (1988)

A good introduction to the basics of electric circuits is found in

- *Physics*, Robert Resnick, David Halliday, and Kenneth Krane,
  John Wiley and Sons, Fourth Edition (1992)
  - Chap.31 *Capacitors and Dielectrics*
  - Chap.32 *Current and Resistance*
  - Chap.33 *DC Circuits*
  - Chap.38 *Inductance*
  - Chap.39 *AC Circuits*

2.1 Voltage, Resistance, and Current

Let's start at the beginning. Figure 2.1(a) shows the run-of-the-mill DC current loop. It is just a battery that provides the electromotive force V which drives a current i through the resistor R. This is a cumbersome way to write things, however, so right off the bat we will use the shorthand shown in Fig. 2.1(b). All that ever matters is the *relative* voltage between two points, so we specify everything relative to the "common" or "ground". There is no need to connect the circuit loop with a line; it is understood that the current will flow from the common point up to the terminals of the battery.

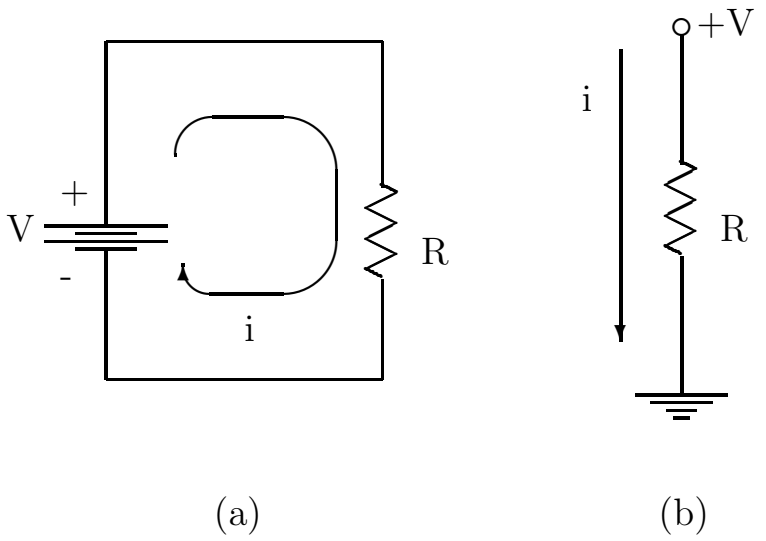


Figure 2.1: The simple current loop (a) in all its glory and (b) in shorthand.

2.1.1 Loop and Junction Rules

The concept of electric potential is based on the idea of electric potential energy, and energy is conserved. This means that the total change in electric potential going around the loop in Fig. 2.1(a) must be zero. In terms of Fig. 2.1(b), the "voltage drop" across the resistor R must equal V. It's actually a pretty trivial statement when you look at it that way.

This is a good time to remind you of the definition of resistance, namely R is just the voltage drop across the resistor divided by the current through the resistor. In other words, the voltage drop through a resistor R is equal to iR where i is the current through it. In terms of the simple loop in Fig. 2.1, $V = iR$. The SI unit of resistance is Volts/Amps, also known as the Ohm (Ω).

Just about all the resistors you will care about in this course obey *Ohm's Law*, which just states that the resistance R is independent of the current i. In fact, we nearly always use the symbol R to mean a constant value of a resistance, that is, a resistor that obeys Ohm's Law.

Electric current is just the flow of electric charge ($i \equiv dq/dt$, to be precise), and electric charge is conserved. This means that when there is a "junction" in a circuit, like the one shown in Fig. 2.2, the sum of the currents flowing into the junction must equal the sum of the currents flowing out. In the case of Fig. 2.2, this rule just implies that $i_1 = i_2 + i_3$. It doesn't matter whether you specify the current flowing in or out, so long as you are consistent with this rule. Remember that current can be negative as well as positive.

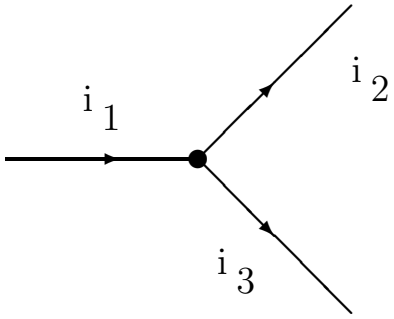


Figure 2.2: A simple three wire circuit junction.

These rules and definitions allow us to determine the resistance when resistors are connected in series, as in Fig. 2.3(a), or in parallel, as in Fig. 2.3(b). In either case, the voltage drop across the pair must be iR, where i is the current through set. For two resistors R_1 and R_2 connected in series, the current is the same through both, so the voltage drops across them are iR_1 and iR_2 respectively. Since the voltage drop across the pair must equal the sum of the voltage drops, then $iR = iR_1 + iR_2$, or

$$R = R_1 + R_2 \qquad \text{Resistors in Series}$$

If R_1 and R_2 are connected in parallel, then the voltage drop across each are the same, but the current through them is different. Therefore $iR = i_1 R_1 = i_2 R_2$. Since $i = i_1 + i_2$, we have

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} \qquad \text{Resistors in Parallel}$$

Remember that whenever a resistor is present in a circuit, it may as well be some combination of resistors that give the right value of resistance.

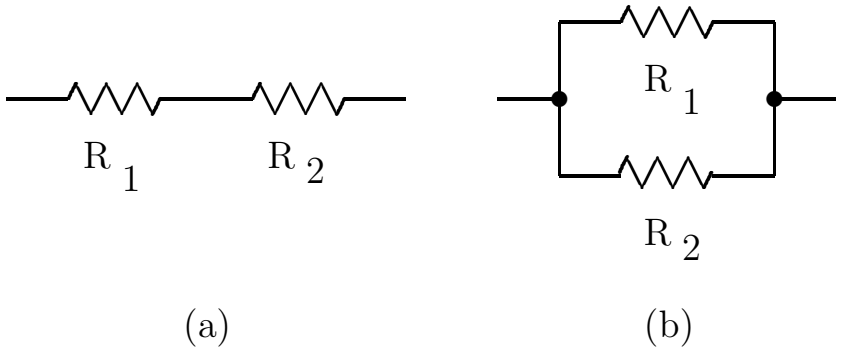


Figure 2.3: Resistors connected (a) in series and (b) in parallel.

2.1.2 The Voltage Divider

A very simple, and very useful, configuration of resistors is shown in Fig 2.4. This is called a "voltage divider" because of the simple relationship between the voltages labeled V_{OUT} and V_{IN}. Clearly $V_{IN} = i\,(R_1 + R_2)$ and $V_{OUT} = i\,(R_2)$, where i is the current through the resistor string. Therefore

$$V_{OUT} = V_{IN}\frac{R_2}{R_1 + R_2} \tag{2.1}$$

That is, this simple circuit divides the "input" voltage into a fraction determined by the relative resistor values. We will see lots of examples of this sort of thing in the laboratory.

Don't let yourself get confused by the way circuits are drawn. It doesn't matter which directions lines go in. Just remember that a line means that all points along it are at the same potential. For example, it is common to draw a voltage divider as shown in Fig. 2.5. This way of looking at it is in fact an easier way to think about an "input" voltage and an "output" voltage.

2.2 Capacitors and AC Circuits

A capacitor stores charge, but does not allow the charge carriers (i.e. electrons) to pass through it. It is simplest to visualize a capacitor as a pair

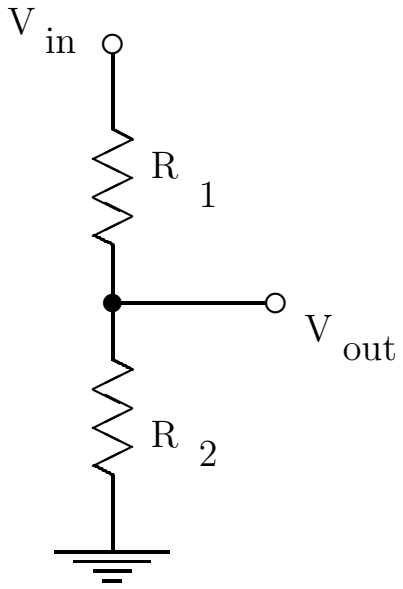


Figure 2.4: The basic voltage divider.

of conducting plates, parallel to each other and separated only by a small amount. Some capacitors (called "parallel plate capacitors") are actually constructed this way, but the kind used in circuits are usually little ceramic disks with a bulge in the middle and two wire leads sticking out.

If a capacitor has a potential difference V across its leads and has stored a charge q on either side, then we define the *capacitance* $C \equiv q/V$. It is easy to show for a parallel plate capacitor C is a constant value independent of the voltage. It is not so easy to do this in general, but it is still true for the most part. The SI unit of capacitance is Volts/Amperes, also known as the Farad (F). As it turns out, one Farad is an enormous capacitance, and laboratory capacitors typically have values between a few microfarads (μF) down to a few hundred micromicrofarads ($\mu\mu$F) or picofarads (pF). People who work with circuits a lot are likely to refer to a picrofarad as a "puff".

It is pretty easy to figure out what the effective capacitance is if capacitors are connected in series and in parallel, just using the above definitions and

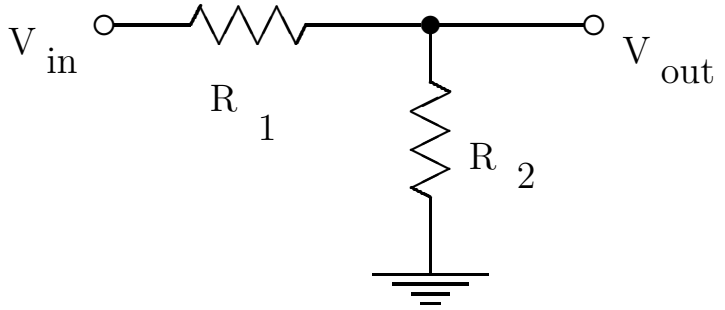


Figure 2.5: An alternate way to draw a voltage divider.

the rule about the total voltage drop. The answers are

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} \qquad \text{Capacitors in Series}$$

and

$$C = C_1 + C_2 \qquad \text{Capacitors in Parallel}$$

That is, just the opposite from resistors.

Now let's think about what a capacitor does in a circuit. Let's take the resistor R_2 in the voltage divider of Fig. 2.4 and replace it with a capacitor C. This is pictured in Fig. 2.6. The capacitor does not allow any charge carriers to pass through it, so the current $i = 0$. Therefore the voltage drop across the resistor R is zero, and V_{OUT}, the voltage across the capacitor C, just equals V_{IN}.

What good is this? We might have just as well connected the output terminal to the input! To appreciate the importance of capacitors in circuits, we have to consider voltages that change with time.

2.2.1 DC and AC circuits

If the voltage changes with time, we refer to the system as an AC circuit. If the voltage is constant, we call it a DC circuit. AC means "alternating current" and DC means "direct current". These names are old and not very descriptive, but everyone uses them so we are stuck with them.

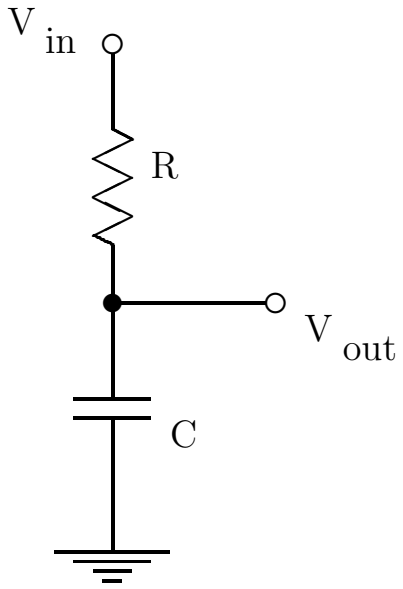


Figure 2.6: A voltage divider with a capacitor in it.

Let's go back to the voltage divider with a capacitor, pictured in Fig. 2.6, and let the input voltage change with time in a very simple way. That is, take

$$V_{IN}(t) = 0 \qquad \text{for } t \leq 0 \tag{2.2}$$

$$= V \qquad \text{for } t > 0 \tag{2.3}$$

and assume that there is no charge q on the capacitor at $t = 0$. Then for $t > 0$, the charge $q(t)$ produces a voltage drop $V_{OUT}(t) = q(t)/C$ across the capacitor. The current $i(t) = dq/dt$ through the divider string also gives a voltage drop iR across the resistor, and the sum of the two voltage drops must equal V. In other words

$$V = V_{OUT} + iR = V_{OUT} + R\frac{dq}{dt} = V_{OUT} + RC\frac{dV_{OUT}}{dt} \tag{2.4}$$

and $V_{OUT}(0) = 0$. This differential equation has a simple solution. It is

$$V_{OUT}(t) = V\left[1 - e^{-t/RC}\right] \tag{2.5}$$

Now it should be clear what is going on. As soon as the input voltage is switched on, current flows through the resistor and the charge carriers pile up on the input side of the capacitor. There is induced charge on the output side of the capacitor, and that is what completes the circuit to ground. However, as the capacitor charges up, it gets harder and harder to put more charge on it, and as $t \to \infty$, the current doesn't flow anymore and $V_{OUT} \to V$. This is just the DC case, where this circuit is not interesting anymore.

The value RC is called the "capacitive time constant" and it is the only time scale we have in this circuit. That is, statements like "$t \to 0$" and "$t \to \infty$" actually mean "$t \ll RC$" and "$t \gg RC$". The behavior of the circuit will always depend on the time relative to RC.

So now we see what is interesting about capacitors. They are sensitive to currents that are changing with time in a way that is quite different from resistors. That is a very useful property that we will study some more, and use in lots of experiments.

The time dependence of any function can always be expressed in terms of sine and cosine functions using a Fourier transform. It is therefore common to work with sinusoidally varying functions for voltage and so forth, just realizing that we can add them up with the right coefficients to get whatever time dependence we want in the end. It is very convenient to use the notation[1]

$$V(t) = V_0 e^{\imath\omega t} \tag{2.6}$$

for time varying (i.e. AC) voltages, where it is understood that the voltage we measure in the laboratory is just the real part of this function. The angular frequency $\omega = 2\pi\nu$ where ν is the frequency, that is, the number of oscillations per second.

This expression for $V(t)$ is easy to differentiate and integrate when solving equations. It is also a neat way to keep track of all the phase changes signals undergo when the pass through capacitors and other "reactive" components. You'll see and appreciate this better as we go along.

[1] If you're not familiar with complex numbers, see Appendix C.4.

2.2.2 Impedance

Now is a convenient time to define *impedance*. This is just a generalization of resistance for AC circuits. Impedance, usually denoted by Z, is a (usually) complex quantity and (usually) a function of the angular frequency ω. It is defined as the ratio of voltage drop across a component to the current through it, and just as for resistance, the SI unit is the Ohm. For "linear" components (of which resistors and capacitors are common examples), the impedance is not a function of the amplitudes of the voltage or current signals. Given this definition of impedance, the rules for the equivalent impedance is the same as for resistance. That is, for components in series, add the impedances, while if they are in parallel, add their reciprocals.

The impedance of a resistor is trivial. It is just the resistance R. In this case, the voltage drop across the resistor in phase with the current through it since $Z = R$ is a purely real quantity. The impedance is also independent of frequency in this case.

Things get to be more fun with capacitors. In this case the voltage drop $V = V_0 e^{\imath\omega t} = q/C$ and the current $i = dq/dt = \imath\omega C \times V_0 e^{\imath\omega t}$. Therefore, the impedance is

$$Z(\omega) = \frac{V(\omega, t)}{i(\omega, t)} = \frac{1}{\imath\omega C} \tag{2.7}$$

Now the behavior of capacitors is clear. At frequencies low compared to $1/RC$, i.e. the "DC limit", the impedance of the capacitor goes to infinity. (Here, the value of R is the equivalent resistance in series with the capacitor.) It does not allow current to pass through it. However, as the frequency gets much larger than $1/RC$, the impedance goes to zero and the capacitor acts like a short since current passes through it as if it were not there. You can learn a lot about the behavior of capacitors in circuits just by keeping this in mind.

There is an important lesson here. Between any two conductors, there is always some capacitance. Therefore, no matter how well some circuit is designed, there will always be some "stray" capacitance around, however small. Consequently, the circuit will always fail above some frequency because effective shorts appear throughout. You can only keep the stray capacitance so small, especially in integrated circuit chips where things are packed tightly

together, and this is a practical limitation for all circuit designers.

2.2.3 The Generalized Voltage Divider

We can easily generalize our concept of the voltage divider to include AC circuits and reactive (i.e. frequency dependent) components like capacitors. (We will learn about another reactive component, the inductor, shortly.) The generalized voltage divider is shown in Fig. 2.7. In this case, we have

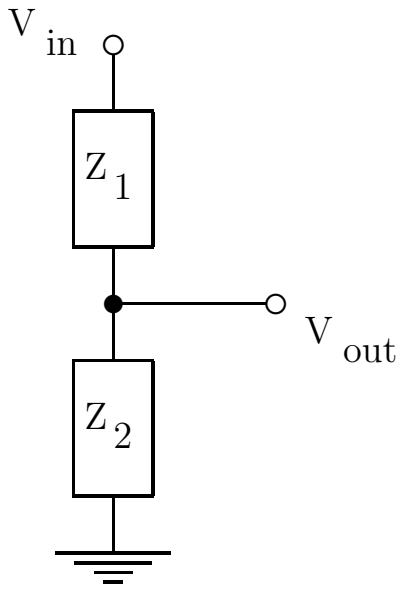


Figure 2.7: The generalized voltage divider.

$$V_{OUT}(\omega, t) = V_{IN}(\omega, t)\frac{Z_2}{Z_1 + Z_2} = V_{IN}(\omega, t)\, g e^{\imath\phi} \tag{2.8}$$

where we take the liberty of writing the impedance ratio $Z_1/(Z_1 + Z_2)$, a complex number, in terms of two real numbers g and ϕ. We refer to $g = |V_{OUT}|/|V_{IN}|$ as the "gain" of the circuit, and ϕ is the phase shift of the output signal relative to the input signal. For the simple resistive voltage divider shown in Fig. 2.4 and Fig. 2.5, we have $g = R_1/(R_1 + R_2)$ and $\phi = 0$.

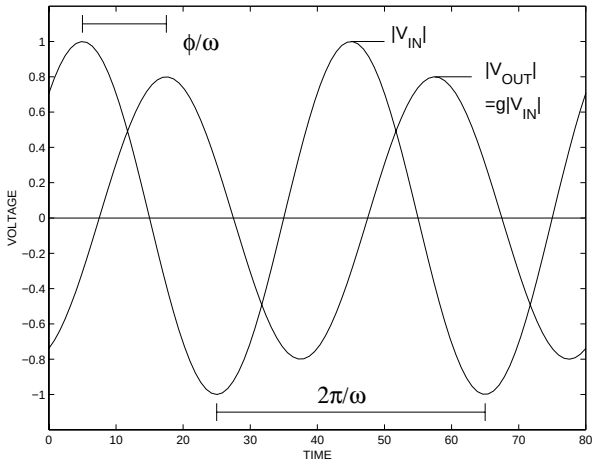


Figure 2.8: Input and output voltages for the generalized voltage divider.

That is, the output signal is in phase with the input signal, and the amplitude is just reduced by the relative resistor values. This holds at all frequencies, including DC.

The relative phase is an important quantity, so let's take a moment to look at it a little more physically. If we write $V_{IN} = V_0 e^{\imath\omega t}$, then according to Eq. 2.8 we can write $V_{OUT} = gV_0 e^{\imath\omega t+\phi}$. Since the measured voltage is just the real part of these complex expressions, we have

$$\begin{aligned} V_{IN} &= V_0 \cos(\omega t) \\ V_{OUT} &= gV_0 \cos(\omega t + \phi) \end{aligned}$$

These functions are plotted together in Fig. 2.8. The output voltage crests at a time different than the input voltage, and this time is proportional to the phase. To be exact, relative to the time at which V_{IN} is a maximum,

$$\text{Time of maximum } V_{OUT} = -\frac{\phi}{2\pi} \times T = \frac{\phi}{\omega}$$

where $T = 2\pi/\omega$ is the period of the voltage fluctuations. This time lag can make all the difference in the world in many circuits.

Now let's consider the voltage divider in Fig. 2.6. Using Eq. 2.8 we find

$$V_{OUT} = V_{IN} \frac{\frac{1}{\imath\omega C}}{R + \frac{1}{\imath\omega C}} = V_{IN} \frac{1}{1 + \imath\omega RC}$$

The gain g of this voltage divider is just $(1+\omega^2R^2C^2)^{-1/2}$ and you can see that for $\omega = 0$ (i.e. DC operation) the gain is unity. For very large frequencies, though, the gain goes to zero. The gain changes from unity to zero for frequencies in the neighborhood of $1/RC$. We have said all this before, but in a less general language.

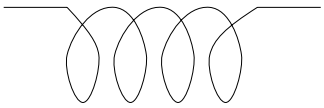
However, our new language tells us something new and important about V_{OUT}, namely the phase relative to V_{IN}. Using equations C.2 and C.3, we find that

$$\frac{1}{1+\imath\omega RC} = \frac{1-\imath\omega RC}{1+\omega^2R^2C^2} = \frac{1}{(1+\omega^2R^2C^2)^{1/2}}e^{\imath\phi}$$

In other words, the output voltage is phase shifted relative to the input voltage by an amount $\phi = -\tan^{-1}(\omega RC)$. For $\omega = 0$ there is no phase shift, as you should expect, but at very high frequencies the phase is shifted by $-90°$.

2.3 Inductors

Just as a capacitor stores energy in an electric field, an inductor stores energy in a magnetic field. An inductor is essentially a wire wound into the shape of a solenoid. The symbol for an inductor is



At first, you might think "A wire is a wire, so what difference could it make to a circuit?" They key is in the magnetic field that is set up inside the coil, and what happens when the current *changes*. So, just as with a capacitor, inductors are important when the voltage and current change with time, and the response depends on the frequency.

The inductance L of a circuit element is defined to be

$$L = \frac{N\Phi}{i}$$

where N is the number of turns in the solenoid and Φ is the magnetic flux in the solenoid generated by the current i. The SI unit of inductance is the Tesla·m^2/Ampere, or the Henry (H).

Now if the current i through the inductor coil is changing, then the magnetic flux is changing and this sets up a voltage in the coil that resists the change in the current. The magnitude of this voltage drop is

$$V = \frac{d\,(N\Phi)}{dt} = L\frac{di}{dt}$$

If we write $V = iZ$, where Z is the impedance of the inductor, and $V = V_0 e^{\imath\omega t}$, then $V = (L/Z)(\imath\omega)V$ or

$$Z = \imath\omega L \tag{2.9}$$

We can use this impedance to calculate, for example, V_{OUT} for the generalized voltage divider of Fig. 2.7 if one or more of the components is an inductor.

You can now see that the inductor is, to large extent, the opposite of a capacitor. The inductor behaves as a short (that is, just the wire it is) at low frequencies, whereas a capacitor is open in the DC limit. On the other hand, an inductor behaves as if the wire were cut (an open circuit) at high frequencies, but the capacitor is a short in this limit. You can make interesting and useful circuits by combining inductors and capacitors in different combinations.

One particularly interesting combination is the series LCR circuit, combining one of each in series. The impedance of such a string displays the phenomenon of "resonance". That is, in complete analogy with mechanical resonance, the voltage drop across one of the elements is a maximum for a certain value of ω. Also, as the frequency passes through this value, the relative phase of the output voltages passes through 90°. If the resistance R is very small, then the output voltage can be enormous, in principle.

2.4 Diodes and Transistors

Resistors, capacitors, and inductors are "linear" devices. That is, we write $V = iZ$, where Z is some (complex) number, which may be a function of frequency. The point is, though, that if you increase V by some factor, then you increase i by the same factor.

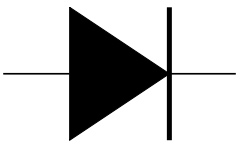
Diodes and transistors are examples of "nonlinear" devices. Instead of talking about some imedance Z, we instead consider the relationship between V and i as some (nonlinear) function. What's more, a transistor is an "active" device, unlike resistors, capacitors, inductors, and diodes which are "passive". That is, a transistor takes in power from some voltage or current source, and gives an output that combines that input power with the signal input to get a response. As you might guess, transistors are very popular signal amplifiers, although they have lots of other uses as well.

Instead of covering the world of nonlinear devices at this time, we will just discuss some of their very basic properties. We will describe their operation in some more detail when we use them in specific experiments, since they can be used in a large variety of ways.

You might know that in the old days, many of these functions were possible with vacuum tubes of various kinds. These have been almost completely replaced by solid state devices based on semiconductors.

2.4.1 Diodes

The symbol for a diode is



where the arrow shows the nominal direction of current flow. An ideal diode conducts in one direction only. That is, its $V - i$ curve would give zero current i for $V < 0$ and infinite i for $V > 0$. (Of course, in practice, the current i is limited by some resistor in series with the diode.) This is shown in Fig. 2.9(a).

A real diode, however, has a more complicated curve, as shown in Fig. 2.9(b). The current i changes approximately exponentially with V, and becomes very large for voltages above some forward voltage drop V_F. For most cases, a good approximation is that the current is zero for $V < V_F$ and unlimited for $V > V_F$. Typical values of V_F are between 0.5 V and 0.8 V.

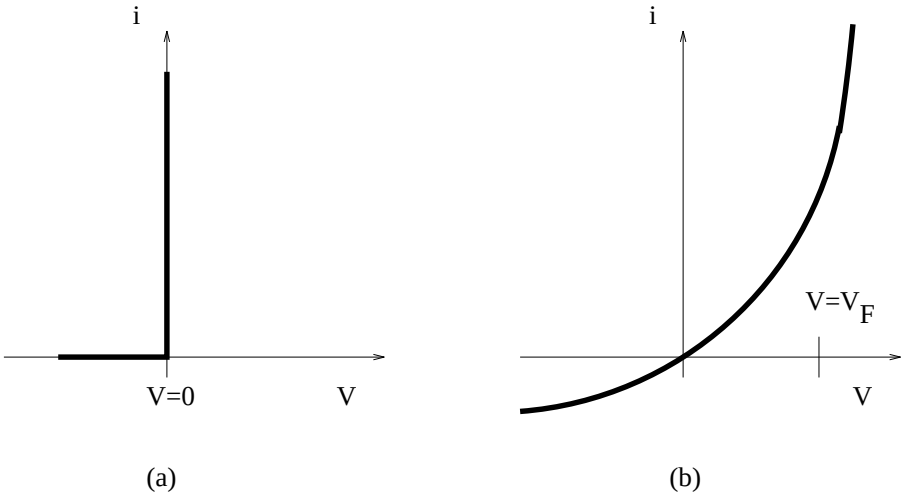


Figure 2.9: Current i versus voltage V for (a) the ideal diode and (b) a real diode.

Diodes are *pn* junctions. These are the simplest solid-state devices, made of a semiconductor (usually silicon). The electrons in a semiconductor fill an energy "band" and normally cannot move through the bulk material, so the semiconductor is really an insulator. If electrons make it into the next energy band, which is normally empty, then they can conduct electricity. This can happen if, for example, electrons are thermally excited across the energy gap between the bands. For silicon, the band gap is 1.1 eV, but the mean thermal energy of electrons at room temperature is $\sim kT = 1/40$ eV. Therefore, silicon is essentially an insulator under normal conditions, and not particularly useful.

That's where the p and n come in. By adding a small amount (around 10 parts per million) of specific impurities, lots of current carriers can be added to the material. These impurities (called dopants) can precisely control how current is carried in the semiconductor. Some dopants, like arsenic, give electrons as carriers and the doped semiconductor is called n-type, since the carriers are *n*egative. Other dopants, like boron, bind up extra electrons, and current is carried by "holes" created in the otherwise filled band. These holes act like positive charge carriers, so we call the semiconductor p-type. In either case, the conductivity increases by a factor of $\sim$ 1000 at room temperature. this makes some nifty things possible.

So now back to the diode, or *pn* junction. This is a piece of silicon, doped *p*-type on one side and *n*-type on the other. Electrons can only flow from *p* to *n*. That is, a current is carried only in one direction. A detailed analysis gives the $i - V$ curve shown in Fig. 2.9(b). See Dunlap for more details.

If you put voltage across the diode in the direction opposite to the direction of possible current flow, that is called a "reverse bias". If you put too much of a reverse bias on the diode, i.e. $V < -V_R^{MAX}$, it will break down and start to conduct. This is also shown in Fig. 2.9(b). Typical values of V_R^{MAX} are 100 V or less.

2.4.2 Transistors

Transistors are considerably more complicated than diodes[2], and we will only scratch the surface here. The following summary closely follows the introduction to transistors in *The Art of Electronics*. For details on the underlying theory, see Dunlap.

A transistor has three terminals, called the collector, base, and emitter. There are two main types of transistors, namely *npn* and *pnp*, and their symbols are shown in Fig. 2.10. The names are based on the dopants used in the semiconductor materials. The properties of a transistor may be summarized in the following simple rules for *npn* transistors. (For *pnp* transistors, just reverse all the polarities.)

1. The collector must be more positive than the emitter.
2. The base-emitter and base-collector circuits behave like diodes. Normally the base-emitter diode is conducting and the base-collector diode is reverse-biased.
3. Any given transistor has maximum values of i_C, i_B, and V_{CE} that cannot be exceeded without ruining the transistor. If you are using an transistor in the design of some circuit, check the specifications to see what these limitng values are.

[2]The invention of the transistor was worth a Nobel Prize in Physics in 1956.

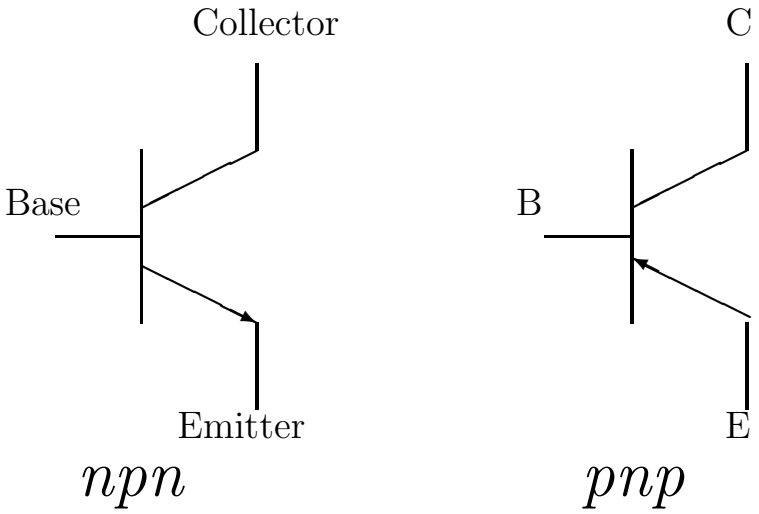


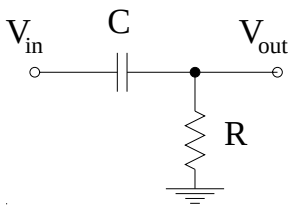
Figure 2.10: Symbols for *npn* and *pnp* transistors.

4. When rules 1-3 are obeyed, i_C is roughly proportional to i_B and can be written as $i_C = h_{FE} i_B$. The parameter h_{FE}, also called β, is typically around 100, but it varies a lot among a sample of nominally identical transistors.

Obviously, rule 4 is what gives a transistor its punch. It means that a transistor can "amplify" some input signal. It can also do a lot of other things, and we will see them in action later on.

2.5 Exercises

1. Consider the following simple circuit:

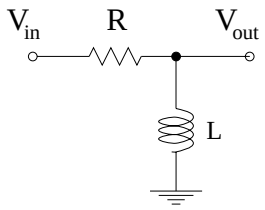


Let the input voltage V_{IN} be a sinusoidally varying function with amplitude V_0 and angular frequency ω.

a. Calculate the gain g and phase shift ϕ for the output voltage relative to the input voltage.

b. Plot g and ϕ as a function of ω/ω_0 where $\omega_0 = 1/RC$. For each of these functions, use the combination of linear or logarithmic axes for g and for ϕ that you think are most appropriate.

2. Consider the following simple circuit:

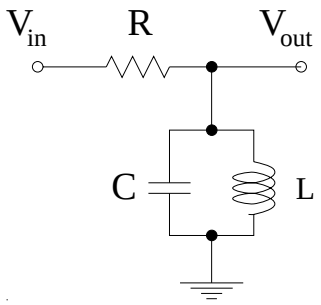


Let the input voltage V_{IN} be a sinusoidally varying function with amplitude V_0 and angular frequency ω.

a. Calculate the gain g and phase shift ϕ for the output voltage relative to the input voltage.

b. Plot g and ϕ as a function of ω/ω_0 where $\omega_0 = R/L$. For each of these functions, use the combination of linear or logarithmic axes for g and for ϕ that you think are most appropriate.

3. Consider the following not-so-simple circuit:



a. What is the gain g for very low frequencies ω? What is the gain for very high frequencies? Remember that capacitors act like dead shorts and open circuits at high and low frequencies, respectively, and inductors behave in just the opposite way.

b. At what frequency do you suppose the gain of this circuit is maximized? Use your intuition, and perhaps some of Chapter 38 in Resnick, Halliday, and Krane.

c. Using the rules for impedance and the generalized voltage divider, determine the gain $g(\omega)$ for this circuit and show that your answers to (a) and (b) are correct.

4. Suppose that you wish to detect a rapidly varying voltage signal. However, the signal is superimposed on a large DC voltage level that would damage your voltmeter if it were in contact with it. You would like to build a simple passive circuit that allows only the high frequency signal to pass through.

a. Sketch a circuit using only a resistor R and a capacitor C that would do the job for you. Indicate the points at which you measure the input and output voltage.

b. Show that the magnitude of the output voltage equals the magnitude of the input voltage, multiplied by

$$\frac{1}{\sqrt{1+\frac{1}{\omega^2 R^2 C^2}}}$$

where ω is the (angular) frequency of the signal. You may use the expression for capacitor impedance we derived in class.

c. Suppose that R=1kΩ and the signal frequency is 1MHz=10^6/sec. Suggest a value for the capacitor C.

Ch 3

Common Laboratory Equipment

There are lots of different kinds of laboratory equipment. In fact, there are too many to cover in any detail, and you will learn about specific pieces of equipment as you do the experiments. However, there are certain kinds of equipment common to nearly all experiments, and we will talk about these in this chapter. As you might imagine, all of this equipment is related to generating or measuring voltage.

I don't know of any book that covers the specific sorts of things in this chapter. If you are interested in some specific piece of equipment, however, a good place to check is with the manufacturer or distributer of a product line. You can typically get good documentation for free, and not always in the form of the company's catalog.

3.1 Wire and Cable

Connections between components are made with wires. We tend to neglect the importance of choosing the right wire for the job, but in some cases it can make a big difference.

The simplest wire is just a strand of some conductor, most often a metal like copper or aluminum. Usually the wire is coated with an insulator so that it will not short out to its surroundings, or to another part of the wire itself. If the wire is supposed to carry some small signal, then it will likely need to be "shielded", that is covered with another conductor (outside the insulator) so that the external environment doesn't add noise somehow. One popular type of shielded wire is the "coaxial cable" which is also used to propagate "pulses".

3.1.1 Basic Considerations

Don't forget about Ohm's law when choosing the proper wire. That is, the voltage drop across a section of wire is still $V = iR$, and you want this voltage drop to be small compared to the "real" voltages involved. The resistance $R = \rho \times L/A$ where L is the length of the wire, A is its cross sectional area, and ρ is the resistivity of the metal. Therefore, to get the smallest possible R, you keep the length L as short as practical, get a wire with the largest practical A[1], and choose a conductor with small resistivity. Copper is the usual choice because it has low resistivity ($\rho = 1.69 \times 10^{-8}$ Ωcm) and is easy to form into wire of various thicknesses and shapes. Other common choices are aluminum ($\rho = 2.75 \times 10^{-8}$ Ωcm) which can be significantly cheaper in large quantities, or silver ($\rho = 1.62 \times 10^{-8}$ Ωcm) which is a slightly better conductor although not usually worth the increased expense.

The resistivity increases with temperature, and this can lead to a particularly insidious failure if the wire has to carry a large current. The power dissipated in the wire is $P = i^2R$, and this tends to heat it up. If there is not enough cooling by convection or other means, then R will increase and the wire will get hotter and hotter until it does serious damage. This is most common in wires used to wind magnets, but can show up in other high power applications. A common solution is to use very low gage (i.e. very thick) wire, that has a hollow channel in the middle through which water flows. The water acts as a coolant to keep the wire from getting too hot.

[1] Wire diameter is usually specified by the "gage number". The smaller the wire gage, the thicker the wire, and the larger the cross sectional area.

The wire insulator must also withstand the temperature increase, and whatever else the outside environment wants to throw at it. It may be necessary, for example, to immerse part of a circuit in liquid nitrogen, and you don't want the insulator to crack apart. It should not be hard to find a conductor and insulator combination that will suit your purpose.

3.1.2 Coaxial Cable

A coaxial cable is a shielded wire. The name comes from the fact that the wire sits inside an insulator, another conductor, and another insulator, all in circular cross section sharing the same axis. A cutaway view is shown in Fig. 3.1. Coaxial cable is used in place of simple wire when the signals are

Figure 3.1: Cutaway view of coaxial cable.

very small and are likely to be obscured by some sort of electronic noise in the room. The outside conductor (called the "shield") makes it difficult for external electromagnetic fields to penetrate to the wire, and minimizes the noise. This outside conductor is usually connected to ground.

A second, and very important, use of coaxial cable is for "pulse transmission". The wire and shield, separated by the dielectric insulator, act as a kind of waveguide and allows short pulses of current to be transmitted with little distortion from dispersion. Short pulses can be very common in the laboratory, in such applications as digital signal transmission and in radiation detectors. You have to be aware of the "characteristic impedance" of the cable when you use it in this way.

Coaxial cable has a characteristic impedance because it transmits the

signal as a train of electric and magnetic fluctuations, and the cable itself has characteristic capacitance and inductance. The capacitance and inductance of a cylindrical geometry like this are typically solved in elementary physics texts on electricity and magnetism. The solutions are

$$C = \frac{2\pi\epsilon}{\ln(b/a)} \times \ell \quad \text{and} \quad L = \frac{\mu}{2\pi} \ln\left(\frac{b}{a}\right) \times \ell$$

where a and b are the radii of the wire and shield respectively, ϵ and μ are the permittivity and permeability of the dielectric, and ℓ is the length of the cable. It is very interesting to derive and solve the equations that determine pulse propagation in a coaxial cable, but we won't do that here. One thing you learn, however, is that the impedance seen by the pulse (which is dominated by high frequencies) is very nearly real and independent of frequency, and equal to

$$Z_c = \sqrt{\frac{L}{C}} = \frac{1}{2\pi}\sqrt{\frac{\mu}{\epsilon}} \ln\left(\frac{b}{a}\right) \tag{3.1}$$

This "characteristic impedance" is always in a limited range, typically $50\Omega \leq Z_c \leq 200\Omega$, owing to natural values of ϵ and μ, and to the slow variation of the logarithm.

You have to be careful when making connections with coaxial cable, so that the characteristic impedance Z_c of the cable is "matched" to the load impedance Z_L. The transmissions equations are used to show that the "reflection coefficient" Γ, defined as the ratio of the current reflected from the end of the cable to the current incident on the end, is given by

$$\Gamma = \frac{Z_L - Z_c}{Z_L + Z_c}$$

That is, if a pulse is transmitted along a cable and the end of the cable is not connected to anything ($Z_L = \infty$), then $\Gamma = 1$ and the pulse is immediately reflected back. On the other hand, if the end shorts the conductor to the shield ($Z_L = 0$), then $\Gamma = -1$ and the pulse is inverted and then sent back. *The ideal case is when the load has the same impedance as the cable. In this case, there is no loss at the end of the cable and the full signal is transmitted through.* You should take care in the lab to use cable and electronics that have matched impedances. Common impedance standards are 50Ω and 90Ω.

3.1.3 Connections

Of course, you will need to connect your wire to the apparatus somehow, and this is done in a wide variety of ways. For permanent connections, especially inside electronic devices, solder is usually the preferred solution. You won't typically make solder joints in the undergraduate laboratory, unless you are building up some piece of apparatus. It is harder than you might think to make a good solder joint, and if you are going to do some of this, you should have someone show you who has a decent amount of experience. Another type of permanent connection, called "crimping", squeezes the conductors together using a special tool that ensures a good contact that does not release. This is particularly useful if you can't apply the type of heat necessary to make a good solder joint. Again, you are unlikely to encounter this in the undergraduate laboratory.

Less permanent connections can be made using terminal screws or binding posts. These work by taking a piece of wire and inserting it between two surfaces which are then forced together by tightening a screw. You may need to twist the end of the wire into a hook or loop to do this best, or you may use wire with some sort of attachment that has been soldered or crimped on the end.

If you keep tightening or untightening screws, especially onto wires with hand made hooks or loops, then the wire is likely to break at some point. Therefore, for temporary connections, it is best to use alligator clips or banana plugs, or something similar. Again, you will usually use wires with this kind of connector previously soldered or crimped on the end.

Coaxial cable connections are made with one of several special types of connectors. Probably most common is the "Bayonet N-Connector", or BNC, standard, including male cable end connectors, female device connectors, and union and T-connectors for joining cables. In this system, a pin is soldered or crimped to the inner conductor of the cable, and the shield is connected to an outer metal holder. Connections are made by twisting the holder over the mating connector, with the pin inserting itself on the inner part. Another common connector standard, called "Safe High Voltage" or SHV, works similarly to BNC, but is designed for use with high DC voltages by making it difficult to contact the central pin unless you attach it to the

correct mate.

For low level measurement you must be aware of the thermal electric potential difference between two dissimilar conductors at different temperatures. These "thermoelectric coefficients" are typically around 1 μV/°C, but between Copper and Copper-Oxide (which can easily happen if a wire or terminal has been left out and is oxidized) it is around 1 mV/°C.

3.2 DC Power Supplies

A lot of laboratory equipment needs to be "powered" in one way or another. Unlike the typical 100 V 60 Hz AC line you get out of the wall socket, though, this equipment usually requires some constant DC level to operate. One way to get this constant DC level is to use a battery, but if the equipment draws much current the battery will die quickly. More often we use DC "power supplies" to get this kind of constant DC level. The power supply in turn gets its power from the wall socket.

Power supplies come in lots of shapes, sizes, and varieties, but there are two general classes. These are "voltage" supplies or "current" supplies, and the difference is based on how the output is regulated. Since the inner workings of the power supply has some effective resistance, when the power supply has to give some current, there will be a voltage drop across that resistance and that will affect how the power supply works. In a "voltage regulated" supply, the circuitry is designed to keep the output voltage constant (to within some tolerance), regardless of how much current is drawn. (Typically, there will be some maximum current at which the regulation starts to fail. That is, there is a maximum power that can be supplied.) Most electronic devices and detector systems prefer to have a specific voltage they can count on, so they are usually connected to voltage regulated supplies.

A "current regulated" supply is completely analagous, but here the circuitry is designed to give a constant output current in the face of some load on the supply. Such supplies are most often used to power magnets, since the magnetic field only cares about how much current flows through the coils. This is in fact quite important for precise magnetic fields, since the coils tend

to get hot and change their resistance. In this case, $V = iR$ and R is changing with time, so the power supply has to know to keep i constant by varying V accordingly. In many cases, a simple modification (usually done without opening up the box) can convert a power supply from voltage regulation to current regulation.

The output terminals on most power supplies are "floating". That is, they are not tied to any external potential, in particular to ground. One output (sometimes colored in red) is positive with respect to the other (black). You will usually connect one of the outputs to some external point at known potential, like a common ground.

You should be aware of some numbers. The size and price of a power supply depends largely on how much power it can supply. If it provides a voltage V while sourcing a current i, then the power output is $P = iV$. A very common supply you will find around them lab will put out several volts and a couple of amps, so something like 10 W or so. Depending on things like control knobs and settings to computer interfacing, they can cost anywhere from $50 up to a few hundred. So-called "high voltage" power supplies will give several hundred up to several thousand (or more!) volts, and can source anywhere from a few μA up to 100 mA, and keep the voltage constant to a level of better than 100 mV. Still, the power output of such devices is not enormously high, typically under a few hundred watts. The cost will run into thousands of dollars. Magnet power supplies, though, may be asked to run something like 50 A through a coil that has a resistance of, say, 2 Ω. In this case, the output power is 5 kW, and that is a force to be reckoned with. Realize, of course, that these are all round numbers just to give you some idea of what you'll see around the laboratory.

3.3 Waveform Generators

You might think there are things called "AC" power supplies, analagous to the DC supplies we've just discussed. Well there are, but we don't call them that because in general (and certainly for the equipment you will see in this course) they don't supply much power. Instead, we talk about "Waveform Generators" which produce an output voltage signal that varies in time.

Later, in Experiment 5 we will combine a waveform generator with a DC power supply to make an AC power supply, and we will talk more about that then.

The function $V(t)$ can be anything from a simple sine wave to an arbitrary function you program into the device, but increased flexibility can cost a lot of money. Most waveform generators, though, do have at least sine waves, square waves, or triangle waves, and can vary the frequency over a wide range. Low frequencies are pretty easy to get, but for very high frequencies (above a MHz or so) things get much harder because of stray capacitance giving effective shorts. (See section 2.2.1.) You can also vary the voltage amplitude and offset over several volts.

Sometimes instead of wanting a "wave" output, you need a "pulse". That is, a signal that is high for some short period of time, with another coming after a much longer time. Most waveform generators can accomodate your wishes either by providing an explicit "pulse" output, or by allowing you to change the symmetry of the waveform so that the "0 to π" portion of the wave is stretched or compressed relative to the "π to 2π" portion.

3.4 Meters

So now that you know how to get some voltage, including time varying ones, and how to connect these voltages using wire and cable, you have to think about how to measure the voltage you create. The simplest way to do this is with a "meter", particularly if the voltage is DC. (Most meters do provide you with AC capability, but we won't go into the details here.) An excellent reference on the subject of meters is given in the "Low Level Measurements Handbook", published by Keithley Instruments, Inc. (If you want a copy, call them at (216)248-0400 and they will probably send you one for free.) As you might imagine, Keithley sells meters.

In the old days, people would use either voltmeters, ammeters, or ohmmeters to measure voltage, current, or resistance respectively. These days, although you still might want to buy one of these specialized instruments to get down to very low levels, most measurements are done with "Digital Mul-

timeters", or DMM's for short. (In fact, some DMM's are available now that effectively take the place of the most sensitive specialized meters.) Voltage and resistance measurements are made by connecting the meter in parallel to the portion of the circuit you're interested in. To measure current, you have to put the meter in series.

Realize that DMM's work by averaging the voltage measurement over some period of time, and then displaying the result. This means that if the voltage is fluctuating on some time scale, these fluctuations will not be observed if the averaging time is greater than the typical period of the fluctuations. Of course the shorter the averaging time a meter has (the higher the "bandwidth" it has), the fancier it is and the more it costs.

Most of the applications in this course do not involve very low level measurements, but you should be aware of a simple fact just the same. Meters have some effective input impedance, so they will (at some level) change the voltage you are trying to measure. For this reason, voltmeters and ohmmeters are designed to have very large input impedances (many MΩ to as high as several GΩ), while ammeters "shunt" the current through a very low resistance and turn the job into measuring the (perhaps very low) voltage drop across that resistor.

3.5 Oscilloscopes

An oscilloscope measures and displays voltage as a function of time. That is, it plots for you the quantity $V(t)$ on a cathode ray tube (CRT) screen as it comes in. This is a very useful thing, and you will use oscilloscopes in nearly all the experiments you do in this course. A good reference is "The XYZ's of Oscilloscopes", published by Tektronix, Inc., probably the world's largest manufacturer of oscilloscopes.

The simple block diagram shown in Fig.3.2 explains how an oscilloscope works. The voltage you want to measure serves two purposes. First, after being amplified, it is applied to the vertical deflection plates of the CRT. This means that the vertical position of the trace on the CRT linearly corresponds to the input voltage, which is just what you want. The vertical scale on the

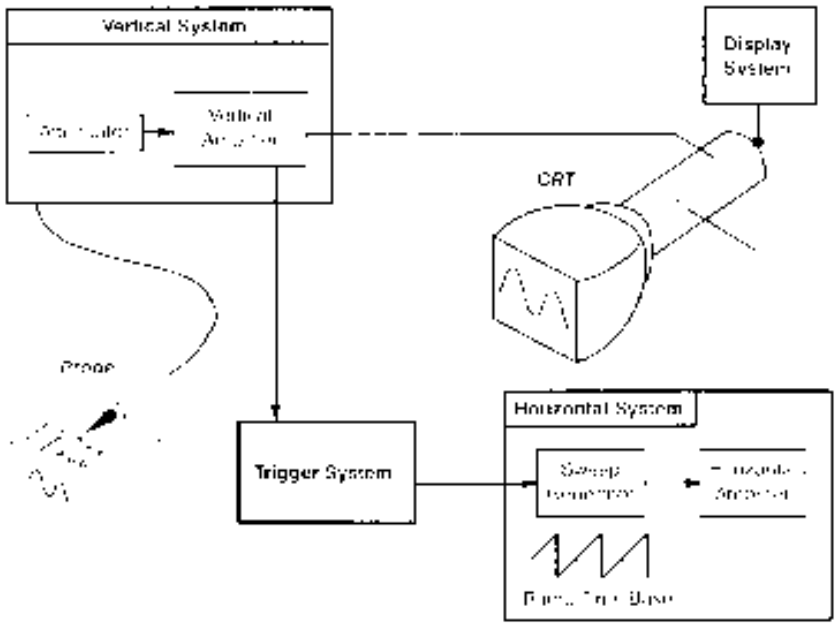


Figure 3.2: Block Diagram of an Oscilloscope

CRT has a grid pattern that lets you know what the input voltage is.

3.5.1 Sweep and Trigger

The horizontal position of the trace is controlled by a "sweep generator" whose speed you can control. However, for repetitive signal shapes, you want the signal to "start" at the same time for ever sweep, and this is determined by the "trigger" system. The place on the screen where the trace starts is controlled by a "horizontal position" knob on the front panel. One kind of trigger is to just have the scope sweep at the line (i.e. 60 Hz) frequency, but this won't be useful if the signals you're interested in don't come at that frequency. Another kind of simple trigger is to have the trace sweep once whenver the voltage rises or falls past some level, i.e a "leading edge" trigger. There is usually a light on the front panel that flashes when the scope is triggered.

Oscilloscopes almost always have at least two input channels, and it is possible to trigger on one channel and look at the other. This can be very useful for studying coincident signals or for measuring the relative phase of two waveforms. In any case, the trigger "mode" can either be "normal", in which case there is a sweep only if the trigger condition is met, or "auto" where the scope will trigger itself if the trigger condition is not met in some period of time. Auto mode is particularly useful if you are searching for some weak signal and don't want the trace to keep disappearing on you.

3.5.2 Input Voltage Control

You have several controls on how the input voltage is handled. A "vertical position" knob on the front panel controls where the trace appears on the screen. You will find one of these for each input channel. The input "coupling" can be set to either AC, DC, or ground. In AC mode, there is a capacitor between the input connector and the vertical system circuit. This keeps any constant DC level from entering the scope, and all you see is the time varying (i.e. AC) part. If you put the scope on DC, then the constant voltage level also shows up. If the input coupling is grounded, then you force the input level to zero, and this shows you where zero is on the screen. (Make sure that the scope is on "auto" trigger if you ground the input, otherwise you will not see a trace!)

Sometimes, you also get to choose the input impedance for each channel. Choosing the "high" input impedance (usually 1 MΩ) is best if you want to measure voltage levels and not have the oscilloscope interact with the circuit. However, the oscilloscope will get a lot of use looking at fast pulsed signals transmitted down coaxial cable, and you don't want an "impedance mismatch" to cause the signal to be reflected back. (See Sec.3.1.2.) Cables with 50Ω characteristic impedances are very common in this work, so you may find a 50Ω input impedance option on the scope. If not, you should use a "tee" connector on the input to put a 50Ω load in parallel with the input.

3.5.3 Dual Trace Operation

By flipping switches on the front, you can look at either input channel's trace separately, or both at the same time. There is obviously a problem, though, with viewing both simultaneously since the vertical trace can only be in one place at a time.

There are two ways to get around this. One is the *alternate* the trace from channel one to channel two and back again. This gives complete traces of each, but doesn't really show them to you at the same time. If the signals are very repetitive and you're not interested in fine detail, this is okay. However, if you really want to see the traces at the same time, select the *chop* option. Here, the trace jumps back and forth between the channels at some high frequency, and you let your eye interpolate between the jumps. If the sweep speed is relatively slow, the interpolation is no problem and you probably can't tell the difference between *alternate* and *chop*. However, at high sweep speed, the effect of the chopping action will be obvious.

3.5.4 Bandwidth

You should realize by now that high frequency operation gets hard, and the oscilloscope gets more complicated and expensive. Probably the single most important specification for an oscilloscope is its "bandwidth", and you will see that number printed on the front face right near the screen. The number tells you the frequency at which a sine wave would appear only 71% as large as it should be. You cannot trust the scope at frequencies approaching or exceeding the bandwidth. Most of the scopes in the lab have 20 MHz or 60 MHz bandwidths. A "fast" oscilloscope will have a bandwidth of a few hundred MHz or more. You will find that you can the sweep speed over a large range, but never much more than $(\text{Bandwidth})^{-1}$. The "vertical sensitivity" can be set independently of the sweep speed, but scopes in general cannot go below around 2 mV/division.

3.5.5 XY Operation

On most oscilloscopes, if you turn the sweep speed down to the lowest value, one more notch puts the scope in the XY display mode. Now, the trace displays channel one (X) on the horizontal axis and channel two (Y) on the vertical. For periodic signals, the trace is a lissajous pattern from which you can determine the relative phase of the two inputs.

Oscilloscopes are also used this way as displays for various pieces of equipment which have XY output options. Thus, the oscilloscope can be used as a plotting device in some cases.

3.6 Digitizers

Computers have become common in everyday life, and the experimental physics laboratory is no exception. In order to measure a voltage and deal with the result in a computer, the voltage must be *digitized*. The generic device that does this is the Analog-to-Digital Converter or ADC. ADC's come in approximately an infinite number of varieties and connect to computers in lots of different ways. We will cover the particulars when we discuss the individual experiments, but for now we will review some of the basics.

3.6.1 ADC's

Probably the most important specification for an ADC is its resolution. We specify the resolution in terms of the number of binary digits ("bits") that the ADC spreads out over its measuring range. The actual measuring range can be varied externally by some circuit, so the number of bits tells you how finely you can chop that range up. Obviously, the larger the number of bits, the closer you can get to knowing exactly what the input voltage was before it was digitized. A "low resolution" ADC will have 8 bits or less. That is, it divides the input voltage up into 256 pieces and gives the computer a number between 0 and 255 which represents the voltage. A "high resolution" ADC has 16 bits or more.

High resolution does not come for free. In the first place, it can mean a lot more data to handle. For example, if you want to histogram the voltage being measured with an 8 bit ADC, then you need 256 channels for each histogram. However, if you want to make full use of a 16 bit ADC, every histogram would have to consume 65536 channels. That can use up computer memory and disk space in a hurry. Resolution also affects the *speed* at which a voltage can be digitized. Generally speaking, it takes much less time to digitize a voltage into a smaller number of bits, than it does for a large number of bits.

There are three general classes of ADC's, which I refer to as *Flash*, *Peak Voltage Sensing*, and *Charge Integrating* ADC's. A Flash ADC, or "waveform recorder", simply reads the voltage level at its input and converts that voltage level into a number. They are typically low resolution, but run very fast. Today you can get an 8 bit Flash ADC which digitizes at 100 Mhz (i.e. one measurement every 10 ns). This is fast enough so that just about any time varying signal can be converted to numbers so that a true representation of the signal can be stored in a computer.

To get better resolution, you need to decide what it is about the signal you are really interested in. For example, if you only care about the maximum voltage value, you can use a peak sensing ADC which digitizes the maximum voltage observed during some specified time. Sometimes, you are interested instead in the area underneath some voltage signal. This is the case, for example, in elementary particle detectors where the net charge delivered is a measure of the particle's energy. For applications like this, you can use an integrating ADC which digitzes the net charge absorbed over some time period, i.e. $\frac{1}{R}\int_{t_1}^{t_2} V(t)dt$, where R is the resistance at the input. For either of these types, you can buy commercial ADC's that digitize into 12 or 13 bits in 5 μs or longer, but remember that faster and more bits costs more money.

Don't forget that one of your jobs as an experimenter will be to calibrate (or otherwise know) how to convert the number you get from an ADC into an actual voltage or charge value. You will need to do this for some of the experiments in this course.

3.6.2 Other Digital Devices

The opposite of an ADC is a DAC, or Digital-to-Analog Converter. Here the computer feeds the DAC a number depending on the number of bits, and the DAC puts out an analog voltage proportional to that number. The simplest DAC has just one bit, and its output is either "on" or "off". In this case, we refer to the device as an "output register". These devices are a way to control external equipment in an essentially computer-independent fashion.

In many cases, you want to digitize a time interval instead of a voltage level. In the old days, this was a two step process involving a device called a "Time-to-Analog Converter" (TAC), followed by an ADC. Nowadays, both these functions are packaged in a single device called a TDC. The rules and ranges are very similar as for ADC's.

Devices known as "latches" or "input registers" will take an external logic level, and digitize the result into a single bit. These are useful for telling whether some device is on or off, or perhaps if something has happened which the computer should know about. For the latter, the computer interface circuit has to be able to interrupt what the computer is doing to let it know that something important happened on the outside.

3.6.3 Dead Time

Why should you care how fast an ADC, or some other device, digitizes? Obviously, the faster the device works, the faster you can take data. In fact, this can be the limiting factor for many kinds of high sensitivity experiments.

When a device is busy digitizing, it cannot deal with more input. We refer to the cumulative time a device is busy as "dead time". Suppose τ is the time needed to digitize an input pulse, and R_0 is the (presumably random) rate at which pulses are delivered to the digitizer. If R_m is the *measured* rate, then in a time T the number of digitized pulses is $R_m T$. The dead time incurred in time T is therefore $(R_m T)\tau$, so the number of pulses lost is $[(R_m T)\tau]R_0$. The total number of pulses delivered ($R_0 T$) must equal the number digitized

plus the number lost, so

$$R_0 T = R_m T + R_m T \tau R_0$$

and therefore

$$R_m = \frac{R_0}{1 + \tau R_0} \tag{3.2}$$

$$\text{or} \quad R_0 = \frac{R_m}{1 - \tau R_m} \tag{3.3}$$

The "normal" way to operate a digitizer is so that it can keep up with the rate at which pulses come in. In other words, the rate at which it digitizes $(1/\tau)$ should be much greater than the rate at which pulses are delivered, that is $\tau R_0 \ll 1$. Equation 3.2 shows that in this case, $R_m \approx R_0$, that is, the measured rate is very close to the true rate, which is just what you want. Futhermore, an accurate correction to the measured rate is given by Eq. 3.3 which can be written as $R_0 = R_m(1 + \tau R_m)$ under normal operation.

On the other hand, if $\tau R_0 \gg 1$, then $Rm \approx 1/\tau$. That is, the digitizier measures a pulse and before it can catch its breath, another pulse comes along. The device is "always dead", and the measured rate is just one per digitizing time unit. Essentially all information on the true rate is lost, because the denominator of Eq. 3.3 is close to zero. You would have to know the value of τ very precisely in order to make a correction that gives you the true rate.

3.7 Digital Oscilloscopes

The digital oscilloscope is a wonderful device. Instead of taking the input voltage and feeding it directly onto the deflection plates of a CRT (Fig. 3.2), a digital oscilloscope first *digitizes* the input signal using a Flash ADC, stores the waveform in some internal memory, and then has other circuitry to read that memory and display the output on the CRT. At first glance, that may sound silly, since we get the same result but in a much more roundabout way. They key, however, is that we have the voltage stored as numbers, and the

internal computer in the digital oscilloscope can do just about anything with the numbers.

Even though it works very differently from analog oscilloscopes, digital scopes have controls that make them look as much like analog scopes as possible. The same terminology is used, and just about any function that is found on an analog scope will also be found on a digital one.

Digital oscilloscopes are relatively new, and in this case Tektronix does not have a corner on the market. In our lab, for example, we use the oscilloscopes by LeCroy and by Hewlett-Packard. For the models we have, the LeCroy scopes are the most powerful.

3.7.1 The LeCroy 9310 Digital Oscilloscope

Our laboratory is equipped with LeCroy model 9310 and 9310A oscilloscopes. The bandwidth of the 9310 is 300 Mhz and digitizes into 10k channels, while the 9310A is 400 Mhz into 50k. (Both have a maximum digitizing rate of 100 Msamples/sec.) These scopes differ in other minor ways, but both are equipped with a complete mathematics library (including Fast Fourier Transform) and a PC-compatible 3.5-inch floppy disk drive for data storage.

We chose these oscilloscopes partly because of how straightforward it is to use them. Given experience with analog oscilloscopes, you will have no trouble using these much more sophisticated devices. You can do a lot by using very few of the features.

Most of the controls are menu driven, and allow you to do any one of a number of things with the data. Very simply, you can stop the scope at any time and consider the last trace it threw up on the CRT. Using the cursors, you can read on the screen the values of $V(t)$ to within the resolution of the ADC (8 bits). You can also read the time scale, so you can do a better job estimating signal periods and frequencies.

The real power of the oscilloscope is realized with the internal math software, allowing you to do much more complicated things with the data. You can take functions of the trace, such as $\log[V(t)]$ to see if $V(t)$ is consistent

with an exponential decay. You can even take the Fourier transform of the voltage as it comes in, and measure the amplitude and phase of the different Fourier components.

For anything but the simplest data taking, you should use the floppy drive to store traces for further analysis. Start with an empty, or just formatted, 1.5 Mb (HD) 3.5-inch floppy, and follow these steps:

- Insert the disk into the drive (mounted on top of the scope).
- Press the "Utilities" button to bring up that menu. Select "Floppy disk utilities" under that menu.
- You will be asked to "Reread" the disk, again from the menu.
- For the first time you use the disk on that particular scope,
  - Press "Perform disk format". You will be asked to confirm that by pressing it again.
  - Press "Copy template to disk". This puts a *template* file on the disk that identifies the properties of that particular oscilloscope.

  At this point, the disk has a directory file on it which containes the template file. All storage operations go to that directory.
- Press "Return" three times to clear all the menus.

To store a particular trace, it is a good idea to make sure the oscilloscope is "Stopped", i.e., no longer updating traces. Then bring up the "Waveform store" menu, and choose the trace (1, 2, or A-D) you want to store, as well as the medium you want to store it to ("Disk"). A file with a name like sc1.000 will be written to the disk, where the name stands for "store channel 1" (if you in fact chose to store the trace corresponding to channel 1) and the extension keeps track of the number of times you stored that channel. These traces are binary files that must be decoded elsewhere, most easily on the PC in the laboratory.

The files produced by the oscilloscope are in binary format to save space. Remember, the default saves 10k real numbers (50k for the 9310A) plus

additional information for each trace. To convert these binary files to ascii information, you need the program 94TRAN which is supplied by LeCroy. The use of binary files in general is described in a readme document, also supplied by LeCroy. These files are kept in the LeCroy subdirectory on the general use PC in our lab.

As described in readme, the basic way to translate the file to a list of ascii values (representing the voltage value for each point of the trace) is through the command

94TRAN -t*file.tpl* -o*file.lis* *file.abc*

where *file.tpl* is the template file, *file.lis* is the output file, and *file.abc* is the binary file created by the oscilloscope. For more detail, you can also type "94TRAN -h" for help. To get different information about that trace, use a particular "format" specification file. (See readme.) For example, if you want to get all the parameter settings of the scope when that trace was saved, use the file all.fmt:

94TRAN -t*file.tpl* -o*file.lis* -fall.fmt *file.abc*

This is a good way to check the way the scope was setup, but it is a good idea to write down the important things in your logbook when you take your data, as some of the different parameter names are pretty cryptic. The readme file has details on how to write your own format files, if you want.

Once the data is in ascii form, you can use anything you want to analyze it. For example, you might use MATLAB, as described in Sec. 1.4.3 and elsewhere in these notes.

3.8 Computer Interfaces

We've talked about digitizing devices like ADC's on a very elementary scale, and also more sophisticated digital instruments like oscilloscopes and multimeters. In the end, you want to get the data collected by these devices into a

computer. What's more, you want the computer to be able to control these devices. The connection between the computer and the external device is done through an "interface". There are a huge number of different kinds of interfaces.

The architecture of an interface falls into one of two categories. A *serial* interface is the simplest. Here the computer communicates one bit at a time with the outside world. The external device responds to a particular pattern of one's and zero's, and so does the computer. Data is transferred between the two one bit at a time as well. The connection is almost always done through a standard RS-232 serial line, the same way a keyboard is attached to the computer. Lots of pieces of this scheme are standard, such as the communications software and even the connectors, and this is a big advantage. The problem, of course, is communication rate. A fast serial line runs at 19,200 bits per second (the "baud rate"), and at this speed it would take over two minutes to read all 10K, 8 bit data points in one trace of the LeCroy 9310. If you are willing to give up the nice, standard features of an RS-232 connection, you can go faster but the interface hardware and software is more complicated.

In order to go faster, the serial architecture is abandoned altogether, and one goes to a *parallel* type of interface. In this scheme, many bits are transferred at the same time over a parallel set of wires. The wires are connected through some kind of plug-in card directly to the "backplane" of the computer, and this really speeds things up. The software for a particular computer can be rather simple as well. Unfortunately, you lose the ability to have some kind of standard interface because there are lots of different kinds of computers out there, so both hardware and software can be very different. Even the IBM/PC and its look-alikes have at least two distinct backplane architectures.

One way people have tried to bridge the gap between fast-but-specific parallel interfaces and standard-but-slow serial interfaces is to build parallel "middleman" interfaces, and hope that they become popular enough to be an industry standard. That is, a device like an ADC or meter might be designed to connect to the middleman, and computer interfaces would also be designed to connect to it as well. This potentially gives you more freedom of choice, assuming that people out there provide you with lots of choices on

both the device side and on the computer side. Of course, the middleman costs money by itself, so this solution is generally more expensive. Some examples of this type of interface are the following:

GPIB or "General Purpose Interface Bus". Also known as the IEEE-488 standard, or as HPIB by people at Hewlett Packard corporation, this has become quite popular in recent years. It uses an ASCII code to communicate, very similar to most serial line communication systems, but uses a 24-pin connector allowing data to be transferred in parallel at some level. It can transmit up to 1 MByte per second, within this communication protocol.

CAMAC or "Computer Automated Measurement And Control". This standard has been around for a long, long time, and many people are hooked on it because they've already purchased lots of devices that connect to it. It uses a rigid protocol called the Dataway for communication and data transfers can be quite fast and flexible. Programming in CAMAC is rather difficult, however, and people usually end up buying commercial CAMAC software for their favorite computer.

FASTBUS. This architecture was developed originally as a modern replacement for CAMAC, particularly for very high rate and high density applications. It is being used heavily at several large modern laboratories. However, it is rather costly and its popularity has been somewhat limited.

VME or "Versa Module Europa". Developed by a consortium of commercial companies, VME maps device locations directly into computer memory and is designed for high speed, computer intensive applications. Data transfer is very efficient, and the speed is around 20 MBytes per second. It is becoming increasingly popular, particularly in Europe.

3.9 Exercises

1. An electromagnet is designed so that a 5 V potential difference drives 100 A through the coils. The magnet is an effective inductor with an inductance L of 10 mH. Your laboratory is short on space, so you put the DC power supply across the room with the power cables along the wall. You notice that the meter on the power supply has to be set to 6 V in order to get 5 V at the magnet. On the other hand, you are nowhere near the limit of the supply, so it is happy to give you the power you need.

Is there any reason for you to be concerned? Where did that volt go, and what are the implications? If there is something to be concerned about, suggest a solution.

2. You are given a low voltage, high current power supply to use for an experiment. The manual switch on the power supply is broken. (The power supply is kind of old, and it looks like someone accidently hit the switch with a hammer and broke it off.) You replace the switch with something you found around the lab, and it works the first time, but never again. When you take it apart, the contacts seem to be welded together, and you know it wasn't that way when you put it in. What happened? *(Hint: Recall that the voltage drop across an inductor is Ldi/dt, and assume the switch disconnects the circuit over 1 msec or so.)*

3. The following table is from the Tektronix Corp. 1994 catalog selection guide for some of their oscilloscopes:

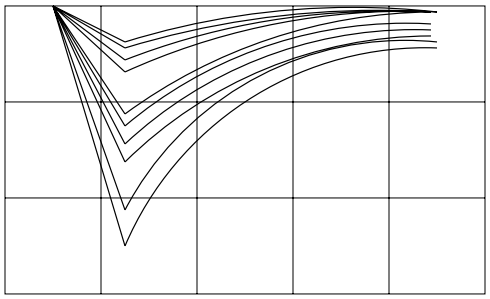
Model	Bandwidth	Sample Rate	Resolution	Time Bases
2232	100 MHz	100 MS/s	8 bits	Dual
2221A	100 MHz	100 MS/s	8 bits	Single
2212	60 MHz	20 MS/s	8 bits	Single
2201	20 MHz	10 MS/s	8-bits	Single

You are looking at the output of a waveform generator on one of these oscilloscopes. The generator is set to give a ± 2 V sine wave output. If the sine wave period is set at 1 μsec, the scope indeed shows a 2 V amplitude.

However, if the the period is 20 nsec, the amplitude is 1 V. Assuming the oscilloscope is not broken, which one are you using?

4. You want to measure the energies of various photons emitted in a nuclear decay. The energies vary from 80 keV to 2.5 MeV, but you want to measure two particular lines that are separated by 1 keV. If you do this by digitizing the output of your energy detector, at least how many bits does your ADC need to have?

5. Pulses emitted randomly by a detector are studied on an oscilloscope:



The vertical sensitivity is 100 mV/div and the sweep rate is 20 ns/div. The bandwidth of the scope is 400 MHz. The start of the sweep precedes the trigger point by 10 ns, and the input impedence is 50Ω.

a. Estimate the pulse risetime. What could you say about the risetime if the bandwidth were 40 MHz?

b. Estimate the trigger level.

c. These pulses are fed into a charge integrating ADC, also with 50Ω input impedence. The integration gate into the ADC is 100 ns long and precedes the pulses by 10 ns. Sketch the spectrum shape digitized by the ADC. Label the horizontal axis, assuming $\frac{1}{4}$ pC of integrated charge corresponds to one channel.

d. The ADC can digitize, be read out by the computer, and reset in 100 μs. Estimate the number of counts in the spectrum after 100 sec if the average pulse rate is 1 kHz. What is the number of counts if the rate is 1 MHz?

6. A detector system measures the photon emission rate of a weak light source. The photons are emitted randomly. The system measures a rate of 10 kHz, but the associated electronics requires 10 μsec to register a photon, and the system will not respond during that time. What is the true rate at which the detector observes photons?

Ch 4

Experiment 1: The Voltage Divider

Now's a good time to make some measurements based on what you've learned so far. We will do some simple things with the voltage divider circuit, including both resistors and capacitors.

Circuits are most easily put together on a "breadboard". This is a flat, multilayered surface with holes in which you stick the leads of wires, resistors, capacitors, and so on. The holes are connected internally across on the component pads, and downward on the power pads. You can play around with a DMM and measure the resistance between different holes to convince yourself of the connections.

Don't forget to write everything down in your log book!

4.1 The Resistor String

Use a DMM to measure the voltage across the terminals of one of the small DC power supplies. Switch the DMM to measure the current out of the terminals. Do you suspect the supply is voltage or current regulated?

Connect two 1 KΩ resistors in series on the breadboard, and then connect the terminals of the power supply to each end of this two-resistor string. Once again, measure the current across the output of the terminals. Also, measure the current through the string. (You will have to change the way you connect the leads of the DMM.)

Now connect two more 1 KΩ resistors in series with the others. Move the connections from the power supply so that once again it is connected to each end of the string. Repeat your voltage and current measurements.

Explain what you have seen so far. Compare the results to Ohm's law. Is the power supply voltage or current regulated? How well? Can you estimate the equivalent internal resistance of the power supply?

Now measure the voltage drop across each of the four resistors. Compare the result to what you expect based on the voltage divider relation. Use your data and Ohm's law to measure the resistance of each of the resistors. (Do you need to remeasure the current through each resistor?) Compare the resistance values you measure with the nominal value.

Remove the DC power supply and replace it with a waveform generator. Set the waveform to a sine wave. Use an oscilloscope to compare the voltage (as a function of time) across the resistor string from the waveform generator with the voltage across one of the resistors. Put each of these into the two channels of the oscilloscope, and trigger the scope on the channel corresponding to the waveform generator output. Look at both traces simultaneously (on either *chop* or *alternate*) and compare the relative amplitudes of the "input" sine wave across the string, and the "output" sine wave across the single resistor.

Discuss what you've measured. You may want to try any number of variations on this theme. For example, put some of the resistors in parallel or series and see what you get. Remember that Ohm's law should always be valid and you can verify that anywhere you want in your circuit. Also remember that the power supply supplies power. If you hook up some other resistor to the circuit, use what you've learned to calculate the power $P = i^2R = V^2/R$ dissipated in that resistor and make sure it does not exceed the resistor's power rating.

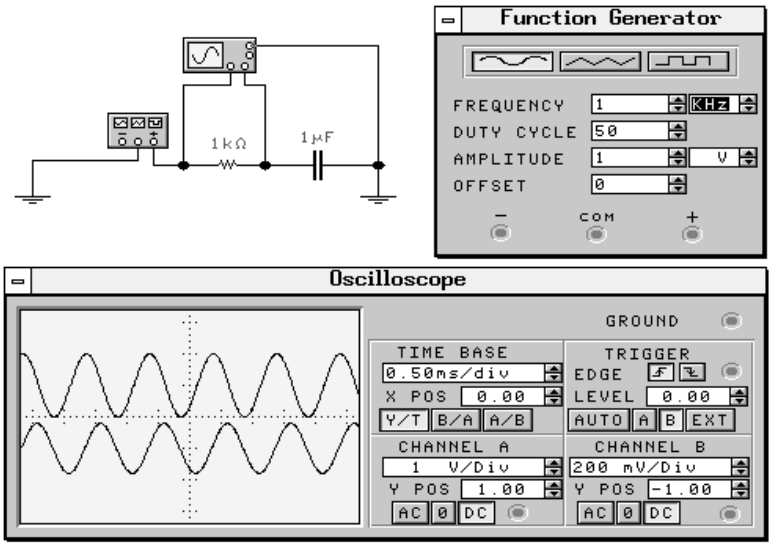


Figure 4.1: Measuring gain and relative phase on an oscilloscope.

4.2 Adding a Capacitor

Now connect a resistor and capacitor in series. Choose a resistance R and capacitance C so that the inverse time constant $1/RC$ is well within the frequency range of the waveform generator and the oscilloscope.

Just as you did for the the resistor string, measure the amplitude of the voltage across either the resistor or capacitor, relative to waveform generator signal applied across the front and back of the pair. (You should take care to set the DC offset of the waveform generator to zero using the oscilloscope to measure the offset relative to ground.) Do this as a function of frequency, spanning well on either side of $1/RC$. Also measure the phase of the output sine wave, relative to the input sine wave. Figure 4.1 shows how to make these measurements on the oscilloscope CRT, using the circuit shown. Refer to Fig. 2.8 for interpreting the input and output waveforms in terms of gain and phase.

It would be a good idea to set your frequency values logarithmically in-

stead of linearly. That is, instead of setting frequencies like

$$\nu_{LO},\ \nu_{LO}+\Delta\nu,\ \nu_{LO}+2\Delta\nu,\ldots,\ \nu_{HI}$$

use something like

$$\nu_{LO},\ f\times\nu_{LO},\ f^2\times\nu_{LO},\ldots,\ \nu_{HI}$$

Make a clear table of your measurements and plot the gain (i.e. the relative amplitudes) and the relative phase as a function of frequency. Think about how you want to scale the axes. (Making both axes linear is the worst choice.)

Don't forget that you measure frequency ν, but most of the relations we've derived are in terms of the angular frequency $\omega = 2\pi\nu$.

Compare your results to the calculated gain and phase difference. Adding this to the plot would be a good idea. Do you expect the same thing whether you were measuring the voltage across the capacitor or the resistor? You can test this by changing the position of the oscilloscope probes in Fig. 4.1.

A sample of data and calculation is plotted in Fig. 4.2. This plot was produced using MATLAB using the following commands:

```
load vdcap.dat
omega=vdcap(:,1);
gain =vdcap(:,2);
phase=vdcap(:,3);
R=1.453E3;
C=0.1E-6;
omegaf=logspace(1.5,7.5);
gainf =1./sqrt(1+(omegaf.*R*C).^2);
phasef=(180/pi)*atan(omegaf.*R*C);
subplot(2,1,1)
loglog(omega,gain,'o',omegaf,gainf)
axis([1E2 1E7 2E-4 2])
xlabel('Angular Frequency (Hz)')
ylabel('Gain')
subplot(2,1,2)
semilogx(omega,phase,'o',omegaf,phasef)
```

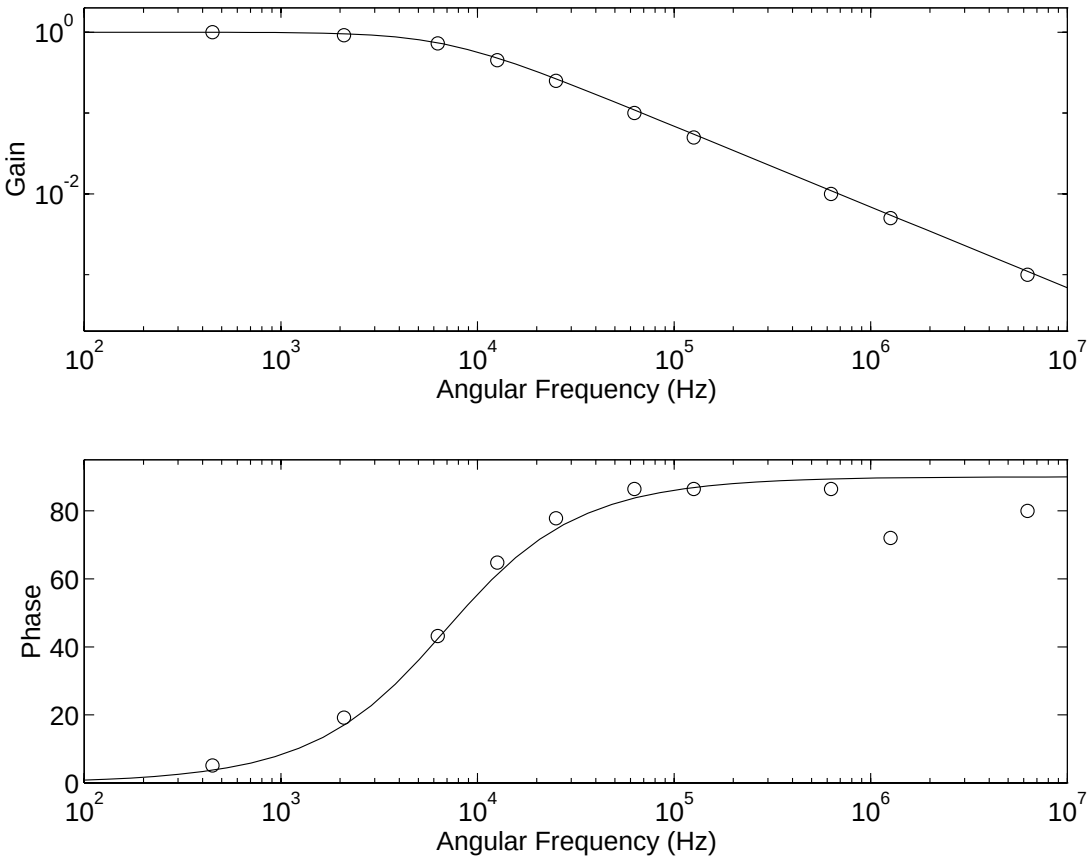


Figure 4.2: Sample of data on gain and phase shift with an RC voltage divider.

```
axis([1E2 1E7 0 95])
xlabel('Angular Frequency (Hz)')
ylabel('Phase')
print -dps vdcap.ps
clear all
```

The angular frequency, gain, and phase were all calculated separately and stored in the ascii file vdcap.dat in three columns. The curves were calculated using the known values of the resistor (1.453 kΩ) and capacitor (0.1 μF). Some more advanced plotting commands were used here, to make log-log and semilog plots, and to put two plots on a single page.

4.3 Response to a Pulse

Use the waveform generator as a pulse generator and study the output using your RC voltage divider circuit. Compare the input and output pulse shapes as a function of the width Δt of the pulse. What happens if $\Delta t \gg RC$? What about $\Delta t \ll RC$?

Ch 5

Experiment 2: The Ramsauer Effect

This is a simple and elegant experiment in quantum mechanical scattering. You will show that when electrons at one particular energy impinge on xenon atoms, they pass right through as if the atom was not there.

The experiment is described in detail in the following references:

- *Demonstration of the Ramsauer-Townsend Effect in a Xenon Thytratron*,
  Stephen G. Kukolich, American Journal of Physics **36**(1968)701

- *An Extension of the Ramsauer-Townsend Effect in a Xenon Thyratron*,
  G. A. Woolsey, American Journal of Physics **39**(1971)558

For more information on the physics associated with quantum mechanical matter wave transmission, see

- *Introduction to the Structure of Matter*,
  John J. Brehm and William J. Mullin, John Wiley and Sons (1989), Chapter Five

- *Introductory Quantum Mechanics*, Richard L. Liboff, Second Edition, Addison Wesley (1992), Section 7.8
- *Quantum Physics*, Robert Eisberg and Robert Resnick, John Wiley and Sons, Second Edition (1985), Chapter Six
- *Does the Spherical Step-Potential Well Exhibit the Ramsauer-Townsend Effect?*, R. C. Greenhow, American Journal of Physics **61**(1993)23

The concepts of mean free path and cross section, and how they pertain to the motion of particles in a gas, are described very well in

- *Physics*, Robert Resnick, David Halliday, and Kenneth Krane, John Wiley and Sons, Fourth Edition (1992), Chapters 23 and 24

You may also want to consult Appendix B.

5.1 Scattering from a Potential Well

The Ramsauer effect (sometimes called the Ramsauer-Townsend Effect) demonstrates the difference between classical mechanics and quantum mechanics, in the simple problem of a particle "scattering" from a potential energy well. We mainly consider the problem in one dimension, but make a few comments about the three dimensional case.

5.1.1 Transmission past a One Dimensional Well

Figure 5.1 summarizes the situation.[1] A particle is incident from the left,

[1]The textbooks by Brehm&Mullin, Liboff, Eisberg&Resnick, and others all treat this or similar cases at appropriate levels of detail. Other cases include the potential "barrier" as opposed to the "well", and the "step" function where the height of the potential energy changes abruptly.

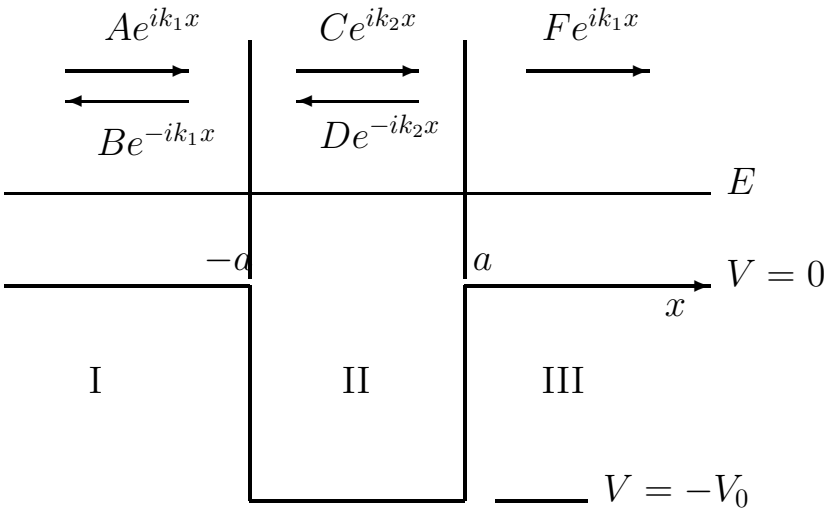


Figure 5.1: A particle incident on a potential energy well.

where the potential energy is zero. Its total mechanical energy (i.e. kinetic plus potential energy) is E, which is constant with time.

Let's first consider what happens classically. Conservation of energy determines the motion through the equation

$$\frac{p^2}{2m} + V(x) = E \tag{5.1}$$

The function $V(x)$ is zero everywhere except for $-a \leq x \leq a$ where it is equal to $-V_0$. The particle is incident from the left and has a momentum $p = +\sqrt{2mE}$. It maintains this momentum until it gets to the well at $x = -a$, where its momentum abruptly changes to $p = +\sqrt{2m(E + V_0)}$. Next it continues to the right hand edge of the well where the momentum changes back to $p = +\sqrt{2mE}$. Finally, the particle continues on its way to the right forever.

The basic idea of quantum physics, however, is that particles can behave as waves with a wavelength $\lambda = h/p$, where Planck's constant $h = 6.626 \times 10^{-34}$ J sec$=4.14 \times 10^{-15}$ eV sec. The motion of the particle is governed by the wave function $\psi(x)$ with the quantity $\psi^*(x)\psi(x)dx$ interpreted as the probability of finding the particle between x and $x + dx$. The wave function

is determined by solving Schrödinger's wave equation

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} + V(x)\psi(x) = E\psi(x) \tag{5.2}$$

where $\hbar \equiv h/2\pi$.

Equation 5.2 is easy to solve. The quantity $E - V(x)$ is positive everywhere, so we can write it as

$$\begin{aligned} \frac{d^2\psi(x)}{dx^2} &= -k^2\psi(x) \qquad (5.3)\\ \text{where} \quad \frac{\hbar^2k^2}{2m} &= E + V_0 \quad \text{for} \quad -a \le x \le a \\ \text{and} \quad \frac{\hbar^2k^2}{2m} &= E \quad \text{elsewhere.} \end{aligned}$$

You've seen this equation lots of times before. The first time was probably when you studied the harmonic oscillator, and learned that the solution is either $\sin(kx)$ or $\cos(kx)$ with the appropriate integration constants. We will use complex numbers (see Appendix C.4) to write the solution instead as

$$\psi(x) = Ae^{+\imath kx} + Be^{-\imath kx} \tag{5.4}$$

Let's stop here for a moment and think about this. Remember that $\psi(x)$ is supposed to represent the wave that is the particle. The wavelength λ of this wave, by Eqn. 5.4 is just $2\pi/k$. Therefore the requirements on k listed in Eq. 5.3 just state that

$$\frac{\hbar^2k^2}{2m} = \left(\frac{h}{\lambda}\right)^2\frac{1}{2m} = \frac{p^2}{2m} = E - V(x)$$

which is just conservation of energy all over again. The Schrödinger equation is just a statement of conservation of energy for a wavy particle.

Now let's go back to the wave function $\psi(x)$ in Eq. 5.4 and see what it implies about the particle's motion. The time dependence of the wave is given by $e^{-\imath\omega t}$, so the term proportional to $e^{+\imath kx}$ represents a wave moving to the right and $e^{-\imath kx}$ is a wave moving to the left. Divide the $x-$axis into

three regions, namely regions I ($x \leq -a$), II ($-a \leq x \leq a$), and III ($x \geq a$). (See Eqn. 5.1.) We have $k = k_1$ in regions I and III, and $k = k_2$ in region II, where k_1 and k_2 are defined in Eq. 5.3. We write $\psi(x)$ for each of the three regions as

$$\begin{aligned} \text{Region I}: \quad \psi_I(x) &= Ae^{+\imath k_1 x} + Be^{-\imath k_1 x} \\ \text{Region II}: \quad \psi_{II}(x) &= Ce^{+\imath k_2 x} + De^{-\imath k_2 x} \\ \text{Region III}: \quad \psi_{III}(x) &= Fe^{+\imath k_1 x} \end{aligned}$$

We do not include a leftward moving wave in region III since we assume there are no more changes in potential past the well so the particle cannot turn around and come back.

There is already a key difference between the classical and quantum mechanical treatments. The solution allows for some portion of the incident wave to be "reflected" from the well. That is, B need not be zero, and in fact generally is not. This is clearly different from the classical case where the particle would always travel on past the well, albeit with greater momentum for the time it is in the well.

Now the wave function and its first derivative must be continuous everywhere. This allows us to determine relations between A, B, C, D, and F by matching $\psi(x)$ and $\psi'(x)$ at $x = \pm a$. These four conditions give us

$$\begin{aligned} Ae^{-\imath k_1 a} + Be^{+\imath k_1 a} &= Ce^{-\imath k_2 a} + De^{+\imath k_2 a} \\ \imath k_1 Ae^{-\imath k_1 a} - \imath k_1 Be^{+\imath k_1 a} &= \imath k_2 Ce^{-\imath k_2 a} - \imath k_2 De^{+\imath k_2 a} \\ Ce^{+\imath k_2 a} + De^{-\imath k_2 a} &= Fe^{+\imath k_1 a} \\ \imath k_2 Ce^{+\imath k_2 a} - \imath k_2 De^{-\imath k_2 a} &= \imath k_1 Fe^{+\imath k_1 a} \end{aligned}$$

These are four equations in five unknowns. A fifth relation would just determine the normalization of the wave function, but we won't bother with this here.

Let's calculate the probability that an incident particle makes it past the well. The amplitude of the incident wave is A and the amplitude of the transmitted wave is F. Therefore, the transmission probability T is given by

$$T = \frac{|F|^2}{|A|^2} = \frac{F^* F}{A^* A} = \left(\frac{F}{A}\right)^* \left(\frac{F}{A}\right)$$

It is pretty easy to solve for F/A using the above relations. Solve the first two for A in terms of C and D by eliminating B. Then solve the last two to get C and D in terms of F. The result is

$$\frac{A}{F} = e^{2\imath k_1 a}\left[\cos(2k_2 a) - \frac{\imath}{2}\frac{k_1^2 + k_2^2}{k_1 k_2}\sin(2k_2 a)\right]$$

which leads to

$$\frac{1}{T} = 1 + \frac{1}{4}\left[\frac{k_1^2 - k_2^2}{k_1 k_2}\right]^2 \sin^2(2k_2 a)$$

where $k_1 = \sqrt{2mE}/\hbar$ and $k_2 = \sqrt{2m(E+V_0)}/\hbar$. We can therefore write

$$\frac{1}{T} = 1 + \frac{1}{4}\frac{V_0^2}{E(E+V_0)}\sin^2(2k_2 a) \tag{5.5}$$

The reflection coefficient $R = |B|^2/|A|^2$ can also be calculated in the same way. Can you think of a simpler way to do this, having already calculated T?

The transmission coefficient T is plotted as a function of E/V_0 in Fig. 5.2, for a $a = 10\hbar/\sqrt{2mV_0}$. The transmission probability is unity *only* at certain values of the incident kinetic energy E. This is wholly different from the classical case where transmission would always occur.

Consider the physical interpretation of the points where T reaches unity. This is when $\sin^2(2k_2 a) = 0$ or $k_2 a = n\pi/2$ where n is any integer. However, $k_2 = 2\pi/\lambda_2$ where λ_2 is the wavelength of the particle while it is in the well. Therefore, the condition for $T = 1$ is $n(\lambda/2) = 2a$. That is, *there is perfect transmission past the well only when an integral number of half-wavelengths fits perfectly inside the well.* (Note that the width of the well is $2a$.)

5.1.2 Three Dimensional Scattering

Of course, the experiment we will do involves scattering in three dimensions and we have only worked things out for the one dimensional case. The generalization to three dimensions is the "spherical" well for which $V(r) = -V_0$ for $r \leq a$, but zero elsewhere. The analysis of this case is somewhat

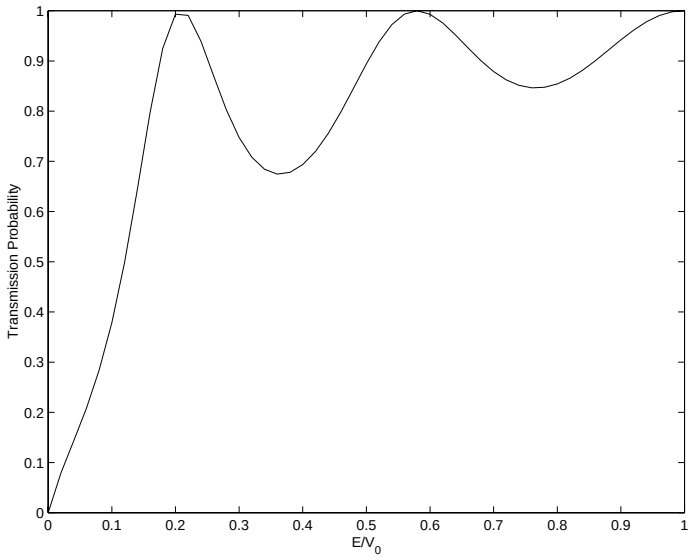


Figure 5.2: Transmission probability for a square well. The barrier width is chosen so that $k_2^2 a^2 = 100(1 + E/V_0)$.

more complicated, and we won't treat it here. Nevertheless, the essential point still remains, namely that the well becomes invisible to the incident particle when $ka = n\pi/2$ where $\hbar^2 k^2/2m = E + V_0$.

When we talk about scattering in three dimensions, the language becomes a bit specialized. In particular, we talk about the scattering "cross section" which measures the probability that an incident particle scatters from some target. In this experiment, you observe the *total* cross section (as opposed to a *differential* cross section) which measures the probability that the particle scatters into any direction at all. For classical scattering of a point particle from a "hard sphere" of radius a, the total cross section is just given by the cross sectional area of the sphere, namely πa^2.

When the well becomes transparent to the incident particle, the total cross section vanishes. Analysis of the three dimensional case shows that when $ka = n\pi/2$, the cross section passes through a *resonance*. That is, the phase of the scattered wave, relative to the incident wave, passes through $90°$.

There is an important difference between the one-dimensional and three-dimensional cases. This is that only the first resonance, i.e. when the condition $ka = \pi/2$ is met, is clearly visible. *Therefore, you expect to see only one dip in the cross section in this measurement.*

The paper by Greenhow provides some interesting, although somewhat advanced, reading. You should review the material on three dimensional scattering in books like Brehm and Mullin or Liboff before getting into it in detail. Greenhow actually analyzed the case of the perfect spherical well and shows that the Ramsauer effect is generated but only in a restricted way. The real potential of the xenon atom, of course, is considerably more complicated than a spherical well, but this nevertheless serves as a convenient and worthwhile approximation.

5.2 Measurements

Your measurements are very similar to those originally performed by Ramsauer, that is, you will be scattering electrons from xenon gas atoms. The procedure we use is based closely on the experiment described by Kukolich. The idea is shown schematically in Fig. 5.3, using a figure borrowed from Kukolich. Electrons are released by a hot filament, and made to accelerate to some energy E by a voltage V, so that $E = eV$ where $e = 1.602 \times 10^{-19}$ C. Electrons which scatter from the xenon atoms in their path move off in some direction and likely hit the "shield", a conductor which transports the electrons back to ground potential. On the other hand, the electrons which make it through without scattering eventually strike the "plate" which also conducts the electrons back to ground. You will determine the behavior of the scattering cross section by measuring the plate current relative to the shield current as a function of V. A large (small) scattering cross section therefore corresponds to a small (large) plate current.

The actual setup is diagrammed in Fig. 5.4. The acceleration and scattering take place in a xenon-filled electron tube called a 2D21 thyratron. You make connections to the various internal components through pins (numbered in Fig. 5.4) on the tube. For your convenience, the tube plugs into a socket wired to a labeled panel with banana plug connectors. Electrons

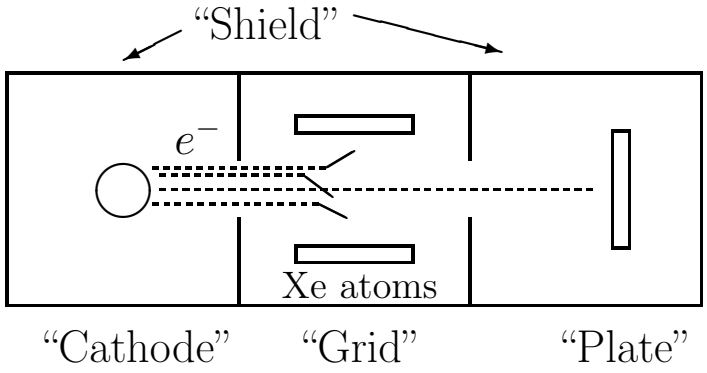


Figure 5.3: Schematic diagram of Ramsauer Effect apparatus. Electrons are accelerated towards the plate, where they are collected if they do not scatter from Xenon atoms. Otherwise, they are collected by the shield or grid.

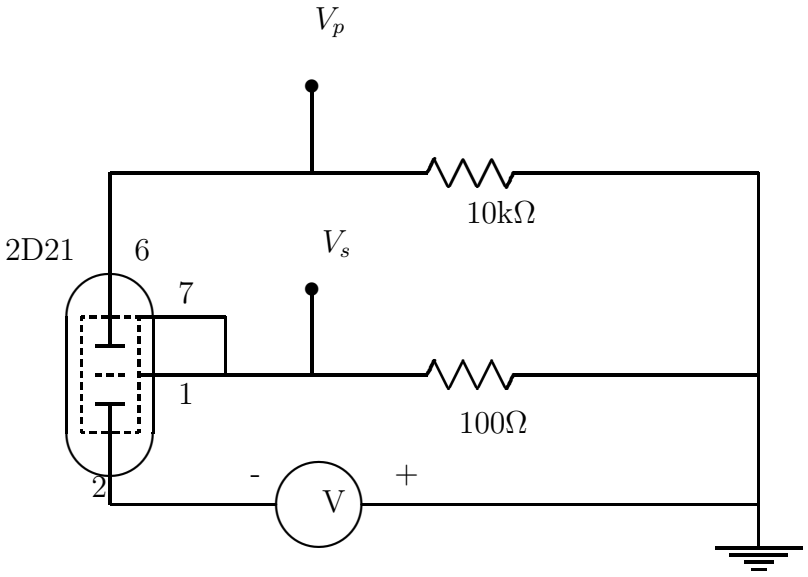


Figure 5.4: Setup used to measure the Ramsauer Effect.

that are captured by the shield or the plate are returned to ground through the resistors on the respective circuit, and you determine the shield or plate currents from the voltage drop across these resistors. These resistors are in a breadboard, and you should consider different values for them and test that the currents you deduce are the same. Since the plate current is typically much less than shield current, you generally want the plate resistor to be much larger than the shield resistor so that their voltage drops are comparable. Suggested starting values are 10 kΩ and 100 Ω for the plate and shield resistors respectively.

5.2.1 Procedure

The data taking procedure is straightforward. First, you need to heat the cathode filament in the thyratron so that it emits electrons. This is done using a standard laboratory DC voltage supply and a high current voltage divider to send a specific current through the filament. The filament is connected to pins 3 and 4 of the thyratron, and you get the right current with a voltage of about 4 V. Adjust the voltage divider and power supply so that you get 4 V before connecting to the pins. Too much voltage can damage the filament and the tube becomes useless. You should monitor this voltage throughout the data taking procedure to make sure it does not change.

Measure the voltages at the plate (V_p) and at the shield (V_s) as a function of the applied voltage V. Adjust V through the voltage divider connected to another DC voltage supply. You should vary V in relatively small steps between 0 and around 5 V. You should find that the plate current i_p passes through a maximum of 0.15 μA or so for $V \sim 1$ V. This is the Ramsauer Effect. The plate current is a maximum because the scattering cross section has gotten very small allowing a large number of electrons to pass through the xenon gas and strike the plate. Sample data, taken from Kukolich, is shown by the open points and solid line in Fig. 5.5.

In order to get quantitative results, some more work needs to be done. First, you must realize that the thyratron is a pretty weird electron accelerator. As you change the value of V, the electric field lines inside change and the probability that electrons get to the plate will certainly change, regard-

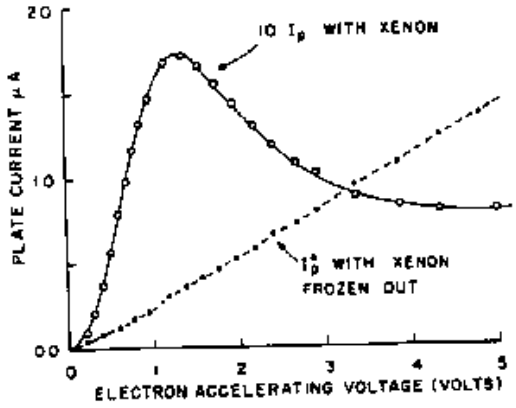


Figure 5.5: Sample of raw data, taken from Kukolich.

less of whether or not there is gas inside. In fact, how do you know for sure that the maximum in the plate current corresponds at all $e^- - Xe$ scattering? Symbolically, the plate and shield currents are related as

$$i_p(V) = i_s(V) f(V) \left[1 - P_{SCAT}(E)\right]$$

where P_{SCAT} is a function of the electron energy and should become small at the Ramsauer Effect resonance, and $f(V)$ is a geometrical factor depending on the accelerating voltage and the details of the thyratron. The problem is that you do not know beforehand how to separate the effects of $f(V)$ and $P_{SCAT}(E)$.

However, you can easily separate these effects using your apparatus. *After turning off the filament voltage and letting the filament cool down*, dunk the top of the tube in liquid nitrogen. This freezes out the xenon and reduces the bulb pressure to a negligible level. Repeat the measurements above and since $P_{SCAT} = 0$, you determine $f(V)$ from

$$f(V) = \frac{i_p^*(V)}{i_s^*(V)}$$

where the i^* indicate measurements taken with the xenon removed. Figure 5.5 also plots i_p^* as a function of V.

5.2.2 Analysis

When analyzing your data, realize that the electrons are accelerated by the potential difference between the the negative terminal of the power supply (V) and the the shield (V_s) and that V_s changes with V. Their energy is therefore given by

$$E = e\,(V - V_s + \text{corrections})$$

where there are still some additional corrections. (These are studied in more detail in Sec. 5.3.) These corrections amount to about 0.4 V which should be added in before calculating E.

Plot P_{SCAT} as a function of the incident electron momentum $p = \sqrt{2mE}$ where m is the electron mass. To compare with the figures in Kukolich, realize that they ignore any extraneous factors and compute "momentum" simply as $\sqrt{V - V_s}$, also ignoring any other corrections.

Different experiments show that the radius of the xenon atom is around 4 Å. Calculate the well depth of the xenon atom potential, assuming that it is approximated by a spherical well with this radius. Does this sound reasonable to you?

Because the electrons scatter, the electron beam intensity diminishes exponentially as a function of the distance traveled, that is $I(x) = I_0 e^{-x/L_{SCAT}}$ where L_{SCAT} is the "mean free path" through the gas in the tube. Insomuch as the plate current measures the beam intensity at the plate, the scattering probability P_{SCAT} is related to the mean free path by

$$e^{-L/L_{SCAT}} = 1 - P_{SCAT}$$

where L is the distance through the tube to the plate. For the 2D21 thyratron, $L = 0.7$ cm.

This information can be used to estimate the scattering cross section σ since it is related to the mean free path by $L_{SCAT} = 1/\rho_n\sigma$ where ρ_n is the number of xenon atoms per unit volume. Determine ρ_n from the ideal gas law[2] using the quoted pressure of 0.05 Torr for the 2D21 at room

[2]The ideal gas law says that $\mathcal{P}v = NkT$ where N is the total number of atoms in the volume v, hence $\rho_n = N/v$. See for example, Resnick, Halliday, and Krane.

temperature. Compare the calculated cross section on and off resonance with the "geometric" cross section πa^2 where a is the radius of the xenon atom.

To summarize, you can calculate the following quantities from your data:

- The approximate well depth V_0 of the xenon atom.
- The scattering probability, which can be compared to the literature.
- The scattering cross section, on and off resonance.

5.3 Advanced Topics

As discussed by Kukolich, there is a discrepancy between the observed value of V where the minimum cross section occurs, and that found in the literature. He attributes this to a 0.4 V contact potential, but Woolsey shows that this is in fact both from the contact potential and from the thermal energy of the electrons when they emerge from the filament. You can show this in the same way as Woolsey. You will use the same apparatus as for the "standard" measurements above, but with some simple rearrangements.

The filament of the 2D21 is made of barium oxide and the shield is made of nickel. Since the nickel has the higher work function of the two, there is a contact potential difference that causes electrons to spontaneously flow from the filament to the shield, even if $V - V_s$ is zero. Therefore, the actual energy of the electrons is somewhat higher than you would expect from $V - V_s$ alone. Call that contact potential difference V_c.

There is another reason that the electrons are higher energy than you would first expect. The filament is hot, so the electrons have some thermal energy when they are emitted. As dictated by statistical mechanics, this thermal energy is not one single value but instead is distributed over a range of energies. The appropriate distribution function is the Maxwell-Boltzmann distribution which says that the number of electrons with energy E_{TH} is proportional to $e^{-E_{TH}/kT}$, where T is the temperature of the filament. The average energy of the electrons is $\overline{E_{TH}} = 3kT/2$. (See for example, Resnick, Halliday, and Krane.)

So, the incident energy of the electrons is given by

$$E = e(V - V_s + V_c + \bar{V}) \tag{5.6}$$

where $e\bar{V} = \overline{E_{TH}}$ represents the average effect of the thermal electron distribution.

Now the issue is, how do we measure V_c and $\bar{V}$? The key is to realize that, when the xenon in the thyratron tube is frozen out, the plate current will behave like (see Woolsey)

$$i_s^* = i_0 e^{-3V_{RET}/2\bar{V}} \tag{5.7}$$

where V_{RET} is a "retarding" voltage between the shield and the cathode. That is, as V_{RET} increases, it makes it harder for electrons to get to the shield. The fact that i_s^* is a finite value (equal to i_0) when there is no potential difference between the shield and filament ($V_{RET} = 0$) just indicates that electrons still flow to the shield due to their thermal energy. As the retarding voltage is increased, the shield current goes down exponentially. This continues *until* the retarding voltage equals the contact voltage, after which the current decreases even more rapidly due to space charge saturation at the cathode. See Woolsey for more details.

The procedure is therefore straightforward. With the top of the thyratron dunked in liquid nitrogen as before, reverse the polarity of V by switching around the connections. As you increase V from zero, record the shield voltage V_s. (You may need to find a more precise voltmeter than the standard DMM's used in the lab.) If you plot $i_s^* = V_s/R_s$ versus $V + V_s$ on semilog paper, then the slope of the line gives you $\bar{V}$ according to Eq. 5.7. At some value of V, the data will abruptly change and i_s^* will fall more rapidly. At this value of V, you determine $V_c = V + V_s$. This is shown in Fig. 5.6 which is taken from Woolsey's paper.

Take several measurements of this type. Try changing the cathode filament voltage by a volt or so around the standard value of 4 V. This will change the temperature of the filament, so it should change the slope accordingly. The contact potential, on the other hand, should be unaffected. Use measurements of this type to determine V_c and $\bar{V}$, and to estimate their uncertainties. Use your results and Eq. 5.6 to reanalyze the Ramsauer effect. How does this affect your determination of the well depth? What about the cross section determination?

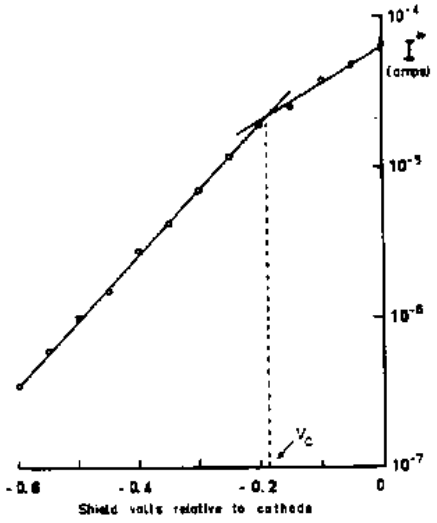


Figure 5.6: Sample of data with reversed polarity, taken from Woolsey.

Ch 6

Experimental Uncertainties

Before we go on to do more experiments, we need to learn one of the most important things there is about making measurements.

Every measurement yields some number. Of equal (and sometimes greater) importance is the *uncertainty* with which we know how close that number approximates the "right" answer. In this chapter, we will learn the basic facts about estimating and reporting experimental uncertainties.

Sometimes people refer to experimental "error" when they mean experimental uncertainty. This is unfortunate, since "error" implies that a mistake was made somewhere, and that is not what we are talking about here. This terminology is pretty well ingrained into the jargon of experiments, though, so you might as well get used to it.

When an experimenter quotes the result of a measurement, the uncertainty in that result should also be quoted. As you will see, the measurement result will give a sort of "central value" of some quantity, call it Q, and the uncertainty gives some idea of how far on either side of Q you have to go to hit that true value. We write the uncertainty in Q as δQ, and quote the result of the measurement as

$$Q \pm \delta Q$$

You should always get used to writing down your results this way.

We will discuss some of the basics of uncertainties and statistical analysis in this course. In particular, the concepts you need to carry out the experiments will be outlined, and they are covered rather well in

- *Practical Physics*, G. L. Squires, Third Edition
  Cambridge University Press (1991)

However, it is a good idea to have a more thorough reference on this stuff. There are a lot of books out there, but I recommend

- *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*,
  John R. Taylor, University Science Books (1982)

We will also discuss using MATLAB for some of the numberical manipulations commonly used for determining uncertainty. Refer to Sec. 1.4.3 for the basics on MATLAB, including the main references

- *The Student Edition of* MATLAB, Prentice Hall (1994)
- *Numerical Methods for Physics*,
  Alejandro Garcia, Prentice-Hall (1994)

6.1 Systematic and Random Uncertainties

There are two kinds of experimental uncertainty, namely *Systematic* and *Random* Uncertainty. Sometimes it can be hard to tell the difference because their meanings are not always precisely defined. I will give you some convenient ways to think about them, but as with all things in Experimental Physics, your intuition will get better with experience.

Systematic uncertainty comes from not knowing everything there is to know about your experiment. If you could precisely duplicate the conditions every time you make a measurement, then your systematic uncertainty would

be zero. However, it is impossible to precisely duplicate things. The room temperature will be different, the positions of other people in the room or the building will have changed, and the phase of the moon is not the same, to name just a few. Another possibility is that your measuring instrument is only accurate to some level, and this may be the most important systematic uncertainty. All of these things can affect your measurement at some level, and one of your jobs is to try and estimate how big the effect can be.

Some guidelines are in order for estimating systmatic uncertainty. In very many cases, one thing in particular may dominate the systematic uncertainty. Try to find out what that thing is, and estimate how much it may have changed your result. That would be an estimate of your systematic uncertainty. You can go further, perhaps, and figure out how much it actually might have changed things. It would cause the central value to shift, and then you would apply a "correction" to your result. How well can you make that correction? Answer that question, and you can get another estimate of your systematic uncertainty. Of course, if you want to make your experiment more and more precise, the approach is to identify the sources of systematic uncertainty and reduce their effect somehow.

Random uncertainties are different. At their most fundamental level, they come from the chance fluctuations of nature, although in many cases, the system is so complicated that you will observe fluctuations that might as well be random. The point is, you cannot account for random uncertainty, other than to calculate how big it is. The key to random uncertainties is that if you make many measurements of the same quantity, then the random fluctuations will average to zero over many trials. Obviously, then, the way to reduce random uncertainty is to make lots of measurements. Because of their random nature, this source of experimental uncertainty can be estimated quite precisely. More on that soon.

Let's try a simple example. Suppose you want to measure the resistance of a 500 foot roll of 32 gage aluminum wire. You just hook up your DMM to the ends of the wire on the spool, and measure the resistance. There is some uncertainty associated with how long the wire actually is, so you measure many spools of wire to get an idea of how big the random fluctuations are. However, there is also an uncertainty associated with the precision of the DMM. No matter how many measurements you make, that systematic

uncertainty will always be present.

Let's get more precise about these things.

6.2 Determining the Uncertainty

Remember that by their nature, systematic and random uncertainties are treated differently. In particular, you can only *estimate* the systematic uncertainty. We'll discuss some ways to do that, but as with just about everything in Experimental Physics, practice makes perfect. On the other hand, you can deal with random uncertainties in well defined ways, and we'll go through those.

6.2.1 Systematic Uncertainty

Try looking for systematic uncertainties in two places. First, consider the accuracy of your measuring instruments. This includes meters, clocks, rulers, digitizers, oscilloscopes, and so on. How precisely can you read the device in the first place? If a ruler is graduated in 1 mm increments, for example, you can't measure the length of something much better than that. Does you clock tick off in seconds? If so, it is hard to argue that you could measure the time it takes something to happen any more precisely. Also keep in mind the manufacturer's specifications. How accurately does your oscilloscope measure voltage? How well do they guarantee the conversion of charge to digits in a charge integrating ADC?

The second thing to keep in mind is the effect external factors have on your measurement. For example, suppose you are trying to precisely measure the length of something with a carefully graduated metal ruler, but the room temperature is fluctuating in a $\pm 5^{\circ}C$ range. The length of the ruler is given by $L = L_0 + \alpha(T - T_0)$, where α is the metal's thermal expansion coefficient. Therefore, the actual length L of your sample will only be known to a precision of $\alpha \cdot (\pm 2.5^{\circ}C)$ due to this systematic uncertainty. There are an infinite number of examples of this sort of thing.

Don't make the mistake of assuming you will figure all these things out when you are analyzing your experiment! Record anything you suspect might be important. Try to find out what you can about your instruments as well.

6.2.2 Random Uncertainty

The idea of random uncertainty is that the uncertainty will average away with a large number of trials. Consequently, you would expect the average value of a number of measurements to closely approximate the true value, at least within the limit of any systematic uncertainty. This in fact is the case, and we will talk more about it when we discuss statistical analysis in a later chapter. However, if the average approximates the true value, how do we calculate the magnitude of the random uncertainty? Let's make some definitions, and then I will tell you what to interpret as the random uncertainty.

Suppose you make n measurements of a quantity x, and the result is the list of numbers x_1, x_2,... x_n. We define the *mean* $\bar{x}$, also written as $\langle x \rangle$, of the measurements to be

$$\bar{x} = \langle x \rangle = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \text{Mean Value} \tag{6.1}$$

That is, $\bar{x}$ is just the average value of x from the measurements. The *variance* σ^2 of the measurements is defined to be

$$\sigma_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \text{Variance} \tag{6.2}$$

and obviously has something to do with how far the values fluctuate about the mean value. (Don't worry about the $n-1$ in the dominator instead of just n. We'll discuss this later as well.) The quantity σ_x (the square root of the variance) is called the *standard deviation*. You can show that the variance can also be written as

$$\sigma_x^2 = \frac{n}{n-1} \left(\bar{x^2} - \bar{x}^2 \right) \tag{6.3}$$

This form is particularly useful for programming computers, since you can calculate both $\bar{x^2}$ and $\bar{x}$ within the same loop.

Now as we discussed above, we interpret the mean $\bar{x}$ as our best approximation to the "true value" of x. Furthermore, we interpret the standard deviation σ_x as the uncertainty in *each* measurement x_i. On the other hand, as we will show in Sec. 6.3.1, the uncertainty in the *mean* value of the x_i, as it approximates the true value of x, is given by

$$\sigma_{\bar{x}} = \sigma_x/\sqrt{n} \tag{6.4}$$

So, when you report the result of a series of measurements of x, you write

$$\bar{x} \pm \sigma_{\bar{x}}$$

That is, the random uncertainty in the measured value is $\sigma_{\bar{x}}$.

Don't forget that these formulas apply only to *random* uncertainties, and do not apply to systematic uncertainties. You can always minimize the random uncertainty by taking lots of measurements and averaging them together. However, if systematic uncertainties dominate, then the total uncertainty in the measurement will be *bigger* than that given by (6.4).

6.2.3 Using MATLAB

MATLAB can be very useful for your data analysis needs. Given a list of numbers read into a vector array `x` (see Sec. 1.4.3), you can easily determine, for example, an array `xsq` corresponding to the squares of these elements:

```
xsq=x.^2;
```

The "." before the exponentiation symbol indicates that the operation is to be performed element-by-element, as opposed to calculating the square of a matrix. This notation is used for all element-by-element operations.

The program also has simple functions available which directly calculate many of the quantities needed here. For example,

```
n=length(x);
```

```
xsum=sum(x);
xbar=mean(x);
sigx=std(x);
```

return the number of elements in the array x, the sum of the values, the mean of the values, and the standard deviation of the values. Various other functions return the maximum (max) value, minimum (min) value, median (median) value, and the product of the elements (prod). In MATLAB language, for example, the standard deviation can also be calculated from the sequence of commands

```
n=length(x);
xbar=mean(x);
xsig=sqrt(sum((x-xbar).^2)/(n-1));
```

This should return precisely the same value you would get using the std function.

Of course, this is just the tip of the iceberg. We will point out the most relevant functions as we go along, but don't forget there are lots more that we won't mention. Consult the MATLAB User's Guide for more information.

6.3 Propagation of Errors

If you measure some value x with an uncertainty δx, but you are interested in some quantity q which is a function of x, i.e. $q = q(x)$, then what is the corresponding uncertainty δq? For example, suppose the gain g of an amplifier depends on voltage V as $g = AV^n$. If the voltage is known to within δV, how well do we know g?

Suppose things are more complicated and q is a function of two independently measured quantities x and y, $q = q(x, y)$. An example might be determining the temperature T from a gas bulb thermometer with volume v and pressure P, through the ideal gas law $T = Pv/NR$. How do you deter-

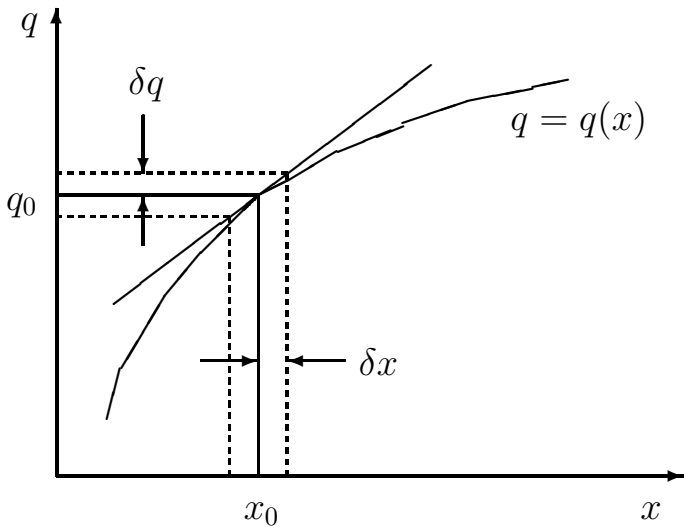


Figure 6.1: Propagation of errors for a single independent variable.

mine the uncertainty in T from the uncertainty in P and v (or, in general, δq from δx and δy)?

All this is accomplished through "propagation of errors". This phrase is so ingrained in the scientific community, that I won't bother substituting "uncertainty" for "error". In any case, the prescription is straightforward.

Let's consider the single variable case first. Figure 6.1 schematically plots the quantity $q = q(x)$ as a function of x. Say the best value for x is x_0. Then, the best value for q is $q_0 = q(x_0)$. As shown in the figure, the uncertainty in x, δx, is related to the uncertainty in q just by the slope of the curve at $x = x_0$. That is,

$$\delta q = \left| \frac{dq}{dx} \right|_{x_0} \times \delta x \tag{6.5}$$

gives the uncertainty in q. The absolute value insures that the result is a positive number.

Now let q be a function of several variables, i.e. $q = q(x, y, \ldots)$. The best value for q is $q_0 = q(x_0, y_0, \ldots)$, and there will be contributions to the

uncertainty δq from each variable, following from Eqn 6.5:

$$\delta q_x = \left|\frac{\partial q}{\partial x}\right|_{x_0} \times \delta x, \qquad \delta q_y = \left|\frac{\partial q}{\partial y}\right|_{y_0} \times \delta y, \qquad \ldots$$

The big question, though, is how to combine the δq_i to get δq? Do we simply add them together, i.e. $\delta q = \delta q_x + \delta q_y + \cdots$? This might seem unfairly large, since if x fluctuates all the way to its maximum uncertainty so that $x = x_0 + \delta x$, then it is unlikely that y would fluctuate that much as well, and so on. In fact, you might think that if x and y are correlated, then an upward fluctuation in x might imply there is a good chance that y fluctuates downward. In this case, you are temped to use something like $\delta q = |\delta q_x - \delta q_y|$.

In general, there is no clear answer to this question. It depends on the specific nature of the uncertainties, whether they are random or systematic, and whether or not they are correlated with each other. There is, however, *one specific case* where there is a straightforward answer. This is the case where all uncertainties are *random* and *uncorrelated*, and the answer is

$$\begin{aligned} \delta q &= \left[(\delta q_x)^2 + (\delta q_y)^2 + \cdots\right]^{\frac{1}{2}} \\ &= \left[\left(\left.\frac{\partial q}{\partial x}\right|_{x_0} \delta x\right)^2 + \left(\left.\frac{\partial q}{\partial y}\right|_{y_0} \delta y\right)^2 + \cdots\right]^{\frac{1}{2}} \end{aligned} \tag{6.6}$$

In this case, we say that the uncertainties are "added in quadrature".

Even though Eqn. 6.6 only applies to random, uncorrelated uncertainties, it is often used (incorrectly!) in other circumstances. Probably the most dangerous incorrect use is for random uncertainties which are not completely uncorrelated. You should at least convince yourself that the variables x, y, and so forth are independent to at least a good approximation. There is a method which can take into account correlations of random uncertainties, and we will discuss it in a later chapter.

Adding errors in quadrature is almost always incorrect for systematic uncertainties, and you should do the best you can to estimate their net effect. One practice is to quote the random and systematic errors separately, i.e.

$$q = q_0 \pm \delta q|_{RANDOM} \pm \delta q|_{SYSTEMATIC}$$

so you can at least let the reader know their relative contributions.

You should always keep in mind the relative sizes of the terms in Eqn. 6.6. If any of the $(\partial q/\partial x_i)^2 \delta^2 x_i$ are significantly bigger than the rest, then it will dominate the net uncertainty, especially since you add the squares.[1] In this case, you may be able to think of that variable as the only important one, as far as the uncertainty is concerned. Many experiments to measure some quantity more precisely than it has been done before, are based on ideas that can reduce the dominant uncertainty.

6.3.1 Examples: Fractional Uncertainty

We will work out some general formulas for propagating uncertainties. In the cases for more than one variable, we assume that errors add in quadrature.

Power Law of One Variable

Consider the earlier example of gain as a function of voltage, i.e. $g = AV^n$ where we know the voltage V to within $\pm\delta V$. Using Eqn. 6.5 we have

$$\delta g = nAV^{n-1}\delta V$$

Notice however that there is a simpler way to write this, namely

$$\frac{\delta g}{g} = n\frac{\delta V}{V} \tag{6.7}$$

That is, the *fractional uncertainty* in g is just n times the fractional uncertainty in V. This is true for any power law relation $q = \alpha x^\beta$ where α and β are arbitrary constants, that is

$$q = \alpha x^\beta \quad \Rightarrow \quad \frac{\delta q}{q} = \beta\frac{\delta x}{x}$$

[1] Don't be swayed by the notation $\delta^2 x$. It is just a simple and common shorthand for $(\delta x)^2$.

Sum of Two Variables

Consider the general case $q = Ax+By$ where A and B are arbitrary constants. Equation 6.6 tells us that

$$\delta q = \sqrt{A^2\delta^2 x + B^2\delta^2 y} \tag{6.8}$$

In this case, there is no simple form for the fractional error in q.

General Power Law Product

Now look at the general case $q = Ax^m y^n \cdots$. Again using Eqn. 6.6 we have

$$\delta q = \left[\left(mAx^{m-1}y^n\cdots\right)^2 \delta^2 x + \left(nAx^m y^{n-1}\cdots\right)^2 \delta^2 y + \cdots\right]^{\frac{1}{2}}$$

but it is obviously simpler to write

$$\frac{\delta q}{q} = \left[\left(m\frac{\delta x}{x}\right)^2 + \left(n\frac{\delta y}{y}\right)^2 + \cdots\right]^{\frac{1}{2}} \tag{6.9}$$

Knowing the fractional uncertainties in x, y, and so on makes it simple to see if any of them dominate the result.

Two simple but useful cases of Eqn. 6.9 are $q = xy$ and $q = x/y$. In both cases, the fractional uncertainty in q is the sum in quadrature of the fractional uncertainties in x and y.

The Uncertainty in the Mean

Back in Sec. 6.2.2 we just quoted the result for the uncertainty in the mean. We can now derive it using propagation of errors. Start with the definition of the mean value (Eq. 6.1):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

Here n is a constant and we determine the uncertainty in the mean simply by applying Eq. 6.8:

$$\delta\bar{x} = \left[\frac{1}{n^2}\delta^2 x_1 + \frac{1}{n^2}\delta^2 x_2 + \cdots + \frac{1}{n^2}\delta^2 x_n\right]^{\frac{1}{2}}$$

Now the supposition in Sec. 6.2.2 was that the x_i are all separate measurements of the same quantity x, and that the uncertainty in x is given by the standard deviation σ. Therefore, all the terms in this equation are the same, and we have

$$\delta\bar{x} = \left[n\frac{1}{n^2}\sigma^2\right]^{\frac{1}{2}} = \frac{\sigma}{\sqrt{n}} = \sigma_{\bar{x}}$$

which proves Eq. 6.4.

6.3.2 Dominant Uncertainty

If two or more quantities are measured to determine the value of some derived result, their individual uncertainties all contribute to the uncertainty in the final value. If one of the uncertainty in one of those quantities makes the largest contribution to the final uncertainty, we refer to it as the "dominant uncertainty". It is smart to identify the dominant source or sources of uncertainty in an experiment. That's the one you want to learn how to measure better. Doing a better job on the others might be nice, but it won't buy you a significantly more precise result in the end.

The relative precision of each of the quantities is not all that matters. You also need to know how that quantity contributes in the end. Equation 6.9 makes this point particularly clear. If one of the quantities enters with some large exponent, then that exponent amplifies the contribution of its uncertainty. Even though $\delta x/x$ may be smaller than $\delta y/y$, x may dominate the uncertainty in the end if m is much larger than n.

6.4 Exercises

1. You measure the following voltages across some resistor with a three-digit DMM. As far as you know, nothing is changing so all the measurements are supposed to be of the same quantity V_R.

2.31	2.35	2.26	2.22	2.30
2.27	2.29	2.33	2.25	2.29

a. Determine the best value of V_R from the mean of the measurements.

b. What systematic uncertainty would you assign to the measurements?

c. Assuming the fluctuations are random, determine the random uncertainty from the standard deviation.

d. Somebody comes along and tells you that the true value of V_R is 2.23. What can you conclude?

2. (From Squires.) In the following examples, q is a given function of the independent measured quantities x and y. Calculate the value of q and its uncertainty δq, assuming the uncertainties are all independent and random, from the given values and uncertainties for x and y.

a. $q = x^2$ for $x = 25 \pm 1$

b. $q = x - 2y$ for $x = 100 \pm 3$ and $y = 45 \pm 2$

c. $q = x \ln y$ for $x = 10.00 \pm 0.06$ and $y = 100 \pm 2$.

d. $q = 1 - \frac{1}{x}$ for $x = 50 \pm 2$.

3. Police use radar guns to catch speeders. The guns measure the frequency f of radio waves reflected off of cars moving with speed c. This differs from the emitted frequency f_0 because of the Doppler effect:

$$f = f_0 \left(1 - \frac{v}{c}\right)$$

for a car moving away at speed v. What fractional uncertainty must the radar guns achieve to measure a car's speed to 1 mph?

4. The period T of a pendulum is related to its length L by the relation

$$T = 2\pi\sqrt{\frac{L}{g}}$$

where g is the acceleration due to gravity. Suppose you are measuring g from the period and length of a particular pendulum. You have measured the length of the pendulum to be 1.1325 ± 0.0014 m. You independently measure the period to within an uncertainty of 0.06%, that is $\delta T/T = 6 \times 10^{-4}$. What is the fractional uncertainty (i.e. % uncertainty) in g, assuming that the uncertainties in L and T are independent and random?

5. You have a rod of some metal and you are changing its temperature T. A sensitive gauge measures the deviation of the rod from its nominal length $l = 1.500000$ m. Assuming the rod expands linearly with temperature, you want to determine the coefficient of linear expansion α, i.e. the change in length per degree K, and the actual length l_0 before any temperature change is applied. The measurements of the length deviation Δl as a function of the temperature change ΔT are as follows:

ΔT (K)	Δl (μm)	ΔT (K)	Δl (μm)	ΔT (K)	Δl (μm)
0.8	70	2.2	110	3.6	130
1.0	110	2.6	150	3.8	170
1.2	130	2.8	120	4.2	160
1.6	100	3.0	130	4.4	190
1.8	130	3.4	160	5.0	160

Plot the points and draw *three* straight lines through them:

- The line that best seems to go through the points.
- The line with the largest reasonable slope.
- The line with the smallest possible slope.

Use your own estimates by eye to determine these lines. (Don't use a fitting program.) Use the slopes and the intercepts of these lines to determine $\alpha \pm \delta\alpha$ and $l_0 \pm \delta l_0$.

6. Suppose you wish to measure the gravitational acceleration g by using something like the "Galileo" experiment. That is, you drop an object from some height h and you know that the distance it falls in a time t is given by $\frac{1}{2}gt^2$. For a given experimental run, the fractional uncertainty in h is $\delta h/h = 4\%$ and the fractional uncertainty in t is $\delta t/t = 1.5\%$. Find the fractional uncertainty in g from this data, assuming the uncertainties are random and uncorrelated.

7. You want to measure the value of an inductor L. First, you measure the voltage V across a resistor R when 1.21 ± 0.04 mA flows through it and find $V = 2.53 \pm 0.08$ V. Then, you measure the decay time τ in an RC dircuit with this resistor and a capacitor C and get $\tau = RC = 0.463 \pm 0.006$ msec. Finally, you hook the capacitor up to the inductor and measure the oscillator frequency $\omega = 1/\sqrt{LC} = 136 \pm 9$ kHz. What is the value of L and its uncertainty?

8. A simple pendulum is used to measure the gravitational acceleration g. The period T of the pendulum is given by

$$T = 2\pi\sqrt{\frac{L}{g}}\left(1 + \frac{1}{4}\sin^2\frac{\theta_0}{2}\right)$$

for a pendulum initially released from rest at an angle θ_0. (Note that $T \to 2\pi\sqrt{L/g}$ as $\theta_0 \to 0$.) The pendulum length is $L = 87.2 \pm 0.6$ cm. The period is determined by measuring the total time for 100 (round trip) swings.

a. A total time of 192 sec is measured, but the clock cannot be read to better than ± 100 ms. What is the period and its uncertainty?

b. Neglecting the effect of a finite value of θ_0, determine g and its uncertainty from this data. Assume uncorrelated, random uncertainties.

c. You are told that the pendulum is released from an angle less than $10°$. What is the systematic uncertainty in g from this information?

d. Which entity (the timing clock, the length measurement, or the unknown release angle) limits the precision of the measurement?

9. The β-decay asymmetry, A, of the neutron has been measured by Bopp, *et.al.*, Phys.Rev.Lett. **56**(1986)919 who find

$$A = \frac{2\lambda(1-\lambda)}{1+3\lambda^2} = -0.1146 \pm 0.0019$$

This value is perfectly consistent with, but more precise than, earlier results. The neutron lifetime, τ, has also been measured by several groups, and the results are not entirely consistent with each other. The lifetime is given by

$$\tau = \frac{5163.7\ sec}{1+3\lambda^2}$$

and has been measured to be

918 ± 14 *sec* by Christenson, *et.al.*, Phys.Rev.D**5**(1972)1628,

881 ± 8 *sec* by Bondarenko, *et.al.*, JETP Lett. **28**(1978)303,

937 ± 18 *sec* by Byrne, *et.al.*, Phys.Lett. **92B**(1980)274, and

887.6 ± 3.0 *sec* by Mampe, *et.al.*, Phys.Rev.Lett. **63**(1989)593.

Which, if any, of the measurements of τ are **consistent** with the result for A? Which, if any, of the measurements of τ are **inconsistent** with the result for A? Explain your answers. A plot may help.

Ch 7

Experiment 3: Gravitational Acceleration

This is a conceptually simple experiment. We will measure the value of g, the acceleration due to gravity, from the period of a pendulum. The main point is to determine g and understand the uncertainty. If you measure it precisely enough, you can see the effect of the Earth's shape. You can also convince yourself that Einstein's "Principle of Equivalence" is valid.

The physics and technique are straightforward, and can be found in just about any introductory physics textbook. Most of the interesting stuff is neatly collected in

- *Handbook of Physics*, E. U. Condon and Hugh Odishaw, McGraw Hill Book Company, Part II, Chapter 7, pg.57-59

7.1 Gravity and the Pendulum

According to lore, Galileo first pointed out that all objects fall at the same acceleration, independent of their mass. This is pretty much true, at least near the surface of the earth. We understand this simply in terms of Newtonian

mechanics, which says that

$$\vec{F} = m\vec{a} \tag{7.1}$$

and Newtonian gravity, which says that

$$F = G\frac{mM_E}{R_E^2} \tag{7.2}$$

where m is the mass of the object, M_E is the mass of the earth, and R_E is the radius of the earth, which we assume is much larger than the height from which the object is dropped. In other words, the acceleration a due to gravity near the earth's surface, which we call g, is

$$g = G\frac{M_E}{R_E^2} \approx 9.8 \text{ m/sec}^2 \tag{7.3}$$

In fact, since the earth is flatter near the poles and therefore closer to the center of the earth, there is some variation with lattitude. At sea level, one finds $g = 9.780524$ m/sec^2 at the equator, and $g = 9.832329$ m/sec^2 at the poles, a fractional difference of about one half of one percent.

There are practical, as well as philosophical, reason to know the value of g with high precision. For example, oil exploration can exploit small changes in the gravitational acceleration due to underground density changes. Consequently, there has been a lot of work over the years aimed at high quality measurements of g. Until very nifty techniques based on measuring the rate of free fall using interferometry came into being[1], the pendulum was the best method. We will explore that technique in this laboratory.

A sketch of the physical pendulum and its approximation as a simple pendulum are shown in Fig. 7.1. For a precise measurement of g it is important to realize that no pendulum is truly "simple", so we'll start with the physical pendulum. The rotational inertia $I \equiv \int r^2 dm$ is defined around the pivot point, and L is the distance from the pivot to the center of mass. Newton's Second Law in terms of the swing angle θ is

$$\tau = I\frac{d^2\theta}{dt^2}$$

where the torque is

$$\tau = \left|\vec{W} \times \vec{L}\right| = MgL\sin\theta$$

[1]See *Practical Physics*, G. L. Squires, Third Edition, Cambridge (1985)

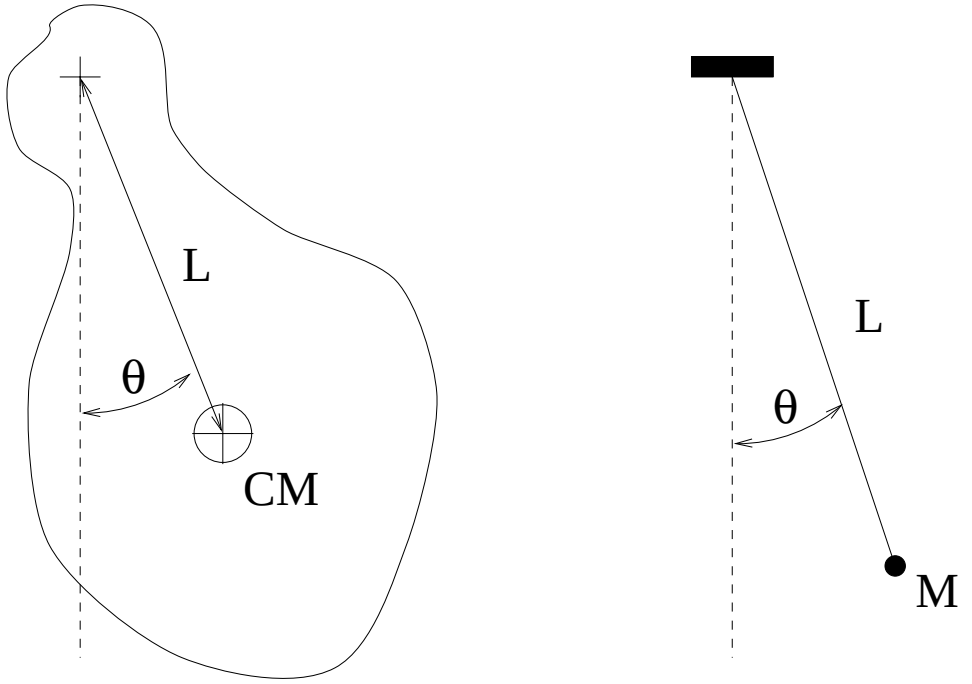


Figure 7.1: Physical and simple pendula. The physical pendulum realizes the size and mass distribution with a rotational inertia I about the pivot point. If approximated as a simple pendulum, i.e. a point mass suspended on a massless string, then $I = ML^2$.

so the equation of motion can be written

$$\frac{d^2\theta}{dt^2} + \frac{MgL}{I}\sin\theta = 0 \tag{7.4}$$

This is generally solved in the "small angle approximation", that is, by setting $\sin\theta \approx \theta$. In this case, we are reduced to simple harmonic motion with angular frequency

$$\omega \equiv \left(\frac{MgL}{I}\right)^{\frac{1}{2}} \quad \text{Physical Pendulum} \tag{7.5}$$

The approximation as a simple pendulum just sets $I = ML^2$, so we have

$$\omega \equiv \left(\frac{g}{L}\right)^{\frac{1}{2}} \quad \text{Simple Pendulum} \tag{7.6}$$

One goal of this experiment is measure the pendulum period precisely enough to see a departure from the small angle approximation. This departure can be calculated theoretically. We find a "first integral of the motion" by first multiplying Eq. 7.4 by $d\theta/dt$

$$\frac{d\theta}{dt}\frac{d^2\theta}{dt^2} + \omega^2\frac{d\theta}{dt}\sin\theta = 0$$

then rearranging the derivatives to get

$$\frac{d}{dt}\left[\frac{1}{2}\left(\frac{d\theta}{dt}\right)^2 - \omega^2\cos\theta\right] = 0$$

which implies that

$$\frac{1}{2}\left(\frac{d\theta}{dt}\right)^2 - \omega^2\cos\theta = \text{constant}$$

The constant[2] can be determined by assuming the pendulum is released from rest ($d\theta/dt = 0$) at an angle θ_0, i.e. constant $= -\omega^2\cos\theta_0$. Therefore

$$\left(\frac{d\theta}{dt}\right)^2 = 2\omega^2(\cos\theta - \cos\theta_0)$$

[2]This constant can in fact be expressed in terms of the total mechanical energy.

and so

$$\int_0^{\theta_0} \frac{d\theta}{(\cos\theta - \cos\theta_0)^{\frac{1}{2}}} = \omega\sqrt{2}\left(\frac{T}{4}\right) \tag{7.7}$$

where the period is T and we realize it takes one-fourth of a period to move to the vertical position from the point of release.

This integral cannot be solved analytically, but we can make use of some mathematical trickery and expand it in powers of θ_0. Since $\cos x = 1 - 2\sin^2(x/2)$ we can rewrite Eq. 7.7 as

$$\int_0^{\theta_0} \frac{d\theta}{\left[\sin^2(\theta_0/2) - \sin^2(\theta/2)\right]^{\frac{1}{2}}} = \frac{\omega T}{2}$$

and then make a change of variables to $\sin x = \sin(\theta/2)/\sin(\theta_0/2)$ which leads us to

$$\int_0^{\pi/2} \frac{dx}{\left[1 - \sin^2(\theta_0/2)\sin^2 x\right]^{\frac{1}{2}}} = \frac{\omega T}{4}$$

Now we can easily expand the integrand in powers of $\sin^2(\theta_0/2)$

$$\frac{1}{\left[1 - \sin^2(\frac{\theta_0}{2})\sin^2 x\right]^{\frac{1}{2}}} = 1 + \frac{1}{2}\sin^2(\frac{\theta_0}{2})\sin^2 x + \cdots$$

and carry out the integral term by term.

The result is

$$T = \frac{2\pi}{\omega}\left[1 + \frac{1}{4}\sin^2\frac{\theta_0}{2} + \cdots\right] \tag{7.8}$$

where ω is given by Eq. 7.5 or Eq. 7.6. The small angle approximation is clearly recovered as $\theta_0 \to 0$. The second term in Eq. 7.8, which we might call the "first order correction", is small but you should be able to confirm it in this experiment.

7.1.1 Principle of Equivalence

Einstein realized that there was some cheating going on when we derived Eq. 7.3 using Eq. 7.1 and Eq. 7.2. The mass M of the object in question

is used in two very different ways, and we just assumed they were the same thing without asking why. In Eq. 7.1, Newton's Second Law, mass is just the proportionality constant that connects acceleration, a precisely defined kinematic quantity, with a new and more mysterious quantity called force. In Eq. 7.2, Newton's Law of Gravity, we use M to mean the quantity that gives rise to a "gravitational force" in the first place. We should actually write the two masses differently, i.e. "inertial mass" M_I for Newton's Second Law, and "gravitational mass" M_G for Newton's Law of Gravity.

We should therefore reduce the physical pendulum to the simple pendulum by writing $I = M_I L^2$ whereas the torque is more properly written as $\tau = M_G g L \sin\theta$. The period for the simple pendulum, in the small angle approximation, becomes

$$T = 2\pi \left(\frac{L}{g} \frac{M_I}{M_G} \right)^{\frac{1}{2}}$$

You might then ask, "Is the gravitational mass the same as the inertial mass for all materials?" and test the answer by measuring the period for pendulum bobs made from different stuff. Clearly if you are going to test whether Einstein was right or not, you must be prepared to make as accurate a measurement as possible.

The best limit[3] on $|M_I - M_G|/M_I$ was obtained by Eric Adelberger and collaborators at the University of Washington. They obtained $|M_I - M_G|/M_I < 10^{-11}$ using a torsion balance. Early in this century, however, a limit of $< 3 \times 10^{-6}$ was obtained with a simple pendulum.

7.2 Measurements and Analysis

The technique is simple and straightforward, but you have to take some care because the point is to make precise measurements.

Set up a pendulum by hanging a massive bob from a flexible but inelastic

[3]See *Gravitation and Spacetime*, Hans C. Ohanian and Remo Ruffini, Second Edition, Norton (1994)

line. You want to keep your "physical" pendulum as "simple" as possible, so make sure the line is very lightweight and the bob is small and massive. (You still will have to be careful when you determine the pendulum "length".) The length of line determines the period, so pick something convenient. A couple of meters is a good place to start.

Timing the period precisely is very important. Set up the pendulum so the bob swings close to the floor or table top. Put a mark on the surface under the bob when it is motionless. You'll use this mark to time the period as the pendulum swings past it.

Set the pendulum in motion and use the digital stopwatch to time the period. The stopwatch reads in 0.01 second intervals and the period will likely be a couple of seconds. That is, you would immediately have a systematic uncertainty of $\sim 0.05\%$ by timing one swing. That's not good enough, since we are trying to measure g to 0.1% or so, which means we need to know the period at least twice as well, or 0.05%. However, you can easily reduce the systematic uncertainty by a factor of 10 by timing 10 swings instead of only one.

Figure 7.2 histograms the period as determined[4] in several runs of ten swings each. (This analysis is done in MATLAB simply by entering the measurements into an array, defining another array to set the histogram bins, using the command hist to sort the data, and the command stairs to make the plot.) There is some scatter in the measurements which probably comes from human response time in starting and stopping the stopwatch. We will treat this scatter as a random uncertainty, that is we can take the period as the average of all these N measurements with an uncertainty given by the standard deviation divided by $\sqrt{N}$. That is, the period is determined to be $T = 2.7994 \pm 0.0009$, a 0.03% measurement.

Determine the length of the pendulum as best you can. Assign an uncertainty to the length, and calculate g from Eq. 7.6 and $T = 2\pi/\omega$. Determine δg, the uncertainty in g, by propagating the errors from the length L and period T. Is your result for $g \pm \delta g$ clearly within the established polar and equatorial values?

[4]Data taken by Jason Castro, Shaker High School Class of 1996.

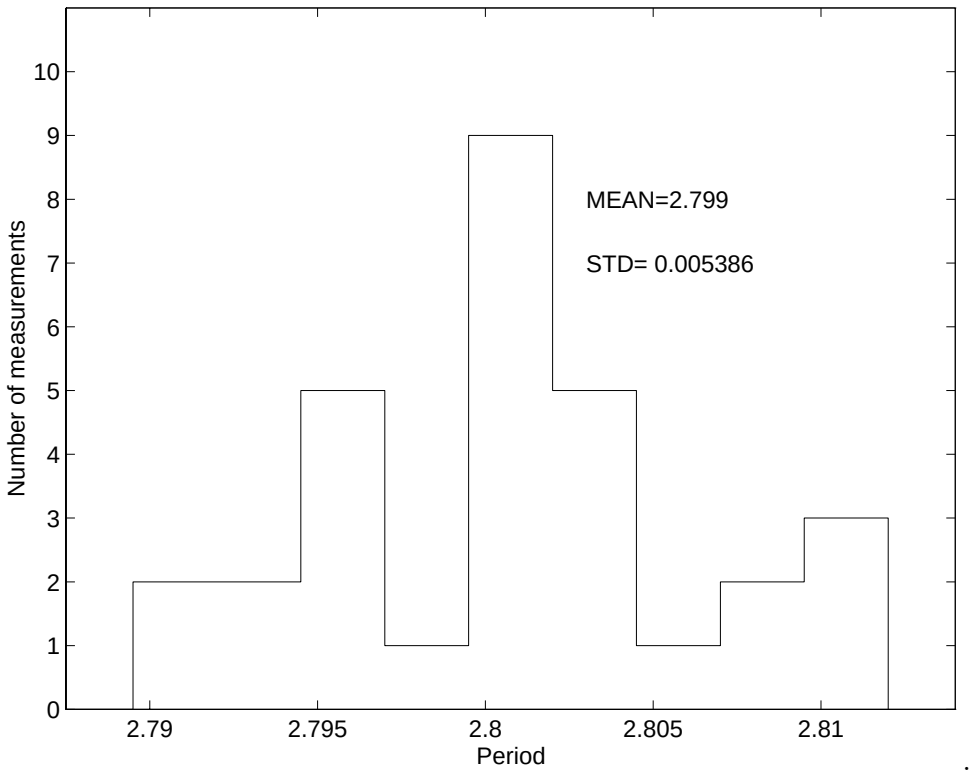


Figure 7.2: Histogram of several measurements of the pendulum period, each made by timing ten swings and dividing by ten to reduce the systematic uncertainty from reading the stopwatch.

Try to confirm the first order correction in Eq. 7.8 by changing the angle θ_0 and plot $T \pm \delta T$ as a function of $\sin^2(\theta_0/2)$. You will have to use an angle θ_0 that causes a correction significantly larger than your measurement uncertainty. You can measure θ_0 accurately enough just be putting a ruled scale on the floor or table top, and use trigonometry to turn the point at which you release the pendulum into an angle θ_0. Do you determine a straight line with the correct slope?

Ch 8

Experiment 4: Dielectric Constants of Gases

This experiment measures the dielectric constant of some gases. This is a simple physical property of materials, and in this case it can be related to the way the electron charge is distributed in atoms or molecules that make up the gas. The technique is simple, and is an instructive way to measure quantities that differ from each other by only a small amount.

The basic physics involved is rather straightforward. For a good basic discussion of the fundamentals, you might review

- *Physics*, Robert Resnick, David Halliday, and Kenneth Krane, John Wiley and Sons, Fourth Edition (1992)
  - Chap.22-24 *The Ideal Gas Law*
  - Chap.31 *Capacitors and Dielectrics*
  - Chap.38 *Electromagnetic Oscillations*

A fine discussion of the electronic properties of gases and how they give rise to the dielectric constant can be found in

- *The Feynman Lectures on Physics*, R. Feynman, R. Leighton, and M. Sands, Addison Wesley (1964), Vol.II Chapt.11

The measurement will be made using the "beat method" of measuring frequency. This is discussed in

- *Practical Physics*, G. L. Squires, Third Edition
  Cambridge University Press (1991), Sec.6.6

You will also likely use some edition of the *Handbook of Chemistry and Physics* to look up dielectric constants, ionization potentials, and dipole moments for various gases.

Note also that an experiment rather similar to this one is described in Y. Kraftmakher, Am. J. Phys. **64**(1996)1209.

8.1 Electrostatics of Gases

The physics associated with this experiment is pretty simple. It has to do with how charge can be stored in a capacitor, and how the material inside the capacitor changes the amount of charge that can be stored. After some review, we will get into specifics for the case where the material inside the capacitor is a gas.

Let's review the traditional definition of the dielectric constant. We'll start with a capacitor, pictured as a pair of parallel plates, separated by some distance that is small compared to their size. Assume first that the space in between the plates is a vacuum. If the capacitor is charged up to some voltage V by a battery and a charge $\pm q_0$ is stored on the two plates ($+q_0$ on one and $-q_0$ on the other), then the capacitance is defined to be

$$C_0 = q_0/V$$

Now suppose that the space between the plates is filled with some (non-conducting) material. It turns out that if the capacitor is charged to the

same voltage V, then more charge ($\pm q$) can be stored on the plates. In other words, the capacitance increases to

$$C = q/V$$

The increase in the capacitance defines the *dielectric constant* κ through

$$\kappa = \frac{C}{C_0} > 1 \tag{8.1}$$

Obviously, κ also measures the increased stored charge if the plates are kept at constant potential, i.e. $\kappa = q/q_0$.

The dielectric constant κ is a property of the material and does not depend on the capacitor geometery or the voltage. This is not at all obvious from these simple definitions, but we won't go into it in any more detail here.

So why does the charge on the capacitor plates increase when the material is inserted? The reason is that although the atoms or molecules that make up the material are electrically neutral, the positive and negative charges in them are somewhat independent. When they are inside the electric field of the capacitor, the negative charges tend to point towards the positive capacitor plate, and vice versa. This cancels out some of the electric field. However, if the plates are kept at constant voltage, the total electric field inside must remain unchanged. Therefore there is a buildup of charge on the plates, and the capacitance increases.

When the positive and negative charges "line up" in this way in an atom or molecule, it obtains a "dipole moment". For point charges of $\pm q$ separated by a distance x, the dipole moment $p = qx$. (See Resnick, Halliday, and Krane.) If the charge is not concentrated at a point, but has some distribution in space (as for an atom or molecule), then the dipole moment comes from integrating the charge distribution, weighted by the position.

The dielectric constant κ can be directly related to the atomic or molecular dipole moment p. The electric field inbetween the plates of the capacitor is $E = \sigma/\epsilon_0$ where $\sigma = q/A$ is the charge per unit area on the plates. There are always some "free" charges supplied by the voltage source, but with the dielectric in place there are also some "polarization" charges from the effect of the dipole moments. The key is to realize that the polarization charge per

unit area is just given by the net dipole moment per unit volume, called P. (See the *Feynman Lectures*.) Therefore, the electric field inside the capacitor is given by

$$E = \frac{\sigma}{\epsilon_0} = \frac{\sigma_{FREE} - \sigma_{POL}}{\epsilon_0} = \frac{\sigma_{FREE} - P}{\epsilon_0}$$

so that

$$\sigma_{FREE} = \epsilon_0 E \left(1 + \frac{P}{\epsilon_0 E}\right)$$

Equation 8.1 then implies that

$$\kappa = 1 + \frac{P}{\epsilon_0 E} \tag{8.2}$$

The task, then is to relate the individual atomic or molecular dipole moments to the net dipole moment per unit volume. How we do this depends on where those dipole moments come from, and there are two ways that can happen.

Some molecules have permanent dipole moments. They make up the class called *polar* dielectrics. This happens because the atoms that make up the molecules are arranged in some asymmetric pattern and the atomic nuclei cause the charge to be redistributed in some way. The most common example is the water molecule H_2O, where the atoms form a triangular shape with the oxygen at the vertex. A permanent dipole moment forms along the line passing through the oxygen nucleus and which bisects the two hydrogen nuclei. It is hard to calculate the magnitude of the dipole moment, but you can look it up in the *Handbook of Chemistry and Physics* and you find $p(H_2O) = 1.85$ Debye$= 6.17 \times 10^{-30}$ C·m. This is more or less typical of most polar molecules, with values ranging from about a factor of ten smaller to a factor of ten larger.

Atoms and most molecules, however, have their electric charge symmetrically distributed and do not have permanent electric dipole moments. They can nevertheless have dielectric properties because the electric field between the capacitor plates *induces* an electric dipole moment in them. These materials are called nonpolar dielectrics, and their behavior is considerably different from polar dielectrics. The action with a nonpolar dielectric inside the capacitor plates is shown schematically in Fig. 8.1, taken directly from Resnick, Halliday, and Krane, which shows the effect on the electric field.

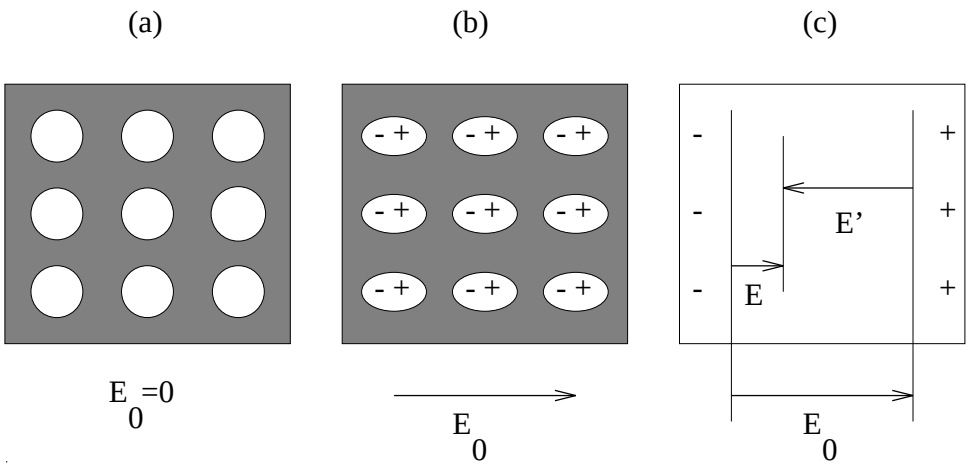


Figure 8.1: (a) A slab of nonpolar dielectric material. The circles represent neutral atoms of molecules. (b) An external electric field $\vec{E}_0$ displaces the positive and electric charges in the atom and induces a dipole moment. These displaced charges induce charges of the opposite sign on the capacitor plates (c), increasing the stored charge in order to keep $\vec{E}_0$ unchanged.

Let's first estimate the dielectric constant of some gas made of nonpolar atoms. We will use a very simple model, namely where the electron is bound to the atom by some imaginary spring with spring constant k. When the electron is placed in an electric field E, there is an electric force on it of magnitude eE. This causes the electron to be displaced a distance x where it is counterbalanced by the spring force kx. It will make more sense to express the spring constant k in terms of the electron mass m and the angular frequency of the simple harmonic oscillations ω_0, namely $k = m\omega_0^2$. Therefore $m\omega_0^2 x = eE$, and the atomic dipole moment is

$$p = ex = \frac{e^2 E}{m\omega_0^2} \tag{8.3}$$

Before we go further, it is instructive to estimate the size of this dipole moment. Estimate ω_0 by assuming that $\hbar\omega_0 = h\nu_0$ is the energy needed to ionize the atom. (This is a real seat-of-the-pants estimate!) It takes something like 10 eV to ionize an atom, so take $\omega_0 = 10$ eV$/\hbar = 1.52 \times 10^{16}$/sec. Let's also take a relatively high electric field, say 100 V across a capacitor with a 1 mm gap, or $E = 10^5$ V/m. Then we find $p = 1.22 \times$

10^{-35} C·m. You certainly expect, therefore, that the dielectric constant for a nonpolar gas should be a lot smaller than for a gas made of polar molecules. There are, however, other important differences as we shall soon see.

Anyway, let's continue and estimate the dielectric constant for the nonpolar gas. The dipole moment per unit volume is just $P = Np$ where N is the number of atoms or molecules per unit volume and p is given by Eq. 8.3. Equation 8.2 then gives

$$\kappa = 1 + \frac{Np}{\epsilon_0 E} = 1 + \frac{Ne^2}{\epsilon_0 m\omega_0^2} \qquad \textbf{Nonpolar Gas} \tag{8.4}$$

We approximate N from the ideal gas law, namely $N = \mathcal{P}/kT = 2.4 \times 10^{25}/\text{m}^3$ at room temperature ($T = 300$ K) and atmospheric pressure ($\mathcal{P} = 1.01 \times 10^5$ N/m^2). Using the same seat-of-the-pants estimate, we find that κ is very close to unity, in fact $\kappa - 1 = 3.3 \times 10^{-4}$. This is surprisingly close to what is actually measured, especially for such a very simple estimate. Keep in mind, however, how $\kappa - 1$ depends on N and ω_0^2.

Lastly, we will briefly derive the dielectric constant for a polar gas. At first, we suspect that it should be a lot larger because the dipole moment is so much bigger, but it isn't quite as simple as that. The permanent dipoles do indeed tend to line up along the electric field, but they are thermally agitated and don't stay aligned very long because they are always bumping into each other. See the *Feynman Lectures* or a book on statistical mechanics if you want to go through the derivation, but for now I will just quote the result

$$P = \frac{Np^2E}{3kT} = N \times p \times \left(\frac{pE}{3kT}\right)$$

This makes good sense qualitatively. The effective dipole moment of the polar molecule, i.e. P/N is just the permanent dipole moment p reduced by the factor $pE/3kT$ which measures the electrostatic energy of dipole alignment (i.e. pE, roughly) with respect to the thermal energy of the molecules (kT, roughly). This reduction factor is significant. For example, water vapor ($p = 6.17 \times 10^{-30}$ C·m) at room temperature in a 10^5 V/m electric field has a reduction factor of 5×10^{-5} bringing it more in line with nonpolar dielectrics. Putting this together into an expression for the dielectric constant gives

$$\kappa = 1 + \frac{Np^2}{3\epsilon_0 kT} \qquad \textbf{Polar Gas} \tag{8.5}$$

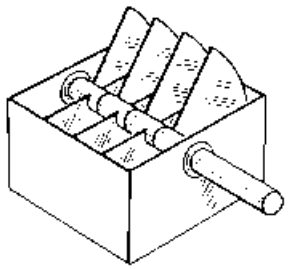


Figure 8.2: Sketch of an old fashioned variable parallel plate capacitor.

Note that as for nonpolar gases, $\kappa - 1$ is proportional to N and therefore proportional to the pressure. However for polar gases, it is a strong function of temperature. On the other hand, the dielectric constant for polar gases shows no dependence on the ionization potential (i.e. $\hbar\omega_0$) of the molecules.

8.2 Measurements

You should realize something right away. For gases, typical values for the dielectric constant κ are very close to unity. In fact, $\kappa - 1$ will be on the order of 10^{-4} or so. If you were to measure κ directly, therefore, you would need a fractional experimental uncertainty $\delta\kappa/\kappa \approx 10^{-5} = 0.001\%$ in order to get a 10% measurement of $\kappa - 1$. This would be hard!

The trick is to come up with a way to measure $\kappa - 1$ directly. We will do this by first relating κ to the frequency of electromagnetic oscillations, and then by learning how to measure the *difference* of two such frequencies.

The heart of the experiment is a variable parallel plate capacitor, the kind that had been used to tune the frequency in old fashioned radios. A sketch of such a thing is shown in Fig. 8.2. The relative surface area of the plates is changed by turning the knob which moves half the plates past the half alternately between, and this is how the capacitance is "tuned". The space between the plates is usually filled with air, but you will be able to introduce various different gases in that space, as well as evacuate it. The capacitance depends, of course, on what is between the plates because of the

dielectric constant of the material. The capacitance C of the capacitor can be changed either by tuning it (which makes a big change) or by changing the gas between the plates (a small change.)

This capacitor is put in series with an inductor (with inductance L), forming an "LC Oscillator". (See Halliday, Resnick, and Krane.) The current in this circuit, as well as the voltage across either the capacitor or the oscillator, varies sinusoidally like $\cos \omega t$ where $\omega = 1/\sqrt{LC}$. Changing the capacitance, then, changes the angular frequency ω. Still, however, it is very hard to measure the dielectric constant by introducing a gas inbetween the plates and remeasuring the frequency, since the change in frequency would be very small.

Instead of measuring the frequency directly, we will measure how much it changes using the "method of beats". Suppose you have two signals, call them y_1 and y_2, both with the same amplitude A but with different angular frequencies ω_1 and ω_2. If those two signals are added, you find

$$\begin{aligned} y_1 + y_2 &= A\left(\cos \omega_1 t + \cos \omega_2 t\right) \\ &= \left[2A \cos\left(\frac{\omega_1 - \omega_2}{2} t\right)\right]\left[\cos\left(\frac{\omega_1 + \omega_2}{2} t\right)\right] \end{aligned}$$

If $\omega_1 \approx \omega_2$, then the addition signal oscillates with a angular frequency $\bar{\omega} \approx \omega_1 \approx \omega_2$, *but with an amplitude that itself oscillates with a very low frequency* $|\omega_1 - \omega_2|/2$. These slow oscillations in the amplitude are called "beats", and it is not hard to build a circuit that gives an output signal which oscillates with the beat frequency of two input signals.

Okay, so the beat frequency measures the difference between two numbers (ω_1 and ω_2) which are very close to each other. That is essentially what you want, namely to measure the difference between the angular frequency with the capacitor in vacuum ($1/\sqrt{LC}$) and the angular frequency with the capacitor in gas ($1/\sqrt{\kappa LC}$). The problem, though, is that if the capacitor is in vacuum, you don't have the signal with it filled with gas, and vice versa! How can you get the two signals you want at the same time?

The solution to this problem is to have an external reference frequency, and measure the change in the LC oscillator frequency relative to the reference. In the experiment setup, the external oscillator is packaged in a box,

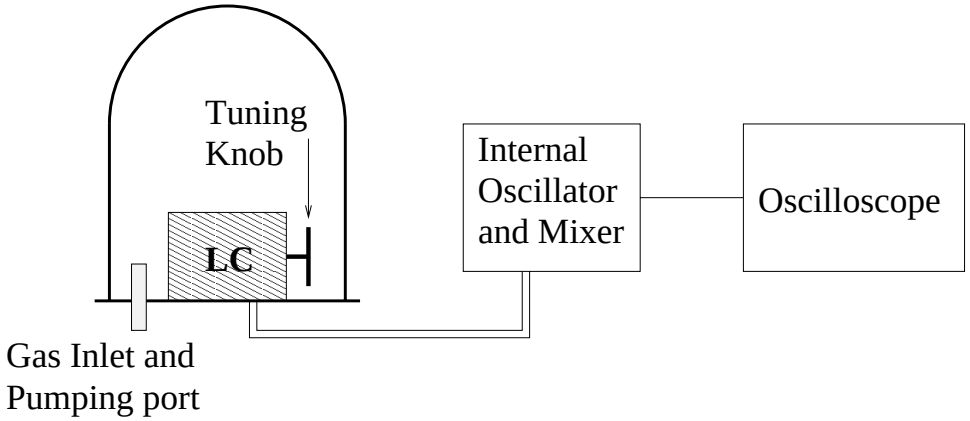


Figure 8.3: Setup for measuring the dielectric constant of a gas.

together with the circuit which forms the difference signal. You connect the LC oscillator signal into the box, and the output is a signal whose angular frequency is the *difference* (as opposed to half the difference) of the two angular frequencies of the input.

8.2.1 Procedure

The setup is shown in Fig. 8.3. The LC oscillator is inside another box which sits inside a bell jar. With the bell jar removed, you can tune the capacitance by adjusting the knob on the side of the box. *It is important to tune the capacitor so that the LC oscillator gives very closely the same frequency as the reference signal in the external box.* Do this by carefully turning the knob and watching the output difference signal on an oscilloscope. The object is to make the difference signal frequency as small as possible after the bell jar is replaced and then evacuated. This will likely take some trial and error. Also, be careful that the frequency of the LC oscillator is always less than the reference. That is, keep your eye on the difference frequency when you pump out the bell jar; the frequency will change, but you don't want it to go through zero.

You might notice that the difference frequency changes drastically while you're trying to measure it. In fact, you'll find that as you bring some object,

like your hand, near the bell jar, you can change the difference frequency at will. The reason is that you can disturb the electric field, and hence the capacitance, of the variable capacitor without actually touching it. You can't change the capacitance by very much, but you're going through all this so you can detect small changes in capacitance! For this reason, it is a good idea to cover the bell jar with a grounded, conducting shell that shields the capacitor from external sources of noise. This is pretty easily accomplished using a large sheet of aluminum foil, connected with a wire to a ground point.

Once you've tuned the capacitor, you need to measure the reference frequency $\nu_0 = \omega_0/2\pi$. In fact, you cannot get at the internal oscillator that provides it, but you can easily measure the frequency of the LC oscillator at this point, just by hooking up its output leads to the oscilloscope. If you tuned the capacitor perfectly, then this would be exactly equal to the reference frequency.

Now evacuate the bell jar. Using the valves connected to the filling hose, let in air or one of the gases inside the pressurized gas bottles. Let it in a little at a time, and measure the difference frequency for each pressure. The pressure gauge that we are using measures pressure $\mathcal{P}$ in inches of mercury, where one atmosphere is 30 inches. Record the pressure and difference frequency at each setting.

8.2.2 Analysis

Don't forget that the angular frequency ω is $2\pi\nu$ where ν is the frequency you measure using the oscilloscope. Let's make some definitions.

- ω_0 is the angular frequency of the external oscillator, that is, the reference frequency that you measured earlier.
- C_0 is the capacitance of the variable capacitor when the space inbetween the plates is evacuated.
- The dielectric constant $\kappa = 1 + \chi$, where χ is called the *electric susceptibility.*

Realize the χ is a small number in this experiment. Also realize that it is a function of the pressure $\mathcal{P}$. In fact, according to Eqs. 8.4 and 8.5, χ is proportional to $\mathcal{P}$.

We can write that

$$\omega_0 = \frac{1}{\sqrt{LC_0}} + \delta\omega_0$$

where $\delta\omega_0$ represents the (small) difference between the reference frequency and the evacuated frequency of the LC oscillator. You should be able to have tuned the capacitor so that $\delta\omega_0/\omega_0 \approx 0.1\%$ or smaller.

Measuring the difference frequency allows you to determine $\omega_0 - \omega$, where ω is the angular frequency of the LC oscillator, whether or not there is some gas between the plates. Therefore

$$\begin{aligned}
\omega_0 - \omega &= \frac{1}{\sqrt{LC_0}} - \omega + \delta\omega_0 \\
&= \frac{1}{\sqrt{LC_0}}\left(1 - \frac{1}{\sqrt{\kappa}}\right) + \delta\omega_0 \\
&\approx \omega_0\left(1 - \frac{1}{\sqrt{\kappa}}\right) + \delta\omega_0
\end{aligned}$$

where we write $1/\sqrt{LC_0} = \omega_0$. We can get away with this because it multiplies another very small number, namely

$$1 - \frac{1}{\sqrt{\kappa}} \approx \frac{1}{2}\chi$$

This introduces an uncertainty in χ on the order of $\delta\omega_0/\omega_0$, and it is unlikely that this small uncertainty will dominate the measurement.

If you plot your measurements as $\omega - \omega_0$ versus $\mathcal{P}$ or $\mathcal{P}/\mathcal{P}_{ATM}$, then you should get a straight line. In fact, the slope of the line should give you $\omega_0\chi_{ATM}/2$ if plotted against $\mathcal{P}/\mathcal{P}_{ATM}$. (The $y-$intercept of the line tells you how closely you tuned the capacitor to the reference frequency.) Draw the best straight line that you can through your data points, and call this the best value for the slope. Also draw lines with the largest and smallest slopes you think are reasonable. Use these lines to estimate the uncertainty in your measurement. An example is shown in Fig. 8.4. How does this uncertainty

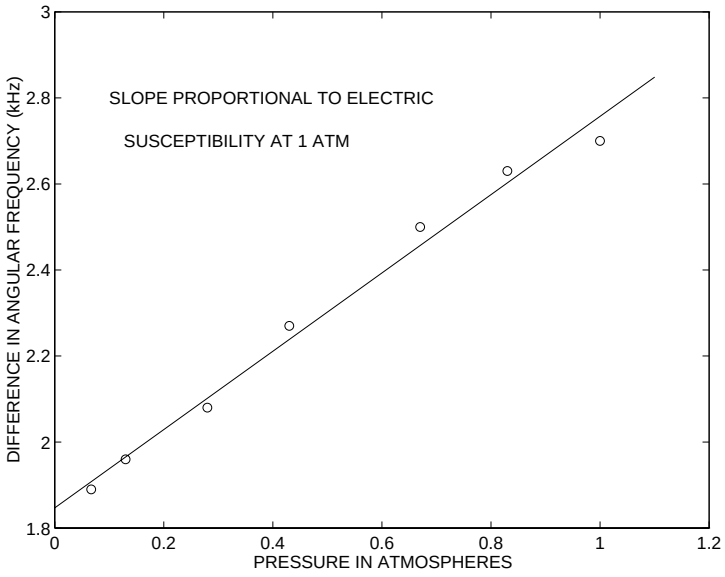


Figure 8.4: Sample Data for the Dielectric Constant

compare with that from the imprecise tuning of the variable capacitor? How does your value compare with "book" values for the dielectric constant? Is it within your experimental uncertainty?

Do this with a few different gases, including the air. Check to see if there is a correlation between the ionization potential of the gas and the dielectric constant. You might check the humidity in the atmosphere on the day you do the measurement. What effect does moisture in the air have on your result? Remember that unlike N_2 and O_2, water is a *polar* molecule.

8.3 Advanced Topics

Get a bottle of Helium gas to try. The electrons in a He atom are very tightly bound; it takes 24.5 eV of energy to remove an electron. Use this, and the discussion of nonpolar dielectric constants, to estimate the dielectric constant of Helium. Compare this to your measurements. If you are unable to determine a value for the He dielectric constant, see if you can determine

an "upper limit" for it. That is, if you see no slope in your plot of $\omega_0 - \omega$ versus $\mathcal{P}$, how big a slope would you be able to put through the data? Is this upper limit consistent with the book value?

Check the literature for possible polar gas molecules that you could measure. Be careful, since a lot of such molecules make explosive gases it would probably be wise to pick something less dangerous! Try to vary the temperature by cooling or heating the outside of the bell jar. Do the same with a nonpolar gas like N_2 or CO_2. Can you at least approximately verify the temperature dependence in Eqs. 8.4 and 8.5?

Ch 9

Statistical Analysis

We continue our discussion of uncertainties. In this chapter we will be talking about random uncertainties only, although some of the techniques we will develop (like curve fitting) can be applied in more general cases. As before, refer to the books by Squires or Taylor for more details:

- *Practical Physics*, G. L. Squires, Third Edition
  Cambridge University Press (1991)
- *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*,
  John R. Taylor, University Science Books (1982)

This chapter also begins a more serious discussion about datat analysis, especially using computers. As before, we will make use of MATLAB for all examples, and again I refer you to the following documentaion:

- *The Student Edition of* MATLAB, Prentice Hall (1994)
- *Numerical Methods for Physics*,
  Alejandro Garcia, Prentice-Hall (1994)

See Sec. 1.4.3 for more details.

9.1 The Mean as the Best Value

Let's introduce the subject by reconsidering something we took for granted. If we measure some quantity x a whole bunch of times, then we assumed that the best approximation to the "true" value of x was the mean of x, called $\bar{x}$. (See Eq. 6.1). This is in fact true, and we can prove it.

Let A be the value that best approximates the true value of x. Assume that we have n measurements of x, called x_i, and that each measurement has a standard deviation uncertainty σ. Consider the quantity $\chi^2(A)$ defined as

$$\chi^2(A) = \sum_{i=1}^{n} \frac{(x_i - A)^2}{\sigma^2}$$

The conjecture is that A is the value that *minimizes* χ^2. This actually makes some sense since if A gets too far away from all the values, then χ^2 gets very big. We will put this conjecture on firmer ground at the end of this chapter when we talk about the Gaussian Distribution, but for now let's take it at face value.

So, let's minimize $\chi^2(A)$ with respect to A. That is,

$$\frac{d\chi^2}{dA} = \frac{2}{\sigma^2} \sum_{i=1}^{n} (x_i - A)(-1) = 0$$

which implies

$$\sum_{i=1}^{n} x_i - A \sum_{i=1}^{n} (1) = 0$$

or

$$A = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

And there it is! The best approximation to the true value of x, i.e. A, is given by the mean of x.

From the definition of the standard deviation, it is clear that the minimum value of χ^2 is

$$\chi^2_{MIN} \equiv \chi^2(\bar{x}) = n - 1$$

When we generalize the definition of χ^2, the mean value will be very useful when evaluating data. We'll come back to this a few more times in this chapter.

This allows us to make a very useful generalized definition of the mean, called the *weighted average*. If the measurements of x do *not* all have the same standard deviation uncertainty, it doesn't make sense to just take a straight average of all of them. Instead, the values with the smallest uncertainties should be worth more, somehow. In this case,

$$\chi^2(A) = \sum_{i=1}^{n} \frac{(x_i - A)^2}{\sigma_i^2}$$

where σ_i is the standard deviation uncertainty of x_i. We can determine A in exactly the same way, namely by setting $d\chi^2/dA = 0$. We find that

$$A = \bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad \text{Weighted Average} \tag{9.1}$$

where the "weights" $w_i \equiv 1/\sigma_i^2$. Obviously, if all the weights are equal, then Eq. 9.1 reduces to Eq. 6.1. The uncertainty in the weighted average can be derived using propagation of errors, just as we did in the unweighted case. You find

$$\sigma_{\bar{x}} = \left[\sum_{i=1}^{n} w_i \right]^{-1/2} \tag{9.2}$$

There are no built-in functions like mean in MATLAB for the *weighted* average, but it is pretty simple to either do it from the command line, or write an appropriate m-file to carry this out. It could be done, for example, almost entirely within the sum command. See the discussion in Sec. 6.2.3.

What about the minimum value of χ^2 for a weighted average? That depends on the weights, i.e. the individual uncertainties assigned to the measurements x_i. However, if the various data points $x_i \pm \sigma_i$ are indeed consistent with a single "true" value, then you expect that $\chi^2_{MIN} \approx (n-1)$.

9.2 Curve Fitting

You will very, very often want to test your data against some model. Plotting your data in a suitable way can help you do that, as we've discussed. Sometimes, however, you need to be more precise. In particular, if the model depends on some parameters, you want to vary those parameters so that the model "fits" your data. This would give you the "best" value for those parameters. Of course, you also want to know with what uncertainty your data determines those parameters.

This is the subject of "curve fitting". We can develop it just by following our prescription for showing that the best value for some quantity is given by the mean of the measured values. To be sure, we are developing the technique known as the "method of least squares", since the object is to minimize the sum of the squares of the deviations between the data and the fitting function. There are in fact other techniques, such as the principle of maximum likelihood and multiple regression, that may actually be better suited to some class of problems, but we won't be discussing them here.

9.2.1 Straight Line Fitting

Let's start with the simplest generalization of our prescription for the mean. Whereas the mean is a one parameter "fit" to a set of values x_i, we now consider a two parameter fit to a set of points (x_i, y_i). In particular, the model is a straight line of the form

$$y = a_0 + a_1 x$$

and our job is to find the best values of a_0 and a_1, and their uncertainties, as determined by the data. If we were to fix $a_1 = 0$, then we should get $a_0 = \bar{y}$.

For now we assume all the values y_i have the same uncertainty we call σ_y, and we ignore any uncertainties in the x_i. The χ^2 function is defined just as before, namely

$$\chi^2(a_0, a_1) = \sum_{i=1}^{n} \frac{[y_i - y(x_i)]^2}{\sigma_y^2} = \sum_{i=1}^{n} \frac{(y_i - a_0 - a_1 x_i)^2}{\sigma_y^2}$$

which we minimze just as before, namely

$$\begin{aligned}\frac{\partial\chi^2}{\partial a_0} &= -\frac{2}{\sigma_y^2}\sum_{i=1}^{n}\left[y_i - a_0 - a_1 x_i\right] = 0 \\ \frac{\partial\chi^2}{\partial a_1} &= -\frac{2}{\sigma_y^2}\sum_{i=1}^{n}\left[y_i - a_0 - a_1 x_i\right] x_i = 0\end{aligned}$$

which leads to a pair of equations for a_0 and a_1,

$$\begin{aligned}a_0 n + a_1\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ a_0\sum_{i=1}^{n} x_i + a_1\sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i\end{aligned} \tag{9.3}$$

From now on, we will drop the limits $i = 1$ and n from the summation signs because it gets too crowded. The solutions for a_0 and a_1 are simple:

$$\begin{aligned}a_0 &= \frac{\left(\sum x_i^2\right)\left(\sum y_i\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{\Delta} \\ a_1 &= \frac{n\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{\Delta}\end{aligned} \tag{9.4}$$

where

$$\Delta \equiv n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2$$

So, if you draw a line $y = a_0 + a_1 x$ over a plot of your (x, y) data, the line will pass near all the points, assuming that a straight line was a good approximation in the first place. You would likely derive some physical quantities from the values of a_0 and a_1.

Remember when we defined the standard deviation (Eq. 6.2)? Instead of dividing by n, we divided by $n-1$. We won't try to prove it, but the reason is that $\bar{x}$ is not the "true" value of x, but rather just our best estimate for x. Therefore, the uncertainty is actually a slight bit larger than it would have been if we used the true value, and this shows up by dividing by $n-1$ instead of n. If n gets to be very large, then $\bar{x}$ is very close to the true value and $n-1$ is very close to n so this is at least consistent.

Now when we fit to a straight line, we have the same problem. That is, our data determine the best values for a_0 and a_1, not the true values. In this

case, however, there are *two* free parameters, not one as for the simple mean. The standard deviations σ_y are therefore given by

$$\sigma_y^2 = \frac{1}{n-2} \sum (y_i - a_0 - a_1 x_i)^2$$

and the minimum value of χ^2 is $n-2$. The number of data points (n) minus the number of free parameters (2 for the straight line fit and 1 for the simple mean) is called the number of "degrees of freedom".

Of course, we need to know the uncertainties in a_0 and a_1 as well. Equations 9.4 give a_0 and a_1 in terms of things we know the uncertainties for, namely the y_i. Therefore, just use propagation of errors to get what we want. The result is

$$\sigma_{a_0}^2 = \frac{\sigma_y^2 \sum x_i^2}{\Delta} \quad \text{and} \quad \sigma_{a_1}^2 = \frac{n\sigma_y^2}{\Delta}$$

and so the result of your fit should be reported as $a_0 \pm \sigma_{a_0}$ and $a_1 \pm \sigma_{a_1}$.

If the individual points do *not* all have the same uncertainty, but instead are $(x_i, y_i \pm \sigma_{y_i})$, then the generalization is straightforward. We have

$$\chi^2(a_0, a_1) = \sum w_i (y_i - a_0 - a_1 x_i)^2$$

where $w_i = 1/\sigma_{y_i}^2$. The rest follows in the same way as above, and the equations are listed in both Squires and Taylor. They are also written out in Garcia, including programs in MATLAB and in FORTRAN. (See the next section.)

Using MATLAB to fit straight lines.

Straight line fitting is so common a problem that MATLAB has a built-in function for this. The function `polyfit(x,y,1)`, where `x` and `y` are arrays of the same length, returns a two-dimensional array which contains the slope and intercept of the best-fit straight line. (The third argument, `1`, is a simple extension of this function. I will explain it shortly.) Furthermore, the function `polyval(p,x)` returns the best fit functional values, i.e. the approximation to the `y`, for the slope and intercept in `p` as returned by `polyfit`. The following series of commands

```
p=polyfit(x,y,1);
f=polyval(p,x);
plot(x,y,'o',x,f)
```

fits the points to a straight line, and then plots the data points themselves along with the fit.

For unequally weighted points, you can't really use polyfit. However, it is a simple matter to program such a thing using MATLAB, and in fact Garcia has already provided us with an m-file which does this. It is called linreg.m, and it is available via anonymous FTP from The Mathworks[1], or from me. I also reproduce linreg.m in Fig. 9.1.

9.2.2 Fitting to Linear Functions

If you wanted to fit data to a parabola, i.e. $y = a_0 + a_1 x + a_2 x^2$, you could follow the same procedure as for a straight line. You would get three equations in the three unknowns a_0, a_1, and a_2, and solve them as before. However, there is something more profound going on.

There is an entire class of functions that can be fit this way. You might suspect as much if write Eq. 9.3 as

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Shades of Math III! This is called a "system of linear equations", and there are very general ways of solving these things.

Any function of the form

$$y = b_1 f_1(x) + b_2 f_2(x) + \cdots + a_m f_m(x)$$

can be fit using the procedure as for a straight line. The straight line, of course, is the case where $m = 2$, $f_1(x) = 1$, and $f_2(x) = x$. The parabola,

[1]You can get lots of free software like this from The Mathworks. Check out their World Wide Web address at http://www.mathworks.com. Software is available through their FTP site at ftp.mathworks.com.

```
function [a_fit, sig_a, yy, chisqr] = linreg(x,y,sigma)
% Function to perform linear regression (fit a line)
% Inputs -
%   x - Independent variable
%   y - Dependent variable
%   sigma - Estimated error in y
% Outputs -
%   a_fit - Fit parameters; a(1) is intercept, a(2) is slope
%   sig_a - Estimated error in the parameters a()
%   yy - Curve fit to the data
%   chisqr - Chi squared statistic
N = length(x);
temp = sigma .^ (-2);
s = sum(temp);
sx = sum(x .* temp);
sy = sum(y .* temp);
sxy = sum(x .* y .* temp);
sxx = sum((x .^ 2) .* temp);
denom = s*sxx - sx^2;
a_fit(1) = (sxx*sy - sx*sxy)/denom;
a_fit(2) = (s*sxy - sx*sy)/denom;
sig_a(1) = sqrt(sxx/denom);
sig_a(2) = sqrt(s/denom);
yy = a_fit(1)+a_fit(2)*x;     % Curve fit to the data
chisqr = sum( ((y-yy)./sigma).^2 );  % Chi square
return;
```

Figure 9.1: Garcia's program linreg.m for computing the weighted least squares linear fit to data using MATLAB.

and in fact any polynomial function, is just larger values of m and successive powers of x. The functions $f(x)$, of course, don't have to be power laws, but any function you like. These are called "linear functions" because they are linear in the free parameters. In general, they are *not* linear in x, so don't get these two uses of the word "linear" confused.

The solution for a general linear fitting problem takes the form

$$\begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1m} \\ F_{21} & F_{22} & \cdots & F_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ F_{m1} & F_{m2} & \cdots & F_{mm} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} \sum f_1(x_i)y_i \\ \sum f_2(x_i)y_i \\ \vdots \\ \sum f_m(x_i)y_i \end{bmatrix}$$

where $F_{jk} = \sum_{i=1}^{n} f_j(x_i)f_k(x_i)$. This reduces the job to solving an $m \times m$ system of linear equations. All you really need to do is setup the matrix. (In fact, the matrix is symmetric since $F_{jk} = F_{kj}$.)

These days there are plenty of computer programs that can do the matrix algebra for you. Of course, unless you have one of your own favorites, I recommend you use MATLAB. You might recall that MATLAB actually stands for MATrix LABoratory, and it is in fact very well suited for doing all sorts of linear algebra problems, just like this one. You need to construct the matrix F_{jk}, which is pretty simple to do, and then let MATLAB solve the matrix equation using the "\" operator.

The function polyfit, for example, actually uses general matrix manipulation to solve the linear fit problem for a general n-dimensional polynomial. That is what the third argument is about. As described in the User's Manual, the MATLAB function call

$$\mathsf{p} = \mathsf{polyfit}(\mathsf{x}, \mathsf{y}, \mathsf{n});$$

returns the coefficients p_i of the function

$$f(x) = p_1 x^n + p_2 x^{n-1} + \cdots + p_n x + p_{n+1}$$

which best fits the data points (x, y). The call

$$\mathsf{f} = \mathsf{polyval}(\mathsf{p}, \mathsf{x});$$

returns the function $f(x)$ evaluated at the same $x-$values as the data points.

For polynomial least squares fitting where the points are not all equally weighted, Garcia has provided another m-file called pollsf.m. If you intend to write your own linear least-squares fitting code, it would be a good idea to examine Garcia's technique.

9.2.3 Nonlinear Fitting

If you want to fit your data to some nonlinear function, that is, a function that is nonlinear in the free parameters, then the problem is harder. The approach is still the same, namely form the χ^2 function and minimize it with respect to the free parameters, but there are no general formulas. This minimization job can of course be done numerically, but when the number of free parameters gets large, that can be easier said than done.

As you might imagine, MATLAB contains the ability to do nonlinear fitting through numerical minimization. The functions fmin and fmins minimize functions of one or more than one variable, respectively. They are pretty easy to use, but be careful of the pitfalls. You need to have a reasonable starting point defined, and then feed fmin or fmins the χ^2 function you want to minimize. You can pass arguments to the χ^2 function through the arguments of fmin or fmins. There is lots of additional stuff in the MATLAB Optimization Toolbox, which is devoted to all sorts of minimizing and maximizing problems. This toolbox, however, is not part of the Student Edition of MATLAB, but it is available on the RCS version of the program.

Another computer program that is very popular for minimizing functions in general (but is almost always used to fit data to some curve), is called MINUIT and is available from the CERN Program Library. The program is continuously updated, but an older version is described in a paper by F. James and M. Roos. *Computer Physics Communications*, **10** (1975)343. You will likely come across other numerical minimization programs in analysis packages on just about any flavor of computer.

Sometimes, though, you get lucky. If you can "linearize" a nonlinear function, then a simple redefinition of variables turns the job into something

simple. This is not unlike finding the right way to plot data so that a simple curve is what you expect. (See Sec. 1.3.2.) It's best to illustrate this with an example.

One nonlinear functional form you run into a lot is the simple exponential, namely

$$y = Ae^{-\lambda x}$$

This is easily linearized, just by taking the natural log,

$$\ln y = \ln A - \lambda x$$

which can be fit to a straight line. Be careful, though, about the individual uncertainties. Even if the points y_i all have the same uncertainty σ_y, this will not be true for the straight line you are fitting. In this case, the points $\ln y_i$ will have uncertainty $\sigma_{\ln y_i} = (\sigma_y/y_i)$, and you need to use the weighted averages when computing the free parameters in the fit.

9.2.4 χ^2 as the Goodness of Fit

We'll conclude this section on curve fitting with a few more words about χ^2. This quantity is actually quite important in advanced statistical theory. If you really want to learn more about it, look at Taylor's book and some other texts, but for now just realize a simple way to use it.

Let's suppose you've taken your data and analyzed it by fitting it to a straight line or perhaps some more complicated function. You've included the individual point uncertainties in the fit, using the formulas like Eq. 9.4 after including all the weights. You graph the fitted function along with the data, and it comes close to most of the points so you figure you've done things correctly. Is there any way you can be more confident of the result? Is there some measure of how good the fit really is? Maybe you need to use a function that is slightly more complicated, and the additional terms are telling you something important about the physics, or about your experiment?

Recall that if all points x_i have the same uncertainty, then the minimum value of χ^2 is identically equal to $n - 1$. If the points do not have the same uncertainty, then I said that you expect χ^2 to be around the same value *if*

the individual points and their uncertainties are consistent with measuring a single value. The same can be said for a straight line fit. That is, if χ^2 is around $n-2$, then the fit is pretty good, and you probably don't need to go looking around for other sources of uncertainty.

In general, if the quantity $\widetilde{\chi}^2$ (sometimes called the "reduced χ^2"), defined as χ^2 divided by the number of degrees of freedom, is approximately unity, then the fit is "good". (Recall that the number of degrees of freedom is defined as the number of data points, minus the number of free parameters, or "constraints".) If the uncertainties are truly random, then you can even interpret the probability of data being given by your model. See Taylor.

9.3 Covariance and Correlations

Let's return to the discussion about "Propagation of Errors" in section 6.3. In particular, we talked about how to combine the various contributions δq_x, δq_y,... to get the net uncertainty δq. We listed a few possible choices, namely

$$\begin{aligned} \delta q &\stackrel{?}{=} |\delta q_x| + |\delta q_y| + \cdots \\ \text{or} \quad \delta q &\stackrel{?}{=} |\delta q_x - \delta q_y \pm \cdots| \\ \text{or} \quad \delta q &\stackrel{?}{=} \left[(\delta q_x)^2 + (\delta q_y)^2 + \cdots \right]^{1/2} \\ \text{or} \quad \delta q &\stackrel{?}{=} \text{Something Else} \end{aligned}$$

I told you that in the one specific case of random, uncorrelated uncertainties, the answer is clear and it is the third choice where we add uncertainties in quadrature. We are going to go a little further now, and look at the issue of *correlated*, but still random, uncertainties.

We are following the discussion as described in Taylor's book. We'll just deal with $q(x,y)$, that is, a function of two variables only. Since we are working with random uncertainties only, the best value is equal to the mean. If we expand $q(x,y)$ about the mean values of x and y we have

$$\begin{aligned} q_i &= q(x_i, y_i) \\ &= q(\bar{x}, \bar{y}) + \frac{\partial q}{\partial x}(x_i - \bar{x}) + \frac{\partial q}{\partial y}(y_i - \bar{y}) \end{aligned}$$

and we can determine the mean of q from

$$\begin{aligned}\bar{q} &= \frac{1}{n}\sum_{i=1}^{n} q_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[q(\bar{x},\bar{y}) + \frac{\partial q}{\partial x}(x_i-\bar{x}) + \frac{\partial q}{\partial y}(y_i-\bar{y})\right] \\ &= q(\bar{x},\bar{y})\end{aligned}$$

where the second and third terms are identically zero just from the definition of the mean. This actually proves, for the case of random uncertainties, our assertion in Sec. 6.3 that the best value of q is the function evaluated at the best values of x and y.

The standard deviation of q, σ_q, is determined from

$$\begin{aligned}\sigma_q^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(q_i-\bar{q})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left[\frac{\partial q}{\partial x}(x_i-\bar{x}) + \frac{\partial q}{\partial y}(y_i-\bar{y})\right]^2 \\ &= \left(\frac{\partial q}{\partial x}\right)^2\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] + \left(\frac{\partial q}{\partial y}\right)^2\left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2\right] \\ &\qquad + 2\left(\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\right)\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]\end{aligned}$$

You should immediately recognize the first two terms in brackets as the definitions of σ_x^2 and σ_y^2. By making the definition

$$\sigma_{xy} \equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \tag{9.5}$$

we can write

$$\sigma_q^2 = \left(\frac{\partial q}{\partial x}\right)^2\sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2\sigma_y^2 + 2\left(\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\right)\sigma_{xy} \tag{9.6}$$

The quantity σ_{xy} is called the "covariance". You should recall that σ_x^2 and σ_y^2 are called the "variance" of x and y. This is nearly the same as Eq. 6.6,

which defines the net random uncertainty when the uncertainties in x and y are uncorrelated, except for the extra term involving the covariance σ_{xy}.

Consider what happens to σ_{xy} when the uncertainties are uncorrelated, that is, when measurements of x and y are independent. This means that if x fluctuates to more than $\bar{x}$, it is equally probable for y to fluctuate to more or to less than $\bar{y}$. Therefore, the sum in Eq. 9.5 will tend to zero, and Eq. 9.6 is identical to Eq. 6.6. A nonzero covariance, however, can make Eq. 9.6 give a larger or smaller answer than Eq. 6.6 depending on the sign of σ_{xy}.

It is interesting to note that $|\sigma_{xy}| \leq \sigma_x \sigma_y$. Therefore,

$$\sigma_q \leq \left|\frac{\partial q}{\partial x}\right| \sigma_x + \left|\frac{\partial q}{\partial y}\right| \sigma_y$$

In other words, this is an upper limit for the uncertainty of random errors when the variables may or may not be correlated.

Unfortunately, Eqn. 9.6 is practically useless when analyzing experiments, especially in the basic laboratory. You will see many ways to estimate or determine the uncertainty σ_x for quantities x, but it is not so clear how to do the same for the covariance σ_{xy}. One of the easiest way to use the covariance, however, is to determine whether or not two variables x and y are correlated at all. For a set of points (x, y), we can define the *coefficient of linear correlation*

$$r \equiv \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2\right]^{1/2}} \tag{9.7}$$

Obviously, $-1 \leq r \leq 1$ where $r = 0$ implies there is no correlation between x and y. If $r = 1$, then x and y are perfectly linearly correlated, in otherwords $y = mx + b$ where $m > 0$. If $m < 0$, but there is still a strict linear relationship between x and y, then $r = -1$ and we say x and y are perfectly *anti*-correlated. This is all quite simple to prove since for all x_i and y_i, you must have $(y_i - \bar{y})/(x_i - \bar{x}) = m$ if $y = mx + b$.

Calculating the linear correlation coefficient for a set of pairs of numbers can be quite useful. It can tell you if it likely or not that one variable depends on another, and by how much. In many cases where you expect that two

Table 9.1: Example of (x, y) Correlation Data

#	GPA	$K	#	GPA	$K	#	GPA	$K
1	1.0	32	6	1.8	40	11	3.0	48
2	1.2	40	7	2.2	44	12	3.0	68
3	1.0	48	8	2.2	56	13	3.4	60
4	1.4	48	9	2.6	48	14	3.6	52
5	1.6	40	10	2.8	56	15	3.8	64

things might be independent, you can use the correlation coefficient to show that this is or is not so, at least to some level.

We can illustrate this with a simple example. Table 9.1 gives a made-up list of annual salaries in K$, for students who graduate with the given cumulative Grade Point Average after four years of college. You're studying hard, and you'd like to make sure that good grades pay off in the long run, at least in terms of earning potential! So, you ask, does the data indicate a correlation?

A good thing to do is plot the data. This is shown in Fig. 9.2. There certainly seems to be a trend in the data. You calculate the correlation coefficient from Eq. 9.7 and you find $r = 0.74$, which is a rather large value. There certainly seems to be a big correlation between grades and salary. You go back to studying.

These manipulations are all quite simple in MATLAB. Given a set of (x, y) data points, a few operations using the sum and sqrt functions will determine the value of r in Eq. 9.7. You might also look over the functions cov which calculates the covariance matrix, and corrcoef which calculates correlation coefficients. Also, the plot in Fig. 9.2 was made using MATLAB.

In this example, the plot shows a pretty clear correlation, and the correlation coefficent is actually quite close to unity. In many cases, the data is not so clearly correlated from the plot, and you rely heavily on the statistical analysis (i.e. the correlation coefficient) to tell if there is something behind the data. You might want to look up some socially popular correlation coefficients, like that between the serum cholesterol level in the body and the

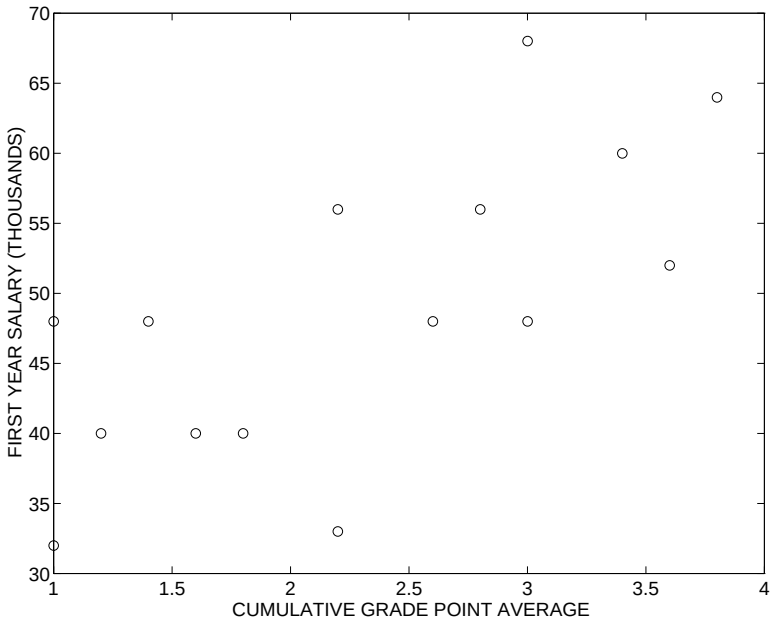


Figure 9.2: Example of (x, y) Correlation Data

incidence of heart disease. You might be surprised.

In a strict sense, the correlation coefficient can be used to tell you the *probability* that the variables are correlated or not. This of course is only valid if the fluctuations are purely statistical. Tables are available (see Taylor, for example) that allow you to look up these probabilities.

9.4 Distributions

Statistical analysis is based on probability. The entire formal basis of random uncertainty is based on "probability distributions", or "distributions" for short. Of course, leaving this until the end is rather cockeyed of me, but for the most part you don't need the formal descriptions of distributions to carry out the necessary analyses. However, for dealing with purely random situations, such as radioactive counting, you get valuable information from the underlying distributions.

We start by describing the only "true" random probability distribution, the *Binomial Distribution*. We then extend it to its key approximations, the *Poisson Distribution* and the *Gaussian* or *Normal Distribution*.

9.4.1 The Binomial Distribution

The binomial distribution is the basis for the study of random uncertainties. It is based on a very simple principle of random statistical probabilities that you are all familiar with.

Let's throw dice.

You have, say, five dice in your hands. You throw them all at once. If everything is random, the probability that none of the five dice shows a "one" is $(5/6)^5$, that is, the probability that any particular die has something other than a "one" (i.e. 5/6) raised to the number of dice that you threw. Similarly, the probability that all the dice show a "one" is $(1/6)^5$.

What is the probability that any single one of the dice shows a "one", and the rest show something other than "one"? The probability of one of them showing a "one" is 1/6 and the probability of the other four showing something else is $(5/6)^4$, but this is not the whole story. There are *five* different ways I can combine the five dice so that one shows a "one" and the others don't. Therefore, the probability of one showing a "one" and the others showing something else is $5\,(1/6)\,(5/6)^4$.

Similarly, the probability that two of the dice show "one's" and the other three do not is $10\,(1/6)^2\,(5/6)^3$. The factor of 10 comes from the fact that there are 10 different ways you can combine five dice so that two show a "one" and the other three don't. The only real trick is calculating the factor of 10. Let's look at this first.

This factor is really the number of combinations of n things taken ν at a time, where $n = 5$ and $\nu = 0, 1, \ldots, n$ in the above examples. We will use the notation $\left(\begin{array}{c} n \\ \nu \end{array} \right)$ for this factor. To calculate it, imagine you have a handful of n things, and you want to take out ν of them and set them on the table, in

any order. The number of ways to pick the first one is n. The number of ways to pick the next is $n-1$ since there is one less than when you started. So, the number of ways to pick them all is $n \times (n-1) \times (n-2) \cdots (n-\nu+1)$, and now there are ν of these things on the table. However, the order doesn't matter. Since there are $\nu!$ different permutations of the ν things, this procedure over counts by $\nu!$. Therefore,

$$\begin{aligned} \left(\begin{array}{c} n \\ \nu \end{array} \right) &= \frac{n \times (n-1) \times (n-2) \cdots (n-\nu+1)}{\nu!} \\ &= \frac{n!}{\nu!(n-\nu)!} \end{aligned} \tag{9.8}$$

There are some simple identities like

$$\left(\begin{array}{c} n \\ n-\nu \end{array} \right) = \left(\begin{array}{c} n \\ \nu \end{array} \right)$$

and

$$\left(\begin{array}{c} n \\ n \end{array} \right) = \left(\begin{array}{c} n \\ 0 \end{array} \right) = 1$$

and

$$\left(\begin{array}{c} n \\ 1 \end{array} \right) = n$$

Now let's go back to probabilities and throwing dice. Let p be the probability of a single success (namely 1/6 in the above example). In that case, $1-p$ (namely 5/6) is the probability of a single failure. Let n be the number of trials, i.e. $n = 5$ throws of the dice. Then, the probability $b_{n,p}(\nu)$ of ν successes in a total of n trials is

$$b_{n,p}(\nu) = \left(\begin{array}{c} n \\ \nu \end{array} \right) p^\nu (1-p)^{n-\nu} \tag{9.9}$$

This is the *binomial distribution*. It describes the probability distribution of a purely random set of events. It is called the "binomial" distribution because of the similarity between Eq. 9.9 and the equation for the binomial expansion:

$$(p+q)^n = \sum_{\nu=0}^{n} \left(\begin{array}{c} n \\ \nu \end{array} \right) p^\nu q^{n-\nu}$$

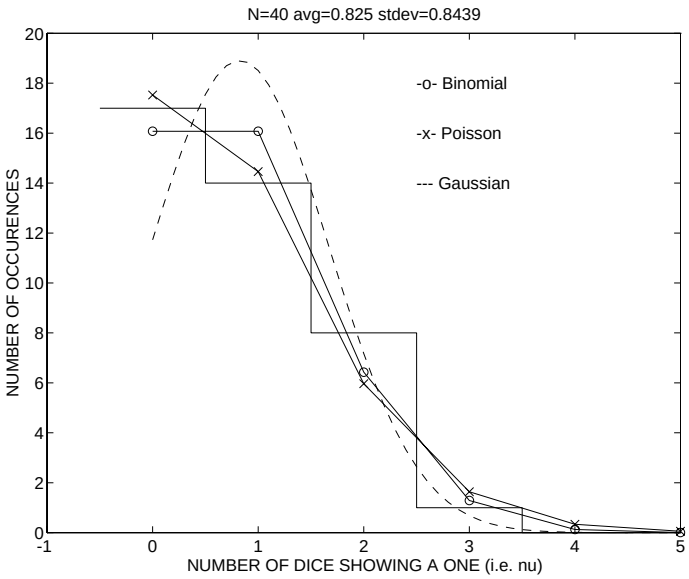


Figure 9.3: Data and probability distributions for throwing five dice.

This expression tells you something very important about the binomial distribution, which is true for all probability distributions, namely

$$\sum_{\nu=0}^{n} b_{n,p}(\nu) = 1 \tag{9.10}$$

That is, the binomial distribution is normalized to one. This just says that the sum of the probabilities of all the possible individual outcomes is unity, which of course it must be.

Let's take a look at what Eq. 9.9 is telling us. Assume you are doing an experiment, like throwing five dice. Each time you throw the five dice, you record the number of "one's" that come up. (We've called this number ν.) You do this 40 times, and make a histogram of the values of ν. This is histogrammed[2] in Fig. 9.3. Also plotted as a solid line in Fig. 9.3 is the prediction of Eq. 9.9, i.e. $40 \times b_{5,\frac{1}{6}}(\nu)$. (We will talk about the other curves soon.) You can see that the experiment pretty clearly matches the prediction. You can imagine that if instead of 40 times, you ran the experiment a very large number of times, then the match would be much closer. *The probability*

[2]Thanks to Jim Cronen and John Karcz for helping take this data.

distribution represents the outcome of the experiment if it were performed an infinite number of times. If the outcome of the experiment is governed solely by random statistics, then the binomial distribution describes the outcome. If there is a deviation from the binomial distribution, then the experiment is *not* governed by purely random statistics.

We can derive the mean $\mu = \bar{\nu}$ and standard deviation σ for the binomial distribution. It is not hard to do (see Taylor), but we just quote the results here:

$$\text{Mean} \quad \mu \equiv \sum_{\nu=0}^{n} \nu b_{n,p}(\nu) = np \tag{9.11}$$

$$\text{Variance} \quad \sigma^2 \equiv \sum_{\nu=0}^{n} (\nu - \mu)^2 b_{n,p}(\nu) = np(1-p) \tag{9.12}$$

Equation 9.11 is easy to interpret. It just says that the mean is the number of trials times the probability of individual success. In other words, if you throw five dice and the probability to get a "one" is 1/6, then on the average you expect to get just under one (5/6, to be exact) "one" each time you throw them.

We can check these equations against the data plotted in Fig. 9.3. The mean and standard deviation of the data is shown on the plot and are $\mu = 0.825$ and $\sigma = 0.8439$. The uncertainty in the mean is $\sigma/\sqrt{40} = 0.132$, so we can write $\mu_{DATA} = 0.825 \pm 0.133$ and $\sigma_{DATA} = 0.844$. The predictions, assuming the data is governed by the binomial distribution, are $\mu_{RANDOM} = 5 \times (1/6) = 0.833$ (from Eq. 9.11) and $\sigma_{RANDOM} = \sqrt{5 \times (1/6) \times (5/6)} = 0.833$ (from Eq. 9.12). The agreement is quite good.

In practice, we actually rarely use the binomial distribution directly. This is because it is expressed in terms of quantities that we have no direct measure of, namely n and p, so we can only use it if we know what n and p are supposed to be. Instead, we generally use one of two approximations to the binomial distribution which can be expressed in terms of measured quantities, namely μ and σ, where we assume that the measured values are good approximations to the true values for the binomial distribution. We discuss these two approximations in the remaining sections.

9.4.2 The Poisson Distribution

The *Poisson distribution* is based on the limit where $p \to 0$ and $n \to \infty$, but where the mean $\mu = np$ remains fixed. This limit is clearly very useful, since in many cases you can think of a very large number of chances for something to happen (i.e. n very large), but the probability of any one of those chances being successful is very small (i.e. p very small).

For example, the probability that you will get hit by any particular raindrop in a rain shower is very small, but there are a huge number of raindrops that can potentially hit you. The result is that some average number of raindrops does hit you and you get wet. The prototypical example of the Poisson distribution is radioactive decay, a truly random process, where the number of atoms that may decay in any time interval is enormous (something like Avogadro's number), but their individual chance to decay is tiny. The result is some average decay rate.

So let's take this limit of the binomial distribution, given by Eq. 9.9. First, since n is very large, we know that $n \approx (n-1) \approx (n-2) \cdots \approx (n-\nu+1)$ and so

$$\lim_{n\to\infty} \left(\begin{array}{c} n \\ \nu \end{array} \right) p^\nu = \frac{n^\nu}{\nu!} p^\nu = \frac{\mu^\nu}{\nu!}$$

Now consider what happens to $(1-p)^{n-\nu} \approx (1-p)^n = \left[(1-p)^{-1/p}\right]^{-\mu}$ as p goes to zero. The answer should[3] be well known to you, namely

$$\lim_{p\to 0}(1-p)^{-1/p} = \lim_{p\to 0}(1+p)^{1/p} = e$$

This gives us the expression for the Poisson distribution

$$P_\mu(\nu) = \frac{\mu^\nu}{\nu!} e^{-\mu} \tag{9.13}$$

As promised, the result does not depend on either n or p, but only on the measurable quantity μ. This is a big advantage.

[3] If you have been taught calculus in a formal way, you might have learned instead that $\ln(x) \equiv \int_1^x (1/t)dt$ and that e^x is the inverse function of $\ln(x)$, and that $e = e^1$. In that case, the above expression for e is a theorem that you may not have seen.

Figure 9.3 also plots the Poisson distribution approximation to the binomial distribution, comparing it to the data. Even for these modest values of $n(= 5)$ and $p(= 1/6)$, the result is quite close to the correct result, i.e. the binomial distribution.

Notice a glaring example of how the Poisson distribution is only an approximation. Figure 9.3 might have been extended to $\nu = 6$, that is, the probability that six dice show a "one", even though you only rolled five! This is meaningless for the binomial distribution (Eq. 9.9), but the formula for the Poisson distribution (Eq. 9.13) is perfectly calculable in this case.

The Poisson distribution allows you to estimate, for example, the probability that you get no successes ($\nu = 0$) when the mean is known. If, on the average, you get hit by 3.7 raindrops in one second, then the probability that in any particular second you don't get hit by any raindrops is

$$P_\mu(0) = \frac{\mu^0}{0!}e^{-\mu} = e^{-\mu} = e^{-3.7} = 2.5\%$$

Of course, calculating the probability for any number is just about as straightforward.

Note that the Poisson distribution depends only on μ and not on σ. In fact, Equations 9.11 and 9.12 allow you to write *in the limit of the Poisson distribution*,

$$\sigma = \sqrt{\mu} \tag{9.14}$$

That is, if the Poisson distribution describes your data, then the standard deviation of the distribution is determined only by the mean.

Consider the power of Eq. 9.14. If in some unit of time T, you measure N events, say the number of radioactive atoms that decay, then your best measure of the mean number of events you expected during that time T is just N, assuming you make no other measurements. Since you expect the result to be governed by the Poisson distribution (it is a truly random process, satisfying the limiting conditions), then your best measure of the uncertainty in N is just $\sqrt{N}$. Therefore the decay rate you measure is given by $R = N/T$ and the uncertainty is $\delta R = \sqrt{N}/T$. (We ignore any systematic uncertainty from the measurement of T.) The fractional uncertainty in the rate, $\delta R/R = \sqrt{N}/N = 1/\sqrt{N}$ gets smaller as N increases. Consequently, if

you want to measure the rate with twice as small a random uncertainty, you need to have N be four times as large, so you need to measure the number of counts over a time period of $4T$, or four times as long.

Don't confuse the rate R and the number of counts N in the above example or in problems like it. The discrete number N is what is distributed according to the Poisson distribution, not the normalized rate R. A common, but completely incorrect, mistake is to figure $\sqrt{R}$ for the uncertainty in R. This would imply that collecting more data does not improve the precision of your result.

9.4.3 The Gaussian Distribution

Both the binomial distribution and it's exact approximation in the limit of large n and small p, the Poisson distribution, share one inconvenient feature. They are functions of an *integer* we've called ν that translates into the number of "successes" of some particular type. In most cases, though, you are working with some variable that is not discrete, even though it may have its roots in some discrete quantity. In other words, you are working with a transformed quantity that ends up being a continuous, real number, and it is inconvenient, if not innacurate, to interpret the distribution of this variable in terms of a discrete distribution.

The *Gaussian distribution*, also known as the *normal distribution*, provides a way around this. In fact, the Gaussian distribution contains the basis for all the use of "squares" in definitions of things like the standard deviation (Eq. 6.2), adding errors in quadrature (Eq. 6.6), and the χ^2 function. It is an approximation to the binomial distribution that holds in the limit of large n and any p, but it is not an "exact" approximation because it is not valid over the entire region, no matter how large n becomes.

We won't go through it all here, but to arrive at the Gaussian distribution from the binomial distribution, you convert the factorial function to a function of a real variable using Stirling's approximation,

$$x! = \sqrt{2\pi} x^{x+\frac{1}{2}} e^{-x}$$

which is valid in the limit of large x. Note that x doesn't have to be an

integer on the right hand side, and that is the essence of why you can now use any real value for ν. What you discover is that

$$\lim_{n\to\infty} b_{n,p}(\nu) \approx \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(\nu-\mu)^2}{2\sigma^2}\right] \equiv f_{\mu,\sigma}(\nu) \tag{9.15}$$

so long as $|\nu - \mu|$ doesn't get too much larger than a few σ. In other words, as long as you are relatively near the mean, the Gaussian distribution is a good approximation to the true distribution for random variables. Note that $f_{\mu,\sigma}(\nu)$ is still normalized to unity (i.e. $\int_{-\infty}^{\infty} f_{\mu,\sigma}(\nu)d\nu = 1$) as any good probability distribution must be, so it overestimates the probability for $(\nu - \mu) \ll 0$ and underestimates it for $(\nu - \mu) \gg 0$.

The Gaussian distribution is similar to the Poisson distribution in that there is no longer any dependence on n or p. However, unlike the Poisson distribution, the Gaussian distribution depends independently on the mean μ and standard deviation σ. You have to know them both. You can't derive one from the other.

The Gaussian approximation to our dice-throwing example is also shown in Fig. 9.3. It is a rather poor approximation in this case, but not only is n not very large, the distribution is also somewhat narrow, so this is not a place where you expect such an approximation to do a good job. Try an experiment yourself where you throw a larger number of dice (say $n = 10$) and an individual success is defined to be either a "four", "five", or "six" on any of the dice. Thus $p = 1/2$ in this case, and you can investigate how well the Gaussian approximation holds in the region of the peak.

Since the outcome of some event involves the product of the probabilities of the individual events that make it up, and because Eq. 9.15 involves an exponential of a square of the deviation from the mean, it is not too hard to see how prescriptions like "adding in quadrature" arise for random events described by the Gaussian distribution. For more details, see the book by Taylor.

9.5 Data Analysis With MATLAB

We conclude this chapter with some final remarks about MATLAB as a tool for data analysis. Up to this point we have talked about the basics (Sec. 1.4.3); how to read in data from external sources like a digital oscilloscope (Sec. 3.7.1); using MATLAB for simple manipulations (Sec. 6.2.3); using MATLAB to fit data to straight lines (Sec. 9.2.1) as well as other linear (Sec. 9.2.2) and non-linear (Sec. 9.2.3) functions; and some notes on determining covariance and correlation (Sec. 9.3).

We will finish up with the procedure used to create Fig. 9.3. Other examples will be imbedded in the individual experiments.

The m-file which produced Fig. 9.3 is shown in Fig. 9.4. If you recall, this is from a simple experiment of throwing five dice at a time, and recording the number of dice which show a "one" on each throw. The plot shows the number of times zero, one, two, three, four, or all five dice came up with a "one". The dice were tossed 40 times.

The array dice corresponds to each of these six cases. The array freq is the result of the experiment, that is, the number of times (i.e. frequency) that each of these cases occurred. After making sure the next plot will clear the page (hold off), we make a stairs plot of the data, centered on the halfway points between the integers. (This just makes for a cleaner looking plot.) Labels are added to the x and y axes, and we prepare to add more things to this plot (hold on).

First we determine the number of throws, average number of dice with a "one", and standard deviation about this mean directly from the data. These values are added to the plot using the title command.

We then calculate the three distributions that will be plotted on top of the data. Notice that the binomial and Poisson distributions use the gamma function to find the factorial, i.e. $\Gamma(n+1) = n!$. Instead of calculating the Gaussian function on the course integer grid, we instead define it on an array x which goes from zero to five in tenths. (Note the use of the MATLAB built-in value pi= π.) In all three cases we normalize the distribution to the total number of tosses.

```
%
% Enter the data points and plot as a histogram
dice=[ 0  1 2 3 4 5];
freq=[17 14 8 1 0 0];
hold off;
stairs(dice-0.5,freq);
xlabel('NUMBER OF DICE SHOWING A ONE (i.e. nu)');
ylabel('NUMBER OF OCCURENCES');
hold on;
%
% Get statistical info about the data and add to the plot
N=sum(freq);
avg=sum(freq.*dice)/N;
sig=sqrt(sum(freq.*(dice-avg).^2)/(N-1));
title(['N=',num2str(N),' avg=',num2str(avg),' stdev=',num2str(sig)])
%
% Calculate distribution functions
% Binomial:
bnomcoef=gamma(5+1)./(gamma(dice+1).*gamma(5-dice+1));
bnomdist=N*bnomcoef.*(1/6).^dice.*(5/6).^(5-dice);
% Poisson:
poisdist=N*(avg.^dice)*exp(-avg)./gamma(dice+1);
% Gaussian:
x=[0:0.1:5];
gausdist=N*exp(-(x-avg).^2/(2*sig^2))/(sig*sqrt(2*pi));
%
```

Figure 9.4: MATLAB commands for calculating and plotting data and distributions.

```
% Add them to the plot
plot(dice,bnomdist,'o',dice,poisdist,'x',x,gausdist,'--')
plot(dice,bnomdist,'-',dice,poisdist,'-')
text(2.5,18,'-o- Binomial')
text(2.5,16,'-x- Poisson')
text(2.5,14,'--- Gaussian')
%
% Make PostScript file of the plot
print -dps distribs.ps
%
% Display the integrals
disp('  Areas under distributions:')
disp('  Binomial   Poisson  Gaussian')
disp([sum(bnomdist),sum(poisdist),trapz(x,gausdist)])
%
```

Figure 9.4 continued.

We then add these three distributions to the figure. The plot command is used twice, once to draw the distributions with symbols, or a dashed line in the case of the gaussian, and then to connect the symbols with straight lines. A legend is added to the plot with the text command; this is one of the many higher level graphics functions that are available with MATLAB, but which should not be important for nearly all the applications you will need in this course.

A postscript file is created of this figure. In fact, the file distribs.ps is what is included into this LaTeX document using the PSFIG macro.

Finally, the m-file puts some numbers out on the terminal screen which tell a little more about the distributions. That is, it calculates the area under the distribution curves and prints out the following:

```
  Areas under distributions:
  Binomial   Poisson  Gaussian
   40.0000   39.9913   33.4232
```

Note that we use the sum command to do the integral over the discrete

distributions, but the trapz command to integrate the gaussian distribution using the trapezoidal rule. Only in the case of the binomial distribution, which is an exact distribution not involving any approximations, do we get exactly the right answer, namely the total number of tosses of the dice. The Poisson distribution is very close to the right answer, only really making the approximation to the value of e. The gaussian distribution, however, misses by quite a bit, mainly because so much of the integral of this continuous function is for $x \leq 0$.

The Student Edition of MATLAB User's Guide contains an introduction to the program followed by a detailed writeup on each command or function, listed in alphabetical order. At the end of these detailed writeups there are a list of similar or related commands, and this is an excellent way to extend your knowledge of MATLAB.

9.6 Exercises

1. Formally prove some things that we just glossed over in the text.

a. Prove Eq. 9.1, that is, the standard definition for the weighted average is the value which minimizes χ^2.

b. Use propagation of errors to derive the uncertainty in the weighted average, i.e., Eq.9.2.

2. Recall problem #4 from Chapter 6. Use the method of least squares to fit the data for Δl as a function of ΔT to a straight line. Use the fitted slope and the uncertainty to determine the coefficient of linear expansion α. (You can use any program you might have handy, but if you just give the answer then you can't get any partial credit.) Also calculate the uncertainty $\delta\alpha$. This can be tedious by hand, but you can have partial credit if you at least show what needs to be done to calculate it. Are hand estimates just as good as a fitting program? What are the relative advantages or disadvantages?

3. Let's suppose you have some peculiar dice which each have 10 faces. The faces are numbered from 0 to 9. You throw *eight* of these dice at a time

and record which numbers land face down on the table. You repeat this procedure (i.e. throwing the dice) 50 times.

a. For how many throws do you expect there to be exactly three dice landing with either face 1 or face 5 landing face down?

b. What is the average number of dice you expect to land with either face 1 or face 5 down, for any particular throw? What is the standard deviation uncertainty in this number?

c. Use the Poisson approximation to calculate the same number as in (a).

d. Use the Gaussian approximation to calculate the same number as in (a).

You may want to review the material in the notes concerning Fig. C5.2.

4. A radioactive source emits equally in all directions, so that the intensity falls off like $1/r^2$ where r is the distance to the source. You are equipped with a detector that counts only radioactivity from the source, and nothing else. At $r = 1$ m, the detector measures 100 counts in 10 seconds.

a. What is the count rate, and its uncertainty, in counts per second?

b. What do you expect for the *fractional* uncertainty in the count rate if you count for 100 seconds instead of 10?

c. Based on the original 10 second measurement, predict the number of counts you should observe, and its uncertainty, if the detector is moved to a distance of 2 m and you count for one minute.

5. Suppose you are using a Geiger counter to measure the decay rate of a radioactive source. With the source near the detector, you detect 100 counts in 25 sec. To measure the background count rate, you take the source very far away and observe 25 counts in 25 sec. Random counting uncertainties dominate.

a. What is the count rate (in counts/sec) *and its uncertainty* when the source is near the Geiger counter?

b. What is the count rate (in counts/sec) *and its uncertainty* when the source is far away?

c. What is the net count rate (in counts/sec) *and its uncertainty* due to the source alone?

d. Suppose you want to reduce the uncertainties by a factor of 10. How long must you run the experiment?

6. An experimenter is trying to determine the value of "absolute zero" in degrees Celsius using a pressure bulb and a Celsius thermometer. She assumes that the pressure in the bulb is proportional to the *absolute* temperature. That is, the pressure is zero at absolute zero. She makes five measurements of the temperature at five different pressures:

Pressure (mm of Hg)	65	75	85	95	105
Temperature (°C)	-21	19	41	93	129

Use a straight line fit to determine the value of absolute zero, and its uncertainty, from this data.

7. Fit the following (x, y) values to a straight line...

x=	2.5	63	89	132	147
y=	406.6	507.2	551.3	625.5	651.7

...and plot the data points and the fitted line.

a. Does it look like a straight line describes the data well?

b. Study this further by plotting the *deviations* of the fit from the data points, i.e. ydev=y-yfit. What does this plot suggest?

c. Try fitting fitting the points to a quadratic form, i.e., a polynomial of degree 2. Is this fit significantly better than the straight line?

8. The following results come from a study of the relationship between high school averages and the students' overall average at the end of the first year of

college. In each case, the first number of the pair is the high school average, and the second is the college average.

78,65	80,60	85,64	77,59
80,56	82,67	81,66	89,78
87,71	80,66	85,66	87,76
84,73	87,63	74,58	91,78
81,72	91,74	86,66	90,68

a. Draw a scatterplot of the college average against the high school average.

b. Evaluate the correlation coefficient. Would you conclude there is a strong correlation between the grades students get in high school and the grades they get in their first year of college?

9. Using the data in Table 10.1, draw a scatterplot of electrical conductivity versus thermal conductivity for various metals. (Electrical conductivity is the inverse of electrical resistivity.) Calculate the linear correlation coefficient.

10. Graph the ratio of the Poisson distribution to the Gaussian distribution for mean values $\mu = 2$ and for $\mu = 20$. Use this to discuss where the Gaussian approximation to the Poisson distribtuion is applicable. Repeat the exercise, but comparing the Gaussian approximation directly to the Binomial distribution with $p = \frac{1}{2}$.

Ch 10

Experiment 5: Resistivity of Metals

If you apply a voltage across two points on a metal, then a current flows through the metal. We take it for granted that electrons moving through the metal are carrying that current.

In fact, the physics behind current carriers in metals is far from trivial. In this experiment we will explore some of that physics. What's more, we will do it with a novel technique that measures the *resistivity* of the metal, a property only of the type of material and independent of the size or shape of the conductor. This technique in fact can make measurements of the sample without actually touching it, and has found a lot of use in modern applications.

We introduce this experiment at a time when we've covered the basics of electronics and uncertainty analysis, and you will see lots of both while making these measurements. The circuits will make use of diodes and transistors, and you will take data using a digital oscilloscope. The data will lend itself to analysis using some of the curve fitting formulas we've derived.

The physics we will cover, including Faraday's Law which is intimately connected to the technique, can be found in the following books:

- *Physics*, Robert Resnick, David Halliday, and Kenneth Krane, John Wiley and Sons, Fourth Edition (1992)
  - Chapter 32, Sections 1-5: Resistivity
  - Chapter 36: Faraday's Law
- *Introduction to the Structure of Matter*, John J. Brehm and William J. Mullin, John Wiley and Sons (1989) Section 12-3, Especially Pg.593
- *Solid State Physics: An Introduction for Scientsts and Engineers*, Ronald F. Brown, El Corral Bookstore, California Polytechnic State University, Chapter 5 Sec. 1

The technique used in this experiment is based on the following paper:

- *Eddy-Current Method for Measuring the Resistivity of Metals* C.P. Bean, R.W. DeBlois, and L.B. Nesbitt Journal of Applied Physics **30**(1959)1976

10.1 Resistance and Faraday's Law

First we'll look over the definition of electrical resistivity and how it is related to resistance. We'll include a discussion about how the resistance might be expected to change as a function of temperature. Second, we talk briefly about the technique developed by Bean and collaborators which uses Faraday's law of induction to measure the resisitivity of a sample.

10.1.1 Resistance and Resistivity

Let's start with the assumption that Ohm's law is valid, that is, $V = iR$ where R is independent of voltage or current. Consider the idealized resistor pictured in Fig. 10.1. The resistor has a length L and a cross sectional area A. A voltage drop V is applied across the ends of the resistor. A current i

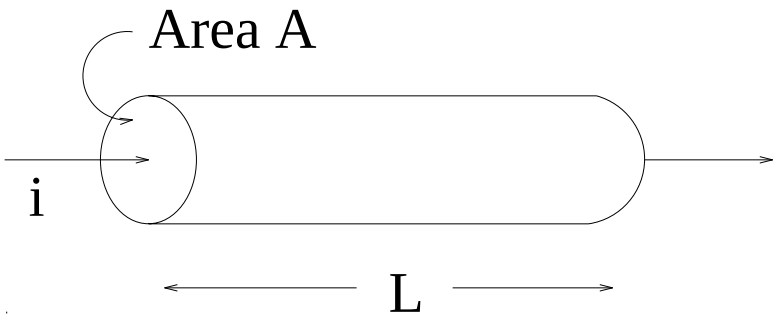


Figure 10.1: An idealized resistor.

of electrons flows from one end to the other, against a resistance R which is due to the electrons interacting somehow with the atoms of the material.

Consider Ohm's law on a microscopic level. The electric field set up across the ends of the resistor is just $E = V/L$. The electrons which carry the current will be spread out over the area A, so at any point within the resistor you expect a *current density* $j = i/A$. Therefore Ohm's law becomes

$$\begin{aligned} E &= j\rho \\ \text{where} \quad R &= \rho\frac{L}{A} \end{aligned} \tag{10.1}$$

and ρ is the "resistivity" , a property of the material which is independent of the dimensions of the resistor. In fact, Eq. 10.1 can be derived from the theory of electrons in metals, as shown in Brehm&Mullin and in Brown.

The resistivity comes from collisions between the electrons and the atoms of the material, which we now specify will be a metal. In a metal, the electrons are essentially free, so without any collisions they would continually accelerate under and applied field with an acceleration $a = eE/m$ where e and m are the electron charge and mass. However, the collisions cause the electrons to stop and then start up again, until the next collision. If the time between collisions is called τ, then the "drift" velocity v_d is just

$$v_d = a\tau = \frac{eE\tau}{m}$$

Now if there are n electrons per unit volume in the resistor, then a total charge $q = (nAL)e$ passes through the resistor in a time $t = L/v_d$. Therefore

Table 10.1: Electrical and Thermal Properties of Metals

Name	Z	A	Electrical Resistivity ($\mu\Omega$·cm)	Temperature Coefficient (10^{-3}/K)	Thermal Conductivity $\left(\frac{\text{cal}}{\text{cm·K·sec}}\right)$	Θ_D (K)
Al	13	26.98	2.65	4.29	0.53	395
Fe	26	55.85	9.71	6.51	0.18	420
Cu	29	63.55	1.67	6.80	0.94	333
Zn	30	65.38	5.92	4.19	0.27	300
Sn	50	118.69	11.50	4.70	0.16	260
Pb	82	207.19	20.65	3.36	0.083	86
Bi	83	208.98	106.80	-	0.020	118

the current density is

$$j = \frac{i}{A} = \frac{1}{A}\frac{q}{t} = \frac{1}{A}\frac{nALe}{L/v_d} = nev_d$$

in which case we have

$$\rho = \frac{m}{ne^2}\frac{1}{\tau} \tag{10.2}$$

Sometimes people will quote the "conductivity" $\sigma \equiv 1/\rho$ instead of the resistivity.

Electrical resistivities are listed[1] for various metals at room temperature in Table 10.1. Also included are some thermal properties, which we will see are closely related to the resistivity through the underlying physics. One of these is the temperature coefficient of resistivity, defined as $(1/\rho)d\rho/dT$. This quantity is in fact temperature dependent as we shall see, and the quoted numbers should be valid near room temperature.

Clearly, the fundamental physics of resistivity lies in the values for the collision time τ. The interaction of the quantum mechanical electron waves

[1]Values for Z, A, resistivity, and thermal conductivity are taken from the "Review of Particle Properties", Physical Review D50(1994), p.1241-1242. The temperature coefficient of resistivity, and all data for Zn and Bi, is from the "CRC Handbook of Chemistry and Physics", 56th Edition, CRC Press(1975), p.F-166. The Debye temperature is from the "Handbook of Physics", 2nd edition, McGraw-Hill(1967), Part 4, Tables 6.1 and 6.3.

and the quantized lattice of the metal crystal accounts for the collision time in a *pure* metal crystal. If there are impurities, then the scattering will contain an additional contribution. You can write

$$\frac{1}{\tau} = \frac{1}{\tau_{CRYSTAL}} + \frac{1}{\tau_{IMPURITY}}$$

The scattering from the crystal depends crucially on the vibrational energy stored in the crystal lattice, and therefore on temperature. (See Brehm&Mullin and Brown for more details.) The impurity scattering is essentially independent of temperature.

The connection between the crystal lattice scattering and the temperature points to some of the most basic condensed matter physics. Note the close correspondence between the thermal and electrical conduction properties of the metals listed in Table 10.1. This lead to an early connection between the heat capacities of various materials as a function of temperature, and their electrical and thermal conductivities. Using the Debye theory of heat capacities, Grüneisen calculated the quantum mechanical scattering from the residual ion sites in the metal, thus obtaining $1/\tau$. The result is

$$\rho(T) \propto \frac{m}{ne^2}\frac{T}{M_{ION}\Theta_D^2} G\left(\frac{\Theta_D}{T}\right) \tag{10.3}$$

where M_{ION} is the mass of the ion, Θ_D is the Debye Temperature (as determined from the heat capacity), and $G(\Theta_D/T)$ is called the Grüneisen function and is given by

$$G\left(\frac{\Theta_D}{T}\right) = 4\left(\frac{T}{\Theta_D}\right)^4 \int_0^{(\Theta_D/T)} \frac{u^5 du}{(e^u - 1)(1 - e^{-u})} \tag{10.4}$$

where $T_n \equiv T/\Theta_D$ is called the normalized temperature. The resisitivity ρ is proportional to $(T/\Theta_D)G(\Theta_D/T)$ which is plotted in Fig. 10.2. Notice that this function is roughly linear for temperatures greater than $\sim 0.3\Theta_D$ or so, giving you some idea of the range over which the fourth column of Table 10.1 is valid. The slope of this line is the "Temperature Coefficient" listed in the table. Remember that this formula is only valid for the contribution to resistivity from scattering from the crystal lattice.

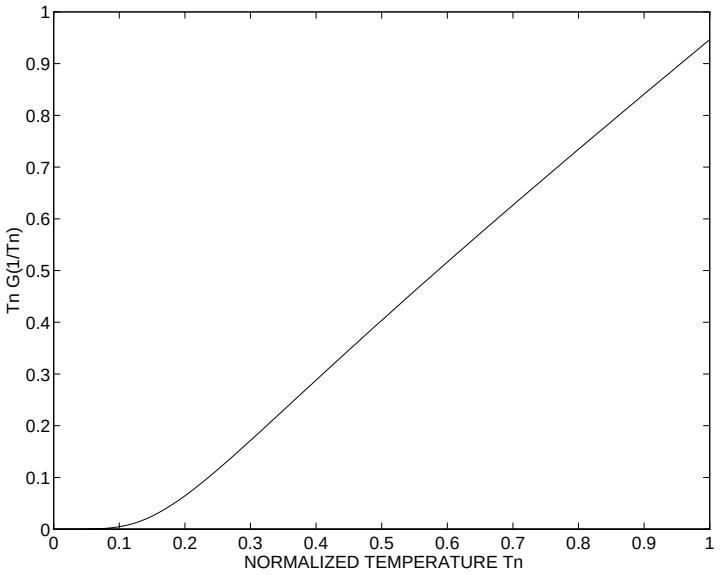


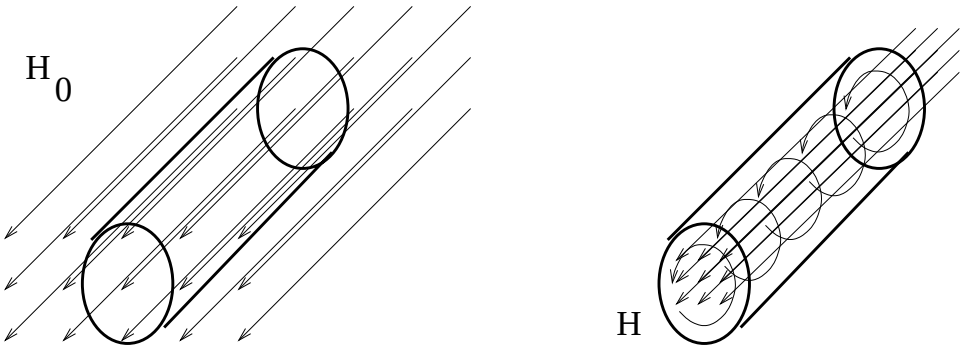
Figure 10.2: The function $(T/\Theta_D)G(\Theta_D/T)$.

10.1.2 The Eddy Current Technique

The technique we use measures resistivity directly, using the method developed by Bean. The idea is based on Faraday's law, which gives the EMF (i.e. voltage) induced in a coil that surrounds a magnetic field which changes with time. That is, you measure a signal $V(t)$ that is proportional to some dB/dt. This magnetic field B comes from the "eddy currents" left in a metallic sample when the sample is immersed in a constant magnetic field which is rapidly switched off.

Figure 10.3 shows how this is done. In Fig. 10.3(a), a cylindrical metallic bar is immersed in a constant magnetic field whose direction is along the axis of the cylinder. We assume the bar is not ferromagnetic, so the magnetic field inside is essentially the same as it is outside. Remember that the bar is filled with electrons which are essentially free to move within the metal.

Now shut the field off abruptly. By Faraday's Law, the electrons in the metal will move and generate a current so that tries to oppose the change in the external magnetic field. These so-called "eddy currents" are loops in the



(a) Field On (b) Field Shut Off

Figure 10.3: The eddy current technique for measuring resistivity. (a) A magnetic field H_0 permeates a cylindrical metal sample. (b) Eddy currents set up when the field is shut off generate a field H of their own. The eddy currents, and therefore H, dcrease with time at a rate that depends on the resistivity.

plane perpendicular to the axis of the sample, and they generate a magnetic field of their own. See Fig. 10.3(b). However, as soon as the external field is gone, there is nothing left to drive these eddy currents, and they start to decay away because of the finite resisitivity of the metal. The time it takes for the currents to decay away is directly related to the resistivity, as we shall see.

We again use Faraday's Law to detect the decaying eddy currents. The magnetic field set up by the eddy currents also decays away with the same time dependence as the currents. Therefore, if we wrap a coil around the sample, Faraday's law says that an induced EMF shows up as a voltage drop around this coil. This voltage drop is our signal, and the rate at which it decays to zero gives us our measure of the resisitivity of the metal sample.

In order to determine the voltage signal as a function of time, one needs to solve Maxwell's equations in the presence of the metal. The derivation is complicated, but outlined in Bean's paper, where a series solution is obtained

by expanding in exponentials. For a cylindrical rod, this series takes the form

$$V(t) \propto \sum_{i=1}^{\infty} \exp(-\lambda_i^2 \alpha t)$$

where α is proportional to ρ and the λ are roots of the zero-order Bessel function, i.e. $\lambda_1 = 2.405$, $\lambda_2 = 5.520$, $\lambda_3 = 8.654$, and so on. Since the λ increase with each term, for long enough times, only the first term is significant because all the rest die away much faster. That is, the falloff of $V(t)$ with time will look like a single exponential if you wait long enough, but will be more complicated at shorter times.

For a cylindrical metal sample where the external magnetic field points along the axis of the cylinder, the result is

$$V(t) = V_0 e^{-t/t_E} \tag{10.5}$$

$$\text{where} \quad t_E = 2.17 \times 10^{-9} \, \frac{\Omega \cdot \sec}{\text{cm}} \frac{R^2}{\rho} \tag{10.6}$$

$$\text{and} \quad V_0 = 10 N \rho H_0 \tag{10.7}$$

where $t = 0$ is the time when the external field is switched off. In this equation, R is the radius of the cylinder, expressed in cm, and ρ is the resistivity of the metal, expressed in Ωcm. Also, N is the number of turns in the detector or "pickup" coil and $H_0 = \mu_0 i n$ (in SI units) for a magnetic field H_0 set up by a solenoid carrying a current i through n turns. *This equation is only valid for times t on the order of t_E or larger.* At earlier times, there are transient terms left over which cause $V(t)$ to fall more rapidly than given by Eq. 10.5.

Bean, et.al., also derive the equivalent expression for a rod of rectangular cross section, instead of for a cylinder. They find

$$V(t) = V_0 e^{-t/t_E}$$

$$\text{where} \quad t_E = 1.27 \times 10^{-9} \frac{\Omega \cdot \sec}{\text{cm}} \frac{1}{\rho} \frac{a^2 b^2}{a^2 + b^2}$$

where the cross sectional dimensions of the rectangle are a and b, both in cm.

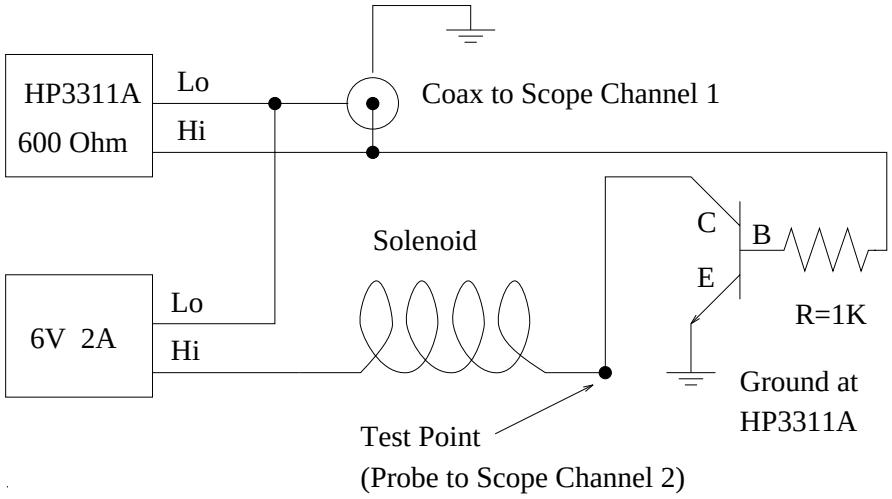


Figure 10.4: Switching circuit for turning the magnetic field on and off. It is a good idea to check the current through the solenoid by measuring the voltage at the testpoint, timed against the HP3311A square wave generator.

10.2 Measurements

The lifetime t_E given by Eq. 10.6 is on the order of tenths of milliseconds. Therefore, the magnetic field must be switched off considerably more rapidly than that. This is hard to to mechanically, so we will resort to an electrical switch, using a transistor.

The circuit which produces the switching magnetic field is shown in Fig. 10.4. A garden variety 6 V/2 A power supply puts current through the solenoid, creating the magnetic field H_0. However, after passing through the solenoid, the current encounters a transistor (321/TIP 122) instead of passing directly back to ground. The lead out of the solenoid is connected to the collector on the transistor, and the emitter is connected to ground. The base is connected through a 1 kΩ resistor to the 600Ω output of the HP 3311A waveform generator. The waveform generator is set to produce a square wave, oscillating between around -10 V to $+10$ V with a period of a few milliseconds. Let's see what this implies for current through the solenoid.

Recall the transistor rules back in Sec. 2.4.2. (You in fact are using an

npn transistor.) First, the DC power supply is connected so that the solenoid is always positive with respect to ground, so the collector is always above the emitter. Second, the base-emitter acts like a conducting diode, so there will be a voltage drop across it of around 0.6 V when it conducts. Also, if there is no current through the base, then the base-collector is reversed biased and no current flows through the transistor, or therefore through the solenoid. That is, the switch is off.

This transistor is actually a "darlington pair" which effectively gives a single transistor with a gain parameter $h_{FE} = \beta = 1000$ or so. Trust me that $V_{CE} = 6$ V does not exceed the specifications. Now when the waveform generator is at +10 V, the current through the base is $i_B \approx 10$ V/1 kΩ=10 mA. This turns the switch on and lets the current flow through the solenoid pretty much as if the transistor wasn't there, so long as $i_C \ll \beta i_B = 10$ A. You might want to measure the resistance in the solenoid coil to make sure it doesn't draw a lot of current, but since you're using a 2 A power supply, it is a good bet that you're in the clear.

So, when the square wave generator is at +10 V, the solenoid conducts. However, when the generator switches to -10 V (or presumably anything less than around 0.6 V), the solenoid and the magnetic field shut off. This is $t = 0$ in Eq. 10.5.

The pickup coil is wound on a separate tube which can be inserted inside the solenoid. You can then insert and remove different metal samples from the inside of the pickup coil. You might think that all you need to do is connect the terminals of the pickup coil to an oscilloscope, and that is pretty much what you do, but there is one complication. The magnetic field shuts off so fast, that the instaneous induced voltage in the pickup coil is very large. That is, Δt is so small that $dB/dt \approx \Delta B/\Delta t$ and so V is enormous. It is so large that it screws up the input circuitry of most oscilloscopes, since they are designed to guard against large voltages.

To fix this problem, the simple circuit shown in Fig. 10.5 is used to connect the pickup coil terminals to the oscilloscope input. The two diodes are arranged so that any current is taken to ground, so long as the voltage is bigger than +0.6 V or smaller than -0.6 V, for $V_F = 0.6$ V. That is, the circuit "clamps" the input to the oscilloscope so that it never gets very big,

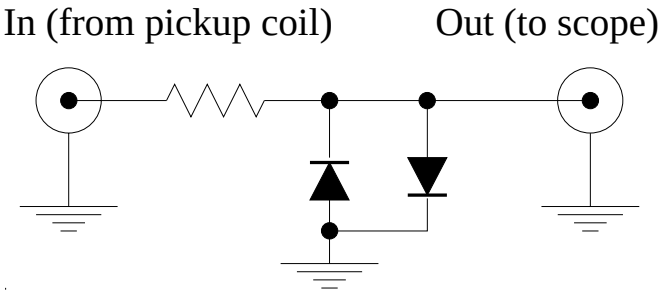


Figure 10.5: Clamping circuit for oscilloscope input.

but still big enough to make the measurement.

Sometimes you will see the signal "ring" just as the switch shuts off. That is, you see the decaying exponential but a rapid oscillation is superimposed on it, and this gets in the way of measuring the decay time. (Your circuit has lots of "loops" each of which is essentially an inductor. Any capacitance somewhere will cause oscillations, but the exact source can be hard to pin down.) If the ringing goes away while the signal is still decaying exponentially, just use the data past the point where the ringing is gone. If you don't have that luxury, attach a resistor in parallel with the scope input. It's best if you can get a variable resistor, and play with the values so that the exponential decay is unaffected but the ringing is thoroughly damped out.

You can use an analog oscilloscope to set up the circuits and convince yourself that things are working correctly, but to take data you should use one of our LeCroy 9310 digital oscilloscopes. *Be sure that the input impedance is high and that the coupling is DC.* Things will look weird otherwise.

10.2.1 Procedure

Before measuring the resistivity, you should convince yourself that you know what the solenoid circuit is doing. Connect up the components according to Fig. 10.4. Note that the collector on the switching transistor is connected to the metal block. *Do not let this block come in contact with ground.* Hook up the output of the waveform generator to one of the input channels of the oscilloscope, and confirm that you see a square wave of the right period and

amplitude. Make adjustments if necessary. It is probably a good idea to keep this signal always in one oscilloscope channel throughout the measurements since it is a simple way to tell when the magnetic field is on (square wave high) or off (low).

Now connect a probe to the junction between the solenoid and the transistor collector. View this on the other channel of the oscilloscope, and confirm that you see what you expect. That is, when the square wave is high, the solenoid is conducting and the voltage at this point should be around +1.2 V, i.e. the sum of the two forward voltage drops for the CB and BE diode equivalents for the transistor. On the other hand, when the square wave is low, the solenoid should not be conducting and there is no voltage drop across it, so the voltage at this junction should be around +6 V, i.e. the voltage of the DC power supply. You can remove this probe now since you will need this oscilloscope channel to make the resistivity measurements.

Next, connect the pickup coil to the clamping circuit and plug it into the second channel of the scope. Don't put any metal sample in just yet. You should see a voltage spike, alternaltively positive and negative, when the magnetic field switches on and off. This is just Faraday's law in its most brutal form. If the diode clamps were not there, the voltage spikes would be so large the protection circuit on the scope input would mess up the signal.

Now you're ready to take some data on resistivity. By this time, you should be using one of the LeCroy 9310 digital oscilloscopes. You can just use the cursor to read the values from the decaying exponential trace, or you can store the data on a floppy disk for later analysis. (See Sec. 3.7.1 for general instructions.)

Take the 5/8 inch diameter aluminum (alloy) cylinder and insert it into the pickup coil tube. Watch the pickup coil signal on the scope as you do this. The effect of the decaying eddy currents should be clear. You may see some transient oscillations of the signal right after the field shuts off, but there should be plenty of time left after these oscillations die away for you to get a smooth curve.

There are several samples for you to try. Don't forget that you can do the same thing with a rectangular bar. It might be a good idea to try that.

10.2.2 Analysis

The analysis is rather straightforward. The simplest thing to do is just stop the trace on the digital oscilloscope, and use the cursors to read out several points. Do this with the 5/8 inch aluminum rod. Tabulate this $V(t)$ data and plot it on semilog paper. Extract the value of t_E and determine the resistivity ρ from Eq. 10.6. Do you get the value you expect? Don't forget that you are dealing with an alloy, not the pure metal. Explain the difference. Make an estimate of the experimental uncertainty in t_E and propagate that uncertainty to the value of ρ. Show whether or not any difference you see is within this experimental uncertainty.

A more complete analysis is best done by saving the data on a floppy disk, and using MATLAB, or some other program, to fit the trace to a decaying exponential. You might refer to sections 1.4.3 and 3.7.1. After transferring the data to a PC and converting it to an ascii file, called sc1.lis, the following MATLAB commands produce the plot shown in Fig. 10.6:

```
fid=fopen('sc1.lis');
a=fscanf(fid,'%f')';
fclose(fid);
chan=[3000:50:7000];
vdat=a(chan);
lvdat=log(vdat);
coef=polyfit(chan,lvdat,1);
efit=exp(polyval(coef,chan));
plot(chan,vdat,'o',chan,efit);
```

The data as stored in sc1.lis is 10002 numbers, that is, the list of voltage values as stored in the trace. This is best read in directly with the fscanf function, yielding an array a that has 10002 elements. Note that we take the transpose of a so that it is a column vector. We don't need all those numbers, so a separate array chan is defined which are the channels we will use. (Note the upper and lower limits are set to 3000 and 7000 respectively, based on the data as taken. Your case will likely be different.) The array vdat contains the voltage values only at the channel values in chan.

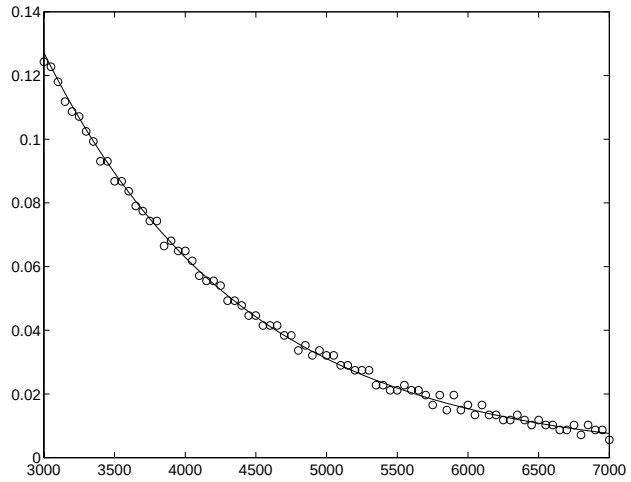


Figure 10.6: Sample data analyzed with MATLAB.

This is a simple case to fit, because we just take the logarithm and then fit to a straight line. Taking the exponential of this straight line gives the function that fits the data, and that is what we plot in Fig. 10.6. The array coef contains the fit parameters which give the exponential decay time. You might also compare your result for V_0 to the expression given in Eq. 10.7.

It is possible to use the digital oscilloscope to extract

$$\frac{d}{dt}\log_{10}\langle V(t)\rangle = -\frac{1}{t_E}\log_{10} e$$

directly from the display, using the "Measure/Parameters" menu after doing some manipulations with the "Math" menu. This number, in the time region where it is a constant, gives you the resistivity ρ through Eq. 10.6. On the other hand, it takes some time to get a good value for the average $\langle V(t)\rangle$, and for some of your measurements (particularly those that involve a temperature change) you might not have enough time.

The main source of systematic uncertainty is likely to come from the times over which you fit the decaying voltage signal. At short times, the decay is not a pure exponential because the transient terms have not all died away, so you want to exclude these times when you fit. At long times, there may be some left over voltage level that is a constant added to the exponential,

and again, a pure exponential fit will be wrong. Try varying the upper and lower fit limits until you get a set that gives the same answer as a set that is a little bit larger on both ends.

You should convince yourself that you are getting consistent results. Use the 1/2 inch alumimum alloy rod and measure it as well. Check to make sure that the decay lifetimes t_E scale like R^2. This should certainly be the case to within the experimental uncertainty you estimate.

You also have several pure ($> 99\%$) aluminum rods, that are 1/2 inch diameter. There are also various rods of copper, one of which is more than 99.9% pure, and a rod of lead which is 99.999% pure.

10.3 Advanced Topics

Having learned how to take and analyze data on resistivity, you can now investigate the temperature dependence. It is best to start simply by comparing the two samples of 1/2 inch diameter aluminum rod, one an alloy and the other a (relatively) pure metal. Vary the temperature by immersing the samples in baths of ice water, dry ice and alcohol, and liquid nitrogen. You can also use boiling water, and if you're really ambitions, hot oil.

These measurements will be tricky. You must remove the sample from the bath and measure the eddy current decay before the temperature changes very much. Probably the best way to do this is to take a single trace right after you insert the sample, stop the oscilloscope, and store the trace to disk. Then, you can analyze the trace offline to get the decay constant.

You might try to estimate how fast the bar warms up by making additional measurements after waiting several seconds, e.g. after saving the trace on the floppy. This would best be done with a sample whose resistivity, and therefore t_E can be expected to change a lot with temperature. Pure aluminum is a good choice.

Remember that the temperature dependence will be much different for the pure metal than for the alloy. Explain why. Try to estimate the contribution

to the mean free path of the electrons due to the impurities.

The other (pure) metals will give different temperature dependences. The Grüneisen function should describe the resistivity as a function of temperature for all pure metals. This is a function, however, of the quantity θ_D/T, where T is the temperature and θ_D is the Debye temperature for that particular metal. Therefore, different metals will give you different functions of T, but they should all give a universal shape when plotted as a function of θ_D/T. For your convenience, the Debye temperatures for various pure metals is given in Table 10.1.

Note that you can calculate the Grüneisen function using MATLAB. A series of statements like

```
u=[x/100:x/100:x];
intg=u.^5./((exp(u)-1).*(1-exp(-u)));
G(m)=4*trapz(u,intg)/x^4;
```

will calculate the value of $G(x = 1/T_n)$. The plot in Fig. 10.2 was made by embedding these statements in a for loop which varied the value of T_n over the appropriate range.

Ch 11

Light Production and Detection

Light is everywhere and we use it all the time. Nature is full of it because light represents a range of wavelenghts in the electromagnetic spectrum that is naturally emitted and absorbed by matter. We use light in many different ways when performing experiments. Furthermore, there is a good deal of physics in studying light itself.

I will be a little loose in my use of the word "light". Light is electromagnetic radiation with wavelength greater than ∼150 nm (=1500 Å) and less than several tens of microns. This range is a bit arbitrary, but is laboratory based. If light has a wavelength shorter than 150 nm or so, it is very hard to reflect from mirrors, and it will not penetrate far in most any material, so nothing is transparent to it. On the other hand, if the wavelength is more than several tens of microns long, then it is on the order of distances that can be seen with the naked eye, and our intuitive feeling for "light" breaks down.

A fine book which describes the use of light in the laboratory (and which contains many other excellent discussions on techniques used in experimental physics) is

- *Experimental Physics: Modern Methods*, by R. A. Dunlap, Oxford University Press (1988); Chapters 8, 9, and 10

Table 11.1: Wavelengths of Visible Light

Color	Wavelength Range (nm)
Red	622-770
Orange	597-622
Yellow	577-597
Green	492-577
Blue	455-492
Violet	390-455

Of course, we can see just a portion of the spectrum which we call "light". This portion ranges from around 400 nm to 800 nm, but the limits depend on your particular set of eyes. As listed by Dunlap and relisted in Table 11.1, the visible spectrum is broken up into the colors of the rainbow in different bands of wavelengths. Light with wavelengths longer than 770 nm is called "infrared" or IR. Light with wavelengths shorter than 390 is called "ultraviolet" or UV.

Recall some common notation. We use λ for wavelength, and typical units are nm (10^{-9} m) or Å (10^{-10} m), although microns (μm) are typically used in the IR. Frequency $\nu = c/\lambda$, and angular frequency $\omega = 2\pi\nu$, are measured in Hz=cycles/sec. The energy of a photon at a particular wavelength is $E = h\nu = \hbar\omega = hc/\lambda$. It is convenient to remember this relation in the form

$$E \text{ (in eV)} = \frac{1239.8}{\lambda \text{ (in nm)}}$$

In the visible part of the spectrum, the energy of a typical photon is ∼few eV and frequencies are several$\times 10^{14}$/sec.

The rest of this chapter discusses sources of light including thermal and line sources as well as lasers, and methods of measuring light intensity. Very important topics such as optical spectroscopy and polarimetry are left to the experiments which depend most heavily on them. Optical interferometry is a very powerful technique, but is not used in our course so we only mention it in passing. Dunlap, however, is a good source for information on all of these.

11.1 Sources of Light

Light production can be traced to the motion of atomic electrons and, to some extent, nuclei. These are charged particles that emit electromagnetic radiation when accelerated, and in nature much of this radiation is in the optical region.

If the motion is that of electrons bound up in individual atoms or molecules, then quantum mechanics tells us that only transitions between well-defined energy states are possible. In that case, the light will have a spectrum consisting of lots of discrete "lines" at wavelengths corresponding to the energies of transition between discrete states. On the other hand, if the radiation is from some sort of collective action of all the atoms, the spectrum will be continuous. Of course, the continuous spectrum can result from lots of very closely spaced energy levels giving discrete lines that are packed together. Sometimes, therefore, the time spectrum shows both discrete and continuous features.

11.1.1 Thermal Radiation

If you make something very hot, it glows. The light it gives off is essentially continuous. It is pretty well described as if it were a black body, that is, as if all the light incident on it were absorbed. The energy emitted by a black body is straightforward to derive based on "cavity radiation". The intensity (i.e. J/m^2 per sec·sr) of emitted radiation between frequencies ν and $\nu + d\nu$ is given by

$$I(\nu)d\nu = \frac{2\pi h}{c^2}\frac{\nu^3}{e^{h\nu/kT} - 1}d\nu \tag{11.1}$$

Recall that the factor $\left(e^{h\nu/kT} - 1\right)$ forces $I(\nu) \to 0$ as $\nu \to \infty$ in agreement with experiment. This factor arises from Planck's hypothesis that light is quantized. It is simple to rewrite this equation in terms of the wavelength λ since $\nu = c/\lambda$ and $d\nu = (c/\lambda^2)d\lambda$. That is

$$I(\lambda)d\lambda = \frac{2\pi hc^2}{\lambda^5}\frac{1}{e^{hc/\lambda kT} - 1}d\lambda$$

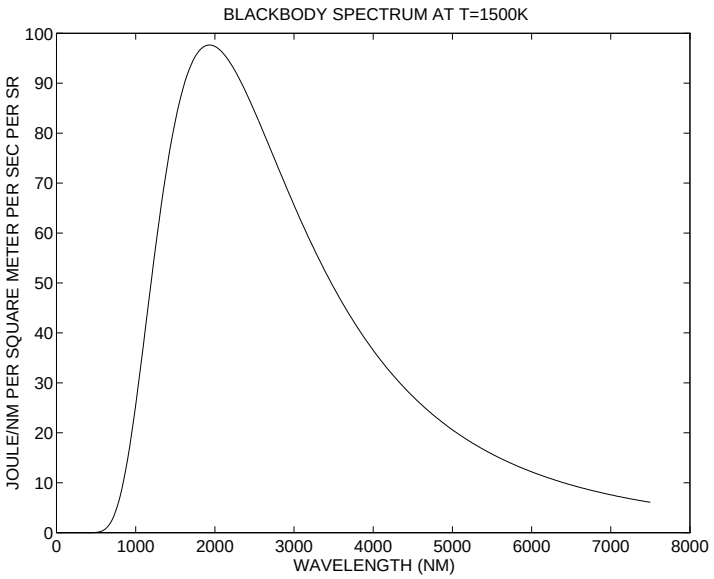


Figure 11.1: Black body radiation spectrum

To make something hot enough to glow, you need to raise its temperature to more than 1500 K or so. Figure 11.1 plots the photon intensity as a function of wavelength λ, at $T = 1500$ K. Only a very small portion of the spectrum at shorter wavelengths extends into the visible region ($\lambda \leq 770$ nm), with most of the photons at the "red" end of the range.

This explains a common phenomenon. As you slowly turn up the heat, you first see a red glow. (You can still get a severe burn even if there is no observable glow, however!) Increasing the temperature further pushes the spectrum to shorter wavelengths (higher energies, so higher frequencies), and the light is no longer biased to the red end. In fact, the entire spectrum is pretty much filled out, and the glow is white. If you increase the temperature further, the object will probably burn up or melt.

This is how an ordinary incandescent light bulb works. It gives off a white light, and can be used to illuminate many things besides this page. If you need light at any particular optical wavelength, shine a light bulb at your experiment and you will get some. Such lamps are very useful in the lab, but the are often limited in intensity. If you need more photons than you can get from a lamp, you have to resort to some discrete line light sources.

11.1.2 Discrete Line Sources

For an partcle sitting in some potential well (such as an electron in an atom), only specific energy states are permitted. This is a consequence of quantum mechanics. If an electron falls into a lower energy state thant the one it is in, the atom emits one photon whose energy is given by the difference in energy between the two levels. It takes around 10 eV to ionize an atom in its ground state, so a zero energy free electron would emit around 10 eV if it fell into the ground state. This is an upper limit to the energy of the photons emitted in atomic transitions, and corresponds to $\lambda = 125$ nm. This is well into the UV, but most photons will be from lower energy transitions and will have longer wavelenghts, so there are typically many transitions in or near the visible region.

The prototypical example of these kinds of transitions are in the hydrogen atom. We put off the explicit discussion of this case until Experiment 6, but the only thing special about one-electron atoms is that they can be solved exactly. Even in atoms with many electrons, these kinds of "electronic" transitions are still of roughly similar (i.e. optical) energies and are very common. Hydrogen, in fact, only has one visible line (in the red) that is not blue or shorter in wavelength. Neon, on the other hand, has dozens of lines in the red and orange, but none in the blue. (Do you recall what a neon light looks like?) Mercury and sodium vapor, in contrast, have several lines throughout the entire region.

The trick to making light from a discrete light sources is to bump electrons out of the atoms. The electrons then fall back into the various holes left behind. You can remove electrons in a variety of ways, but the best is ususally through an electric discharge.

Non-electronic, but nevertheless discrete, transitions are not possible in atoms, but they are possible in molecules. The individual atoms in the molecules can vibrate relative to each other, giving rise to new energy levels similar to those in the harmonic oscillator. The spacing of these energy levels is on the order of a tenth of an eV, so the transitions have $\lambda \approx 10$ μm or so, well into the IR. The atoms may also rotate around one another, giving even more energy levels but this time analagous to the rigid rotor instead of the harmonic oscillator. Here the energy levels are spaced by hundredths of an

eV or smaller, so the emitted photons approach a millimeter in wavelength, which is into the "radio" region and out of the optical.

However, all these rotation and vibrational states are built on top of various electronic excitations, and cause the different electronic transition to become "split" into lines clustering about the central wavelength. These are also covered in more detail in Experiment 6, but for now you should realize how rich they make the spectra of atoms and molecults, even in a tiny portion of the spectrum like the optical region.

11.1.3 Lasers

Lasers are extremely useful light sources that find their way into many applications, including a lot in this course. We won't go through a very detailed explanation of how lasers work, but it is important to understand how they differ from either thermal or standard discrete line sources. Again, Dunlap contains a reasonably complete discussion.

Lasers are examples of discrete line sources. The difference is how the excited electronic state gets de-excited. In standard sources, the excited state decays spontaneously, whenever it feels like it, so the light that comes out is rather random in character. In lasers, however, the de-excitation is "stimulated" and the light is anything but random.

Stimulated decay of an excited state happens whenever a photon with energy very close to the excitation energy passes near the atom. This stimulates the emission of a photons with the same wavelength as the incident photon. Furthermore, the emitted photon will travel in the same direction as the incident photon and will be in phase with it. If these two photons then come upon two more atoms in the same excited state, a total of four photons (all with the same energy and moving in the same direction and with the same phase) are present.

These four photons can produce four more copies of themselves, and so on, and a beam of light is produced. This is called *coherent* light and it is quite intense, among other things, since the strict phase correlation prevents destructive interference among the various photons. This is how we produce

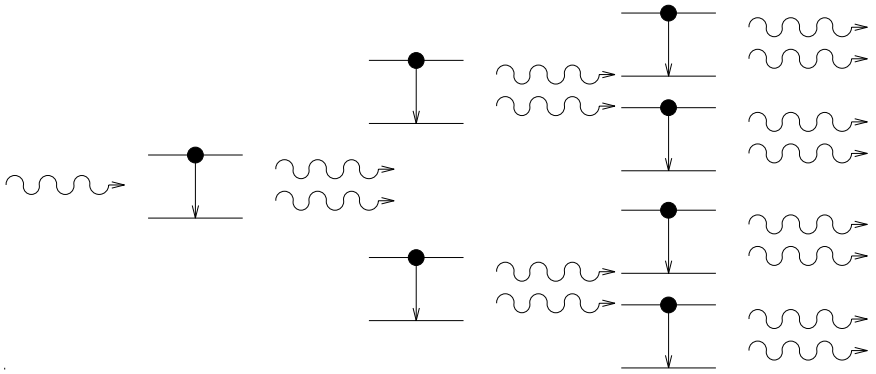


Figure 11.2: Schematic representation of a laser

Light **A**mplification through **S**timulated **E**mission of **R**adiation, or **LASER** for short.

There is a problem, however. For a laser to work, we need more atoms or molecules in the excited state than in the ground state. This can never happen at any finite temperature for matter in thermal equilibrium, so we must find some artificial way of creating this so-called *population inversion*. We can use a lamp or electrical discharge to create excited states, but they will likely decay spontaneously before we can build up a significant number.

The trick is to identify specific states called "metastable" states, in whatever material you can find them in. These metastable states, because of quantum mechanical selection rules, have very long lifetimes, because transition to lower energy states are strongly inhibited. This also means that it is hard to excite the states directly, but we get around that by exciting states *above* the metastable state, and letting them decay to it. Once there are a lot of metastable states populated relative to the ground state, the laser can do its thing. This is shown schematically in Fig. 11.2. This suddenly depletes the metastable state and gives a laser light "pulse", after which the metastable states can be repopulated.

In practice, the lasing[1] material sits in an optically resonant cavity that traps photons inbetween twomirrors for many reflections. This is how the

[1]The acronym LASER has become part of English as many parts of speech, such as the adjective *lasing* material or transition, and as the verb *to lase*.

amplification stage is actually achieved, and the physics of this resonant cavity is actually quite elegant. Experiments in a laboratory optics course can study this in detail. These experiments are not in our course, however, so I won't go into more detail here.

Many types of lasers are commercially available. Probably most common is the He-Ne gas laser, which lases at three lines, one in the visible (632 nm) and two in the IR (1.15 μm and 3.39 μm). The lasing transitions are between different excited states in the Ne atom, and do not include the ground state. This makes it possible to operate continuously, as opposed to pulsed.

Lasers based on semiconductor diodes are becoming very popular and cheap. They operate in the near IR, and even can be tuned over some range of wavelengths, although this feature is likely to give you headaches the first time you try it. Lasers that can be tuned over large wavelength ranges are also available, and generally go by the name of *dye lasers*. However, these are professional devices that are touchy and expensive to operate, so you are not likely to encounter them in a basic experimental physics laboratory course.

11.2 Measuring Light Intensity

If we are going to do experiments with light, we have to learn to measure it. There are several properties of light that can be measured, for example, its intensity, wavelength, or degree of polarization. In this section we discuss ways to measure the intensity, either as energy per unit time or number of photons per unit time.

In order to work with intensity quantitatively, we need to convert it to a voltage level which can be recorded or digitized or whatever. However, the simplest option, namely photographic film, still lets you distinguish "dark" from "light" and has some advantages. We discuss it first.

11.2.1 Photographic Film

Photographic film uses light and chemical reactions to record light intensity. It of course has some obvious drawbacks. For example, it is hard to convert this record into a voltage, although film scanning machines are built for this purpose. Another disadvantage is that it is inconvenient to record large amounts of data this way, unless some fast and efficient scanning method is available. On the other hand, film has some great advantages as well.

First of all, film is economical. You can record light intensity over quite a large area for very little money. Astronomers, for example, photograph large sections of star fields on a single photographic plate giving and accurate and reliable record, all for only a few dollars (in film) per picture.

Secondly, film gives you data that you can easily relate to. Distances between images are true, at least to the extend of your focussing device, and you can remeasure or recheck them easily. There can be an abundance of data on a single photograph, and you can always go back to the same picture if you want to recheck things.

Most importantly, however, film has outstanding position resolution, especially for its price. This resolution is limited by the grain size of the film, and 10 μm is simple to achieve while 1 μm is routine with a little care. What's more, this resolution can be achieved simultaneously over many centimeters of distance. This is almost impossible to achieve with direct electronic means, and can be quite important to astronomers measuring star maps to optical spectroscopists measuring precise wavelengths.

An important tradeoff is between resolution and speed. A film like Kodak Tech-Pan can be used routinely for 1 μm resolution or smaller, but it takes a lot of photons to convert a grain. Thus, such a film is limited to cases of rather large light intensity or where you can afford long exposure times. Somewhat faster films, like Kodak Pan-X, are much faster, and still give resolutions perfectly suitable for most applications.

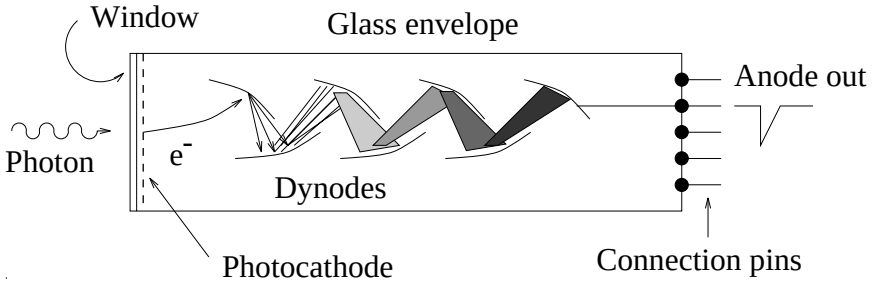


Figure 11.3: How a photomultiplier tube works. The connection pins are used to supply high voltage to the individual dynodes, and to extract the anode output.

11.2.2 Photomultiplier Tubes

The photomultiplier tube (sometime shortened to "phototube" or PMT) is probably the oldest device for converting optical photons directly into electrical signals. It does this with very high efficiency and is very reliable. Some can detect single photons and easily distinguish the signal from background noise. Others are made to measure beams of light. Photomultiplier tubes have been in development for more than 50 years, and have evolved into lots of varieties, some of which are quite sophisticated. The basic operation, though, is quite simple.

The photomultiplier tube is based on two effects, both of which involve the emission of electrons from the surface of materials. The first is the photoelectric effect, where a photon is absorbed by an electron on the material surface. The electron them emerges with some small kinetic energy, thus a photon is "converted" into an electron. The second effect is that when an electron of some moderate energy strikes a surface, some number of electrons are emitted. (This process is called "secondary emission".) Secondary emission is used to multiply the initial electron into a large number of secondary electrons. All of this takes place on surfaces enclosed within an evaculated glass tube, hence the name Photo-Multiplier-Tube.

A schematic photomultiplier tube is shown in Fig. 11.3. The photoelectric effect acts at the front surface, or face, of the PMT, and there one photon is converted into one electron. There is a potential difference of $\sim$100-300 V

between the face and the first "stage" of the tube, and this accelerates the electron. Then this 100-300 eV electron strikes the first stage, it emits more electrons, which are accelerated to the next stage, and so on. These materials which act as stages are called "dynodes" since they act both as acceptors of electrons (i.e. anodes), and emitters of electrons (i.e. cathodes). After several (usually between 6 and 14) stages, a significant number of electrons emerge in place of the incident photon. Electrical connections are made with the outside world by pins which penetrate the glass envelope on the end.

The front window of the PMT is made of glass, or some other transparent material. A thin layer of some optically active material is evaporated on the inner surface of the window. This layer, called the photocathode, is semi-transparent and is usually brownish in color. If the tube breaks and air fills the inside, the photocathode oxidizes away and the brownish color disappears. In this case, the photomultiplier tube will never work again.

A photon incident on the window penetrates it if it can. In fact, glass window tubes become very inefficient in the near UV because photons with wavelengths below 350 nm or so are quickly absorbed in ordinary glass. Special UV transmitting glass is available on some photomultipliers tubes, and this can extend the range down to 250 nm or so. To get further into the UV, special windows made of quartz or CaF_2 are necessary, and the devices become very expensive.

If the photon penetrates the window, it reaches the photocathode and has a chance to eject en electron through the photoelectric effect. Recall that in the photoelectric effect, a photon of energy $h\nu$ gives rise to an electron of kinetic energy K via

$$K = h\nu - \phi$$

where ϕ is called the "work function" and represents the energy needed to remove the electron from the surface. Several different materials are used for photocathodes, but all are designed to have work functions small enough so that optical photons can eject electrons. It is in fact hard to find materials for which ϕ is less than $\approx$2 eV, so photomultipliers become quite insensitive at the red end of the visible spectrum.

The probability that an incident photon ejects an electron from the photocathode is called the "Quantum Efficiency" or QE. It is clearly a function

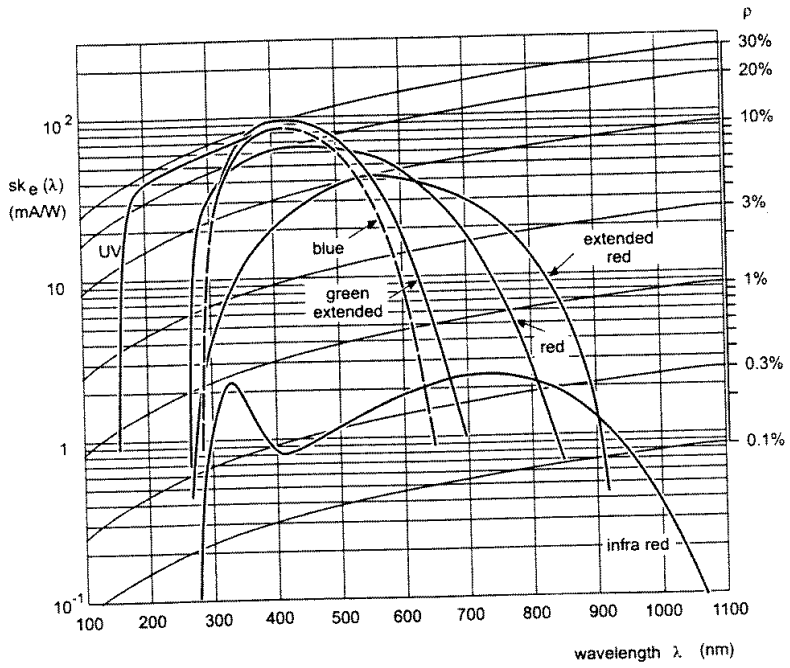


Figure 11.4: Spectral sensitivity (sk_e) and quantum efficiency (ρ) for some photomultiplier tube windows and photocathodes. From the Philips photomultiplier tube handbook.

of wavelength λ, tending to zero both for $\lambda \leq$UV, and $\lambda \geq$red. It is also a function of window and photocathode material for the same reasons. Figure 11.4, taken from the Philips photomultiplier tube handbook, shows the "spectral sensitivity" S in mA/W for various combinations of windows and photocathodes. Manufacturers tend to quote S rather than QE since it is closer to what the actually measure. By shining so much light enery per unit time (P) on the face of the PMT, and measuring the current (i) of electrons coming off the photocathode, they determine

$$S \equiv \frac{i}{P} = \frac{N_{ELECTRON} \times e/t}{N_{PHOTON} \times h\nu/t} = \frac{N_{ELECTRON}}{N_{PHOTON}} \times \frac{\lambda}{hc/e} = QE \times \frac{\lambda}{1.24}$$

where S is written in mA/W and λ is in nm. Curves of constant QE are drawn in on Fig. 11.4. Typical quantum efficiencies are maximum in the blue region and range upwards of 25% or so.

Now let's return to Fig. 11.3 and see how the photomultiplier tube amplifies the signal. The incident photon has ejected an electron with something like an eV of kinetic energy. This electron is accelerated to the first dynode, and strikes it. The dynodes are constructed out of materials that give a significant mean number of electrons out for each that strike the surface. This multiplication factor δ is a strong function of the incident electron energy, and is roughly linear with energy up to a few hundred eV or so for most materials used in PMTs.

There is clearly some randomness associated with the operation of a photomultiplier. The quantum efficiency, for example, only represents the probability that a photon will actually eject an electron. The result is that the output voltage pulse corresponding to an input light signal will have random fluctuations about a mean value. We therefore frequently talk in terms of the "mean number of photoelectrons" N_{PE} that correspond to a particular signal.

Assuming that Poisson statistics dominate, this number will dominate the size of the fluctuations, since the numbers of electrons ejected in subsequent stages will be larger. That is, the fractional RMS width of the signal fluctuations should be given by $\sqrt{N_{PE}}/N_{PE} = 1\sqrt{N_{PE}}$. This can be particularly important if the signal corresponds to a very low light level, i.e. a small value of N_{PE}. In this case, there is a probability $e^{-N_{PE}}$ that there will be no photoelectrons ejected and the signal will go unobserved.

The gain g of a photomultiplier tube is the number of electrons out the back (i.e. at the anode) for a single incident photon. So, for an $n-$stage tube,

$$g = \delta_1 \times \delta_2 \cdots \times \delta_n \approx \delta^n$$

where we tacitly assume that δ is the same at each stage, i.e., all dynodes are identical and the potential difference across each stage is the same. If δ is proportional to V, then these assumptions[2] predict that g is proportional to V^n. Thus if you want to keep the gain constant to 1% in a 10-stage photomultiplier tube, you must keep the voltage constant to 0.1%. This is not particularly easy to do.

[2]These assumptions are almost always wrong. We are using them just to illustrate the general performance of the PMT. For actual gain calculations, you must know the specific characteristics of the PMT.

The accelerating voltage is usually applied to the individual stages by a single external high voltage DC power supply, and a multi-level voltage divider. The voltage divider has output taps connected to each stage through the pins into the tube. This is connected to the circuit which extracts the signal from the anode. The extraction circuit and voltage divider string are housed together in the photomultiplier tube "base", and their design will vary depending on the application. The base is usually some sort of closed box with a socket which attaches to the tube pins. Two examples of base circuits, taken from the Philips photomultiplier tube handbook, are shown in Fig. 11.5. If the signal is more or less continuous, and for example a meter reads the current off the anode to ground, you must use the negative high voltage configuration so that the anode is at (or near) ground. If the output is pulse-like, such as when "flashes" of light, or perhaps individual photons, are detected intermittently, then it is usually best to use the positive high voltage configuration since that leaves the photocathode at ground. In this case, and RC voltage divider at the anode output allows fast pulses to reach the counter, but the capacitor protects the downstream electronics from the high DC voltage.

No matter what circuit is used, either those in Fig. 11.5 or otherwise, you must choose the resistor values carefully. Although the stage voltages only depend on the relative resistor values, you must make sure the average current passing through the divider string is much larger than the signals passing through the PMT. Otherwise, the electrons in the multiplier will draw current through the resistors and change the voltage drop across the stage. Even if this is a small change, it can affect the gain by a lot since the gain depends on voltage to a large power.

On the other hand, you can't make the resistors arbitrarily small so the divider current gets very large, because this would require a big and expensive high current, high voltage DC power supply. What's more, the power dissipated in the divider string, i.e. i^2R, gets to be enormous making things very hot. Tradeoffs have to be made, and always keep your eye on the gain.

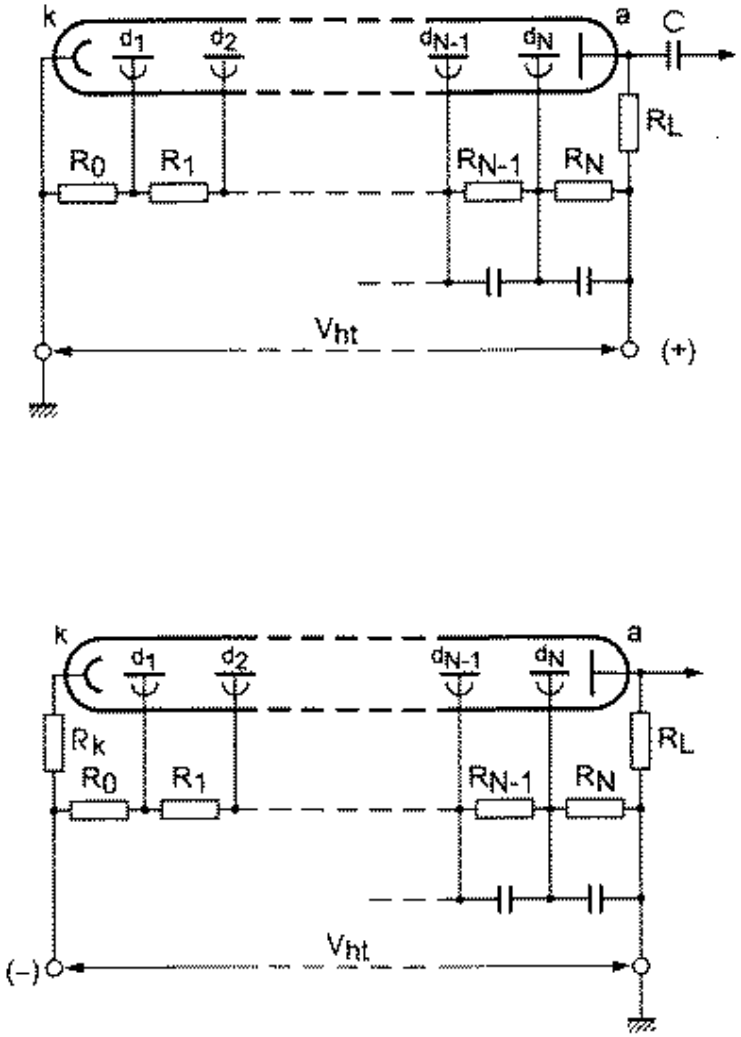


Figure 11.5: Typical photomultiplier base circuits. The upper figure shows connections for a positive high voltage configuration, while the lower shows negative high voltage.

11.2.3 Photodiodes

Photodiodes are an alternative to photomultipliers. Both turn light directly into electrical signals, but there are distinct differences. First, let's learn how photodiodes work.

Recall our discussion about diodes in Sec. 2.4.1. A piece of bulk silicon is essentially an insulator. Only thermally excited electrons can move to the upper, empty energy band to conduct electricity, and there are few of them at room temperature. By adding n-type or p-type dopants, lots more charge carriers can be created, and it's a much better conductor. A piece of silicon doped n on one end and p on the other, a pn junction, only conducts in one direction. If a "reverse" voltage is applied, only a tiny current flows, due to the small number of thermally excited electrons.

A photodiode uses light (photons) to excite more electrons than those excited thermally. This is possible if the photon energy is larger than the band gap. Thus, the "reverse" voltage current would increase if you shine line on the diode. This is the principle of the photodiode.

The actual mechanism is a bit more complicated, because of how excited electrons actually conduct. So, for example, for a given applied voltate, the output current is not very linear with intensity. That is, if you double the light intensity, the output current does not change by quite a factor of two (over the "noise" from the thermal electrons). Furthermore, a photodiode can work if there is *no* applied voltage, reverse or otherwise. This all means that you have to calibrate your photodiode response to some degree if you really want a quantitative measure of the light intensity.

A popular form of photodiode puts a large region of pure, or "intrinsic", silicon inbetween the p and n ends. this increases the active area and decreases the thermal noise current. These photodiodes are called p-i-n or "pin" diodes.

Now let's look at a clear advantage that photodiodes have over phototubes. The energy gap in silicon is 1.1 eV, so photons with wavelengths up to ≈ 1.1 μm can be detected. This is well past red and into the IR. Photomultiplier tubes peter out at around 600 nm (see Fig. 11.4) or so because

of the work function of the photocathode. The band gap of germanium (another popular semiconductor) is 0.72 eV, so germanium photodiodes reach $\lambda \approx 2$ μm.

So, if you need to detect red light, you probably want to use a photodiode, and not a photomultiplier tube.

Another big advantage of photodiodes over photomultiplier tubes is cost. A photomultiplier tube with voltage divider circuitry, high voltage supply, and mechanical assemblies can easily cost upwards of $2000. A photodiode costs around $1, and is very easy and cheap to instrument.

Photodiodes can also be made with very small active areas (say 50 μm across). This along with their low cost makes "photodiode arrays" practical. These are lines of photdiodes, separately instrumented, that measure photon position along the array. Such things are frequently used in spectrographic instruments. A typical example might by 1024 25 μm$\times$2.5 mm photodiodes arranged linearly in a single housing with readout capability. The cost for such a thing is typically several $K.

Of course, photomultipliers have some advantages over photodiodes. The biggest is the relative signal-to-noise[3] ratio. A μW of incident light power gives around a 1 μA signal in a photodiode, but around 1 A in a photomultiplier tube. This big enhancement in signal is due to the large gain ($\sim 10^6$ or more). Thermally excited electrons are plentiful in a photodiode, but rarely does such an electron spontaneously jump off the photocathode in a photomultiplier. Therefore, the noise is a lot larger in a photodiode. Thus, the signal-to-noise ratio is much worse in a photodiode.

So, if you need to detect very low light intensities ("photon counting" for example), you probably want to use a photomultiplier tube, and not a photodiode.

Photomultipliers also give a more linear response, particularly if care is given to the base design. Some of these relative advantages and disadvantages are shown in Tab. 11.2. Another advantage of photodiodes is that they

[3]We will discuss the general concept of noise in a later chapter. For now, just take it at face value. Signal is good. Noise is bad.

Table 11.2: Photomultiplier tubes versus photodiodes.

If you are interested in...	Then your choice should likely be Photomultiplier	Photodiode
Low Cost		$\surd$
Red Sensitivity		$\surd$
Low Intensity	$\surd$	
Linearity	$\surd$	

work in high magnetic fields. Photomultiplier tubes rely on electrons with $\approx 100 - 300$ eV energy to follow field lines to the dynodes. A few gauss magnetic field disturbs the trajectories enough to render the PMT useless. In most cases, magnetic shielding solves the problem, but sometimes this is impractical and photodiodes are used instead.

Finally, we mention that photosensitive transistors, or phototransistors, are also available. They use the natural amplification features of the transistor to get a ~ 100 times larger signal than the photodiode. Of course, the transistor also amplifies the noise, so there is no improvement in the sensistivity at low intensities.

11.3 Exercises

1. Consider blackbody radiation.

a. Show that the wavelength at which the intensity of a blackbody radiator is the greatest is given by "Wien's Displacement Law":

$$\lambda_{MAX}\ (\text{m}) = \frac{2.9 \times 10^{-3}}{T\ (\text{K})}$$

Hint: You will need to solve an equation like $xe^x/(e^x - 1) = A$ for some value A. If $A \gg 1$ then this is trivial to solve, but you can be more exact using MAPLE or MATLAB. In MATLAB you would use the

"function function" fzero to find the place where $f(x) = A(e^x - 1) - xe^x$ crosses zero.

b. Stars are essentially blackbody radiators. Our sun is a "yellow" star because its spectrum peaks in the yellow portion of the visible. Estimate the surface temperature of the sun.

2. A particular transition in atomic neon emits a photon with wavelength $\lambda = 632.8$ nm.

a. Calculate the energy E of this photon.

b. Calculate the frequency ν of this photon.

c. An optical physicist tells you the "linewidth" of this transition is $\Delta\nu = 2$ GHz. What is the linewidth ΔE in terms of energy?

d. Use the Heisenberg Uncertainty Principle to estimate the lifetime Δt of the state which emitted the photon.

e. How far would a photon travel during this lifetime?

f. Suppose the neon is contained in a narrow tube 50 cm long, with mirrors at each end to reflect the light back and forth and "trap" it in the tube. What is the nominal "mode number" for 632.8 nm photons, that is, the number of half-wavelengths that fit in the tube?

g. What is the spacing in frequency between the nominal mode number m, and the wavelength corresponding to the mode $m + 1$?

h. Compare the mode spacing $\delta\nu$ (part G) with the line width $\Delta\nu$.

i. What is this problem describing?

3. Estimate the "transit time" for a typical photomultiplier tube. That is, how much time elapses between the photon ejecting an electron from the photocathode, and the pulse emerging from the anode. Assume the photomultiplier has 10 stages and 2000 V between cathode and anode, divided equally among all stages, and that the dynodes are each separated by 1 cm.

4. Some high quality photomultipliers can detect the signal from a single photoelectron, and cleanly separate it from the background noise. Such a

PMT is located some distance away from a pulsed light source, so that on the average, the PMT detects $\langle N_{PE} \rangle$ photoelectrons. If $\langle N_{PE} \rangle \ll 1$ and N_0 pulses are delivered, show that the number of pulses detected by the photomultiplier is given by $\langle N_{PE} \rangle N_0$.

5. A photomultiplier tube observes a flash of green light from an Ar^+ laser. (Assume the photons have wavelength $\lambda = 500$ nm.) The photomultiplier is a 10-stage Philips tube, with an "green extended" photocathode. The voltages are set so that the first stage has a secondary emission factor $\delta_1 = 5$, while the other nine stages each have $\delta = 2.5$. The laser delivers some huge number of photons to a diffusing system which isotropically radiates the light, and only a small fraction of them randomly reach the photomultiplier. On the average, 250 photons impinge on the window for each flash of the laser.

a. What is the average number of electrons delivered at the anode output of the photomultiplier tube, per laser flash?

b. Assume these electrons come out in a rectangular pulse 20 ns wide. What is the height of the *voltage* pulse as measured across a 50Ω resistor?

c. You make a histogram of these pulse heights. What is the standard deviation of the distribution displayed in the histogram?

d. Suppose the photomultiplier tube is moved four times farther away from the source. For any given pulse of the laser, what is the probability that no photons are detected?

Ch 12

Experiment 6: Atomic Spectroscopy

The formulation of quantum mechanics in the Schrödinger equation is beautiful and elegant. Unfortunately, however, their are very few problems that can be solved exactly (more or less) which correspond to physical systems that actually exist.

The hydrogen atom, or more exactly, atoms with a single electron, are probably the best example of a true, measurable, physical system in which the beauty of quantum mechanics can be tested without resorting to approximation techniques and models. In fact, many of the early successes of quantum mechanics came in the study of atomic spectroscopy, with hydrogen providing some of the most crucial tests of the theory.

Some other solvable problems with the Schrödinger equation are the harmonic oscillator, and the rigid rotor. Although ideal cases of these things are hard to find in nature, good approximations are provided by the study of diatomic molecules.

The energy levels involved in atomic physics typically give rise to transitions in the few eV range. Consequently, the photons that are emitted are in the visible region, and our studies of light production and detection will prove to be useful. In this experiment we will actually use high resolution

spectroscopy to study the photon wavelengths to high precision. For example, the wavelengths are slightly different in the deuterium atom as opoosed to hydrogen because of the larger nuclear mass. You will be able to resolve that difference.

The physics surrounding atomic and molecular spectroscopy is a large field. For a good general reference, I suggest

- *Introduction to the Structure of Matter*,
  John J. Brehm and William J. Mullin, John Wiley and Sons (1989),
  Chapters 3, 7, 9, and 10

- *Quantum Physics*, Robert Eisberg and Robert Resnick,
  John Wiley and Sons, Second Edition (1985),
  Chapters 4 and 7

The experiments we will do are more or less standard in the undergraduate laboratory, and there are several good books available which describe such experiments. A few of these are

- *The Art of Experimental Physics*,
  Daryl W. Preson and Eric R. Dietz, John Wiley and Sons (1991)
  Experiments 12 and 13

- *Physical Chemistry; Methods, Techniques, and Experiments*
  Rodney J. Sime, Saunders College Publishing (1990)
  Experiments 30-33.

- *Experiments in Modern Physics*,
  Adrian C. Melissinos, Academic Press (1966)
  Sections 2.1 through 2.4.

The book by Sime is particularly useful for us since the experiments use the same spectrograph as we have in our laboratory.

12.1 Energy Levels of the Hydrogen Atom

We'll start with the Bohr model of one-electron atoms. For most of the measurements you will make, this model correctly predicts the results, and in fact gives the same answers as the Schrödinger equation. This approach is actually worked out in detail by Brehm and Mullin.

We will derive the energy of the atom, and impose simple quantization rules on the result. Assume for now that the nucleus of the atom is infintely heavy compared to the electron, and that the electron moves nonrelativistically. We will examine these assumptions soon, but they are in fact quite good in general. The total mechanical energy of the electron is

$$\begin{aligned} E &= K + V \\ &= \frac{1}{2}mv^2 - \frac{1}{4\pi\varepsilon_0}\frac{Ze^2}{r} \end{aligned} \tag{12.1}$$

Here m, v, and $-e$ are the electron's mass, velocity, and electric charge, $+Ze$ is the charge on the nucleus, and r is the "orbital radius" of the electron. The potential energy, of course, is just the Coulomb attractive potential between the electron and the nucleus.

Before we get into quantization, let's work with this equation a bit. We can relate the velocity v to the other variables by applying $F = ma$, where F is the Coulomb force and a is the centripetal acceleration. That is

$$\frac{1}{4\pi\varepsilon_0}\frac{Ze^2}{r^2} = m\frac{v^2}{r}$$

which implies that

$$v^2 = \frac{1}{m}\frac{1}{4\pi\varepsilon_0}\frac{Ze^2}{r} \tag{12.2}$$

If we plug this into Eq. 12.1 we get

$$E = \frac{1}{2}\frac{1}{4\pi\varepsilon_0}\frac{Ze^2}{r} - \frac{1}{4\pi\varepsilon_0}\frac{Ze^2}{r} = -\frac{1}{2}\frac{1}{4\pi\varepsilon_0}\frac{Ze^2}{r} \tag{12.3}$$

Now we can impose quantization rules on the energy just by considering the implications for the orbital radius r.

Bohr's quantization rules are simple and elegant. The electron is also a wave, with wavelength $h/p = h/mv$. As the electron moves around in its orbit, the path length must be such that the head of the wave "links up" with the tail and the wave pattern keeps repeating itself. If this wasn't the case, then the wave would interfere with itself as the electron orbited, and the wave would disappear, and so would the electron.

Bohr's quantization condition is, therefore, that the circumference of the orbit be an integral number n of wavelengths of the electron. That is, $2\pi r = n(h/p)$ or

$$r = n\frac{\hbar}{mv} \tag{12.4}$$

where $\hbar \equiv h/2\pi$. This actually has a deeper physical significance. The quantity rp is the angular momentum l of the electron, so this equation in fact says that $l = n\hbar$, that is, the angular momentum is quantized. This is really just the beginning of a very interesting story about angular momentum and quantum mechanics, but we won't cover it here.

Combining Eq. 12.4 with Eq. 12.2 gives us

$$\frac{n^2\hbar^2}{m^2r^2} = \frac{1}{m}\frac{1}{4\pi\varepsilon_0}\frac{Ze^2}{r}$$

and therefore

$$\frac{1}{r} = \frac{m}{n^2\hbar^2}\frac{1}{4\pi\varepsilon_0}Ze^2$$

which is a quantization relation for r that does not depend on v. Finally, insert this expression in Eq. 12.3 and get

$$\begin{aligned} E &= -\frac{1}{2}\frac{1}{4\pi\varepsilon_0}Ze^2 \times \frac{m}{n^2\hbar^2}\frac{1}{4\pi\varepsilon_0}Ze^2 \\ &= -\left[\frac{mZ^2e^4}{2(4\pi\varepsilon_0)^2\hbar^2}\right]\frac{1}{n^2} \end{aligned} \tag{12.5}$$

These are the quantized energy levels of the hydrogen atom. That is, the electron can only have energies described by this formula with $n = 1, 2, 3, \ldots$. These energy levels are plotted in Fig. 12.1. *Notice in particular that the energy levels get closer and closer together near the top of the potential well.* This pattern will be clearly apparent in this experiment.

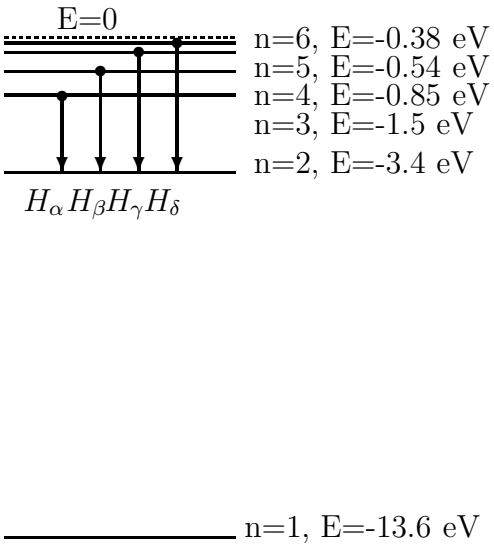


Figure 12.1: Energy levels of the hydrogren atom, and the transitions which make up the Balmer series of visible wavelength lines. Transitions to the $n = 1$ state are also possible, and these lines are called the Lyman series, but the wavelengths are all in the far ultraviolet.

For the hydrogen atom ($Z = 1$), the expression in brackets for Eq. 12.5 works out to be 13.6 eV. This is the energy required to take an electron in the ground state ($n = 1$) and separate it from the nucleus completely ($E = 0$). That is, this is the binding energy of the hydrogen atom. For atoms with more electrons, you imagine that the energy needed to separate the "outermost" electron from the nucleus and the $Z-1$ remaining electrons is about the same value, since the other electrons "shield" the outermost electron from all but unit of charge on the nucleus. In fact, the energy needed to separate one electron from an atom (the "ionization potential") is pretty close to 10 eV for most atoms.

So what's the experiment? We need to add one more ingredient, namely transitions between the energy levels. If you prepare an atom in one of the excited states ($n > 1$), then the electron will make transitions down to the lower states. Each transition emits a photon of energy $h\nu$ which must equal the difference in energy between the initial and final states. If the electron starts from the energy level with $n = n_i$ and ends up with $n = n_f$ then

$$h\nu = E_{n_f} - E_{n_i}$$

We use Eq. 12.5 to rewrite this expression, but the result is usually expressed in terms of the photon wavelength λ instead of the frequency $\nu = c/\lambda$. You find

$$\frac{1}{\lambda} = Z^2 R_\infty \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right) \tag{12.6}$$

where the *Rydberg constant* R_∞ is defined as

$$R_\infty \equiv \frac{me^4}{8\varepsilon_0^2 ch^3} \tag{12.7}$$

The subscript "∞" refers to the assumption that the nucleus is infinitely heavy. The Rydberg constant is actually known quite accurately. It's value is

$$R_\infty = 10973731.534 \text{ m}^{-1}$$

Note that the hydrogen binding energy is just $hcR_\infty = 13.6$ eV.

You will measure several of these wavelengths in this experiment. You should note, however, that only a few of the many combinations are accessible, because you will be detecting essentially visible photons. (See Table. 11.1.) In fact, all of the transitions you will observe correspond to $n_f = 2$,

and the series of lines corresponding to $n_i = 3, 4, \ldots$ is called the Balmer series.[1] The longest wavelength line (the $3 \rightarrow 2$ transition) is red, but all the others are blue or violet. These transitions are also shown in Fig. 12.1.

12.1.1 Corrections

There are various corrections due to the simple Bohr formulas. You will investigate one of these in particular, namely the effect of a finite nuclear mass, and it can be evaluated using a straightforward extension of the simple rules. The other corrections actually led to the breakdown of the Bohr formula, and the development of quantum mechanics based on the Schrödinger equation. You won't likely measure these effects in this experiment, but we will at least mention the physics here.

Finite Nuclear mass

The proton mass m_p is much larger than the electron mass, in fact $m_p/m = 1836$, and all the other nuclei are even heavier. Therefore, the "infinite nuclear mass" assumption is a pretty good one. On the other hand, optical spectroscopy experiments can be very precise (look at all the significant figures on R_∞), so you might think you could see the effect of a finite nuclear mass in this experiment. In fact, you can, and you will measure the difference in spectra for hydrogen (nuclear mass $m_p = 938.3$ MeV/c^2) and for deuterium (nuclear mass $m_d = 1875.6$ MeV/c^2).

The finite nuclear mass means that instead of the electron revolving around the nucleus, both the electron *and* the nucleus revolve about their common center of mass. The quantity r still refers to the distance between the electron and the nucleus, but the orbit radii are actually $r_e = Mr/(m + M)$ and $r_M = m/(m + M)$ for the electron and nucleus respectively. Since the nuclear mass M is much larger than the electron mass m, $r_e \approx r$ and $r_M/r_e << 1$, which is the essence of the infinite mass approximation.

[1] The other series corresponding to other values for n_f are named after the other fellows who discovered the lines, but none of these lines are in the visible.

We now quantize the energy as before, but we need to write E including the kinetic energy of the nucleus, and form the quantization condition using the *total* angular momentum which includes the nuclear contribution. The details are worked out in Brehm and Mullin, and the result is that Eq. 12.6 is unchanged except that the Rydberg constant R_∞ is replaced by

$$R_M = \frac{\mu}{m} R_\infty \tag{12.8}$$

where $\mu \equiv Mm/(M+m)$ is called the reduced mass. The key to measuring this effect is to be able to measure the small difference in wavelengths corresponding to R_p and R_d for hydrogen and deuterium.

Relativistic Effects

By the time the Bohr model was being developed in 1913, people were pretty much convinced that Einstein's theory of special relativity was right. Trying to incorporate relativity into Bohr's model became a problem, and eventually led to its downfall. Let's first estimate how big a problem you expect this to be.

Relativity should become important when the electron speed v approaches the speed of light c, so let's evaluate v/c. Substituting Eq. 12.4 into Eq. 12.2, you find

$$\frac{v}{c} = \frac{1}{m}\frac{1}{4\pi\varepsilon_0} Z e^2 \frac{m}{n\hbar c} = Z\frac{\alpha}{n}$$

where

$$\alpha \equiv \frac{1}{4\pi\varepsilon_0}\frac{e^2}{\hbar c} = \frac{1}{137.036} \tag{12.9}$$

is called the *fine structure constant*. In other words, for the hydrogen atom, the electron speed is always less than 1% of c, so you wouldn't expect relativistic corrections to be very much larger than this. Of course, in heavy (i.e. large Z) one-electron atoms, the velocity can actually get quite large compared to c, and you can't expect the Bohr formula to work well.

I want to take a moment to talk about α. Just about any physicist you know can tell you that "α is around 1/137". It is a very fundamental quantity that you will encounter more and more as you study physics. At

the very least, it helps you remember some of the formulas we've derived. For example, the binding energy of a one-electron atom is just $\frac{1}{2}Z^2\alpha^2 \cdot mc^2$, and the Rydberg constant $R_\infty = \frac{1}{2}\alpha^2(mc/h)$.

It is straightforward to incorporate the effect of special relativity into the one-electron atom, once you've solved the Schrödinger equation for the problem without relativity. This is called *perturbation theory*, and the procedure is outlined in Preston and Dietz, as well as other places. The result is that the individual lines are "split" according to the angular momenta of the initial and final atomic states. This splitting is called *fine structure*, and it amounts to around $\Delta\lambda = 0.1$ Å in the Balmer series.

Spin-Orbit Splitting

Relativistic corrections are only one contribution to the fine structure splittings. The second contribution, which is about the same size as from relativity, is due to the *spin-orbit interaction*.

The electron has some internal angular momentum we call *spin*. This internal angular momentum shows up as a magnetic dipole moment on the electron. In other words, the electron is like a tiny bar magnet. In addition, there is the electron "current" due to the electron orbiting about the nucleus. This current sets up a magnetic field all around the atom, and the electron spin interacts with this magnetic field. This is the spin-orbit interaction, and it is essentially of the form $\vec{\mu} \cdot \vec{B}$ where $\vec{\mu}$ is the spin dipole moment of the electron, and $\vec{B}$ is the magnetic field set up by the electron orbit.

Depending on the relative orientation of $\vec{\mu}$ and $\vec{B}$, which is quantized according to quantum mechanics, the magnitude and sign of the spin-orbit interaction will be different for different angular momentum states. This leads to the spin-orbit splitting. Once again, it is about the same size as relativistic effects, i.e. $\Delta\lambda = 0.1$ Å or so.

Higher Order Corrections

As you look at the spectral lines with higher and higher resolution, you discover more (and smaller) splittings. Each of these has physics associated with them, and some of this physics is very profound.

One example is the so-called *hyperfine splitting*. (See Brehm and Mullin, Sec. 8-12.) This is the interaction between the spin magnetic moments of the electron and the proton, for the hydrogen atom, or for the nucleus in general assuming it has a nonzero spin. The splittings caused by the hyperfine interaction in hydrogen are on the order of 0.01 Å. Transitions between hyperfine levels are in the radio frequency range, and a particular transition in hydrogen ($\lambda = 21$ cm) is famous to radio astronomers who use it to identify hydrogen in hard-to-see regions of the galaxy.

One of the most profound effects in atomic physics is the *Lamb shift*, named for Willis Lamb who discovered it in 1947. Up until that time, a synthesis of quantum mechanics and special relativity, written down by Paul Dirac, was able to correctly predict all the structures observed in the hydrogen spectrum. The Lamb shift, an unexpected splitting at around the 0.01 Å level (relative to optical spectroscopy), was inconsistent with Dirac theory. It's solution turned out to hinge on a new formulation that explicitly includes the radiation field along with electrons, called Quantum Electrodynamics.

12.2 Measurements

Your first goal is to determine the Rydberg constant from the Balmer series in hydrogen. You will do this by measuring the Balmer spectral lines from atomic hydrogen, and fitting the wavelengths to Eq. 12.6. Next, you will calculate the deuteron to proton mass ratio by measuring the isotope shift between the lines of atomic deuterium and hydrogen. You need to make precise measurements of the wavelengths to do all this.

Light from the various elements is produced in a gas discharge tube. Wavelength measurements are made in either one of two ways. We have a

Baird model SB-1 1.5 m grating spectrograph, which exposes a strip of 35 mm film to a dispersed spectrum. A computer controlled table scans the film and measures the grain density as a function of position. The second method uses the Jarrell–Ash 1 m Czerny-Turner Scanning Spectrometer which scans through wavelengths on command. Data is taken directly from a photomultiplier tube as wavelengths pass over a slit as the grating is turned. Results from the two techniques will be compared in Fig. 12.8. Note that the film is darkened at a line position, so less light passes through and the scanner gives a smaller signal when crossing a line. The Jarrell–Ash, however, detects the light directly and the electronics provides something proportional to the light intensity.

The light source works by placing a very high voltage across the ends of the discharge tube, and you will get a nasty shock from it if you are not careful. Always make sure the voltage is off when you change tubes. Be careful not to move the source or any other part of the apparatus while you change discharge tubes, since a small change in position can have a big effect on where the image ends up on the film. Try never to look directly at the discharge tube when it is on, because some of the tubes have rather intense UV light that you cannot see.

12.2.1 Procedure: Baird Spectrograph

A diagram of the Baird spectrograph is shown in Fig. 12.2, and an expanded view of the film window and slit assembly is shown in Fig. 12.3. The discharge tube is controlled through a timer that lets you preset the exposure time. Light from the discharge tube enters through the slit and is focussed onto the concave diffraction grating in the rear. The grating disperses light according to wavelength and focusses it onto the film holder adjacent to the entrance slit. You can adjust the vertical position of the film image by moving the position of the Hartmann slide, and you can put six or so different exposures on the same piece of film. It is also possible to run a HeNe laser beam through the spectrograph for calibration purposes. Furthermore, there is a "reference line" you can superimpose on the film by pressing a button near the film holder.

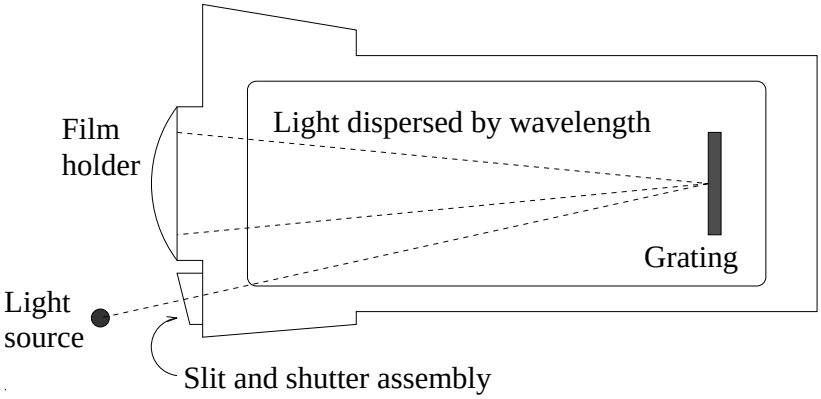


Figure 12.2: Diagram of the Baird SB-1 Grating Spectrograph.

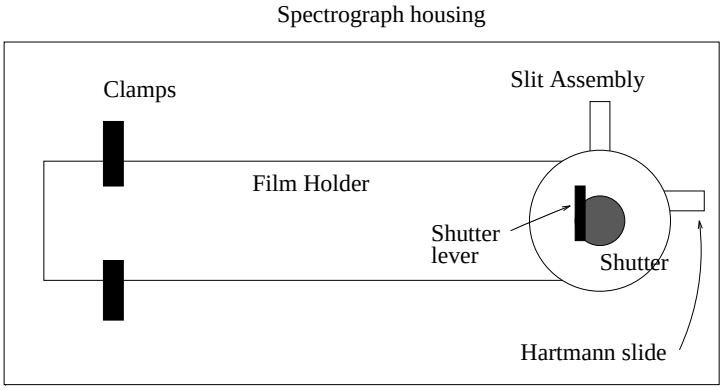


Figure 12.3: Diagram of the window and slit assembly.

The wavelength range is 370 nm to 740 nm, pretty much covering the visible spectrum. The dispersion at the film focus is around 15Å/mm, but you will determine this yourself by calibrating the spectrograph. You can set the slit width to 10 μm, 32μm, or 60μm, and to get the best resolution, you may need to experiment with different widths. The shutter can be operated manually, and you should keep it closed if you are not exposing the source, even if the lights are off.

Before exposing film, you should play around with the roughed-up plexiglas sheet that fits in the film holder. You can see the lines with some care, and you make adjustments of the slit, Hartmann slide, source position, and so on without having to actually develop film.[2] A good source tube to use is neon, which gives you many bright lines mainly in the orange and red. If you use the mercury tube, you will see a strong yellow doublet at 577 nm and 579 nm, and a relatively bright green line at 546 nm.

The film holder uses 10-inch strips of standard 35-mm photographic film, cut from a 100 foot roll. Kodak Tech-Pan is a good film to use, and that is what the laboratory should be stocked with. This film is sensitive to all wavelengths, so the room must be in complete darkness when you are handling the film. With the lights out, remove a length of film from the bulk film winder, and slide it into the film holder. The concavity of the film should match the concavity of the holder. Place the film holder over the window, and secure it with clamps. At this point, light from the discharge tubes may be exposed to the film through the shutter. Spend a film strip or two making measurements with the different slit widths, and varying the exposure time between a few seconds and several minutes. After you develop the film and look at the result, you will be better able to judge the slit widths and exposure times for your "final" set of data. *Hint: Cut the film a bit long so an inch or so hangs out past the blue end of the holder. It will be fully exposed and a good tag of which end is which.*

Remember that you can move the exposure vertically on the film using the various settings of the Hartmann slide. You likely want to include some or all of the following discharge tube combinations in your exposures:

[2]To quote the users manual for the spectrograph, *the Fixed slit Assembly is a very delicate component. Handle it carefully.* If you would like a copy of the users manual, please ask me or the TA.

- Helium+HeNe Laser
- Hydrogen+HeNe Laser
- Mercury+HeNe Laser
- Hydrogen+Helium
- Hydrogen+Mercury
- Hydrogen+Deuterium
- Nitrogen+HeNe Laser

The helium and mercury lines will be used to calibrate the spectrograph, that is, to convert position on the film to wavelength. The HeNe laser serves a similar purpose, giving you a single strong line at 6328 Å, and it might be a good idea to add this line to all your exposures. You might also want to press the reference line button for a moment, to help you orient the film after you've developed it. The hydrogen lines will allow you to determine R_H, and the simultaneous exposure of hydrogen plus deuterium allows you to measure the small difference between these two isotopes, although you will need pretty good resolution to cleanly separate the two. This is one of the things to aim for when you practice in the beginning. Nitrogen will give you a very complicated but interesting spectrum that you can analyze to learn about the diatomic nitrogen molecule. (See Sec. 12.3.)

Developing the film

The chemicals you need to develop the film should all be premixed for you, but specific directions are available in any case. You don't want to get a lot of these chemicals on your skin, so handle them with some care. Remember, all developing must be done in total darkness. We use HC-110 to develop the Tech-Pan film, and it is a good idea to make sure you are using a new (i.e. less than a year old) bottle of it when you are developing.

Set out four developing tays, each containing around a half gallon or so of developer, stop bath, fixer, and photoflo, respectively and in that order. Place the film in the developing tray and agitate for 8-10 minutes at room temperature. This may be longer or shorter if the temperature is cooler or

warmer, but the darkroom stays at a pretty even temperature. Dip the film in the stop bath for 30 seconds to a minute, and then put the film in the fixer and agitate it for another 8-10 minutes. Rinse the film in photoflo and hang it to try. You can take a quick look at it using the light table in the developing room.

Scanning the film

There are three ways you can scan the film to measure the line position. One is simply to use a ruler, although the experimental uncertainty will be pretty large. A second possibility is to use the traveling microscope setup, equipped with a vernier scale to make very precise measurements. The third method uses a computerized film scanner which measures the grain density by shining light through the film into a microscope and onto a slit, behind which is photodiode. The film holder moves by computer control through a stepper motor, and the photodiode is read out at each step. The program lets you scan in "high resolution" (small steps) or "low resolution" (big steps).

The computerized scanner gives you a a digital record of the density all along the film. The data is merely a list of numbers corresponding to the density at each step. It should be simple to identify the peak position, or plot the peak shape using MATLAB. The width of the peak is a measure of the resolution of the instrument unless there is some narrow structure ("splitting") underneath, and we will try to get some physics out of that.

One disadvantage of the computerized scanner is that you cannot move the film vertically in a precise way, so it could spoil the calibration (see below) if you are interested in more than just the dispersion. If, for example, the HeNe line is placed on each exposure, that will make your life a lot easier.

Try to include the HeNe line, or some other reference, in the scan. It is already hard to get all the Balmer lines in one scan, since you are limited by the range of the scanner, but get as many as you can along with a reference line. It is probably best to include the red line and the HeNe (or other reference) line, the green line, and as many blue lines as you can.

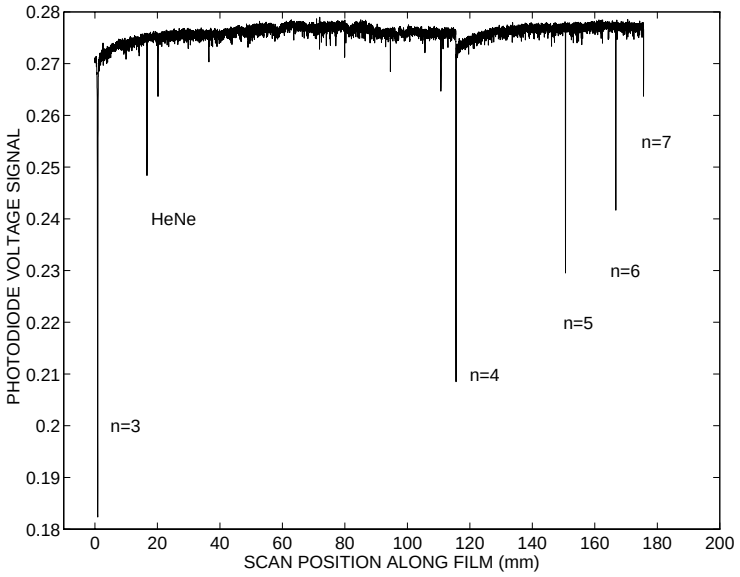


Figure 12.4: Result of a low resolution computerized scan of a hydrogen+HeNe exposure. Because of the limited travel of the scanner, two scans are superimposed by matching the H_β line position.

A low resolution scan of a hydrogen plus HeNe exposure is shown[3] in Fig. 12.4. The first five Balmer lines are clearly visible, as is the HeNe line used for calibration. Instead of the HeNe line, one could include a mercury or helium exposure so that calibration lines are scattered throughout the scan.

12.2.2 Procedure: Jarrell–Ash Spectrometer

Use the Jarrell–Ash scanning spectrometer to measure wavelengths just as you do for the Baird spectrograph. That is, use the same discharge tubes in the same or similar combinations. The Jarrell–Ash has an advantage that it is much higher resolution, so you can see much more fine detail. It is, however, a bit trickier to use.

Figure 12.5 shows a schematic of the spectrometer. The light source sits

[3]Data taken by Marc Crudele, Class of 1996.

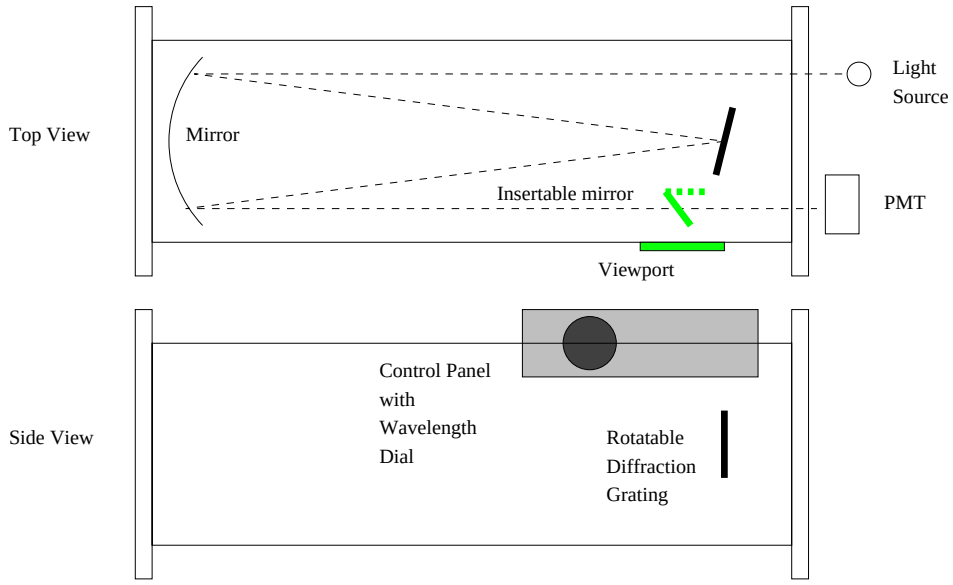


Figure 12.5: Schematic of the Jarrell–Ash scanning spectrometer. The diffraction grating is rotated either with the dial on the control panel, or remotely through the computer control. You can look at the spectral lines with your eye, by inserting the flat mirror and looking at the light on the viewport.

in the front of the device, and light enters through an adjustable slit. You can place one of a variety of wavelength filters in front of the slit, to block out light from other sources and from different directions. Light travels to the back of the spectrometer, reflects from a curved mirror which creates a parallel beam that travels to the diffraction grating. The grating disperses the light according to wavelength, and sends it back to the mirror, which reflects it back to the exit slit where it is detected by the photomultiplier tube. The tilt angle of the grating determines the wavelength which passes through the slit in front of the PMT. By rotating the grating, you scan over wavelengths.

This spectrograph operates in "second order". In other words, the light for a particular wavelength λ is deflected through an angle $\theta_2 = 2\lambda/d$ instead of the "first order" angle $\theta_1 = \lambda/d$. (There are technical reasons why you might want a spectrometer to operate in second order as opposed to first.)

This can lead to some confusion, though, because light of *shorter* wavelengths can show through from *higher* orders, and there are plenty of ultraviolet lines around. For example, suppose you are studying visible light of wavelength λ_0, observed when the grating angle corresponds to $\theta_2 = 2\lambda_0/d$. Then, ultraviolet light of wavelength $\lambda' = 2\lambda_0/3$ will come out at the same position, through third order in the spectrometer.

Remove this confusion by using optical filters between the light source and the input slit. There are two ultraviolet/blue filters available. The 345 nm filter removes all light below this wavelength, so your spectrum is "protected" from 345 nm up to $(3/2) \times 345 = 518$ nm. Using the 455 nm filter you can observe from 455 nm to 683 nm. Using a combination of these two filters, you can therefore cover the full visible spectrum, from violet through red.

You will take your data by letting the computer turn the grating and record the signal level on the PMT. However, it is a good idea to first set things up by looking at the spectral lines by eye. You can do this by inserting the mirror near the PMT (using the "plunger" that comes out the back) which diverts the light to the side. You can then open up the shutter attachment to look into the spectrometer through the ground glass viewport. *Make sure the PMT is turned off when you open up the shutter!* With the lights off, you can see the faint, but sharp, lines of the particular light source you are using. This will help you identify them in the PMT signal.

If you next close the shutter and take the insertable mirror out of the beam, then the light will again pass through the slits into the PMT. Turn the PMT on using the Keithley Model 247 High Voltage power supply. Make sure the power supply is set to zero volts and *negative* polarity, then switch it on, and turn the voltage up to around 500V. The white power cord coming out of the back of the spectrometer powers the amplifiers for the PMT and for the grating stepper motor control. Plug it in. Turn the grating with the hand crank, and measure the voltage signal out of the amplifier with a voltmeter. As the spectral line passes over the slit, you will see a clear increase in the voltage signal. You should "see" the lines this way, just as you could see them with the ground glass plate.

Now get ready to read the signal out with the computer, controlling the grating at the same time. Turn on the computer in the back of the room,

to the left, between the wall and the spectrometer. Start LABVIEW from its file folder in the program manager when the computer comes up. When a registration information dialog box comes up, press "OK", and then press "OK" again when another dialog box comes up. You will now see a blank "vi" (virtual instrument) appear as a grey window. Close this window, and now a dialog box comes up asking for one of three options, namely "new vi", "open vi", or "close LABVIEW". Choose "open vi" and in the "file open" window, double-click on ALPHSCAN.LLB, and then on "visible spectrometer". The program is now ready to go.

Adjust the slit width using the micrometer dial on the front of the spectrometer, and set the starting wavelength position using the hand-crank on the side control panel. Enter the starting wavelength you set on the spectrometer into the "start wavelength" field on the program, and enter the final wavelength (on which the scan will end) in the "finish wavelength" field. Click on the resolution button to read either "high" resolution (about 0.04 Å/step) or "low" resolution (about 1 Å/step). (Low resolution is the best choice when you're first starting out, otherwise it will take a long time to see if you've got things set up correctly.)

When you are ready to start the scan, press the arrow button in the top left corner of the vi. The scan will continue to the "finish" wavelength previously set, and then prompt you for a file name to which you can save the data. If you want to end the scan before the finish wavelength, press the large STOP button in the window, *not* the small octagonal button next to where the start arrow button used to be. If you press the small octagonal button instead of the large STOP button, you will be unable to save the data you just took. Also, the program will lock up the stepper motor so that you will be unable to move the hand-crank.

If for any reason you are unable to move the hand crank, dont force it!!. Press the square black reset button on the alpha stepper board to release the hand crank.

The data that you get will be an intensity as measured by some voltage output from the amplifier circuit on the PMT, as a function of the *nominal* wavelength set on the dial. **The nominal wavelength is just a convenient scale that is close to the true wavelength.** You will have to

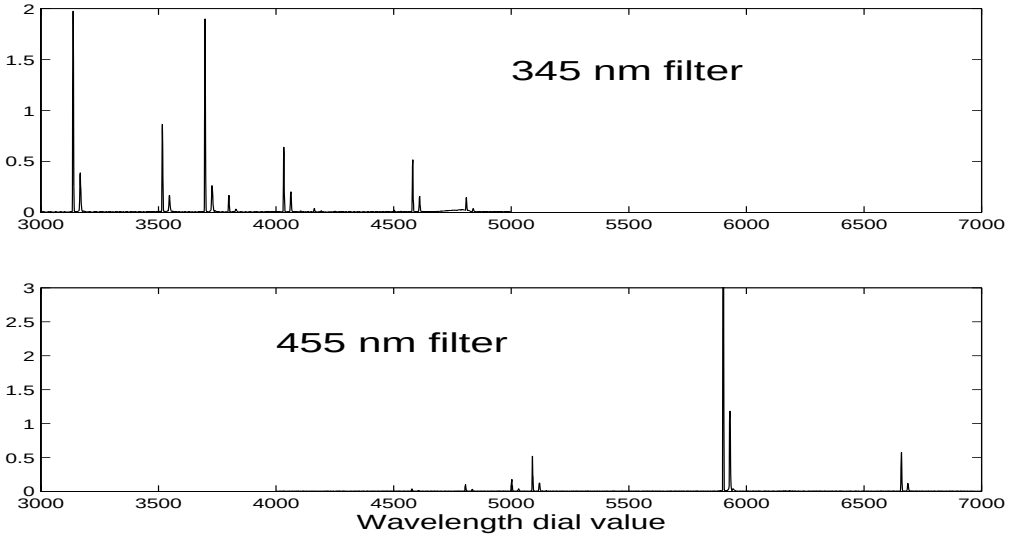


Figure 12.6: Scans of the helium discharge tube using the Jarrell-Ash spectrometer and the two ultraviolet/blue filters. Note that the horizontal axis is the *nominal* wavelength, and must be recalibrated.

determine the true wavelength by using your calibration spectra. If you prefer, you can just plot the data as a function of "step" number, instead of using the nominal wavelength, to keep yourself from getting confused.

Figure 12.6 shows two scans of the helium discharge tube using the Jarrell-Ash spectrometer. One scan uses the 345 nm filter, while the other uses the 455 nm filter. Note that the horizontal axis is the wavelength in Å*as read off the dial, by the computer*. The existence of lines well below 3450 in the scan using the 345 nm filter is clear evidence that the wavelength scale needs to be recalibrated!

12.2.3 Analysis

There are many levels of analysis you can perform, exploring not only the physics of atomic spectra, but also the operation of the spectrograph. The first thing to realize is that the spectrograph is designed so that the line posi-

tion on the film corresponds to the wavelength through a linear relationship, that is,

$$\lambda = Ax + B$$

where x is the position. The quanitity A is called the "dispersion", and for the SB-1 it is supposed to be around 15Å/mm. For the Jarrell–Ash, it will depend on the step size you chose in the program.

You might want to try something very simple first. The relative spacing of the Balmer series lines must be given by Eq. 12.6, with $n_f = 2$. That means that you can just roughly measure the distance between pairs of lines, and check that the ratio of any two distances follow this equation. You can't get any physics out of this because everything cancels except the $\left(\frac{1}{4} - \frac{1}{n_i^2}\right)$ terms, but at least you can check that you can make the n_i assignments correctly.

You can go a little further and check that you are getting around the right value for R_H by using the nominal value for the dispersion. These are things you can do just as you get the data, but to do a more careful job, use scanned data to get the line positions. In this case, you will want to use the helium and mercury lines for calibration.

Precise values for the wavelengths using the helium and mercury discharge tubes are listed in Table 12.1. The values are the "wavelengths in air" as tabulated in the MIT Wavelength tables, by G.R Harrison, et.al. (1969). You can find this book on the main floor reference section of the library under REF QC453.M36 1969. These tables also give the relative intensities of the lines as produced in a discharge tube.

Use the helium and/or mercury data to determine A and B. Determine the uncertainties in A and B from your fit, and propagate these uncertainties through to the physics quantities you derive. How well does A agree with the manufacturers specification? Plot the deviations from the linear fit to see if there is any evidence for nonlinearity. An example[4] of a helium calibration done on the Baird spectrograph is shown in Fig. 12.7. Do you think it would be worthwhile to include a term proportional to x^2 in the calibration? How large you expect the deviations to be for a calibration of the Jarrell–Ash

[4]Data taken by Marc Crudele, Class of 1996.

Table 12.1: Some prominent lines for calibrating the spectrograph. Wavelengths are in Å.

Helium		Mercury	Helium		Mercury
4026.2	Violet			Blue	4916.0
	Violet	4046.6	4921.9	Green	
	Violet	4077.8	5015.7	Green	
	Violet	4339.2		Green	5460.8
	Violet	4347.5		Yellow	5769.6
	Violet	4358.4		Yellow	5790.7
4387.9	Violet		5875.6	Yellow	
4471.5	Violet		6678.1	Red	
4713.4	Blue			Red	6291.3

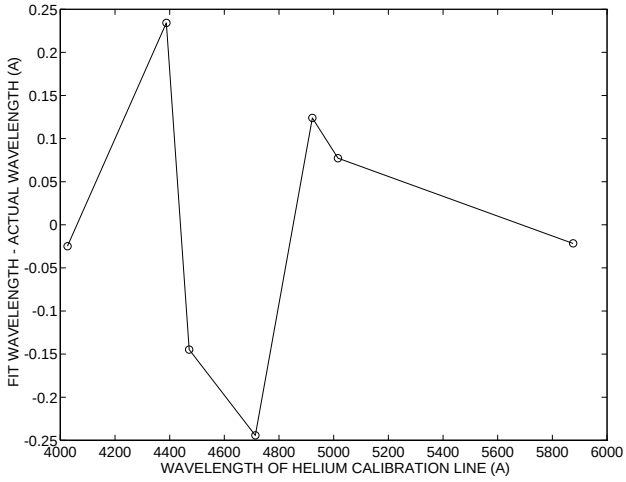


Figure 12.7: Result of a calibration using a helium spectrum on the Baird spectrograph. The helium wavelengths are fitted to a straight line as a function of scanner position. The *deviation* of the fitted wavelength from the actual wavelength is plotted versus the wavelength.

spectrometer? It will be important to carefully determine the position of the maximum voltage signal to determine the wavelength of the line.

The deviations from the fitted wavelengths in Fig. 12.7 give a good indication systematic uncertainty associated with the scan. For this data set, the uncertainty is ~ 0.2Å, or ~ 13 μm. Note that there is a systematic uncertainty of about this size just because of the number of significant figures given in Table 12.1.

Determining R_H and R_D

You can determine the Rydberg constants for hydrogen and deuterium from your calibration and the measured line positions for the two isotopes, and using Equations 12.6 and 12.8. Recall that $n_f = 2$ for the Balmer transitions.

First try the brute force method. Pick one of the lines of either isotope and determine its wavelength from the calibration. Then, make a good guess for the value of n_i for the line you picked. (What color was the line?) You then calculate R from Eq. 12.6. Calculate the uncertainty by propagating the uncertainties in your caibration constants to λ and then to R. Does the accepted value agree with your measurements to within uncertainties? You might try this on more than one line. The brute force method is the most susceptible to uncertainties. For example, it will be hard to measure the calibration "offset" B very precisely because different spectra will be on different exposures. (Your best bet is probably to always include a HeNe laser exposure that you can refer to.)

A different technique eliminates the need for knowing the calibration offset because you can use the difference of two wavelengths, since this difference only depends on the dispersion and not the offset. Use Eq. 12.6 to express $\lambda_i - \lambda_j$ for any *pair* of lines i, j in terms of R. Again, you should try this for a few different pairs of lines. Is the result more precise? (It should be, since you no longer need to include the uncertainty in the offset.) How is the agreement with the accepted value?

Try both of these methods on both hydrogen and deuterium. Compare them to the accepted value. You should come very close. Is this consistent

within your experimental uncertainty?

Hydrogen/Deuterium isotope shift

The values for R_H and R_D differ because M_H and M_D are different, even though both are much larger than the electron mass. You could determine an expression for $(R_H - R_D)/R_H$ in terms of the masses, and compare your individual measurements this way. However, it would be hard to be accurate because R_H and R_D are so close, the difference might be smaller than your combined experimental uncertainty.

A better way is to use the hydrogen and deuterium exposure and measure the splitting $\Delta\lambda = \lambda_H - \lambda_D$ on any of the Balmer lines. You can determine $\Delta\lambda$ very precisely by exposing hydrogen and deuterium together on the same strip of film, for the Baird, or by using a deuterium tube with some hydrogen in it on the Jarrell–Ash, and scanning the double line, as shown[5] in Fig. 12.8. The excellent resolution of the Jarrell–Ash is obvious! Since you already know λ_H to high precision, you can write

$$\begin{aligned} \frac{\Delta\lambda}{\lambda_H} &= 1 - \frac{\lambda_D}{\lambda_H} = 1 - \frac{R_H}{R_D} \\ &= \frac{m}{M_D}\frac{M_D - M_H}{m + M_H} = \frac{1 - M_H/M_D}{1 + m/M_H}\frac{m}{M_H} \\ &\approx \frac{1}{2}\frac{m}{M_H} \approx 3 \times 10^{-4} \end{aligned} \tag{12.10}$$

The splitting is very small indeed! Measure $\Delta\lambda/\lambda_H$ for some of the lines and compare them to this relationship, including uncertainties. Determine an average value with uncertainty for M_D or M_D/M_H, using Eq. 12.10, assuming known values for m and M_H.

[5]Baird data taken by Marc Crudele, Class of 1996. Jarrell–Ash data taken by Steve Irving and Davienne Monbleau, Class of 1999.

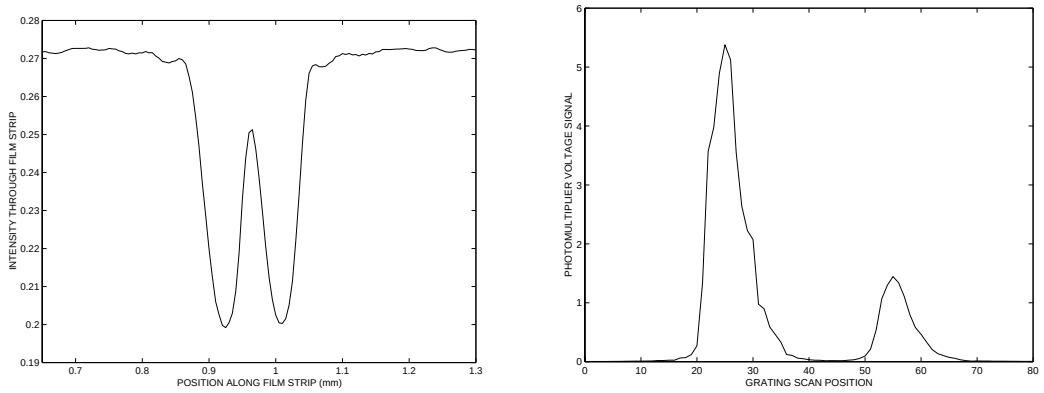


Figure 12.8: Scan of a single Balmer transition line, with both hydrogen and deuterium exposures. The wavelengths differ slightly because of the isotope shift. On the left is a scan of a film strip taken in the Baird spectrograph, exposing hydrogen and deuterium on the same film strip. On the right, a scan using the Jarrell–Ash spectrometer using a deuterium tube with some hydrogen gas inside it as well.

Line widths and splittings

You can also learn about the spectrograph, and some physics as well, by looking at the *widths* of the spectral lines. First, you should try various things to make the individual line widths as narrow as you can. For example, make the input slit width very narrow. The placement of the discharge tube might also be important.

You can quantify the line widths in several ways. Probably the easiest is to locate the "full width at half maximum", that is, the distance between the sides of the line at the points where it is half the maximum value. Plot the widths of the hydrogen lines as a function of wavelength and see if there is a trend. You should keep in mind the effect of "doppler broadening" on the spectral lines. The atoms which emit these photons are moving in lots of random directions with a more or less thermal distribution of energies. If a photon of natural wavelength λ_0 is emitted from an atom moving at velocity component v towards or away from the observer, then its doppler shifted wavelength is just $\lambda = \lambda_0(1 \pm v/c)$. The mean square velocity $\langle v^2 \rangle$ is given

to good approximation by (see Appendix B)

$$\frac{1}{2}M_{ATOM}\langle v^2\rangle = \frac{3}{2}kT$$

where T is the temperature inside the discharge tube. Can doppler broadening explain what you see?

It is probably not possible to see the fine structure splitting in hydrogen (around 0.14 Å) with the Baird spectrograph, but it might be possible with the Jarrell–Ash. [6] In any case, you can measure the line shape. Try to estimate the widest separation you could have for two lines buried inside the single line shape. This would give you an upper limit for the fine structure splitting, and it should be bigger than the expected value.

12.3 Advanced Topics

If you've exposed the nitrogen discharge tube, you likely see something quite different that the other samples. Instead of seeing discrete lines, you should see a series of "bands". That is, lots of lines very close together, but clustered in regularly spaced groups. There should be two major groups like this, one at the red end of the film, and one at the violet end.

You are looking at the deexcitation spectrum of the N_2 *molecule*, not the N atom. This opens up an entirely new area of quantum mechanics, based on different types of potential energies. A thorough treatment is quite involved, but is discussed to some extent in Brehm and Mullin, and in Sime, but a very complete writeup on what's going on is in Preston and Dietz. Here I will just give you some basics, and show you some simple physics you can get out of it.

The nitrogen molecule is your basic diatomic molecule, that is, two nitrogen nuclei separated by some distance, with an electron cloud surrounding them. This object can undergo two basic types of motion. One of these is

[6] An undergraduate laboratory experiment which resolves find structure in the H_α and D_α lines is described in S. Pollack and E. Wong, Amer. Jour. Physics, 39(1971)1386. You might want to use some of their tricks.

a vibration along the axis between the two nuclei, where the separation distance oscillates about some equilibrium value. The other motion is a rotation in space, around the center of mass. Since the two nitrogen nuclei have the same mass, the center of mass is just halfway between the two nuclei.

The equilibrium separation corresponds to the bottom of a potential energy well $V(x)$ where x is the separation distance. Pretty much any potential well can be approxmated by a harmonic oscillator potential, so long as the displacement from equilibrium doesn't get too large. The spring constant k of this approximate harmonic oscillator is given by $k = \frac{1}{2}d^2V/dx^2$, evaluated at the equilibrium point. The solution to the Schrödinger equation for a harmonic oscillator leads to equally spaced energy levels

$$E = \left(n_v + \frac{1}{2}\right)\hbar\omega \qquad \text{Harmonic Oscillator} \tag{12.11}$$

where $\omega = \sqrt{k/\mu}$ and μ is the reduced mass of the system, i.e. $M_N/2$ in the case of the N_2 molecule. The value of n_v must be a nonnegative integer.

The molecule also rotates, and the quantum mechanics of a so-called "rigid rotor" is well defined. It gets complicated because the molecule stretches as its rotation speed increases, but we won't get into that here. The energy levels of the rigid rotor, made of two masses separated by a distance R, are given by

$$E = \frac{\hbar^2}{2I}n_r(n_r + 1) \qquad \text{Rigid Rotor} \tag{12.12}$$

where $I = \mu R^2$ is the rotational inertia, and n_r is a nonnegative integer.

Refer to Brehm and Mullin, for example, and you will find that the spacing of energy levels given by Eq. 12.11 is much smaller than typical atomic energy level spacings, that is, a few eV. What's more, the spacing given by Eq. 12.12 is much smaller than those for the harmonic oscillator, Eq. 12.11.

Don't confuse the different types of spectroscopy used to study these energy levels. Brehm and Mullin, Sec.10-8, does a good job of showing the various ways these excitations are observed, only one of which you are equipped to do with this setup. Transitions between individual vibrational or rotational states *within the same electronic state* correspond to very low energy photons, typically with wavelengths in the far infrared. Your apparatus,

however, is only useful for spectroscopy of visible photons, so you make use of a transition between electronic states which are modified by the presence of vibrational and rotation states built on top of them.

So now let's get back to your piece of film taken with the nitrogen discharge tube. You are looking at the deexcitation of *electronic* energy levels of the molecule, that is, states similar to those found in atoms. However, the big difference is that each of these states correspond to molecular configurations that can undergo their own vibrations and rotations. Therefore, each of these states has a series of states built on top of them, corresponding to the nearly equally spaced vibrational excitations, and each vibrational state has rotational states built on top of it. There are some rules as to which kind of state can decay to another in a lower electronic configuration, but the result is still very complicated as you might imagine. This gets compounded when you realize that the potential well is not really a simple harmonic oscillator, and the rotor is not really rigid. Nevertheless, we can still get something out of it.

Look at the major group of bands near the violet end of the spectrum. This corresponds to the photons emitted from a specific pair of electronic energy levels, and is called the "second positive series" of nitrogen. A diagram of the bands are shown in Fig. 12.9, where it is compared to the mercury spectrum in this region. The figure is taken from Sime. Confirm that your spectrum agrees with this figure. The band spacing, roughly 5 nm according to Fig. 12.9, corresponds to transitions from the more or less equally spaced vibrational levels in one electronic state to those in the lower electronic state. Use this band spacing to determine a rough estimate of the equivalent spring constant k from Eq. 12.11. This lets you draw the shape of the potential well near the equilibrium point. Your next goal is to try and estimate the separation distance of this equilibrium point.

Look closely, using either the traveling microscope or the computerized scanner, at the film *within* the bands. You should see a dense collection of lines which actually make up the band itself. These are transitions between the *rotation* excitations built on top of the electronic+vibrational excitations. Measure the approximate separation between these lines, and use the data to estimate the value of I in Eq. 12.12. From this value of I, you can determine the internuclear separation distance R for the nitrogen molecule. Does this

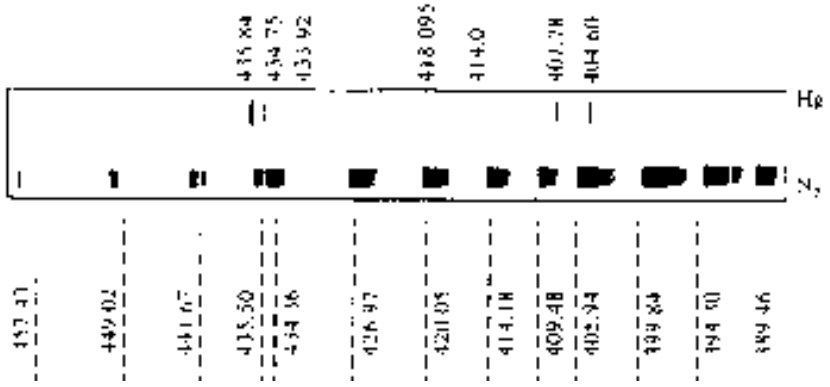


Figure 12.9: The second positive series in the nitrogen molecule, compared to the mercury spectrum, taken with the Baird SB-1 spectrograph. Taken from Sime.

agree with what you expect?

More sophisticated analyses are possible. See Sime or Preston and Dietz for more details.

Ch 13

Noise and Noise Reduction

Noise is one of those odd words scientists use that has no precise meaning. That is, it means different things to different people at different times. I will try to give you the most generic definition of noise, and we can get more specific later.

Let's say you want to measure something. The quantity you are after is called the "signal". Anything that gets in the way of your measurement is called "noise". Signal is good. Noise is bad.

Some kinds of noise are very well defined, and in fact are very fundamental in nature. We will talk about some of these in detail. On the other hand, some noise is just interference of some sort that you can get rid of if you're careful. Sometimes, noise is some empirical property of nature that may have fundamental importance (like $1/f$ noise), but we don't know yet.

There are some general techniques for reducing noise or, equivalently, extracting signal that is obscured by noise. We will discuss some of these.

I recommend the following books for further discussions about signal and noise:

- *Practical Physics*, G. L. Squires, Third Edition
  Cambridge University Press (1991);

Spread through Chapters 6,7,8.

- *Experimental Physics: Modern Methods*, by R. A. Dunlap, Oxford University Press (1988); Mainly Chapter 4, but Chapter 3 talks about Op-Amps.

Squires speaks in more general terms about noise and noise reduction. Dunlap goes into more detail, especially regarding electronic instrumentation, but is still very readable. For a very throrough discussion of noise and how it relates to electronics, see the old stand-by:

- *The Art of Electronics*, by Paul Horowitz and Winfield Hill Second Edition, Cambridge University Press (1989)

13.1 Signal and Noise

Let's not be so vague about noise. The first thing to do is express signal S and noise N in the same units. We can then write the *Signal-to-Noise Ratio* as

$$r = \frac{S}{N}$$

We want r to be large, in some sense, to make a decent measurement.

An alternative way to compare signal and noise has its roots in electronics. It is called the "decibel" and measures the ratio of signal power to noise power:

$$\begin{aligned} dB &\equiv 10 \log_{10} \frac{P_S}{P_N} = 10 \log_{10} \frac{V_S^2}{V_N^2} \\ &= 20 \log_{10} \frac{V_S}{V_N} \end{aligned}$$

(A "bel" would omit the factor of 10.) This definition also finds its way into comparisons of different voltage levels, whether or not they are measuring "signal" and "noise".

13.1.1 Example: Background Subtraction

Before getting into particulars, let's illustrate things with a simple example. We will use what we've learned about random uncertainties to be quantitative about signal and noise. This will lead to our first technique for dealing with noise, namely *signal averaging*. Finally, we'll discuss a common laboratory instrument used for signal averaging, the *multichannel analyzer*.

Suppose you want to see if there is sodium vapor on some distant star. You know that sodium atoms emit distinctive yellow light because of two strong, nearby lines at 589 and 590 nm. You use a prism (or something fancier) to spread out the wavelengths of the starlight, and look for the yellow lines, but there is a problem. You see so much light over all the colors, from the blackbody spectrum of the star, for example, that you can't pick out the lines.

The starlight coming from things other than soidum atoms is called *background*, and it is a kind of noise. It obscures your view and makes it hard to pick out the signal, i.e. the sodium lines. There is much more light given off by the background, even just in the yellow region, than the sodium, so your signal-to-noise (or signal-to-background) ratio is poor.

Signal averaging is the classic way to deal with this problem. Let's talk in terms of intrumentation. Assume you have a photodiode detector that you can move across the prism spectrum from the star. A calibration tells you how to translate the physical position of the photodiode into wavelength. Your measured intensity $M(\lambda)$ is the sum of your signal intensity $S(\lambda)$ and your background intensity $B(\lambda)$, i.e.

$$M(\lambda) = S(\lambda) + B(\lambda)$$

The key point to realize is that $S(\lambda)$ and $B(\lambda)$ have very different shapes. $S(\lambda)$ is a sharply peaked function near $\lambda = 589 - 590$ nm, while $B(\lambda)$ is a smooth function over a large range of wavelengths.

Note that the photodiode itself has some noise, mainly due to thermal fluctuations of electrons across the band gap. This noise shows up as a random uncertainty in the intensity measurement, in addition to background light from the star and other sources. Signal averaging takes this random

statistical noise and reduces it to the point where the signal can be picked out of the background.

So you move your photodiode to a position λ, measure the intensity M and record the value, then step to another λ, measure and record M, step again, and so on. Obviously, you'd like some modern piece of equipment to do this stepping and recording for you automatically. Such a device is generically called a *multichannel analyzer* and you will use different kinds of them in this course.

After one "sweep" through the spectrum, your result looks something like Fig. 13.1(a). The random noise in the photodiode reading makes it impossible to observe any kind of sharp peak near 590 nm. The solution is to sweep more times and average the result. We know (see Eq. 6.4) that the uncertainty in the mean intensity goes down like the square root of the number of sweeps, for each bin of the sweep. Figure 13.1(b) shows the result after 100 sweeps. The flucutations are now 10 times smaller, and the sodium lines are clearer.

Let's be more quantitative about this. The signal S is a small contribution to M as compared to the background B. Therefore, in order to extract S we must have

$$\frac{S}{B} \gg \frac{\delta M}{M}$$

(where δM is just the uncertainty in M) since $M = B(1 + S/B)$ and our assumption is that S/B is small. If you like (and many physicists do) you can talk about a signal-to-noise ratio $r = (S/B)/(\delta M/M)$ as a measure of how the random fluctuations obscure your measurement.

Suppose that after one sweep, you find that $\delta M/M = a$. Averaging n sweeps would still give you about the same value for M, but δM goes down by $\sqrt{n}$ so $\delta M/M = a/\sqrt{n}$. So, if the signal is only 1% as large as the background ($S/B = 0.01$) and one sweep of the multichannel analyzer gives you a 10% measurements of the intensity ($a = 0.1$), then you would need $n = 100$ sweeps just to get to the point where the signal is about as big as the random statistical fluctuations. To make it five times as big, you need $5^2 = 25$ times as many sweeps, or 2500 total.

What is the uncertainty in the signal S that you get out? There are a

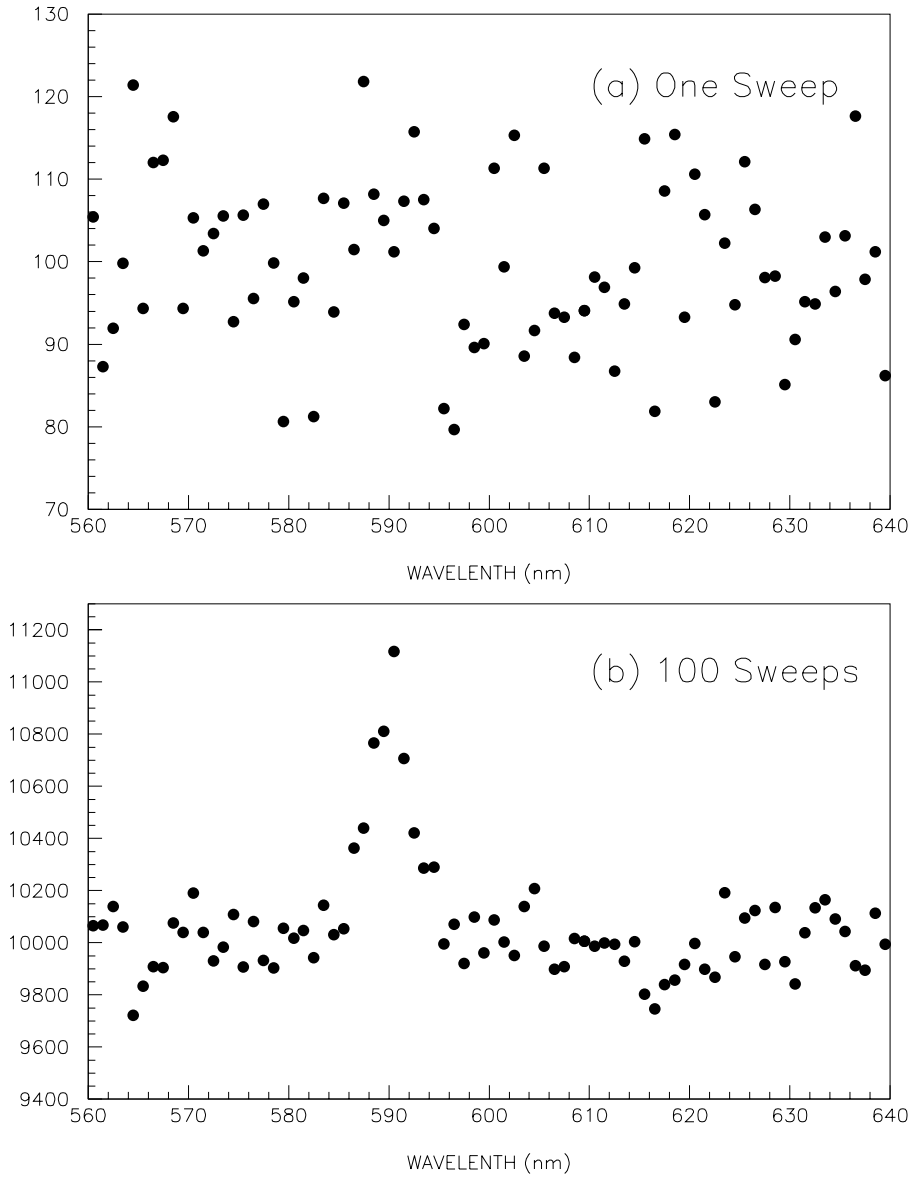


Figure 13.1: Measured intensity versus wavelength for (a) a single sweep of a multichannel analyzer and for (b) 100 sweeps.

few ways to approach this. One is to curve fit $M(\lambda)$ to an assumed signal plus background shape, and use statistical analysis to determine S and δS. A simpler approach, which makes different assumptions, is called *background subtraction*.

Background subtraction is exactly what the name implies. You subtract the background from the measurement, and extract the signal. To determine the background under the signal, you might assume it varies linearly over the signal region, and interpolate for wavelengths above and below 590 nm to get $B(\lambda = 590 \text{ nm})$. This is an independent and (presumably) uncorrelated measurement relative to the signal region, so $S = M - B$ implies that

$$\delta^2 S = \delta^2 M + \delta^2 B \approx 2\delta^2 M$$

since M and B are about the same size. Therefore

$$\frac{\delta S}{S} = \frac{M}{S}\sqrt{2}\frac{\delta M}{M} = \frac{\sqrt{2}}{S/B}\frac{\delta M}{M}$$

and if I want $\delta S/S$ to 10% with $S/B = 0.01$, then I need to get $\delta M/M$ down to 7×10^{-4} or so. Taking enough sweeps can do that in principle, but you don't want systematic uncertainties to creep in. All this may affect aspects of the design of your experiment.

One last point before we leave this example. The sodium line is really *two* distinct lines at 589 and 590 nm. It would sure be convincing if you could resolve these two and be sure you've seen sodium. Various optical techniques will get you getter resolution, but they all will cost you in signal intensity. You can still beat down noise by signal averaging, but you have to judge whether or not it's worth the extra time you have to spend taking data.

13.2 Kinds of Noise

Now we'll talk about some specific physical phenomena that show up as "noise" in various experiments. These phenomena represent natural fundamental limits to the precision with which you can measure things. The nice thing about discussing specific phenomena is that we don't have to speak in

such general terms anymore. The bad thing is that we leave out lots of other kinds of "noise", but experiences like those in Sec. 13.1.1 will give you lots of practice.

Of course, since we're talking about physical phenomena, then these types of "noise" can be "signal" if you're trying to measure them! In fact, you will do just that in Experiment 7.

Again, because we are making a connection to the laboratory, we will end up talking about electronics. The basic principles, though, are applicable at a more fundamental level, and I'll try to point that out.

13.2.1 Shot Noise

The simplest example of fundamental noise is *shot noise*. It is a consequence of the fact that matter is made of discrete units like atoms or electrons. Shot noise is really just an application of simple statistics.

First consider a non-electrical example. Suppose you are sitting in a hut with a tin roof and it starts to rain. The rate of "pings" caused by raindrops on the roof tells you how hard it is raining. That is, $r = N/t$ where N is the number of raindrops falling in time t. Shot noise is just the fluctuations in r caused by Poisson statistics, i.e. $\delta r = \delta N/t = \sqrt{N}/t$. The relative magnitude of the fluctuations, $\delta r/r = 1/\sqrt{N}$, decreases as N gets larger, i.e. the harder it rains.

Now instead of raindrops, let it be electrons moving through space. In this case, a current $i = eN/t$ flows, and if you measure the voltage drop V caused by this current over a resistance R, then $V = iR$. The shot noise in the current is $\delta i = e\sqrt{N}/t$. Therefore, the voltage fluctuations are given by $\delta^2 V = (e^2 N/t^2)R^2 = (ie/t)R^2$ or

$$\frac{\delta V}{V} = \left[\frac{e}{i}\frac{1}{t}\right]^{1/2} \tag{13.1}$$

which can get quite large at low currents.

Having the time t in Eq. 13.1 is inconvenient. Instrumentation responds

as a function of frequency, not time. We use a Fourier transform to go to frequency space. Instead of using the function $h(t)$ which represents a flow of electrons that is always zero except for $0 \leq t' \leq t$ when it is some constant value (say $h(t) = 1$), we want to use its Fourier transform $H(\nu') = \sin \pi\nu' t / \pi\nu'$. We define an *equivalent band width* $\Delta\nu$ so that

$$|H(\nu)|^2_{MAX} \cdot \Delta\nu = \int_0^\infty |H(\nu')|^2 \, d\nu'$$

That is, we replace the real frequency spectrum with a square in frequency space so that the areas are the same. Now $H(\nu)$ reaches a maximum value of t when $\nu = 0$, and the Fourier transforms are related by $\int_{-\infty}^{\infty} H(\nu')^2 d\nu' = \int_{-\infty}^{\infty} h(t')^2 dt' = t$. Therefore

$$t^2 \cdot \Delta\nu = \frac{1}{2} t$$

or $\Delta\nu = 1/2t$. (Note the lower limit of integration changing from $-\infty$ to 0.)

I've glossed over lots of details about this Fourier transform business, but don't be alarmed. I just want you to know it's not black magic. You would certainly expect that $\Delta\nu \sim 1/t$ just by using dimensional reasoning. The factor of $1/2$ comes from a precise treatment.

Anyway, this all lets us write the shot noise voltage as

$$\frac{\delta V}{V} = \left[\frac{2e\Delta\nu}{i}\right]^{1/2} \qquad \text{Shot Noise}$$

over some bandwidth $\Delta\nu$. Note that the *noise power spectrum*, that is the noise power per unit bandwidth, is $(\delta V^2/R)/\Delta\nu = 2eV$ which is *independent of the frequency* ν. For this reason, we say that shot noise is a kind of *white noise*. That is, all frequencies contribute equally to the power spectrum.

Shot noise is generally too small to bother you. Suppose you are measuring voltage using a meter which accepts a bandwidth $\Delta\nu = 10$ kHz. Then, shot noise introduces a 1% fluctuation in the voltage only if the current flowing is less than 32 pA.

13.2.2 Johnson Noise

Johnson noise, like shot noise, comes from a statistical fluctuation. Unlike shot noise, however, it is not as trivial as just random fluctuations in counting the number of electrons in a region of space. Instead, Johnson noise comes from *thermal fluctuations* of electron motion in matter. For this reason, Johnson noise is sometimes called thermal noise.

One of the first phenomena that lead people to believe in molecules and atoms was Brownian motion. this was the observation that specs of dust in a liquid or gas would jitter around randomly if you looked quickly and carefully enough. This jittering happens because the spec is constantly bombarded by molecules in the liquid or gas, each bombardment knocking it one way or another. The motion of the molecules, and so the motion of the spec, is random because it is just due to the thermal energy contained in the liquid or gas. Johnson noise is the same phenomenon, applied to electrons in a resistor instead of the molecules in a liquid or gas.

As the electrons in the resistor jitter around, during any particular time interval there may be more moving towards one end of the reisistor instead of the other end. Therefore, a small net current flows in the resistor during that time interval, giving a small voltage drop. Of course, the net flow during the next time interval is uncorrelated with the previous, and over time this voltage drop averages to zero, i.e. $\bar{V} = 0$. However, the variance $\bar{V^2} - \bar{V}^2 = \bar{V^2}$ is not zero, and the Johnson noise is $\delta V = \sigma_V = \sqrt{\bar{V^2}}$, the root-mean-square (RMS) noise voltage.

We will give a more complete derivation of Johnson noise in Experiment 7, but for now let's try to estimate it from the basic physics. Consider the average noise power $\bar{P} = \bar{V^2}/R$ in the resistor. We expect $\bar{P}$ to be independent of the number of electrons n in the resistor since σ_V should be proportional to $\sqrt{n}$ and according to the microscopic derivation of resistance [1], R is proportional to n. Therefore, you guess that $\bar{P}$ is the energy ϵ per electron, divided by the measurement time t. Statistical mechanics tells you that $\epsilon \sim kT$ where k is the Boltzmann's constant and T is the temperature. Also, you can use the discussion in 13.2.1 to convert the measuring time to band-

[1] See, for example, Resnick, Halliday, and Krane, Chapter 32

width as $1/t = 2\Delta\nu$. So, you would guess that $\bar{P} = \epsilon/t = 2kT\Delta\nu$ and $\delta^2 V = 2kTR\Delta\nu$. This is very close to the right answer. A careful derivation gives this result multiplied by two, hence

$$\delta V = [4kTR\Delta\nu]^{1/2} \qquad \text{Johnson Noise}$$

Note that the power spectrum $\bar{P}/\Delta\nu$ is again independent of frequency. Like shot noise, Johnson noise is another form of white noise.

Johnson noise can be easier to come by than shot noise, but it is still rather small. The Johnson noise in a 100 kΩ resistor at room temperature ($T = 300$ K) measured over a 10 kHz bandwidth is 4.1 μV.

13.2.3 $1/f$ Noise

We breifly mention one last kind of "fundamental" noise, namely $1/f$ noise, also known as "flicker" noise. Unlike shot noise and Johnson noise, $1/f$ noise is not due to fundamental properties of matter, but instead seems to be a part of nature at some more basic level. It is also the dominant source of noise in most setups, after you remove interference from sources with unique frequency spectra.

The name $1/f$ noise comes from it most obvious characteristic. It is not a "white" noise, but the power spectrum instead falls off like (frequency)$^{-1}$, i.e. $1/\nu$ or $1/f$. (Since it tends to lower frequencies, some people prefer to call this "pink" noise.) The magnitude of $1/f$ noise depends on the quality of components used, so there is considerable art in making this contribution as small as possible. Of course, if you can work at sufficiently high frequencies where the $1/\nu$ fall off is enough, you can eliminate this noise that way.

$1/f$ noise seems to be present in nature at all levels, and defies a basic physical description. See Horowitz and Hill for an interesting discussion.

13.3 Noise Reduction Techniques

Getting rid of noise makes your experiment better, and the best way to make your experiment better is to be creative and use common sense combined with experience. On the other hand, there are some standard techniques that we use for noise reduction, and we'll spend some time discussing them.

Keep in mind that noise is just anything that gets in the way of your measurement. Also remember that noise will usually have some specific frequency spectrum, and we will exploit that in some forms of noise reduction. In fact, let's start there.

13.3.1 Frequency filters

If the noise that's bothering you is in some specific range of frequencies, and you can make your measurement in some other range, then a *frequency filter* can do a lot for you. Of course, if you're dealing with white noise, there is no frequency range you can escape to. However, if $1/f$ noise is a problem, work at high frequency, if the physics permits. If the noise is from 60 Hz line interference or some other specific frequency, work above or below that value.

Frequency filters are usually simple circuits (or perhaps their mechanical analogs) that allow only a specific frequency range to pass from the input to the output. You then make your measurement with the output. Of course, you need to be careful of any noise introduced by the filter itself.

We've already worked with a frequency filter back in Chapter 2 and Experiment 1. The circuit shown in Fig. 2.6 is a "low pass" filter. It exploits the frequency dependence of the capacitor impedance $Z_C = 1/\imath\omega C$ to short frequencies much larger than $1/RC$ to ground, and to allow much smaller frequencies to pass. As we showed earlier, the ratio of the output to input voltage as a function of frequency $\nu = \omega/2\pi$ is $(1 + \omega^2 R^2 C^2)^{-1/2}$.

If you want to get rid of noise at low frequency, like $1/f$ noise, you want the opposite of a low pass filter. To make a "high pass" filter, just reverse the

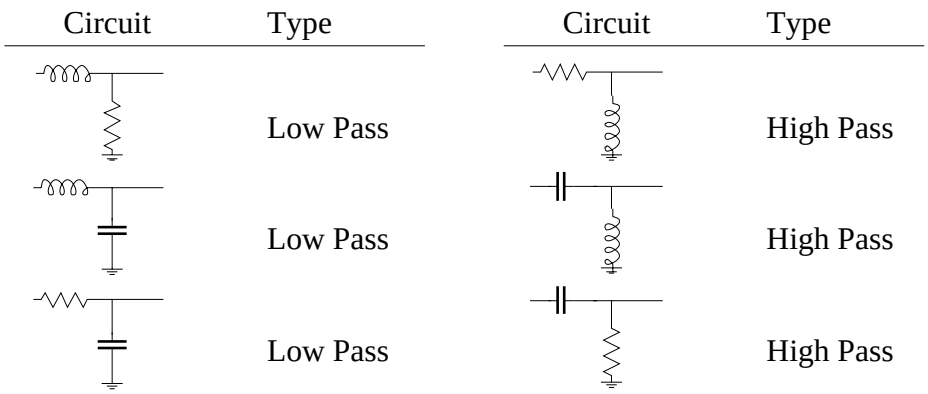


Figure 13.2: Simple passive frequency filters.

resistor and capacitor. Note, however, that you are measuring the voltage across a resistor, and you may need to consider, for example, the Johnson noise it introduces.

You can also use inductors in these simple circuits. Remember that whereas a capacitor is open at low frequencies and a short at high frequencies, an inductor behaves just the opposite. Figure 13.2 shows all permutations of resistors, capacitors, and inductors, and whether they are high or low pass filters.

Suppose you only want to deal with frequencies in a specific range. Then, you want a "bandpass" filter which cuts off at both low and high frequencies, but lets some intermediate bandwidth pass through. Consider the circuit shown in Fig. 13.3. The output voltage tap is connected to ground through *either* a capacitor or an inductor. Therefore, the output will be zero at both low and high frequencies. Analyzing this filter circuit is simple:

$$\frac{V_{OUT}}{V_{IN}} = \frac{Z_{LC}}{Z_R + Z_{LC}}$$

where $Z_R = R$ and $Z_{LC} = \left(Z_L^{-1} + Z_C^{-1}\right)^{-1}$ with $Z_L = 1/\imath\omega L$ and $Z_C = \imath\omega C$. (Note that L and C are connected in parallel.) The result is

$$g = \left|\frac{V_{OUT}}{V_{IN}}\right| = \frac{1}{\left[1 + \frac{R^2}{\omega^2 L^2}(1 - \omega^2 LC)^2\right]^{1/2}}$$

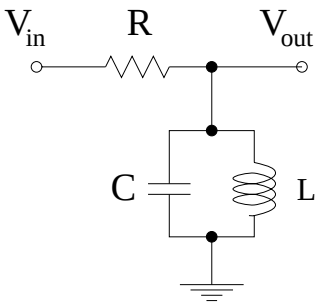


Figure 13.3: A simple bandpass filter.

and as advertised, $g \to 0$ for both $\omega \ll 1/\sqrt{LC}$ and for $\omega \gg 1/\sqrt{LC}$. However, frequencies near $\nu = \omega/2\pi = 1/2\pi\sqrt{LC}$ are passed through with little attenuation. At $\omega = 1/\sqrt{LC}$, $g = 1$ and there is no attenuation at all. Can you see how to build a "notch" filter, or "band reject" filter, that allows all frequencies to pass *except* those in the neighborhood of $\omega = 1/\sqrt{LC}$?

There are an infinite variety of these kinds of circuits. If you want the frequency cut off to be sharper, for example, you could cascade filters. However, this can make problems since you must consider the impedance of a downstream filter when analyzing the circuit. It may be very hard to come up with a "passive" filter that can do what you want. Instead, you may need "active" (i.e. powered) elements in the circuit. We'll get to that in the next section.

13.3.2 Negative Feedback and Operational Amplifiers

One way noise can get in the way of your measurements is by causing things to change when you don't want it. These changes can happen as a function of time, frequency, temperature, etc... To fight this, you want your apparatus to be stable against time, frequency, etc... The most common way to do this is using *negative feedback*.

The idea behind negative feedback is that you take a part of the "output" and *subtract* it away from the "input", causing it to "feed back" to the output and discourage it from changing. The theory of control systems makes extensive use of this idea, applying it to mechanical, thermal, and electrical

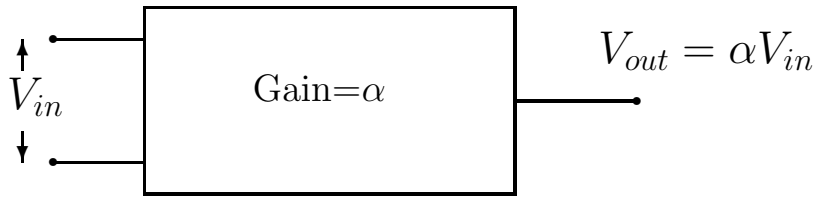


Figure 13.4: A generic amplifier.

systems. I think it is easiest to demonstrate how it works using circuits, and in fact that is how we use it in this course.

We'll build a circuit out of two parts. One part is active and complicated, and is likely to be fraught with all sorts of ugly, noisy instabilities. The other part is passive and very simple, and is rock-solid stable. Combining the two gives us an active circuit that is very stable *in fact rock-solid stable in the ideal limit*, and has lots and lots of uses.

Consider a generic amplifier, like that shown in fig. 13.4, which amplifies the difference voltage between its inputs to give an output voltage. This is the complicated part. It is likely made from lots of transistors connected by a spider web of passive components. Let the gain of the amplifier be α. That is, for the circuit in Fig. 13.4 we have $V_{OUT} = \alpha V_{IN}$. We apply negative feedback by taking some of the output voltage and subtracting it from the input. This is shown in Fig. 13.5. A resistor voltage divider (this is the simple part) is used to take a fraction $\beta = R_2/(R_1 + R_2)$ of the output voltage V_{OUT} and subtract it from the input. The amplifier now does not amplify V_{IN} directly, but instead amplifies $V_{DIF} = V_{IN} - \beta V_{OUT}$. That is,

$$V_{OUT} = \alpha V_{DIF} = \alpha V_{IN} - \alpha\beta V_{OUT}$$

and the net gain g is

$$g = \frac{V_{OUT}}{V_{IN}} = \frac{\alpha}{1 + \alpha\beta} \tag{13.2}$$

Now's here the key point. The generic amplifier is designed so it has enormous gain. That is, α is very, very large. So large, in fact, that $\alpha\beta \gg 1$,

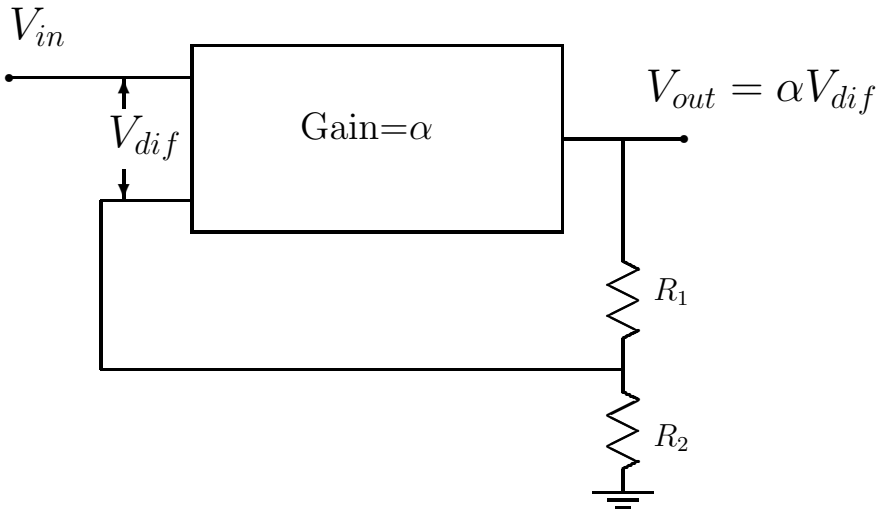


Figure 13.5: A generic amplifier with negative feedback.

no matter how small β is. That means that the gain is

$$g = \frac{1}{\beta} = 1 + \frac{R_1}{R_2} \qquad \text{for } \alpha\beta \gg 1 \tag{13.3}$$

The gain of the system only depends on the ratio of a pair of resistor values, and not on the gain of the generic amplifier. It is hard to get resistor values to change, so this amplifier circuit is very stable. The generic amplifier with gain α, however, is likely to depend a lot on frequency, temperature, and so on.

As you might imagine, commercial versions of the generic amplifier shown in Fig. 13.4 are available in lots of flavors. They are called *operational amplifiers* or *Op-Amps* for short. Instead of a box, they are represented by a triangle, as shown in Fig. 13.6. The two inputs are labeled "+" and "−" for phase considerations, but you can ignore that for this course. The $+V$ and $-V$ terminals are where you apply a voltage source to power the op-amp. It is common to leave these off for schematic circuit diagrams.

Op-amps are cheap. Most cost less than $1, although you can pay a lot if you want special properties. All have very large gain, i.e. α upwards of 10^4 or more, up to some frequency. (Remember that capacitance kills circuits at high frequency because it becomes a short.) The "good ole standby"

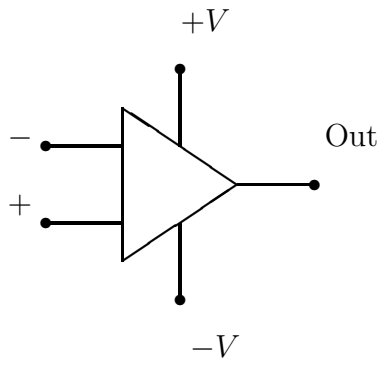


Figure 13.6: Op-Amp notation.

op-amp is the model 741 which is still widely used today. A version of the 741 in standard use today (the LF411) has a gain of at least 88 dB (i.e. $\alpha \geq 2.5 \times 10^4$) and can be used up to frequencies of tens of kHz or more, depending on the feedback circuit. Horowitz and Hill tabulate the properties of your garden variety opamps. (The table covers six pages!) They also tell the interesting story of how op-amps were developed, and why the 741 is such a mainstay.

A common use of op-amps, of course, is just as a negative feedback amplifier. You pick $R_1 \gg R_2$ so that the gain given by Eq. 13.3 is $g \approx R_1/R_2$. For example, to build a stable amplifier with a gain of ≈ 100 up to a kHz or so, you might build the circuit shown in Fig. 13.7.

Another application of op-amps brings us full circule to our discussion of filters. The effective input impedance of an op-amp in negative feedback is huge. That's because even though you apply a voltage V_{IN}, the input to the op-amp is $V_{DIF} = V_{IN} - \beta V_{OUT} \approx V_{IN} - \beta(V_{IN}/\beta) = 0$ so it draws no current. This makes the op-amp ideal for "load buffering". That is, you can use it to make the input to some device (like a filter or perhaps a meter) large enough so you can ignore its effect on the circuit that feeds it.

So, you might build a high pass filter as shown in Fig. 13.8. *All* the output of the op-amp is fed back to the input, so $\beta = 1$ and $g = 1$. However, $Z_{IN} = \infty$ (effectively) because of the op-amp, so all this circuit does is cut off the output of the source for $\omega < 1/RC$ like a good high pass fileter should.

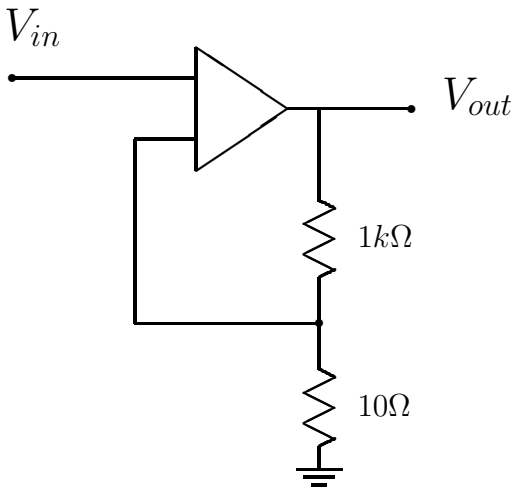


Figure 13.7: An amplifier circuit with gain of 100.

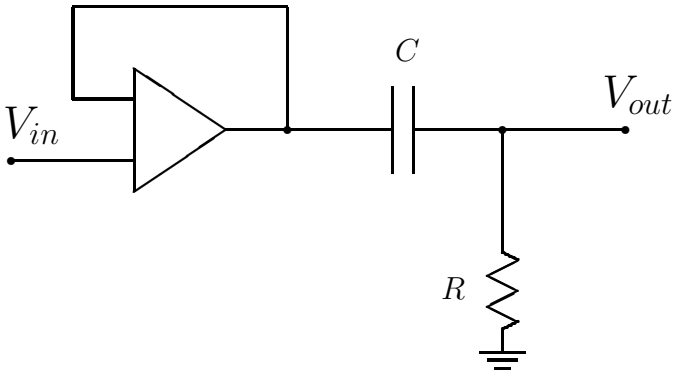


Figure 13.8: A high pass filter with input load buffering.

If the op-amp were not there, you would need to add in the filter input impedance $Z_{FILTER} = R + 1/\imath\omega C$ to the source circuit, and that could really screw things up. See Dunlap for further clever variations on active filters.

13.3.3 The Lock-In Amplifier

We've spent some time talking about filtering out noise because it has a specific frequency or tends toward some frequency range. Suppose instead it is the signal, not the noise, that comes at a specific frequency. We can use that to pick the signal out of the noise. Furthermore, we can be sensitive to the *phase* of the signal as well as its frequency, and that can make a huge improvement. The technique that does all this is called *phase sensitive detection*. The device that you do it with is called a *lock-in amplifier*.

There are two inputs to a lock-in amplifier. One input carries the signal (and the noise). The signal, remember, is varying at some specific frequency which you are aware of. It may be completely buried in noise, however, so you wouldn't see it on an oscilloscope, for example. The other input carries a reference which varies at the frequency of the signal. The signal oscillates because you make it do so, and the way you do that also gives you the reference. For example, your experiment measures a response to a laser, so you turn the laser on and off rapidly with a mechanical chopper. The motor drive for the chopper gives you the reference signal.

The lock-in amplifier takes the reference signal and uses it as a switch. For half the period, the switch is "up" and it lets the signal input pass through it with no change. For the other half, the switch is "down" and it reverses the sign of the signal (i.e multiplies it by -1) before it passes. This is shown in Fig. 13.9. The result of this is a modified signal which is always positive, instead of oscillating around zero like the input signal. A low pass filter takes out the remaining oscillation and lets the *DC* level pass through. This *DC* level is read off a meter, or presented at some output connector, or digitized by some computer, depending on how much money you paid for the lock-in amplifier.

Now consider what happens if the signal is *out of phase* by 90° with

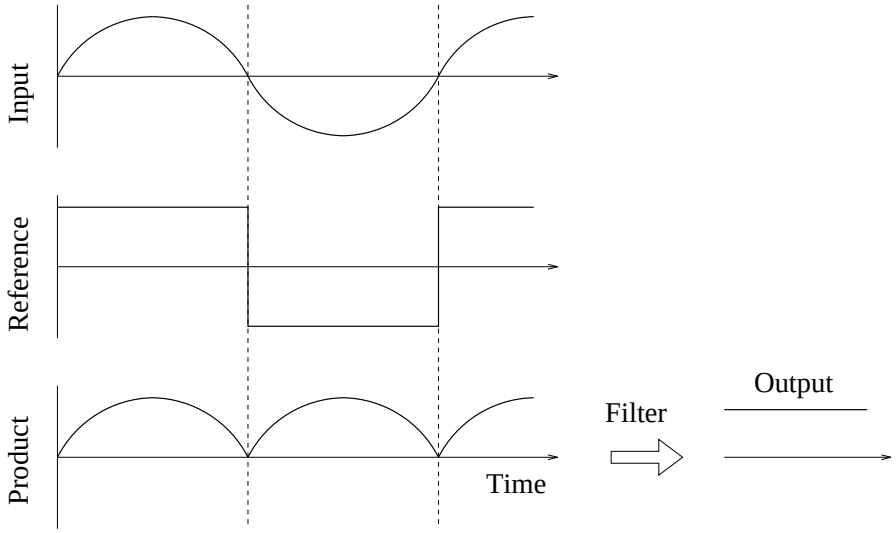


Figure 13.9: The lock-in amplifier acting on an in-phase signal.

respect to the reference. This situation is shown in Fig. 13.10. Now the output of the multiply stage is still something which oscillates about zero. The average *DC* level is zero, and that is the output of the lock-in amplifier.

So, as promised, the lock-in amplifier only detects signals that are *in phase* with the reference. Most lock-in's have a "phase adjustment" knob on the fron that allows you to maximize the output signal. If you have the phase 180° away, the output signal should reverse sign.

Okay, let's see what the lock-in amplifier does to noise that has some frequency other than the frequency of the signal. The answer is obvious. The output of the multiply stage will just be a jumble of noise like the input stage since the reference is essentially just randomly flipping amplitudes. The output of the low-pass filter will *average* to zero over some time determined by the *RC* time constant of the filter.

This is a very powerful technique. It let's you pick out a small signal that may be deeply buried in noise by keying in on its frequency and phase. It is not uncommon to detect a signal even if the signal-to-noise ratio is 1/1000 or worse at the input to the lock-in.

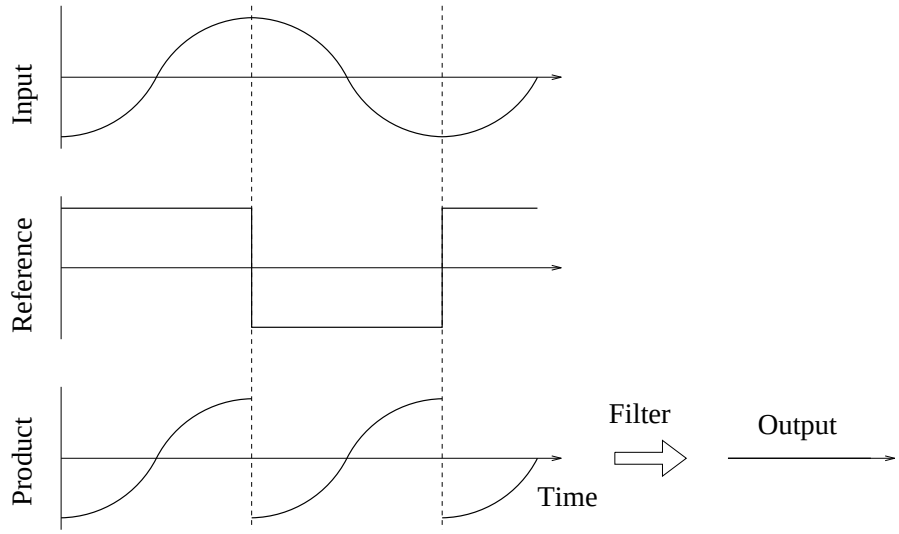


Figure 13.10: The lock-in amplifier acting on an out-of-phase signal.

Modulation Spectroscopy

The lock-in amplifier is actually quite a versatile instrument. One of its uses beyond noise rejection is as a spectroscopy tool. You will use this feature in Experiment 9 on Nuclear Magnetic Resonance, but we'll briefly describe the technique here.

Let's say you have a signal y which is a function of some parameter x. For example, you might have an NMR signal as a function of the large magnetic field which polarizes the sample. Such a thing is graphed in Fig. 13.11.

Now assume the signal is modulated (i.e. made to oscillate) by setting x to some central value x_0 and making it oscillate about x_0 by a small amount Δx. Then the amplitude Δy of the modulated signal is given by

$$\Delta y = \left.\frac{dy}{dx}\right|_{x_0} \Delta x$$

In other words, the output of the lock-in is the derivative of the line shape $y(x)$. It does this, of course, while throwing out any noise that gets in its way. One common technique, described in detail by Dunlap, is to sweep the

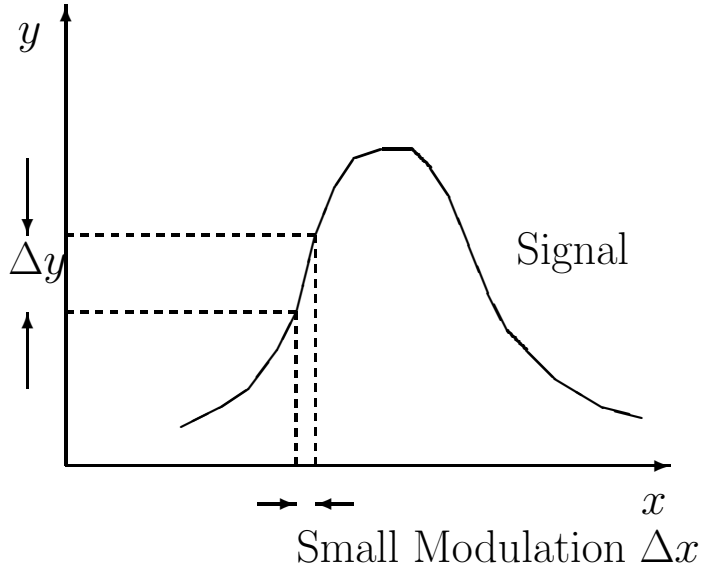


Figure 13.11: Using a lock-in amplifier for modulation spectroscopy.

value of x many times and record the output in a multichannel analyzer. This uses signal averaging to get rid of any remaining noise.

13.4 Exercises

1. A Geiger counter is a device which counts radioactive decays, typically used to find out if something is radioactive. A particular Geiger counter measures 8.173 background counts per second, i.e. this is the rate when there are no known radioactive sources near it. Your lab partner hands you a piece of material and asks you if it is radioactive. You place it next to the Geiger counter for 30 seconds and it registers a total of 253 counts.

a. What do you tell your lab partner?

b. What do you do next?

2. The Tortoise and the Hare have a signal-to-noise problem. A very weak signal sits on top of an enormous background. They are told to determine the signal rate with a fractional uncertainty of 25%, and they decide to solve the problem independently. The Tortoise dives into it and takes data with the setup, and he determines the answer after running the apparatus for a week. The Hare figures she's not only faster than the Tortoise, but smarter too, so she spends two days reducing the background in the apparatus to *zero*, without affecting the signal. She then gets the answer after running the improved setup for one hour. (The Hare really is a lot smarter than the Tortoise, at least this time.)

Assuming Poisson statistics,

a. What is the signal rate?

b. What is the Tortoise's background rate?

3. Consider the passive filters shown in Fig. 13.2.

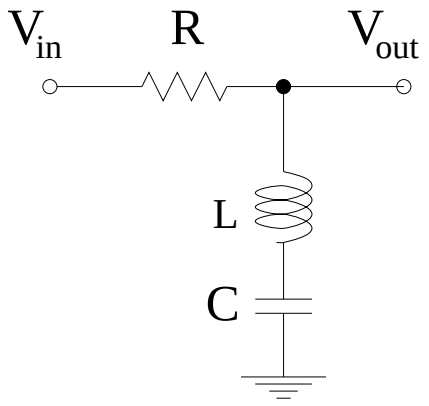
a. Determine the gain as a function of $\omega = 2\pi\nu$ for each filter.

b. Plot the gain as a function of ω/ω_C for the three low pass filters. Define the critical frequency ω_C using the simplest combination of the two components in the circuit, that is, $\omega_C = 1/RC$, $\omega_C = 1/\sqrt{LC}$, or $\omega_C = R/L$. It is probably best to plot all three on the same set of log-log axes.

c. Do the same as (b) for the high pass filters.

d. Can you identify relative advantages and disadvantages for the different combinations of low pass or high pass filter?

4. Consider the following variation on the circuit shown in Fig. 13.3:



a. How does this circuit behave at high frequency?

b. How does this circuit behave at low frequency?

c. Calculate the gain $g = |V_{OUT}/V_{IN}|$ as a function of frequency. What is the behavior for intermediate frequencies?

d. Give an example of where this sort of filter would be useful.

5. A particle detector gives pulses that are 50 mV high when measured as a voltage drop across a 50 Ω resistor. The pulse rises and falls in a time span of 100 ns or less. Unfortunately, there are lots of noisy motors in the laboratory and the ground is not well isolated. The result is that a 10 mV 60 Hz sine wave is also present across the resistor, and adds linearly with the pulses.

a. Draw a simple circuit, including the 50Ω resistor and a single capacitor, that allows the pulses to pass, but blocks out the 60 Hz noise.

b. Determine a suitable capacitance value for the capacitor.

6. You are measuring a quantity Q which is proportional to some small voltage. In order to make the measurement, you amplify the voltage using a negative feedback amplifier like that shown in Fig.C7.5 in the notes.

a. Show that the gain g of the full amplifier circuit can be written as

$$g = g_0 \left[1 - \frac{1}{\alpha\beta} + \mathcal{O}\left(\frac{1}{\alpha^2\beta^2} \right) \right]$$

where $g_0 = 1/\beta$ and $\alpha \gg 1$ is the internal amplifier gain, β is the feedback fraction, and $\alpha\beta \gg 1$.

b. You measure Q with the specific amplifier shown in Fig.C7.7 in the notes. The temperature in the lab fluctuated by 5° F while you made the measurement, and the specification sheet for the op-amp tells you that its gain varies between 2.2×10^4 and 2.7×10^4 over this temperature range. What is the fractional uncertainty in Q due to this temperature fluctuation?

Ch 14

Experiment 7: Johnson Noise

We've spent some time describing things we call "noise". The basic idea is that noise gets in the way of what you want to measure, but here we are going to turn that idea around.

In this experiment, the noise is the signal. That's a flip way of putting it, but we are going to get some physics out of a phenomenon that usually is an annoyance to physicists. The reason we can physics out of it, though, is because it is a very fundamental source of noise. It has to do with the motion of electrons in a conductor, and the heat energy they carry around with them. This is called "Johnson Noise" because it was originally measured by J.B. Johnson. Some people call it "Nyquist Noise", because the phenomenon Johnson measured was first correctly explained by H. Nyquist. I prefer to call it "thermal noise", because that tells you more about where it comes from.

A relatively simple explanation of thermal noise can be found in

- *Random Walk Model of Thermal Noise for Students in Elementary Physics*,
  Richard W. Henry, American Journal of Physics **41**(1973)1361

We will summarize this paper here when we go through a derivation of John-

son Noise. Experiments similar to the one we will do here can be found in

- *Undergraduate Experiment in Noise Thermometry*,
  P. Kittel, W.R. Hackerman, and R.J. Donnelly,
  American Journal of Physics **46**(1978)94
- *An Experiment on Electronic Noise in the Freshman Laboratory*,
  D.L. Livesey and D.L. McLeod,
  American Journal of Physics **41**(1973)1364

Finally, you might want to go back and look at the original work of Johnson and Nyquist. Their papers are actually quite nice.

- *Thermal Agitation of Electricity in Conductors*,
  J.B. Johnson, Physical Review **32**(1928)97
- *Thermal Agitation of Electric Charge in Conductors.*
  H. Nyquist, Physical Review **32**(1928)110

14.1 Thermal Motion of Electrons

We will review the simple model presented in Henry's paper. You might also want to look up the paper by Kittel, et.al. Recall that we went through a simple minded approach in Sec. 13.2.2. A brief review of statistical mechanics and its relation to thermal physics is given in Appendix B.

The model is based on random thermal fluctuations of electrons in a one-dimensional resistor of length L and cross sectional area A. The resistor has resistance R and a voltage drop $V = iR$ is across the ends. The current i, and therefore the voltage V, arises from the thermal fluctuations that allow more electrons to move one way than another in some short time interval t_0.

First, the basics. *On average* no current flows through the resistor, and the average value of V is zero. That is,

$$\langle V \rangle = 0$$

On the other hand, the thermal fluctuations still give rise to a finite voltage as a function of time, in other words $V(t) \neq 0$. Therefore, the variance of V is not zero, that is

$$\sigma_V^2 = \left\langle (V - \langle V \rangle)^2 \right\rangle = \langle V^2 \rangle - \langle V \rangle^2 = \langle V^2 \rangle \neq 0$$

This quantity $\langle V^2 \rangle = \sigma_V^2$ is called the thermal or Johnson noise voltage, and it is what you will measure in this experiment. Let's calculate it in terms of some known quantities.

From Ohm's law and the definitions of current and charge, we can write

$$\begin{aligned} \sigma_V &= \sigma_i R \\ &= \frac{\sigma_q}{t_0} R \\ &= \frac{e\sigma_x / L}{t_0} R \end{aligned}$$

where σ_x is the net x-motion of all the electrons in the measuring time t_0. If we can reduce this to the motion of an individual electron, then we can use a microscopic description of current and resistance. If there is a total of N independent and random electron motions (i.e. "random walks") in time t_0, then

$$\sigma_x = \sqrt{N}\sigma_d$$

where σ_d is the average distance that any single electron moves. Therefore,

$$\sigma_V = \frac{e}{L}\sqrt{N}\frac{\sigma_d}{t_0} R \tag{14.1}$$

Now for the physics. N is the total number of electrons in the resistor times the number of walks in time t_0, so

$$N = (nAL) \times \frac{t_0}{\tau} = \frac{nALt_0}{\tau}$$

where n is the number density of electrons and τ is the time between collisions of a *single* electron. The fluctuation in the motion of a single electron is

$$\sigma_d^2 = \langle d^2 \rangle = \langle v_x^2 \tau^2 \rangle = \langle v_x^2 \rangle \tau^2$$

and this is what we connect to temperature by $\langle E\rangle = \frac{1}{2}m\langle v_x^2\rangle = \frac{1}{2}kT$, where m is the mass of an electron and we note that motion is only in one dimension. (For a review of statistical mechanics, see Appendix B.) The factor k is called Boltzmann's constant and is the basis of the fundamental relationship between temperature and internal energy. Therefore

$$\sigma_d^2 = \frac{kT\tau^2}{m}$$

Finally, we note that (see Resnick, Halliday, and Krane or Expt. 5)

$$\frac{L}{A}\frac{2m}{ne^2\tau} = \frac{L}{A}\rho = R$$

where ρ is the resistivity.[1]

Finally, put this all into Eq. 14.1 to get

$$\begin{aligned} \sigma_V^2 &= \frac{e^2}{L^2}N\frac{\sigma_d^2}{t_0^2}R^2 \\ &= \frac{e^2}{L^2}\frac{nALt_0}{\tau}\frac{kT\tau^2}{mt_0^2}R^2 \\ &= \frac{A}{L}\frac{ne^2\tau}{m}\frac{kT}{t_0}R^2 \\ \text{or}\quad \langle V^2\rangle &= \frac{2kTR}{t_0} \end{aligned} \tag{14.2}$$

As discussed in Sec. 13.2.1, however, it is customary to express the noise using the equivalent bandwidth $\Delta\nu = 1/2t_0$. Therefore, we have

$$\langle V^2\rangle = 4kTR\Delta\nu \tag{14.3}$$

Now in order to measure the voltage V, we will need to amplify or at least process the signal in some way. Let $g(\nu)$ be the gain of this processing circuit at frequency ν. Then the output voltage fluctuations $d\langle V^2\rangle$ integrated over some small frequency range $d\nu$ is given by

$$d\langle V^2\rangle = 4kTRg^2(\nu)d\nu$$

[1]The definition of τ used here differs from that used in Expt. 5 by a factor of two. That is because we are dealing with a single electron. See Resnick, Halliday, and Krane.

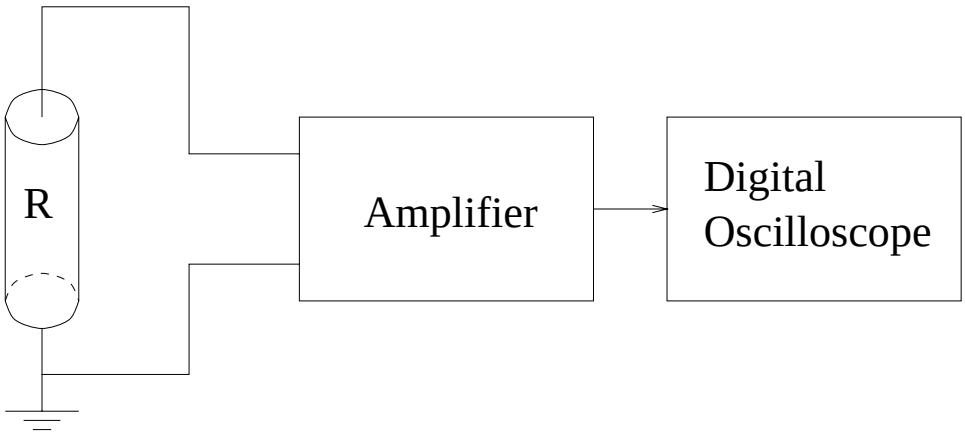


Figure 14.1: Schematic for measuring Johnson noise.

Measurements are made by integrating the signal over a relatively large bandwidth $\Delta\nu$. This bandwidth is typically determined by the gain function $g(\nu)$ which is large only over some finite frequency range. We therefore obtain the expression

$$\langle V^2 \rangle = 4kTRG^2\Delta\nu \tag{14.4}$$

where G and $\Delta\nu$ are constants defined by

$$G^2\Delta\nu \equiv \int_0^\infty g^2(\nu)d\nu \tag{14.5}$$

14.2 Measurements

You will measure the Johnson noise in a series of resistors, and use the result to determine a value for Boltzmann's constant k.

The setup is shown schematically in Fig. 14.1. The voltage across the resistor R is immediately processed by an "amplifier", which essentially multiplies this voltage by a function $g(\nu)$. The output of the amplifier is measured using the LeCroy 9310 digital oscilloscope. You will use the oscilloscope to measure $\langle V^2 \rangle$, given by Eq. 14.4. By changing the value of R (simply by changing resistors), you measure $\langle V^2 \rangle$ as a function of R, and the result should be a straight line. The slope of the line is just $4kTG^2\Delta\nu$, so once

you've calibrated the gain function of the amplifier, you can get k. (You can assume the resistor is at room temperature.)

Let's look a little more carefully at the properties of the amplifier. We will be working in the several tens of kHz range, so to estimate the gain we need, take a bandwidth $\Delta\nu = 10$ kHz. The digital oscilloscope cannot make measurements much smaller than around 0.5 mV, so Eq. 14.4 implies that the nominal gain G must be on the order of 1200 or more to measure the noise in a 1 kΩ resistor. The amplifier also needs to have low noise and good stability itself, if are going to use it on such a small signal. A high gain opamp with negative feedback (see Sec. 13.3.2) sounds like the right solution.

The bandwidth of the amplifier also needs to be considered. In fact, if we are going to do the job right, we want to make sure that all the bandwidth limitations are given by the amplifier, and not by the oscilloscope, for example. That way, we can measure the function $g(\nu)$ of the amplifier stage only. The oscilloscope bandwidth will depend on the timebase used, that is, the time over which the output voltage is averaged and digitized. As long as the oscilloscope's bandwidth is greater than the amplifier's, you will be ok. You ensure this by putting a bandwidth filter on the output of the amplifier. In the beginning, you will use a commercial bandwidth filter with adjustable lower and upper limits.

The first "amplifier" you will use, therefore, is shown in Fig. 14.2. For now the bandwidth filter is just a box with an input and output, and with knobs you can turn. The gain producing part of the amplifier, on the other hand, is essentially a cut and dry application of opamps and negative feedback. In fact, as shown in Fig. 14.2, two such negative feedback loops, are cascaded to get the appropriate gain and input characteristics. The first loop uses a HA5170 opamp and a low gain, while the second stage is higher gain and uses a HA5147.[2] Good starting values to use are $R_1 = 10\Omega$, $R_2 = 100\Omega$, and $R_3 = 2.2k\Omega$. This gives the first stage a gain of 11 and the second stage a gain of 221, times the bandwidth function imposed by the opamps and the bandwidth filter.

All of these components, including your input resistor R (but not the com-

[2]The credit for figuring out the right opamps and amplifier circuit in general goes to Jeff Fedison '94. More details on this circuit design are available.

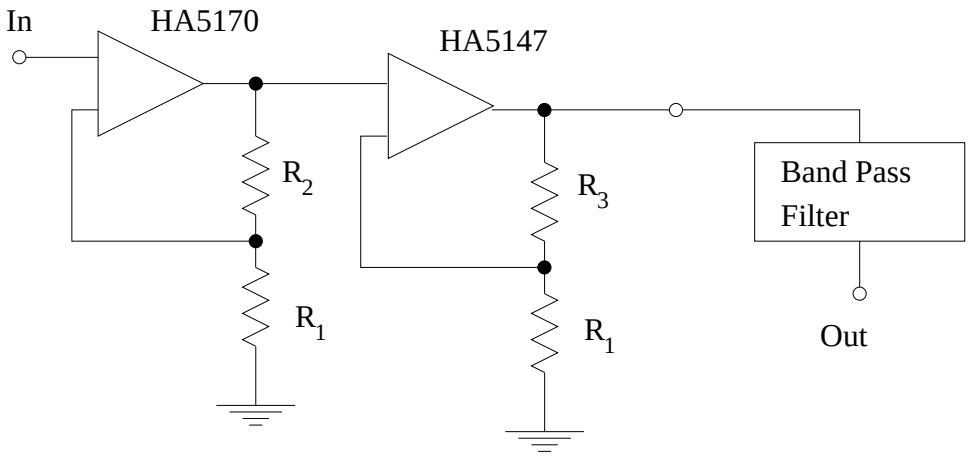


Figure 14.2: Amplifier stage for measurements of Johnson noise.

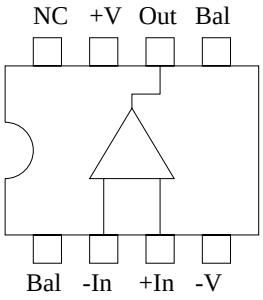


Figure 14.3: Pinout diagram for the opamp chips used in this experiment. We are not using the "Bal" connections. The notation "NC" means "no connection".

mercial bandwidth filter), are mounted on a breadboard so you can change things easily. The pinout diagram for the HA5170 and HA5147 are shown in Fig. 14.3. The opamps are powered by ± 12 V levels applied in parallel with 0.1 μF capacitors to ground, to filter off noise in the power supply. Connections to the breadboard are made using wires soldered to BNC connectors.

14.2.1 Procedure

Set up the circuit shown in Fig. 14.2. Check things carefully, especially if you are not used to working with breadboards. In particular, make sure the 12 V DC levels are connected properly, before you turn the power supply on. The output from the breadboard gets connected to the bandwidth filter, and the output of the bandwidth filter goes into the oscilloscope. The lower and upper limits of the bandwidth filter are not crucial, but 5 kHz and 20 kHz are a reasonable place to start.

First you need to measure the gain of the amplifier/bandwidth filter as a function of frequency. All you really need to do is put a sine wave input to the circuit and measure the output on an oscilloscope. The output should look the same as the input (i.e. a sine wave of the same frequency ν), but the amplitude should be bigger. The ratio of the output to input amplitudes is just the gain $g(\nu)$.

There is a problem, though. You have built an amplifier of very large gain, around 2.4×10^3, and the output amplitude must be less than a few volts so the opamps do not saturate. That means that the input must be less than a couple of millivolts. That is barely enough to see on an oscilloscope, assuming your waveform generator can make a good sine wave with such a small amplitude.

You get around this problem by using the schematic shown in figure 14.4. The waveform generator output passes through a voltage divider, cutting the ampltude down by a known factor. This divided voltage is used as input to the amplifier. It is a good idea to measure the resistor values R_{BIG} and R_{SMALL} using an ohmmeter, rather than trust the color code (which can be off by up to 10%). Pick resistors that give you a divider ratio somewhere between 10 and 100. It is also a good idea to tee the output of the waveform generator and look at it on the oscilloscope along with the amplifier/bandwidth filter output.

Make your measurements of $g(\nu)$ by varying the frequency of the waveform generator, and recording the output amplitude. Of course, you must also record the input (i.e. generator) amplitude, but if you check it every time you change ν, you can be sure it does not change during your mea-

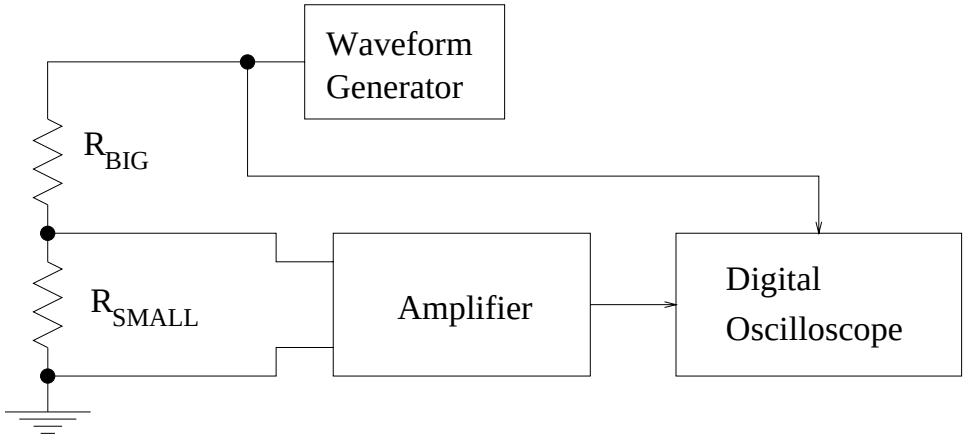


Figure 14.4: Calibration scheme for the noise amplifier.

surement. Measure over a range of frequencies that allows you to clearly see the cutoffs from the bandwidth filter, including the shape as g approaches zero. Also make sure you confirm that the gain is relatively flat inbetween the limits.

An example is shown in Fig. 14.5. The setup used $R_1 = 10\Omega$, $R_2 = 100\Omega$, and $R_3 = 2.2k\Omega$, so the total gain should be 2431, and with bandwidth filter limits at 5 kHz and 20 kHz. The main features seem to be correct, although the filter has apparently decreased the maximum gain a bit.

Now take measurements of the actual Johnson noise as a function of R. Remove the waveform generator and voltage divider inputs, and put the resistor you want to measure across the input to the amplifier. Set the time per division on the oscilloscope so that its bandwidth limit is much larger than the upper frequency you used on the bandwidth filter. For example, if there are 10,000 points[3] (i.e. samples) per trace and you set the scope to 0.2 ms/div, then the time per sample is 0.2 μs since there are ten divisions. The bandwidth is the reciprocal of twice this time (see Sec. 13.2.1) or 2.5 Mhz. If the filter cuts off at 20 kHz, then this would be fine.

It is best to trigger the scope on "line" so you do not bias the input values. Also, the coupling of the input channel should be DC and high impedance

[3]The LeCroy 9310 has 10,000 points but the 9310A has 50,000.

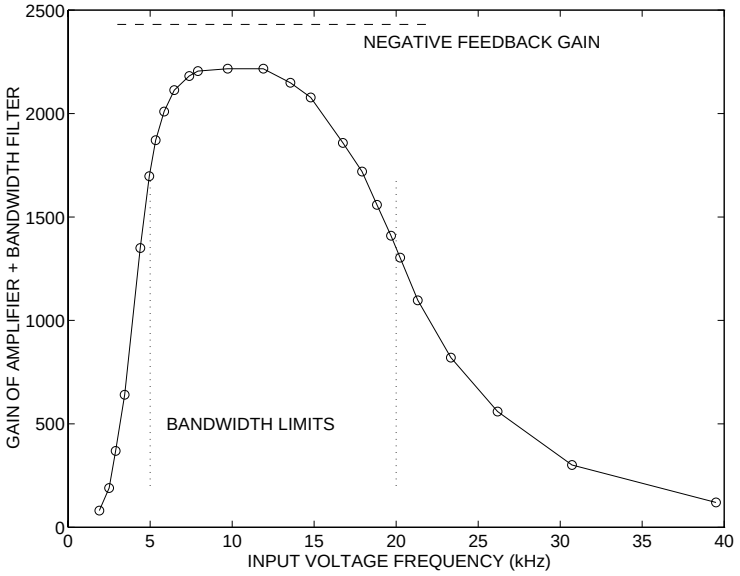


Figure 14.5: Sample of data used to determine $g(\nu)$ for the amplifier followed by the commercial bandwidth filter. The simple negative feedback formula gives a gain of 2431, and the bandwidth filter is set for $\nu_{LO} = 5$ kHz and $\nu_{HI} = 20$ kHz.

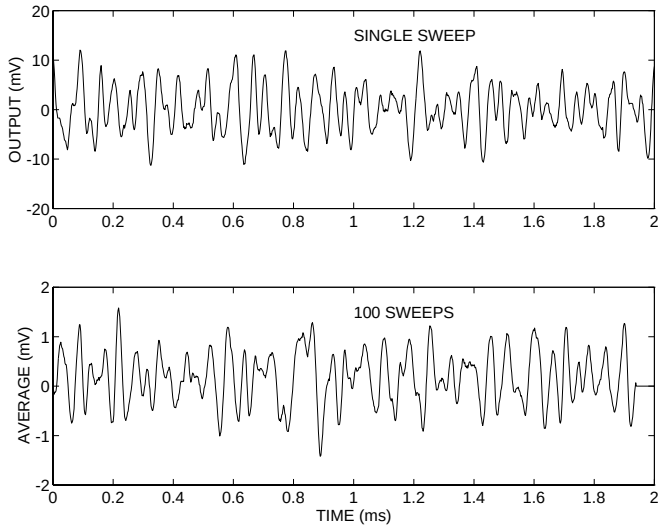


Figure 14.6: Oscilloscope traces of the output of the bandwidth filter, and for 100 traces averaged together by the oscilloscope. Note the difference in the vertical scales.

so the input does not suffer any distortion.

Use resistors with R near zero (10Ω) and up to $R \approx 10\text{k}\Omega$. The oscilloscope trace will look like an oscillatory signal, but that is because you are (likely) using tight bandwidth limits. What would the trace look like if the lower limit was only slightly smaller than the upper limit?

Figure 14.6 shows a single sweep trace on the scope directly from the output of the bandwidth filter, and the average (as done by the scope itself) of 100 traces. The average looks the "same" as the single sweep, but it is 10 times smaller. (Note the difference in the vertical scales.) It is clear, therefore, that the oscillations in the input signal are random in phase, even though they are confined within the limits of the bandwidth filter.

The *PARAMETERS* menu on the scope will allow you to measure lots of things about the trace, and average the values over lots of sweeps to smooth out the effects of the oscillations. You should average over 100 sweeps or so, but if you watch the numbers change as the number of sweeps increases, you can get an idea of how stable they are.

There are three quantities in particular you should record, for each value of R. One is the RMS value, which the oscilloscope handbook tells you is in fact $\sqrt{\langle V^2 \rangle}$, but that is not quite what you want. The amplifier and bandwidth filter have a tendency to add some small DC level to the output, so that the MEAN value $\langle V \rangle$ is not quite zero. Since you are only interested in the fluctuations about the mean, therefore, the quantity you actually want is called SDEV, i.e. the standard deviation. You should probably check to see if these values satisfy the relation $\mathrm{SDEV}^2 = \mathrm{RMS}^2 - \mathrm{MEAN}^2$.

14.2.2 Analysis

The first thing you need to do is determine the value of $\int_0^\infty g^2(\nu) d\nu$. Make a plot of $g^2(\nu)$ as a function of ν and estimate the integral under the curve. You can try to estimate this graphically, but you can easily get an accurate answer using the MATLAB function `trapz` which performs a trapezoidal integration given a list of (x, y) values.

Next make a plot of $\langle (V - \langle V \rangle)^2 \rangle = \mathrm{SDEV}^2$ as a function of R. An example is shown in Fig 14.7. Fit a straight line through these points and determine the slope and intercept at $R = 0$. Get a value for k, and the uncertainty, from the slope, using your gain integral and assuming the resistor is at room temperature. Estimate the contributions to the uncertainty in k from your estimates of the possible temperature range for the resistor, and from the uncertainty in the slope of the straight line.

The intercept of the line is the noise at $R = 0$. You would expect this to be zero if Johnson noise in your input resistor were the only thing going on. The input opamp, however, has some noise of its own, due to internal Johnson noise, shot noise, and so on. The specification sheet for the HA5170 gives an equivalent input noise of around 10 $\mathrm{nV}/\sqrt{\mathrm{Hz}}$. How does this compare to your measurement?

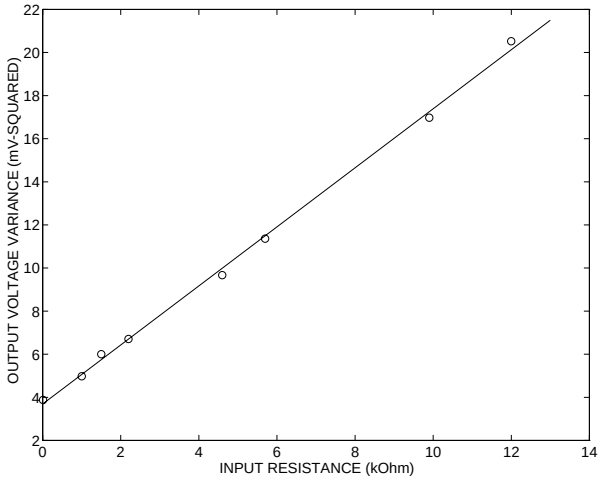


Figure 14.7: Data taken by measuring the standard deviation ("SDEV", using the LRS9310 Parameters menu) of the output voltage signal, as a function of the input resistor value. The variance is just SDEV^2. The slope gives k, while the intercept gives the equivalent input noise voltage, after correcting for the amplifier gain$\times$bandwidth.

14.3 Advanced Topics

This is not an easy experiment. A lot of the concepts are probably new to you, and it would not be surprising if some things didn't work out quite the way you expect.

In fact, the value you got for k in the above analysis is likely a bit high. Furthermore, you likely don't yet have any clear evidence of a systematic effect, other than your not getting the answer you expect. This would indicate some source of noise you've not eliminated.

It is a worthwhile project to repeat the measurements, making changes wherever you can. You might consider changing the resistor values R_1, R_2, and R_3 in Fig. 14.2 to vary the gain of the amplifier. Adjusting the bandwidth limits should also give you different measured quantities, but the same result for k after the calibrations are done. Also, try varying the sample rate of the scope, just make sure that you keep the equivalent scope bandwidth larger than what the bandwidth filter gives.

Beyond these, however, there are some more concrete things that you can do.

14.3.1 Analysis of Traces

Instead of simply using the oscilloscope to determine the standard deviation, use MATLAB and the trace data (as in Fig. 14.6) to get the values and examine their distribution. You can get the data into an array[4] trace by following the procedure outlined in Sec. 3.7.1, and you can use mean(trace) and std(trace) to get the mean and standard deviation. The series of MATLAB commands used to plot the distribution might look like

```
bins=linspace(min(trace),max(trace),50);
[n,x]=hist(trace,bins);
stairs(x,n);
```

The single sweep trace in Fig. 14.6 is plotted this way in Fig. 14.8. The distribution is rather gaussian-like, as you expect, but you could test to see if this is really the case by comparing it to the gaussian with the same mean and standard deviation, and considering the χ^2. It might be that better statistics, or perhaps doing the same thing with the scope-averaged spectrum, will show some peculiarities.

14.3.2 Frequency Spectrum

The LeCroy 9310 digital oscilloscope has the capability of performing a real time Fourier analysis of the input. That means that you can actually demonstrate that the noise spectrum $d\langle V^2\rangle/d\nu$ is indeed "white", that is, independent of frequency. This is straightforward data to take, but will require that you learn more about Fourier analysis to interpret it.

[4]Be careful, though, since the trace can be quite long, say 50,000 values. This will be too big for the Student Edition of MATLAB and you'll need to cut it down by skipping lines when you read it in.

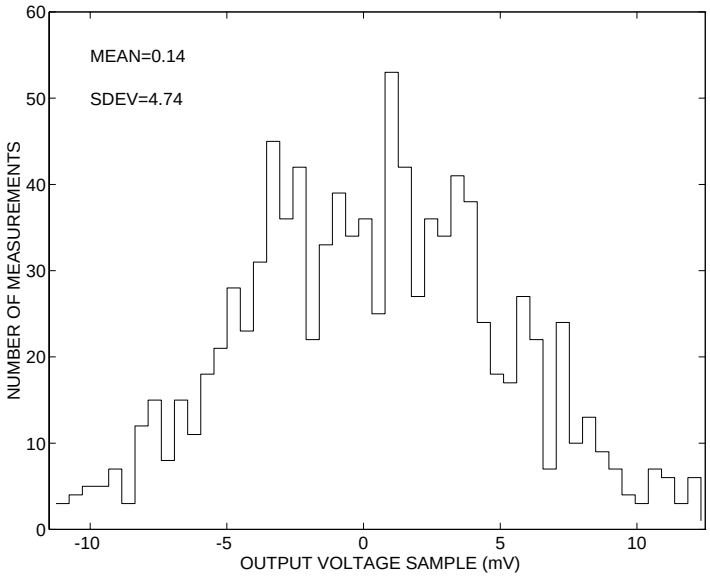


Figure 14.8: Histogram of the individual voltage values from a single sweep trace.

Try the easy part first. Use the oscilloscope to Fourier analyze the input trace by storing the magnitude of the FFT in one of the math traces. Then, form the average of many FFT results and display the result. You should get something that looks somewhat like Fig. 14.9. If the noise spectrum were indeed white, then this should look more or less like the gain function $g(\nu)$. Indeed, that is the general shape, but what are those big spikes at 30 kHz and 53 kHz?! There certainly does seem to be some additional noise getting through.

Attempt to eliminate the noise. Try moving things around or use better shielding on the circuit breadboard. What happens when you adjust the limits on the bandwidth filter?

You can in fact use data like that in Fig. 14.9 to analyze your data directly. For a couple of values of R, you can get $d\langle V^2\rangle/d\nu$ from the FFT function, only using data points where the spectrum is truly white. Your measurements of $g(\nu)$ at those frequencies is then used to determine $\langle V^2\rangle$. Consult the oscilloscope handbook for more details.

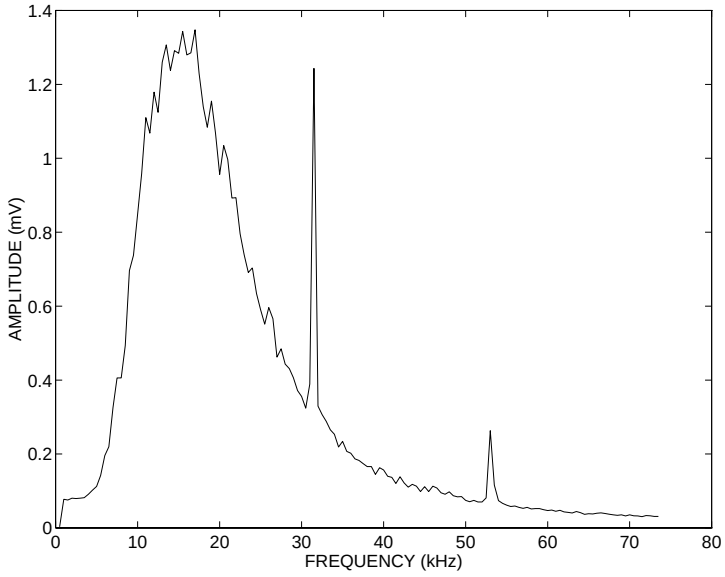


Figure 14.9: Fast Fourier Transform (FFT) of the output from the bandwidth filter. The magnitude of the oscilloscope FFT function was averaged with the result from 100 traces. The large spikes represent some form of noise at those specific frequencies.

14.3.3 Circuit Modifications

One nontrivial circuit modification would be to make your own bandwidth filter. For example, consider the circuit shown in Fig. 13.3[5] Try assembling components that give you reasonable parameters for the gain integral in Eq. 14.5. A simpler kind of filter might simply be two RC filters, one high pass and one low pass, cascaded in series. If you want to do active buffering, though, be careful to use an opamp that works at these frequencies.

Try using a few kΩ resistor as input, but something that is mechanically large and strong enough to take some real temperature change. If you immerse the resistor in liquid nitrogen, for example, it should make a large (and predictable) change in the Johnson noise.

A very nice addition to the experiment would be to put an adjustable current through the resistor R. If you get this current from, say, a vacuum tube diode, then the current would be truly shot noise limited. You would measure the shot noise by varying the current i, and the slope of $\langle V^2 \rangle$ versus i would give you the electron charge e. This is the essence of the experiment described in the paper by Livesey and McLeod, although their goals were much less ambitious.

[5]This, in fact, is what Johnson used in his 1928 paper. You might want to look it up, and compare your results to his.

Ch 15

Experiment 8: The Faraday Effect

Sometimes important physics discoveries are made by trying things mostly for the sake of curiosity. In 1845, Michael Faraday wondered if a magnetic field could affect the polarization of light, despite the fact that nobody really knew what light was or what matter was made from. Faraday did an experiment nevertheless, and discovered that when plane polarized light passes through some material in the presence of a longitudinal magnetic field, the plane of polarization rotates. This the Faraday Effect.

You will measure this effect in this experiment. We take the opportunity to do a neat experiment, making use of the lock-in amplifier to extract the signal from the noise. (You should probably review Sec. 13.3.3.)

Discussions of the Faraday effect and other polarization phenomena can usually be found in advanced undergraduate texts on optics. I suggest

- *Introduction to Optics*, Second Edition
  Frank. L. Pedrotti and Leno S. Pedrotti,
  Prentice Hall (1993), Section 26-6.
- *The Art of Experimental Physics*,
  Daryl W. Preson and Eric R. Dietz, John Wiley and Sons (1991)

Experiment 22.

The experiment done in Preston and Dietz is rather different from the way we do it, but the discussion of the physics is quite thorough. A good introduction to the basics can be found in

- *Physics*, Robert Resnick, David Halliday, and Kenneth Krane, John Wiley and Sons, Fourth Edition (1992)
  - Chap.40 *Maxwell's Equations*
  - Chap.41 *Electromagnetic Waves*
  - Chap.42 *The Nature and Propagation of Light*
  - Chap.48 *Polarization*

These chapters include discussions on the speed of light in matter and circular polarization. An excellent reference on optical techniques, including polarization detection and photodiodes, is given in

- *Experimental Physics: Modern Methods*, by R. A. Dunlap, Oxford University Press (1988); Chapter 10

15.1 Magnetically Induced Optical Rotation

It's worth going back to the beginning, so that's what we'll do. We will quickly come up to speed on the Faraday Effect, but the important concepts are clearest if we look first at the origin of electromagnetic waves. In particular, we will look at their speed of propagation. Then we will look at how media can modify that speed and lead to the Faraday Effect.

15.1.1 Electromagnetic Waves and Polarization

Maxwell's Equations are fundamentally important because together, they predict electromagnetic radiation. As presented in elementary texts (like

Resnick, Halliday, and Krane), Maxwell's Equations *in free space*, i.e. with no matter present, are

$$\oint \vec{E} \cdot d\vec{A} = 0 \qquad \oint \vec{B} \cdot d\vec{A} = 0$$
$$\oint \vec{E} \cdot d\vec{s} = -\frac{d\Phi_B}{dt} \qquad \oint \vec{B} \cdot d\vec{s} = \mu_0 \epsilon_0 \frac{d\Phi_E}{dt}$$

Of course, $\vec{E}$ and $\vec{B}$ are the electric and magnetic field vectors, which are in general functions of space and time. Φ_B and Φ_E are "fluxes", i.e. integrals of the fields over the enclosed regions. The quantities μ_0 and ϵ_0 are the magnetic permeability and electric permittivity *in free space*. We will be returning to this important point.

These equations are written in the "integral" form. All the physics is there, so that's fine, but it is cumbersome to proceed with the equations in this form. It is better to write them in the "differential" form, using some well known theorms of vector integral calculus.

Gauss' theorem relates the integral of a vector field $\vec{F}$ over a closed surface to a volume integral over the region enclosed, i.e.

$$\oint \vec{F} \cdot d\vec{A} = \int \vec{\nabla} \cdot \vec{F} dV$$

Stokes' theorm similarly relates the integral of $\vec{F}$ over the boundary of an open surface to the integral over that surface, i.e.

$$\oint \vec{F} \cdot d\vec{s} = \int \left(\vec{\nabla} \times \vec{F} \right) \cdot d\vec{A}$$

(The fluxes Φ_B and Φ_E in Maxwell's Equations are just integrals over the same areas dictated by Stokes' theorem.) This means that we can collect the integral forms all into integrals over volumes or surfaces, and if we let the integration region get very small, we force the same relations on the integrands. The result is *Maxwell's Equations in Differential Form*:

$$\vec{\nabla} \cdot \vec{E} = 0 \qquad \vec{\nabla} \cdot \vec{E} = 0$$
$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \quad \vec{\nabla} \times \vec{B} = \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t}$$

Now let's find an equation for the electric field vector $\vec{E}$ all by itself. Take the partial derivative with respect to time of the last equation:

$$\frac{\partial}{\partial t}\vec{\nabla} \times \vec{B} = \mu_0\epsilon_0\frac{\partial^2\vec{E}}{\partial t^2}$$

The partial derivative on the left "passes through" and takes the derivative of $\vec{B}$. Then use the third equation to replace this derivative:

$$-\vec{\nabla} \times \vec{\nabla} \times \vec{E} = \mu_0\epsilon_0\frac{\partial^2\vec{E}}{\partial t^2}$$

Now make use of a the "curl-of-a-curl" theorem:

$$\vec{\nabla} \times \left(\vec{\nabla} \times \vec{E}\right) = \vec{\nabla}\left(\vec{\nabla} \cdot \vec{E}\right) - \nabla^2\vec{E} = -\nabla^2\vec{E}$$

where we've used the first equation to eliminate the first term. The end result is the equation for $\vec{E}$ we wanted namely,

$$\nabla^2\vec{E} = \mu_0\epsilon_0\frac{\partial^2\vec{E}}{\partial t^2} \tag{15.1}$$

This is called the *wave equation*, and its solution is profound.

The general solution to the wave equation is simple. Look for a solution of the form

$$\vec{E} = E(z,t)\hat{x} \tag{15.2}$$

that is, where the electric field is a function only of z and t and points in the x-direction. Notice that there can be no z-component to the vector since $\vec{\nabla} \cdot \vec{E} = 0$ would not be satisfied in general. Also, if we just choose the x-direction to be the direction that $\vec{E}$ points towards, we can leave off the y-component with no loss of generality. This electric field vector is said to be *polarized* in the $x-$direction.

Inserting Eq. 15.2 into Eq. 15.1 gives

$$\frac{\partial^2 E(z,t)}{\partial z^2} = \mu_0\epsilon_0\frac{\partial^2 E(z,t)}{\partial t^2}$$

The solution for $E(z,t)$ has the very general form

$$E(z,t) = f(z-ct) + g(z+ct) \tag{15.3}$$

where $f(u)$ and $g(u)$ are *any* arbitrary function and

$$c \equiv \frac{1}{\sqrt{\mu_0 \epsilon_0}} \tag{15.4}$$

Consider what kind of function is $f(z - ct)$. At time $t = 0$ and position $z = z_0$, the function has the value $f(z_0)$. At some later time $t = t_0 > 0$, the function has the *same* value at the position $z = z_0 + ct_0 > z_0$. In other words, $f(z - ct)$ *represents a functional form which moves to the right with a speed* c. Such a moving functional form is called a "wave". Similarly, $g(z + ct)$ is a leftward moving wave. We call c the "speed of light in free space". It is uniquely predicted by Maxwell's Equations.[1]

The wave equation is linear, so we can add any two solutions and the sum is still a solution. Thus, we generally work with sine or cosine solutions for $E(z, t)$, realizing that Fourier analysis gives us the machinery to add them up to be anything we like. Furthermore, we usually work with $\omega = 2\pi\nu$ where ν is the frequency of the wave and $k = 2\pi/\lambda$ where λ is the wavelength, and $\omega/k = \nu\lambda = c$. Thus

$$\vec{E}(z, t) = E_0 \cos(kz - \omega t)\, \hat{x} \tag{15.5}$$

where E_0 is the arbitrary magnitude of the wave. We will only be working with rightward moving waves from here on.

The situation is summarized in Fig. 15.1, where Eq. 15.5 is plotted as a function of z at two different times $t = 0$ and $t = T$. The crest of the wave moves a distance $z = cT$ in time T. The wave is *linearly* polarized, i.e. the vector $\vec{E}$ always points in either the $+x$ or $-x$ direction.

Circular Polarization

We can write Eq. 15.5 for $\vec{E}(z, t)$ as the *sum* of two separate waves, namely

$$\vec{E}(z, t) = \vec{E}_R(z, t) + \vec{E}_L(z, t) \tag{15.6}$$

[1] You should realize that this speed is predicted independent of the reference frame. That is, light has speed c regardless of the motion of an observer relative to the source that emitted the wave. This goes against "Galilean" relativity and is the paradox that led Einstein to deduce the theory of Special Relativity.

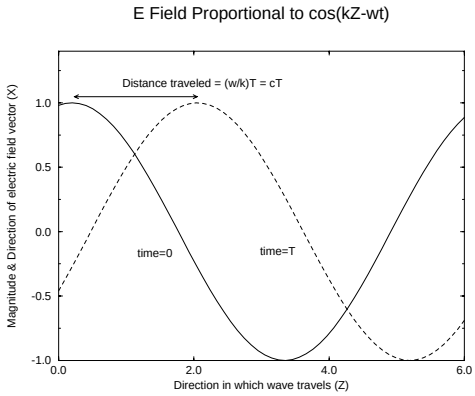


Figure 15.1: A linearly polarized electromagnetic wave.

where

$$\vec{E}_R(z,t) = \frac{E_0}{2}\cos(kz-\omega t)\,\hat{x} - \frac{E_0}{2}\sin(kz-\omega t)\,\hat{y} \qquad (15.7)$$

$$\vec{E}_L(z,t) = \frac{E_0}{2}\cos(kz-\omega t)\,\hat{x} + \frac{E_0}{2}\sin(kz-\omega t)\,\hat{y} \qquad (15.8)$$

Neither of these waves are polarized along a particular direction, but they are polarized in a different sense. Figure 15.2 helps explain this by showing a view of the vectors $\vec{E}_R$ and $\vec{E}_L$ in the x, y plane at $z = 0$ and at two different times $t = 0$ and $t = T$. At time $t = 0$, both $\vec{E}_R$ and $\vec{E}_L$ point in the x-direction. At time $t = T$, $\vec{E}_R$ has rotated through an angle $\theta = \omega T$ towards the $+y$-axis, and $\vec{E}_L$ has rotated through the same angle, but towards the $-y$-axis. That is, as time goes on, $\vec{E}_R$ and $\vec{E}_L$ sweep out circular motions about the z-axis. They are *circularly polarized* waves.

We say that $\vec{E}_R$ has *right-handed* circular polarization, whereas $\vec{E}_L$ is *left-handed*. These names come from the screw sense of the wave as it travels in the $+z$-direction, i.e. out of the page in Fig. 15.2.

This is important. We have shown that a linearly polarized light wave is the sum of a right-handed circularly polarized wave, and a left-handed circularly polarized wave. In the Faraday Effect, we will affect the left and right handed components differently from each other.

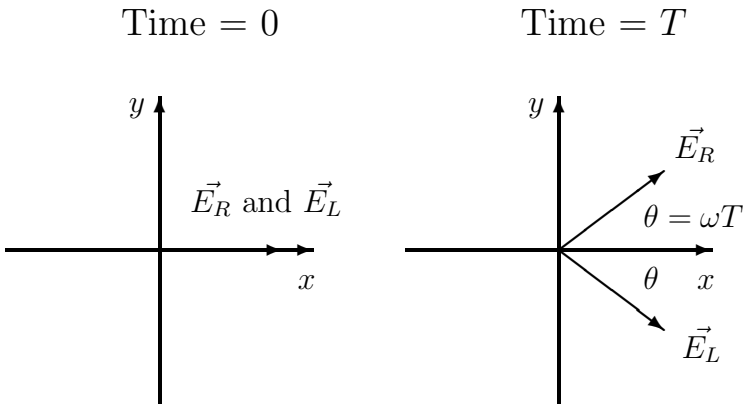


Figure 15.2: The meaning of circular polarization. These diagrams show the directions of the electric field vector components $\vec{E}_R$ and $\vec{E}_L$ for a wave moving in the z-direction, i.e. out of the page. Instead of always pointing one way, the vectors $\vec{E}_R$ and $\vec{E}_L$ rotate around the z-axis. In a time T, they each rotate through an angle $\theta = \omega T$, but in opposite directions. *At any time*, however, the total electric field vector $\vec{E} = \vec{E}_R + \vec{E}_L$ points in the $\pm x$ direction, and is linearly polarized.

15.1.2 Light Propagation in a Medium

Light is just the name we give to electromagnetic waves with wavelengths between ~ 200 nm and ~ 1 μm. See Chapter 11 for more details.

So far we have been dealing only with electromagnetic waves in free space. What happens if we let them propagate in a material? Well, this opens an enormous can of worms, but for light waves the discussion is rather straightforward.

If there are no free charges or currents in the material, Maxwell's Equations are unchanged except that μ_0 and ϵ_0 are replaced with their values μ and ϵ in that material. Recall that $\mu > \mu_0$ and $\epsilon > \epsilon_0$. Therefore the speed of the wave in the material is given by c/n where

$$n = \sqrt{\frac{\epsilon\mu}{\epsilon_0\mu_0}} > 1$$

is called the *index of refraction*. To be sure, both μ and ϵ are functions of

frequency, and are not typically the same as their static (i.e. $\nu = 0$) values. This gives rise to a wavelength dependence of n which is the reason white light can be separated into the colors of the rainbow using a triangular prism.

15.1.3 The Faraday Effect

Dramatic things happen when the index of refraction n is different for waves that have right or left circular polarization. Some materials in nature have this property. (They are called "birefringent".) The Faraday Effect is the observation that an applied magnetic field causes otherwise normal materials to become birefringent. Let's think a little bit about how this could happen, and then see how the effect can be observed.

Consider how ϵ, for example, might change in a material. (You may want to review the discussion in Expt. 4.) This is related to how charge might be "stored" on capacitor plates with the material between the places, and this is directly related to how polarized the atoms in the material become when an electric field is applied. In turn, the degree to which the atoms are polarized depends on how tightly bound are the electrons in the atoms. In other words, if the applied magnetic field affects the binding energy of the electrons in the atom, then you expect it to affect the propagation of light through the material.

Figure 15.3 schematically shows an electron in orbit about the nucleus of an atom with no applied field, and then with a magnetic field $\vec{B}$ both into and out of the page. There is clearly an additional central force on the electron given by $e\vec{v} \times \vec{B}$ when the field is applied, and this will change the binding energy. What's more, for a particular angular momentum state of the electron, this additional force causes the electron to be more tightly bound if $\vec{B}$ is one way, and less tightly bound if it is opposite. This is exactly the kind of effect which would cause ϵ, and so n, to be different for right or left handed circular polarizations.

The effect of the applied $\vec{B}$ field is not large. For an electron in the lowest Bohr orbit of the hydrogen atom (see Expt. 6), the force due to the magnetic

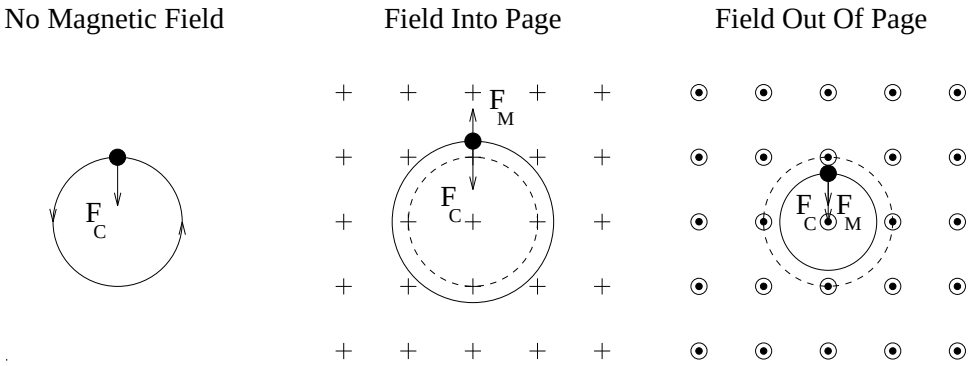


Figure 15.3: Schematic representation of the Zeeman Effect in an atom. The external magnetic field alters the electron orbits, and this affects everything from energy levels to the propagation of light past the atoms. The Coulomb force F_C and the magnetic force F_M are not drawn to scale, in particular $F_C \gg F_M$.

field is

$$F_{MAGNETIC} = evB = e(\alpha c)B = 3.5 \times 10^{-14} \text{ N}$$

for a (relatively large) 0.1 T magnetic field. On the other hand, the coulomb force is

$$F_{COULOMB} = \frac{1}{4\pi\epsilon_0}\frac{e^2}{r^2} = \frac{1}{4\pi\epsilon_0}e^2\left(\frac{mv}{\hbar}\right)^2 = 8.2 \times 10^{-8} \text{ N}$$

or about two million times larger. This effect of magnetic field on the binding energy is called the *Zeeman Effect*, and it can be directly observed in high resolution atomic spectroscopy experiments. In this experiment, we observe it's effect on the propagation of light through the material, i.e. the Faraday Effect.

So now suppose that there are two indices of refraction, n_L and n_R for each of the two circularly polarized waves 15.7 and 15.8. This means that the two components of the linearly polarized wave 15.6 will propagate with different speeds $v_{R,L} = c/n_{R,L}$. Because the wave must be continuous at the entrance and exit boundaries of the material, the frequency ν of the wave must be the same for the wave both inside and outside. Nevertheless, we must still satisfy the relation $\omega/k = c/n$ inside the medium to be consistent with Maxwell's Equations. Therefore, inside the medium, we must have $k_{R,L} = \omega n_{R,L}/c$.

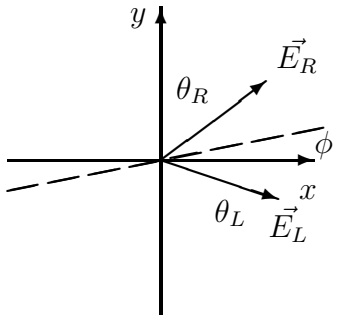


Figure 15.4: The rotation of the plane of linear polarization after different propagation times for the right- and left-handed circular components. Because $\vec{E}_R$ and $\vec{E}_L$ rotate by different amounts, the plane of linear polarization for $\vec{E} = \vec{E}_R + \vec{E}_L$ rotates away from the x-axis by an amount $\phi = (\theta_R - \theta_L)/2$.

Let's look at what happens to the right- and left-handed components by the time they emerge from the other end of the sample. Let the sample have length L. Therefore, the right-handed component comes out at time $t = L/(c/n_R)$ and the direction of $\vec{E}_R$ has rotated an amount (see Fig. 15.2) $\theta_R = \omega L/(c/n_R = kLn_R$ where k is for the wave outside the medium (i.e. the free space value). Similarly, $\vec{E}_L$ rotates by an amount $\theta_L = kLn_L$. Therefore, as shown in Fig. 15.4, the emerging wave is *linearly* polarized but in the plane rotated by an angle

$$\phi = \frac{1}{2}(\theta_R - \theta_L) = \frac{kL}{2}(n_R - n_L) = \frac{\pi}{2}L(n_R - n_L)$$

This is the Faraday Effect. The difference in the propagation speeds for the left- and right-handed components, induced by the external magnetic field, rotates the plane of polarization of light incident on some material medium.

If we want to quantify this further, we need to get deeper into the discussion of light propagation in matter. See Preston&Dietz for more details. It is plausible, and in fact true, that the difference $n_L - n_R$ is proportional to the applied magnetic field B. Therefore, we write the rotation angle ϕ as proportional to both L and to B as

$$\phi = VBL \tag{15.9}$$

Table 15.1: Verdet Constants for Distilled Water

λ (nm)	V (10^{-3}min/Gauss·cm)
Sodium D-Light	13.1
600	12.6
800	7.0
1000	4.4
1250	2.9

where V is called the *Verdet constant*. Clearly V is a function of wavelength, as well as the medium. Values for distilled water at various wavelengths[2] are listed in Table 15.1. Note that sodium D-light is essentially two lines between 589 nm and 590 nm, and that 60 min=1°. You will measure the Verdet constant in this experiment.

15.2 Procedure and Analysis

It requires a lot of power to generate a kG magnetic field, and we will not go that far. Instead, we will use a small but oscillating magnetic field. The size of the effect will be small, but the oscillations will allow us to pick it out of the noise.

It is a good idea to proceed stepwise through this experiment. First, demonstrate that you are detecting polarized light. Next, make a rough determination of the Verdet constant by analyzing the oscilloscope signal. Finally, make a more precise measurement by analyzing the signal with a lock-in amplifier.

The experimental setup is shown in Fig. 15.5. Your main source of polarized light is a HeNe laser. The magnetic field is supplied by a 1026-turn solenoid driven by the amplified signal of a waveform generator, in series with a monitor resistor. After passing through the sample and polarization analyzing filter, the light is detected in a photodiode. Your signal is based

[2]Data from the Handbook of Physics, Second Edition, Edited by Condon and Odishaw, McGraw-Hill, 1967

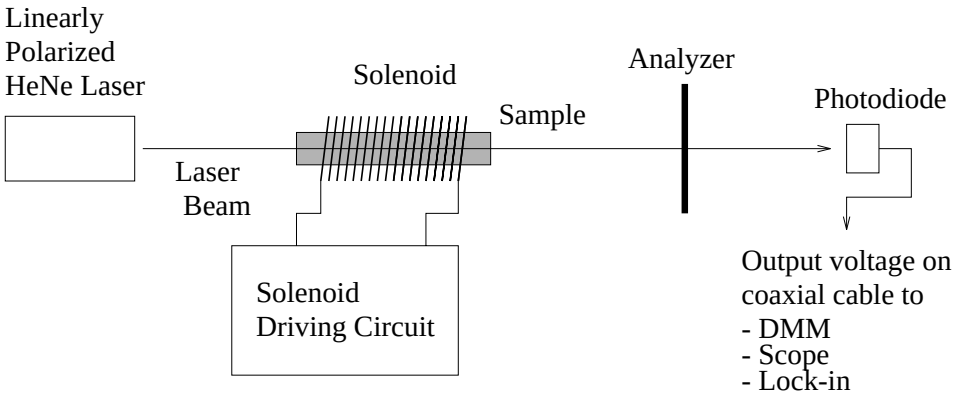


Figure 15.5: Experimental setup used for the Faraday Effect. The photodiode output goes to the DMM for the polarization calibration, and to the oscilloscope or lock-in to measure the Verdet constant.

on the output voltage of the photodiode.

15.2.1 Polarization Calibration

You will measure the polarization rotation angle ϕ by detecting the *change* in intensity of the light as it passes through the polarization analyzer. This intensity is converted to the photodiode voltage, so you need to know the conversion factor.

The analayzer is an etched sheet which filters out the components of the electric field vector that are not parallel to the etched lines. Therefore, the electric field $\vec{E}$ that emerges is reduced in magnitude by a factor of $\cos\phi$, where ϕ is the angle between the analyzer direction and the polarization of the incident electric field $\vec{E}_0$. The light intensity is proportional to the square of the electric field. Consequently, if linearly polarized light of intensity I_0 passes through an analyzer at angle ϕ, then the light emerges with an intensity

$$I(\phi) = I_0 \cos^2 \phi$$

This is called Malus' Law.

You will measure the intensity of the transmitted light by measuring the

output voltage V_D of the photodiode.[3] *The photodiode output is not a linear function of the intensity. See Dunlap.* Therefore, if you plot V_D as a function of the analyzer rotation angle, you will not get the $\cos^2\phi$ function of Malus' Law. On the other hand, V_D is a monotonically increasing function of the intensity, so you should see qualitatively the same thing.

Calibrate the polarization analyzer with the magnetic field off. Measure the photodiode voltage V_D with a DMM as you change the angle of the analyzer. It might be smart to simply confirm that V_D goes to zero if you block the laser light into the photodiode. Make a table of the values and plot the result. Does the plot look like what you expect? What do you think V_D looks like as a function of the intensity I, assuming that Malus' Law is correct?

Your Faraday Effect signal will be a fluctuation in the light intensity that is in time with the fluctuating magnetic field. (The intensity fluctuates because the polarization angle ϕ changes with magnetic field according to Eq. 15.9.) You detect this intensity fluctuation by measuring fluctations in V_D, so you want V_D to change by as much as possible as you change ϕ. Therefore, set the analyzer angle so that the slope $dV_D/d\phi$ is large. Use your calibration data to estimate $dV_D/d\phi$, and its uncertainty, at this angle.

Some sample calibration data is shown in Fig. 15.6. In this case, with the analyzer set at 163°, the change in photodiode voltage with angle is $dV_D/d\phi = 10.3$ mV/deg based on a linear fit to the set point and the points on other side. (The fit and plot were done in MATLAB with polyfit.) You should be able to estimate the uncertainty in $dV_D/d\phi$ by considering how much the slope might be different than the fitted value.

15.2.2 Applying the Magnetic Field

Now it's time to apply the magnetic field and first observe the Faraday effect. The magnetic field is provided by the 1026-turn solenoid coil around the sample, driven by a sinusoidaly varying current. The current is provided

[3] To be sure, the photodiode has an output *current*, which is converted to a voltage by passing it through a resistor.

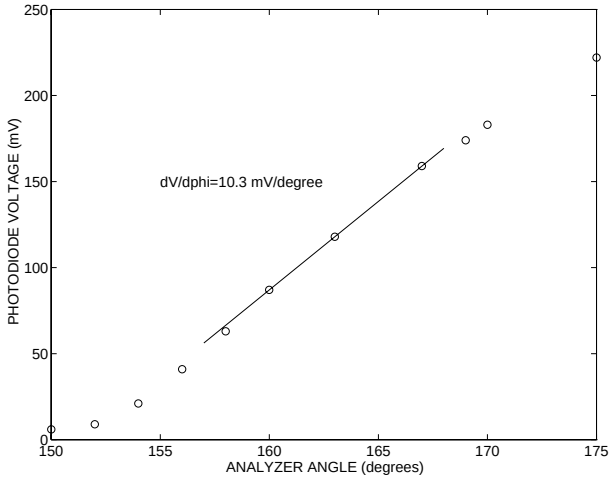


Figure 15.6: Sample polarization calibration data.

by an HP3311A waveform generator (sine wave, 600Ω output) amplified by the Bogen MU10 monaural audio amplifier. The driver setup is shown in Fig. 15.7.

The wave generator provides the input to the audio amplifier at the upper input in the rear panel. The output goes loops through the solenoid coil with a high power resistor R_{COIL} in series. (See Fig. 15.7.) You will determine the current, and so the magnetic field, by measuring the voltage drop across this resistor. *Do not ground either side of the amplifier output signal.*

Use clip leads on a coaxial cable to measure the voltage V_{COIL} across R_{COIL} on an oscilloscope. You want the shape to be a good sine wave with no DC offset and amplitude on the order of 10 V peak to peak. To do this, you have to adjust the amplitude of the HP3311A and the amplification (i.e. "volume") of the audio amplifier appropriately. You will also likely need to adjust the distortion on the amplifier so that the shape is alright.

Now take the photodiode output and connect it to the other channel of the oscilloscope. Set the scope trigger to fire on coil voltage, and look at both channels simultaneously. If the channel on which you measure V_D is DC-coupled, all you should see is a large DC level, corresponding to the mean light intensity on the photodiode. (This DC level should agree with

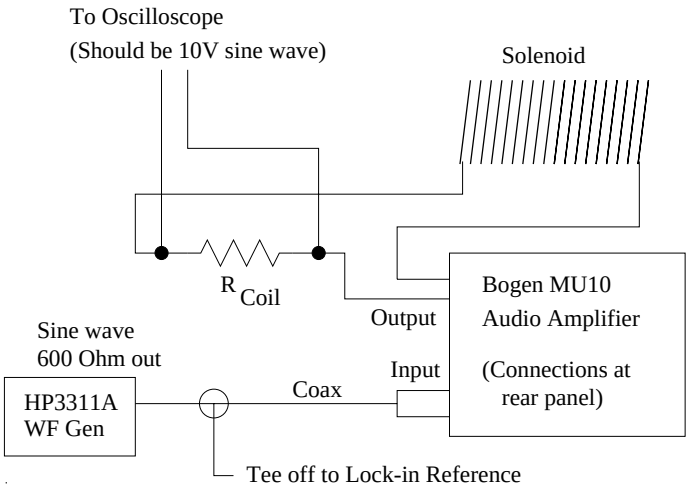


Figure 15.7: The driver circuit used to generate the oscillating magnetic field for measurement of the Faraday Effect and the Verdet constant.

what you measured with the DMM.) The Faraday Effect, on the other hand, shows up as a small oscillation on top of this DC level, in time with the V_{COIL}. You should just be able to see this small oscillation if you set the channel sensitivity to its lowest scale and AC-couple the input so that the large DC level is removed. You can confirm that the amplitude of these small oscillations move up or down with the amplitude of V_{COIL}, which is best adjusted by changing the amplifier gain. You can also confirm that the oscillations disappear if you block the photodiode from the laser. In fact, you can watch the amplitude of the oscillations change (and the phase reverse) if you rotate the analyzer, but remember to either recalibrate to find $dV_D/d\phi$, or set the analyzer back where you had it.

You can now check that you are getting about the right Verdet constant, although it is hard to do a careful job with the small signal you get on the oscilloscope. From Eq. 15.9, you know that the small changes in polarization angle $\Delta\phi$ are related to the changes in magnetic field ΔB through

$$\Delta\phi = V \cdot \Delta B \cdot L_{\text{SAMPLE}} \tag{15.10}$$

and from your calibration, you can convert $\Delta\phi$ to a change in photodiode voltage ΔV_D through

$$\Delta\phi \frac{dV_D}{d\phi} = \Delta V_D \tag{15.11}$$

The magnetic field in a solenoid of length L_{SOLENOID} and $N = 1026$ turns is given by

$$B = \mu_0 i_{COIL} N / L_{\text{SOLENOID}} \tag{15.12}$$

when a current i_{COIL} passes through the coil. By combining Eq. 15.10, Eq. 15.11, and Eq. 15.12 along with $V_{COIL} = i_{COIL} R_{COIL}$, you can obtain an expression for the Verdet constant V in terms of V_D, V_{COIL}, and other quantities which you know or can measure separately.

Make sure you use consistent definitions for V_{COIL} and for V_D. That is, if V_{COIL} is amplitude of the sine wave, then make sure you do the same for V_D.

15.2.3 Using the Lock-In

The lock-in amplifier allows you to measure oscillations in V_D more precisely than with the oscilloscope. Furthermore, the lock-in will remove any noise that is out-of-phase or is at the wrong frequency.

The lock-in is a PARC model 120 with a fixed reference frequency of $\sim$ 100 Hz. It is best used by defining the reference wave externally, but it needs to be close to 100 Hz so that the internal circuit responds correctly. Turn the lock-in mode dial to "SEL.EXT." and set the HP3311A to a frequency near 100 Hz and use a BNC Tee connector to apply the reference input to the lock-in, while the signal is on the way to the audio amplifier. This assures you that you are using a reference signal with precisely the same frequency as your Faraday Effect signal in V_D. The photodiode output should be connected to the lock-in input.

You still need to tune the phase of the lock-in amplifier so that you have maximum sensitifity to the oscillating V_D signal. There are a few ways to do this, but the most instructive is to use the oscilloscope.

1. With the oscilloscope still triggered on the V_{COIL} signal, use the other channel to view the "monitor out" port of the lock-in, with the switch set to "OUT×1", which is the basic output signal of the lock-in. If the time constant is set to a value much smaller than $(100\text{ Hz})^{-1}$ (1 msec

will do), then you should just get the sine wave folded with the reference signal oscillating between ± 1. That is, it should look pretty much like Fig. 13.9 or Fig. 13.10, or something inbetween, depending on the phase setting.

2. Adjust the phase knob so that it looks like Fig. 13.9, that is, symmetric about the cusps, and with the cusp points at ground level. If you flip the relative phase quadrant knob so that the phase is 90° lesser or greater, the trace should look like Fig. 13.10. It should change sign, on the other hand, if you flip by 180°.

3. With the phase adjusted so the output looks like Fig. 13.9, turn the time constant up to 1 sec or so. You can read the monitor out on the DMM, or use the meter on the lock-in. It is probably a good idea to block the light to the photodiode, and adjust the zero-trim so that the lock-in output is zero.

Vary V_{COIL} by adjusting the audio amplifier gain. (You shouldn't touch the waveform generator settings anymore, since it is now serving a dual role as both amplifier input and lock-in reference.) Make a table of V_D as measured with the lock-in and V_{COIL}. *Realize that the value of V_D provided by the lock-in is the RMS value, i.e. $1/\sqrt{2}$ times the amplitude.* Plot V_D versus V_{COIL} and make sure you get a straight line through zero. Either fit to find the slope or average your values of V_D/V_{COIL} to determine the Verdet constant with an uncertainty estimate. A sample of this sort of data is shown in Fig. 15.8.

15.3 Advanced Topics

You can run this experiment with a mercury lamp in place of the laser, and filters are available so that only particular lines in mercury will make it to the detector. This allows you to measure the Verdet constant as a function of wavelength, and the physics is very cool. See Preston and Dietz for an explanation of the physics.

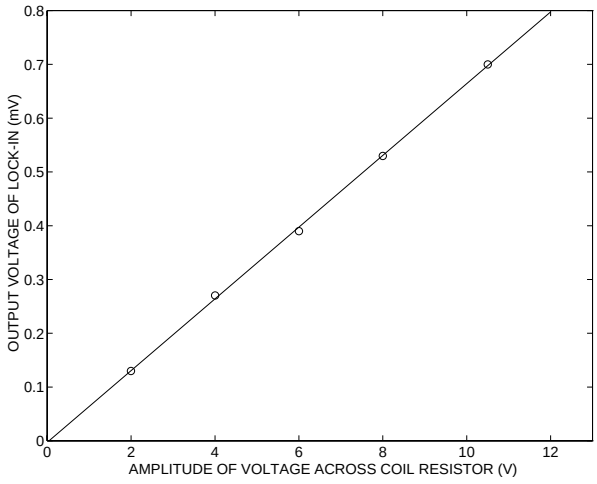


Figure 15.8: Sample data for the Faraday Effect using the lock-in amplifier.

Ch 16

Experiment 9: Nuclear Magnetic Resonance

The discovery of fundamental magnetism, at scales smaller than the atom, were crucial to our understanding of nature. It is easy to see why an atom might have a magnetic dipole moment, since it has electrons orbiting around the nucleus. Orbiting electrons obviously give rise to a loop of current, and this is the simplest way to create a magnetic dipole.

However, it is not so easy to see how the electron could have a magnetic moment all to itself. The same goes for protons or neutrons. The only way to classically understand such an effect would be if these particles were "spinning", but the velocity of the particle at its surface (assuming you have a clear idea of its radius) is so large that it would be faster than c. Therefore, the observation of such magnetism meant that quantum mechanics has to play a role, and the result is physics that cannot be understood at all without it.

In this experiment, you will study the magnetism of the free proton. You will use a technique that allows this magnetism to be very precisely picked out. This technique is known as *Nuclear Magnetic Resonance*, or NMR for short. You may also know it from the medical profession, where it is used to image the location of protons (i.e. water) in the body using "Magnetic Resonance Imaging", or MRI.

The essential physics is covered in

- *Introduction to the Structure of Matter*,
  John J. Brehm and William J. Mullin, John Wiley and Sons (1989),
  Chapter 8

but this book mainly deals with electron spin magnetic resonance. Good references that not only describe experiments done at other universities, but also include the essential physics, are

- *The Art of Experimental Physics*,
  Daryl W. Preston and Eric R. Dietz, John Wiley and Sons (1991)
  Experiment 15, and preceding discussion.
- *Experiments in Modern Physics*,
  Adrian C. Melissinos, Academic Press (1966)
  Chapter 8.

16.1 Nuclear Magnetism and Precession

Recall the behavior of a current loop in an external magnetic field. The loop tends to line up so that the magnetic field it generates along its axis points in the same direction as the external field. See Fig. 16.1. It takes a torque to change the orientation of the loop, that is, to change the angle between the axis and the external field. All forces are conservative, and we define a potential energy

$$U = -\vec{\mu} \cdot \vec{B} \tag{16.1}$$

where $\vec{B}$ is the external field and $\vec{\mu}$ is a vector that points along the axis of the loop, in the direction of the generated field according to the right hand rule. This all follows from simple classical physics if we define the magnitude of $\vec{\mu}$ to be iA where i is the current in the loop and A is the area it encloses. We call $\vec{\mu}$ the *magnetic moment* of the current loop.

Subatomic particles like protons, neutrons, and electrons all have an "intrinsic" angular momentum called *spin*. This is a purely quantum mechanical

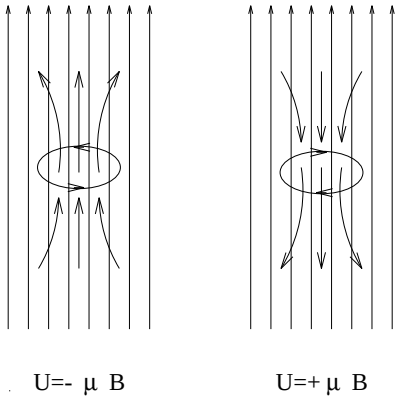


Figure 16.1: A current loop in an external magnetic field.

phenomenon, and in fact the z-component of the spin is in general quantized in units of $\hbar/2$. As with quantum mechanical orbital angular momentum, the total spin vector has magnitude $\hbar\sqrt{\frac{1}{2}\left(\frac{1}{2}+1\right)}$. For protons, neutrons, and electrons, S_z can only be $\pm\hbar/2$, relative to some directions, generally that of some external magnetic field. Furthermore, these particles also have magnetic moments, which is particularly surprising for the neutron since it is electrically neutral.

Around 1933, P.A.M. Dirac devised a relativistically consistent theory of quantum mechanics which accounted nicely for particles with $S_z = \pm\hbar/2$ spin. This theory also predicted that these particles should have magnetic moments given by $q\hbar/2m$ where q and m are the particle's charge and mass. The predicted magnetic moment[1] of the electron is $\mu_B \equiv e\hbar/2m_e$, a unit called the *Bohr magneton*. The prediction for the proton is $\mu_N \equiv e\hbar/2m_p$, called the *nuclear magneton*.

Dirac's theory works quite well for the electron. In fact, $\mu_e = \mu_N$ to within about one part in 1000. What's more, the small correction turns out to be precisely and accurately predicted by Quantum Electrodynamics, an extension of Dirac's theory that quantizes the electromagnetic field as well.

On the other hand, the story is quite different for protons and neutrons.

[1]Our language is somewhat cavalier. We are really only talking about the z-component of the magnetic moment.

The proton magnetic moment is $\mu_p = 2.79\mu_N = (g/2)\mu_N$, where g is called the *gyromagnetic ratio*. For the neutron, $\mu_n = -1.91\mu_N$ whereas the Dirac theory predicts zero. For protons and neutrons bound together into a nucleus, the problem is considerably more complex. The serious disagreement with Dirac's theory points to the fact that protons and neutrons are not really "elementary" particles, but we will not pursue this further. Instead, we will focus on Nuclear Magnetic Resonance, a very precise technique for measuring magnetic moments of protons, neutrons, and nuclei.

Consider a proton suspended in an external magnetic field, taking the place of the current loop in Fig. 16.1. Because of the way its angular momentum is quantized, the proton has two discrete "spin states" with magnetic moments whose z-projections are $\pm(g/2)\mu_N$. Since the z-direction is defined by $\vec{B}$, Eq. 16.1 tells us that the energies of these two spin states are $E_0 \pm (g/2)\mu_N B$, where E_0 is the proton energy in the absence of any magnetic field. The difference in energy between these states is just

$$\Delta E = g\mu_N B \tag{16.2}$$

Transitions between these two states will occur with the emission or absorbtion of photons with energy ΔE. These are in fact very low energy photons, and we tend to think of them in different terms. For a relatively large magnetic field $B = 0.5$ T=5 kG, you find $\Delta E = 8.8 \times 10^{-8}$ eV, and we generally talk about the frequency of the photons which come with these transitions. The photon frequency ν is given by

$$\nu = \frac{\Delta E}{h} = g\frac{\mu_N}{h}B = g\frac{e}{4\pi m_p}B \tag{16.3}$$

Again, for $B = 0.5$ T, we find $\nu = 21.3$ MHz or $\lambda = 14$ m. These are typical of radio waves, and these are called Radio Frequency or RF transitions.

Nuclear magnetic resonance is the application of RF waves that cause protons to jump between these two spin states. There are some things that are very different from other types of quantum mechanical photon emission or absorbtion. These differences come from both the fact that an external magnetic field sets up the energy difference, and from the fact that the energy difference is very small. We will examine these consequences now.

There is clearly a directionality to these states since they are setup by the applied $\vec{B}$ field. Notice that the total proton spin vector is longer than the z-component, so the spin is tipped at some angle relative to the z (or $\vec{B}$) axis. Therefore, the $\vec{B}$ field exerts a torque on the magnetic moment vector, and the spin *precesses* about the z-axis. The precession frequency is straightforward to calculate classically (see Melissinos), and you find that $\nu = \omega/2\pi$ where

$$\omega_{PRECESS} = \frac{g}{2}\frac{e}{m_p}B$$

Notice that the classical precession frequency turns out to be the same as the quantum mechanical transition frequency! This is not an accident, as we will now see by considering what it takes classically to cause a transition from one state to the other. As you look at the proton spin vector, it precesses about the z-axis. Now, picture a time-varying magnetic field $\vec{B}_1$ that is also rotating about the z-axis in the same direction as the precessing spin vector, and lying in the x, y plane. For now, the precession speed of $\vec{B}_1$ is arbitrary. Imagine that you "jump onto" the vector $\vec{B}_1$ and watch the motion of the proton spin vector. In general, the spin vector will move in some complicated way, a combination of its own precession about $\vec{B}$ and your own rotating motion.

Now imagine that your own rotating motion, i.e. the rotation speed of $\vec{B}_1$ exactly matches the precession. What do you see? You think, "the proton spin vector is moving along with me, so it doesn't appear to be moving at all". However, you are attached to another magnetic field, namely $\vec{B}_1$! The spin vector will precess about $\vec{B}_1$ instead. This corresponds to the spin vector going from "up" to "down" and back again in the original, nonrotating frame. *The rotating magnetic field vector $\vec{B}_1$ causes transitions between the two spin states if it rotates at the precession frequency.*

So, we've learned that not only is there a good reason for the classical precession frequency to be equal to the quantum mechanical transition frequency, there is clearly a need (from the classical standpoint surely, and it holds up in a quantum mechanical treatment) for the RF transition "photons" to have a magnetic field component $\vec{B}_1$ rotate in the same direction as the precessing spin. Notice that if the proton is in the other spin state, the precession changes direction and we need a $\vec{B}_1$ component moving in that

direction as well if we are to make transitions the other way. This is easy to accomplish just by using a *linearly* oscillating magnetic field, which is the sum of two fields oscillating circularly in opposite directions. (See the discussion on circular polarization in Sec. 15.1.1.) The experimental apparatus will do exactly this.

What effect does the smallness of ΔE have this measurement. Notice that at room temperature, the thermal energy $kT \approx 1/40$ eV, or $\sim 3 \times 10^5$ larger than ΔE. Consequently, mere thermal collisions will continually excite and deexcite the upper spin state, and at room temperature the number up relative to the number down is $\sim e^{-\Delta E/kT} \approx 1 - \Delta E/kT$. In other words, the difference between the number up and the number down is around three parts per million. With a linearly oscillating RF field $\vec{B}_1$ which has both right and left-handed components, nearly an equal number up and down transitions will occur. This small population difference will therefore account for our signal, and it will be small as well.

16.2 Measurements

The remainder of this chapter is just lifted from the final report written by Paul Bilodeau, Erik Mohrmann, and Dan Stouch from their work with the NMR setup.

The interfaces and connections of each device are described and shown in Fig. 16.2. The signal carried between each device is also shown. The user should be able to perform experiments without too much time wasted on discerning the function of each component and how to change settings.

We'll start with the power supplies. Power supply #1 (PS1) is in a circuit with the electromagnet through two leads on its rear. These leads are attached via hoop clips to screws on the rear surface on the lower crossbeam of the magnet. The only interface that should be touched on PS1 is the current control knob. In our work, the current was at most about 11 A and usually set at 9.5 A. It should be noted that the magnet is rated for 18 A and 15 VDC.

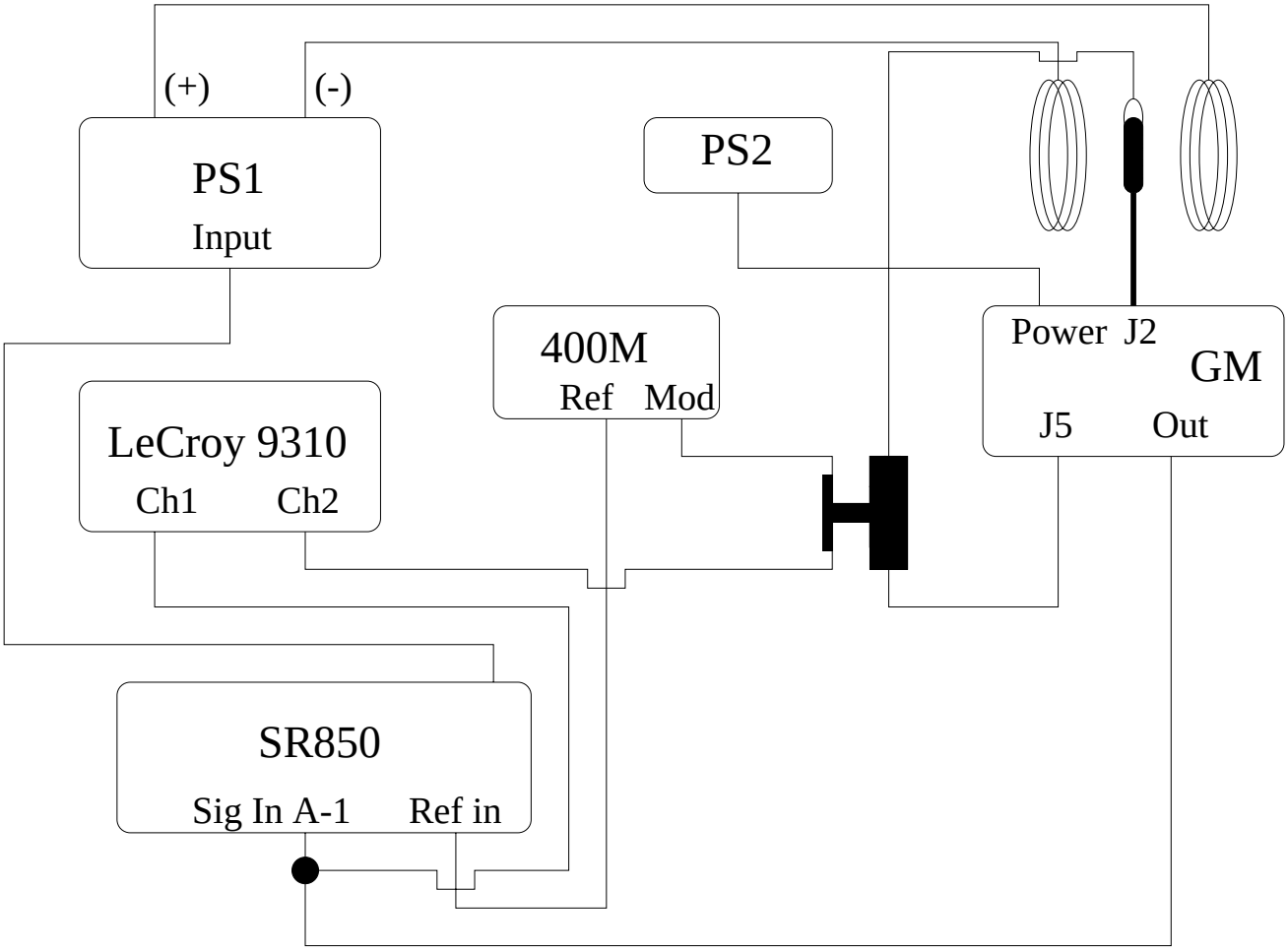


Figure 16.2: Setup for the NMR measurements, including the SR850 lock-in amplifier for signal measurement and storage.

The input jack on the front of PS1 is connected to Auxiliary Output 1, which is on the rear of the Stanford Research Systems SR850 lock-in amplifier. It is through this connection that the SR850 can be programmed to vary the magnetic field. The signal to the the PS1 input jack varies depending on what one wants to do with the equipment. To keep the magnetic field fixed, a constant voltage is sent from the SR850. This is usually 0V, but can be incremented using the AUX OUPUTS menu on the SR850. If a significant current increase is desired, it is easier to adjust the current with the control knob on PS1 and use the SR850 for fine-tuning. For scanning through resonance, the signal can vary through a range set in the AUX OUTPUTS menu with a linear scan.

Power supply #2 (PS2) provides power and a modulated 60 Hz signal to the "gauss meter" (GM). (This is actually the power supply to an old system for measuring magnetic fields with NMR, hence the name.) The power switch and knob labeled "Modulation" are the only interfaces that do anything meaningful. All of the other knobs are for use with the PS2 display screen, which does not work. The "Marker Generator" knob only turns the "Marker Generator" light on and does nothing applicable to the experiment. PS2 connects to the gauss meter through a cable to the port labeled "POW OSC" on the rear of the gauss meter.

The GM's connections are somewhat complex. All of the input and output jacks are on the rear of the device, except for the ouput to the frequency counter. This output is on the front of the GM, at the top left, and is unlabeled. The J2 jack is the interface with the NMR probe that holds the samples between the magnet poles. "SIG OUT" is the output for the NMR signal, and is connected to the A-1 "Signal In" port on the SR850. This signal branches into channel 1 on the LeCroy 9310A digital oscilloscope and is on the order of 200 mV and 60Hz.

J5 is an output that carries the modulated 60 Hz signal generated by PS2. This is 10V peak to peak when the modulation is set at its maximum value (the knob turns beyond the maximum labeled value of 10). The J5 leads are connected across a resistor to the leads of a Helmholtz coil that is mounted in the NMR probe head. The Helmholtz coil is wound such that the plane of its coils are parallel to the plane of the magnet coils. It is the oscillating magnetic field of the Helmholtz coil that allows us to see the resonance as a

periodic signal. The resonance peaks occur at the points of inflection of the modulated sine wave sent through J5.

The signal from J5 is added to the signal coming out of the 400 Hz Modulator (400M) "Mod Out" port. This combined signal is sent to channel 2 on the LeCroy. The "Mod Out" signal is a 375Hz sine wave. With the switch on the 400M set on "Hi" and the knob turned to its maximum value, the amplitude is approximately 2V peak to peak. With the switch set on "Lo", the amplitude drops to approximately 200mV. The knob is for adjusting the amplitude, and the switch changes voltage scales. The "Ref Out" port sends out a slightly warped square wave that is 375 Hz and roughly 2V peak to peak that is not controlled by the knob or the "Hi/Lo" switch.

It is interesting to note that the setting of the "Hi/Lo" switch does have an effect on the results of scans through resonance. When set to "Hi", the signal has a larger amplitude and an upper peak. When set to "Lo", the signal loses the upper peak and its amplitude drops. Clearly, the switch must be set to "Hi" to get a reasonable estimate of the FWHM of the signal to calculate the magnetic field width of the resonance.

16.2.1 Equipment Settings and Parameters

A full understanding of the equipment used in the experiment allows the adjustment of parameters for the nominal taking of data. Of the equipment used in the experiment most of them have some adjustable features, while a few, namely the LeCroy Digital Oscilloscope and the SRS Lock-in Amplifier, have far more features than will be used in this setup.

The M-2 Precision Gauss meter varies the frequency of a driving signal, and then picks up the response of the sample, i.e., resonance. Several of the adjustments on the gauss meter need not be adjusted. The probe knob should remain on "blue", and the RF frequency should remain in the vicinity of three or four.

The generator for the gauss meter has only one knob of interest and, considering the equipment's great age, function. The modulation may be adjusted across the full scale, from zero to ten. As the modulation is increased

the width of the signal decreases, and its sharpness and depth increase. This also moves the frequency at which resonance will be found. Generally it is desirable to keep the modulation as high as possible, but I believe experimentation here may yield some insights into the workings of the apparatus, as is the case with the next piece of equipment, the 400M.

This modulator affects the signals to be processed, as opposed to the modulator on the gauss meter that affects the signal in the magnetic driver (Helmholtz coil). The optimum setting for this is in the 6 to 6.5 range. At lower frequencies under-sampling occurs, and the signal is not processed correctly. When frequencies are adjusted higher over-sampling occurs, and unnecessary noise is allowed to enter the signal. Although most of this can be removed by the lock-in the best bet is to get the cleanest signal to start.

If using a LeCroy Digital Oscilloscope (which is a necessity if T2 decay wiggles are to be observed and/or analyzed) the divisions/sec should be adjusted so that one complete wave of the reference signal is visible on the display. This will allow three resonance peaks to be visible (at the points of inflection of the sine wave), and equalizing their amplitudes assures resonance. Triggering should be set to channel two (the reference channel), since the signal varies too much for reliable triggering. The display to screen for channel two can the be turned off, for it is not helpful in the finding of resonance. Channel one should have the volts/div set so that the signal is cleanly visible, but does not leave the display. The offsets should be set in whatever fashion allows easiest viewing of the signal, since the numerical data here is not pertinent.

A few of the features of the LeCroy that an analog scope does not possess are helpful for this experiment. By pressing the "wave form store" button, and by following the menus as far as location and format are concerned, data can be saved to memory or disk for later retrieval. The "SNGL" button in the upper right of the top left panel allows a single sweep to be taken, and with the LeCroy's precision, allows the discerning of T2 wiggles. Once the signal has been taken to resonance using the oscilloscope one of the alligator clips is removed (drastically changing the signal on the scope) and attention is turned to the SRS Lock-in Amplifier.

The lock-in has a treasure trove of functions and parameters, a large

number of which are useful in this experiment. If all else fails, the manual for the lock-in is very informative and well written. A button by button description is in order for this device.

Reference/Phase. This button brings up a menu that allows setting of the interactions between the reference and signal. For this setup the reference must be set to external, and all of the other parameters can be left at default – a sine wave of 1V rms, and phase shift set to zero degrees.

Input/Filters. Nothing should be adjusted on this menu. It is imperative that the 'source' remain set on a, and the 'coupling' stay set for ac.

Gain/TC. I would encourage experimentation with this menu. Vary the filtering strength, time constant, and sensitivity to receive maximum results. In general best results occur when the time constant is set to 100 ms and the sensitivity is set as small as possible without causing a "chopping" or clamping of the signal. It is very obvious when this occurs because resonance signals will display a flat top instead of a peak. The maximum sensitivity varies depending on the material being sampled (or more accurately, with the incoming signal). For the first run on a particular substance it might be advisable to set the sensitivity at 1V, and then reduce it for further runs. The filtering should probably be left at 12dB, but adjustment here is fine. The minimum filtering gives a very choppy signal, while higher dampening smoothes out the peaks, making the measurements slightly less precise. This uncertainty is most likely overwhelmed by other uncertainties in the experiment, but noise removal is always good.

Output/Offset. This menu is important only if output from the front panel of the lock-in is used, so it is unused in this experiment.

Trace/Scan. This menu can make the difference between no signal and the gorgeous loop one expects. Trace one is currently set to 'X' which gives an account of the resonance scan. This scan is not very useful for what is being

done in this experiment. Trace 2 is set to display 'Y', which is the derivative of the signal, and yields a beautiful trace. By measuring the time from the top peak to the bottom the full width half max may be measured (assuming the signal is a gaussian or similar shape, which it approximates). Sometimes a cleaner signal can be obtained by setting one of the traces (currently 4) to display Y/X. If this is done be sure to adjust the scale under Display/Scale. The lock-in should also be set to store all of the traces, and in can be set to either loop or single mode depending upon the preference of the individual. The sample rate should be kept high (it is currently set at its max, 512 Hz). The scan length may be adjusted, but values of either ten or twenty seconds simplify the math and are appropriate values. Longer scans take too long to go through resonance.

Display/Scale. This menu allows the adjustment of the display so that one may view the data in a way that is meaningful to the eye. The settings here will be adjusted constantly, most notably the scale, which can vary from 10-6 to 100, and the offset, to center the trace on the screen. Information is most easily discernible in the chart format (reading a moving bar graph is a difficult endeavor at best).

Aux Output. Only output one is used in this setup, so the setting on the rest may be ignored. This is the menu that is used to make the actual scans through resonance. It is very important that the output be set to fixed, and the offset to 0.00000 when looking for resonance, otherwise the scan will not be centered around resonance. When a scan is to be taken the alligator clip should be removed at the wire junction, and the output should be shifted to linear (a logarithmic output would be particularly hard to decipher). The best results occur when resonance is scanned through as fast as possible, so the min and max voltages should be set at least 1V apart. The offset must be used to center this interval about zero. Note that this means a negative offset is necessary, otherwise many headaches will result when resonance does not appear!

Math. This menu is almost entirely unused. It is used for finding various fits and other functions of the trace, but these are uninteresting to the current

application.

System Setup. These settings should all be left at default.

Disk. This menu is of use for the saving of data. Pressing this button causes a menu to appear on the screen that may be manipulated using the "soft" buttons. The most useful way to save data is to save it to disk as an ASCII file so that it may be analyzed using Matlab or a similiar program. File names are entered by using the number pad or by using the 'alt' key and the subscripted buttons.

Help. The button for when all else fails. The lock-in has a relatively comprehensive "online" help index. Pressing the help button and then any other button (even one of the buttons whos function changes) will yield a screen which tells the current function and settings available for that button.

The other buttons of interest on the lock-in are the number entry keypad, which can be used to enter values much faster than by using the spinning knob, when applicable, and the row of buttons across the top. The Start/Cont button does just that, it starts a trace or continues one that has been paused. The Pause/Reset button may be used to pause a trace in the middle (note that pausing during the actual sweep though resonance gives lousy results) or by pushing again and then pressing enter the trace is cleares. The cursor button enables the cursor and allows the time values to be read off the display to the thousandth of a second! The remainder of buttons, particularly the 'auto-buttons,' are not of use in this experiment. They either do not need to be adjusted, or give settings that are not useful with the kind of data being explored.

The final piece of equipment that is adjustable is the magnet power source itself. This is set at 5 on the dial, which yields just below 10 amps. If looking for resonance at a low frequency, and low frequency noise is a problem then the current may be increased, thereby increasing the frequency that resonance will be found at; however either very low or very high currents produce poor results. The magnet is rated up to 18 amps, so overpowering the equipment

is not a risk.

16.2.2 Procedure and Analysis

Throughout this section, there will be references to various magnetic fields. B0 refers to the field generated by the large electromagnet. B1 is the oscilatting magnetic field generated by the converted gaussmeter. BT is the magnetic field at the probe where the sample rests.

Setup. With this understanding of the equipment that is in use the technique for taking data may be described. The first thing that must be done is to turn on the magnet and gauss meter power sources at least a half hour before the start of data taking. During this initial period the magnet is warming up at such a rate that the signal drift during a data run can be prohibitive.

Having allowed the system to warm up there are several things that should be double checked before the start of every run. The sample should be placed into the signal probe, and then the probe should be centered in the magnet as exactly as possible, since small variations can have profound effects upon the resultant data. Lining the sample up includes making sure that it is centered front to back as well as top to bottom, and also making sure that the probe is aligned with the magnet; i.e. the coils of the probe are in the same plane as the magnet's poles.

Tuning to Resonance. The modulator should now be checked to see that the switch is in the "low" position and the alligator clips should be checked to see that both are firmly connected. Next comes the task of making sure all of the settings on the lock-in are in proper order. If this is the first run for a certain sample then the sensitivity (under the Gain/TC menu) should be set at 1V. The Aux Output must be set to fixed, and to an offset of 0.0V.

Having checked all of these initial setting the signal can now be tuned to resonance. Using the signal on the oscilloscope adjust the frequency on

the gauss meter until "dips" or "valleys" appear at points sharing the same slope on the reference sine wave. This can then be fine tuned so that the valleys are equidistant. This is most easily done if the scope is set so that one complete reference signal appears showing three of the resonance dips.

The Magnetic Moments. The magnetic moment of the proton can be calculated from measurements of the resonant frequency and the value of BT. Once resonance is achieved, use the frequency counter to measure the frequency and the hand-held gaussmeter to measure BT. After recording your values, you can change B by changing the voltage output of auxillary output #1 under the AUX OUTPUTS menu on the SR850. A good range of voltage outputs is -1 V to 1 V. Make sure that the output is set on fixed voltage. Given BT and the resonance frequency, the magnetic moment can be calculated using Eq. 16.3.

It is easiest to first find the proton resonance with an oil sample. After you have found resonance, record the frequency. The teflon resonance is rather small compared to the oil, so you might want to increase the modulation amplitude. The 19F moment can then be calculated from the proton moment and the ratio of their frequencies.

T2 with the SR850. The spin-spin relaxation time (T2) can be calculated from the width of the resonance peak in Teslas. This can be found by telling the SR850 to scan over a B range and converting the time interval in resonance into a B width.

Once the signal is tuned to resonance one (and only one) of the alligator clips is removed from the signal junction, and the modulator switch should be adjusted to high. The lock-in should have the Aux Output adjusted to linear, and the settings adjusted as detailed above. A data run may then be taken simply by pushing the "Start" button. This can be repeated many times quite simply by clearing the data (push the pause/reset button once or twice and then hit enter), and pushing start again. Data may be saved using the disk button and then following the menu on the screen to input the desired filename and location of save, usually to $3\frac{1}{2}$" disk.

There are two methods of measuring the magnetic field. First, the hand gaussmeter can be used. Place the probe in the "Zero Chamber", set the readout to the highest level, and use the tuning to adjust it to zero. Then adjust the meter to the next most sensitive level and again adjust to zero. Continue doing this until the gaussmeter reads zero (or close thereto) on the lowest level. Then using the Aux Outputs on the lock-in set the voltage to the smallest value that was attained during the data set (this is equal to the starting voltage plus the offset on the linear mode). Measure the magnetic field at the sample and record. The adjust the voltage to the highest level that was reached and repeat. The magnetic field can be assumed to vary in a linear fashion with the current throughout the sweep.

The other possibility is to tune the RF frequency to resonance at the extreme values of B0 and derive the values of B0 at these points using the theoretical value for p.

The difference between initial and final B divided by the scan length gives you the rate of change of B with respect to time, or dB/dt. This multiplied by the time the sample was in resonance gives the B width necessary to calculate T2.

Once the first run is taken it is often useful to alter the sensitivity to the lowest point at which a clean resonance signal is still measured. This point is different for every sample, depending on the signal strength.

T2 From the Wiggles on the LeCroy. The final part of this experiment, determining the T2 relaxation time through the "wiggles" on the oscilloscope, is only possible when using a very uniform magnetic field, and at the moment this precision is not available with the current magnet. Although this is not currently practical due to the inhomogenity of the magnetic field B0 about the sample, the process would entail the following.

Find resonance in an oil sample on the oscilloscope. Adjust the display so that only a single peak can be seen. Using the math functions, take an average of the signal. You would see wiggles at the end of the peak. Record the trace and save it to disk. A plot of ln(wiggle amplitude) vs. time will allow you to calculate T2.

16.3 Advanced Topics

16.3.1 Spin Relaxation Times

16.3.2 Magnetic Moments of Nuclei

Ch 17

Elementary Particle Detection

What is matter "made of"? Matter is made of molecules, and molecules are made of atoms, and atoms are made of electrons and nuclei, and nuclei are made of protons and neutrons. But what are protons and neutrons are made of? There is no simple answer, and they key lies in the interchangeability of matter and energy. It takes more energy to break molecules up into atoms, than to break matter into molecules; more energy to break atoms into electrons and nuclei than to break molecules into atoms; and so on. As you get smaller and smaller, the energy scales get larger and larger. By the time you get to the size of the atomic nucleus, around several fm=10^{-15} m, the energy scale is around an MeV=10^6 eV. When you look at protons and neutrons, distances are even smaller and energies are higher.

Something fundamentally different is happening at these small distances. In fact, you can't break up protons or neutrons any more. If you put in more energy, you make new particles like π's, μ's, and K's. This is the realm of nuclear and particle physics. We will do some experiments in this field, mainly using nuclear physics phenomena since they are well suited to the undergraduate laboratory. We will be working with energies on the scale of MeV, or around a million times larger than what we are used to in atomic physics and optics. Photons have wavelengths a million times smaller, which is much smaller than the size of the atom (but still much larger than the nucleus).

There are some differences in the experimental techniques we use in this field, and the stuff we've talked about so far. For one thing, we will be doing a lot of "particle counting" to understand what goes on. The instruments we use are somewhat different, because they will measure "pulses", where each pulse corresponds to an elementary particle. What's more, the processes are fundamentally random, so the formalism we derived to describe random uncertainties in Chapters 6 and 9 will be particularly useful here.

There are lots of good books around that discuss the techniques used in nuclear and particle physics. I suggest the following:

- *Techniques for Nuclear and Particle Physics Experiments*, Second Edition, by W. R. Leo, Springer-Verlag (1994)
- *Experimental Physics: Modern Methods*, by R. A. Dunlap, Oxford University Press (1988); Chapters 11 and 12.
- *Experiments in Modern Physics*, by A. R. Melissinos, Academic Press (1966); Chapter 5.
- *The Review of Particle Properties*, by the Particle Data Group, Published every two years, the latest version appears in Physical Review D **50**(1994) Page 1173. You can request this book and other materials by sending email to pdg@lbl.gov.

For a basic review of nuclear or particle physics, any modern introductory textbook will do. I recommend

- *Physics*, Robert Resnick, David Halliday, and Kenneth Krane, John Wiley and Sons, Fourth Edition (1992); Chapters 54 through 56.

17.1 Ionizing Radiation

We use the term "radiation" to generically describe the stuff that comes out of a nuclear reaction, but we're really talking about elementary particles. If

there are no particle accelerators around, there are only four types of particles you normally encounter:

> α *particles.* These are ^{4}He nuclei which are ejected by very heavy nuclei like Uranium or Plutonium. Since the ^{4}He nucleus is two protons and two neutrons, α particles have charge $+2e$.
>
> β *particles,* sometimes called β rays. These are electrons or positrons (i.e. antimatter electrons), emitted in some nuclear decays. They have charge $-e$ (electrons) or $+e$ (positrons).
>
> γ *rays.* These are the photons emitted when a nucleus makes a transition from one energy level to a lower level. It is just like the optical photons emitted in atomic transitions, but the energy is higher and the wavelength is much smaller. Photons have no charge, of course.
>
> n*, or neutrons.* Neutrons can be emitted in simple, low energy nuclear reactions. They have no charge.

Other types of particles include protons p, deuterons d, other atomic nuclei, the π and K mesons, and the muon μ, but we won't run into these in this context.

Because the energies are rather high, a collision between an elementary particle and an atom of matter can knock an electron (or perhaps several) out of the atom. That is, the atom is "ionized", and elementary particles at these energies are collectively called "ionizing radiation". *Ionization is the primary principle used for nearly all forms of elementary particle detection.*

Ionizing radiation makes particle detection possible in two ways. First, if the detector material allows the ions and knocked-out electrons to move relatively freely, then an electric field can collect the faster-moving electrons at an anode terminal giving you an electrical pulse to deal with. This is typically used in gaseous detectors. The second technique relies on the fact that some materials, after recapturing electrons to neutralize the ions, emit light in the visible region. In this case, a photomultiplier tube or perhaps a photodiode can be used to turn the light pulse into a detectable signal.

Table 17.1: Atomic and Nuclear Properties of Materials

Material	Z	A	Density ρ (gm/cm^3)	$(-1/\rho)dE/dx\|_{MIN}$ ($MeV/(gm/cm^2)$)
Be	4	9.01	1.85	1.59
C	6	12.01	2.27	1.75
Al	13	26.98	2.70	1.62
Si	14	28.09	2.33	1.66
Fe	26	55.85	7.87	1.45
Cu	29	63.55	8.96	1.40
Sn	50	118.69	7.31	1.26
Pb	82	207.19	11.35	1.12
Sodium Iodide (NaI)			3.67	1.31
Plastic Scintillator			1.03	1.95
Pyrex Glass			2.23	1.70

Let's start with the basics of how elementary particles ionize atoms. First we will discuss the interactions of charged particles with matter. Then we talk about photon interactions, and how the interactions of electrons are closely related to them. We briefly discuss neutron interactions as well.

Table 17.1, taken from the Review of Particle Properties, lists various important properties of materials relevant for the detection of subatomic particles.

17.1.1 Charged Particles

A charged particle is surrounded by an electric field. When a charged particle comes near an atom, the force of this electric field on the atomic electrons can knock an electron out of the atom creating an ion. The larger the charge on the particle, the greater the chance is of creating an ion. Also, the slower the particle moves, the more time it spends in the vicinity of the atom, and again the chance of creating an ion increases.

We express this interaction in terms of the energy lost by a charged par-

ticle as it passes through a certain thickness of matter. Dunlap goes through a classical calculation of this energy loss, but a correct calculation using a quantum mechanical description of the atom is more complicated. The result was first worked out by Bethe, Bloch, and others. The complete formula is written out and discussed in detail by Leo, but a reasonable approximation is

$$-\frac{1}{\rho}\frac{dE}{dx} \approx \frac{Q^2}{\beta^2} \times \frac{1.71 \text{ MeV}}{\text{gm/cm}^2} \tag{17.1}$$

for a particle with charge $\pm Qe$ and velocity βc where ρ is the density of the material and dE/dx is the amount of energy the particle loses per unit path length. We sometimes refer to $(-1/\rho)dE/dx$ as the "stopping power". This equation is valid if the incident energy E_0 is much larger than the energy loss per collision (i.e. β is not too small), and if the particle is not highly relativistic (i.e. β is not too close to unity). The factor of 1.71 actually depends on the material, but it is generally good to around $\pm 10\%$ for the lighter elements.

Figure 17.1, taken from the 1992 edition of the Review of Particle Properties, shows the energy loss given by the full Bethe-Bloch formula. The value of dE/dx is plotted as a *dotted* curve. Notice that at any given momentum, the particles μ, π, K, p, and d, which all have $Q = 1$, lose more energy for the larger mass (i.e. smaller β), but they all tend towards the same value at large momenta where β approaches unity. The α particle, however, has $Q = 2$ and therefore loses four times as much energy as the others.

Figure 17.1 also plots the *range*, or depth of penetration, of the charged particle in the material. It is obviously important to understand the range of an elementary particle in some material if we are to build a detector for that particle, especially one which measures its energy. It is tempting to determine the range by integrating the energy loss (which depends on β and therefore on the incident energy E_0) over all energies, that is

$$\text{Range (gm/cm}^2) \approx \int_{E_0}^{0} \frac{dE}{\frac{1}{\rho}\frac{dE}{dx}}$$

but you have to be careful. This approximation assumes the particle travels in a straight line through the material until it stops, but typically it will be scattered by atomic electrons and jitter around (or "straggle") along the way and the actual path length will be longer than the range.

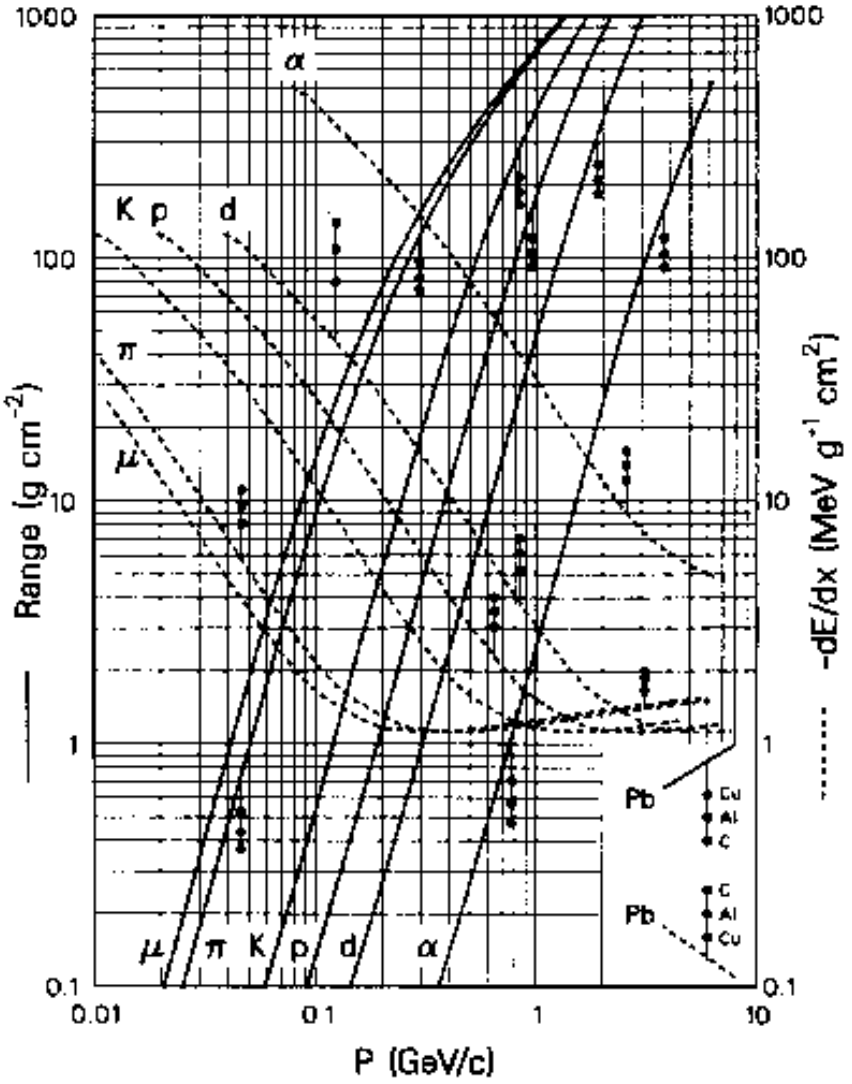


Figure 17.1: Mean range and energy loss in lead, copper, aluminum, and carbon. Taken from *The Review of Particle Properties*.

Straggling is actually not a very large effect for heavy charged particles, but it is very important for electrons, since they can lose a lot of energy in a single collision with an atomic electron. For this reason and others having to do with the low mass of the electron, the energy loss of the electron is actually closely tied to the energy loss processes of photons.

17.1.2 Photons and Electrons

Photons, or $\gamma-$rays, also ionize atoms. However, since they have $Q = 0$, they do so in ways very different from charged particles. There are three main ways that photons interact with matter, and each results in electrons which continue the ionization process as charged particles. The three ways photons interact, as far as nuclear and particle physics is concerned, are the photoelectric effect, the Compton effect, and pair production.

In the *photoelectric effect*, a photon is absorbed by an atom and an atomic electron is ejected. The energy of the electron is equal to the photon energy minus the electrons's binding energy in the atom. The binding energy may or may not be a large fraction of the photon energy, but the point is that the electron has less energy than the original photon. Therefore, some of the energy of the original photon is lost in this process.

In the *Compton effect*[1], the photon collides with an atomic electron and knocks it out of the atom. Instead of being absorbed, however, the photon itself is scattered in another direction. Having given a substantial fraction of its energy to the electron, the photon's energy is reduced. The electron and scattered photon then go their separate ways, each producing more ions.

For particularly high energy photons, *pair production* dominates over, first, the photoelectric effect and then the Compton effect. In this case, the photon disappears and an e^+e^- pair is created in its place. This can only happen in matter because an atom has to be nearby so that momentum is conserved. Obviously, it also can happen only if the photon energy is greater than $2m_ec^2$=1.022 MeV. Then, of course, the e^+ and e^- each produce a number of ions.

[1]See also Experiment 12.

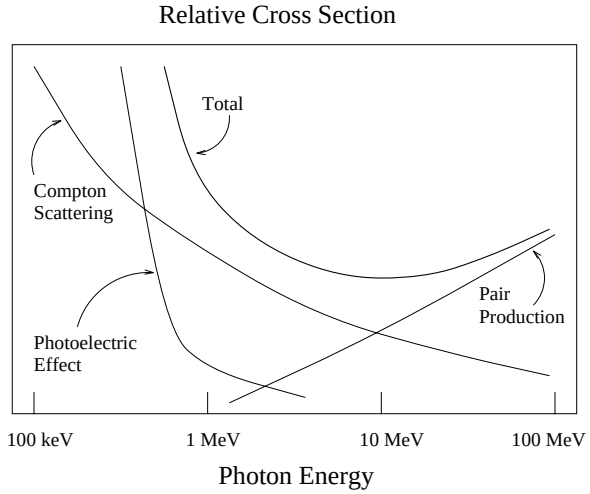


Figure 17.2: Relative cross sections for the photoelectric effect, Compton scattering, and pair production, as a function of energy in a large Z material.

The relative cross sections, i.e. probabilities of interaction, for these three processes is plotted in Fig. 17.2. The relative cross sections actually depend quite a lot on the material (in particular on the atomic number Z), but Fig. 17.2 is typical of large Z atoms.

This cross section is directly reflected in the ability for photons to penetrate matter. It is small enough for the energies here so that the mean free path is rather large. That is, there is a fair probability that a photon passes through a detector-sized piece of material with no interaction. Therefore, unlike for a charged particle, "range" is not well defined. Instead, we talk about the attenuation of photons, i.e.

$$I = I_0 e^{-\mu x} \tag{17.2}$$

where I is the photon intensity after passing through a thickness x of material and I_0 is the incident intensity. The attenuation coefficient μ, the inverse of the mean free path, depends very strongly on energy. That is, it behaves rather like the inverse of the cross section as plotted in Fig. 17.2.

Values for μ/ρ as a function of energy are plotted in Fig. 17.3 for a variety of materials. Notice that μ rises rapidly, up to a few MeV, and is very dependent on the particular material for low photon energies. However,

for energies near 1 MeV we find that $\mu/\rho \sim 1/15$ cm^2/gm, rather independent of the material. That is a 15 gm/cm^2 thick piece of material, i.e. 15 cm of water, 5.6 cm of aluminum, or 1.3 cm of lead, would attenuate a 1 MeV photon beam by a factor of $1/e$. Such calculations are important for both radiation shielding considerations as well as for particle detection.

Realize that the photon does not disappear after it interacts, particularly in the region of a few hundred keV to a few MeV. Compton scattering dominates for these energies, and the scattered photon continues on in some direction, albeit with lower energy than it came in with. Equation 17.2 gives the attenation of photons with the *incident* energy only, and you need to consider the "left over" photons in many applications.

It's clearly important to understand the energy loss interactions of electrons (and positrons) to finish the story. First, electrons are charged particles which lose energy basically as described in Sec. 17.1.1. There is one important difference, however. Unlike any other charged particles, electrons are not much heavier than the atomic electrons they collide with. That means that electrons can lose a lot, or nearly all, of their energy in a single collision, and change their direction drastically. That is, if a moving bowling ball hits a ping-pong ball, the bowling ball is not greatly disturbed and it continues in its path, suffering many collisions until it slows down and stops. On the other hand, if a queue ball hits a stationary billiard ball on a pool table, both balls having the same mass, the queue ball can stop and give up all its momentum to the ball it strikes. This means that fluctuations both in the energy loss and in the range are very large for electrons, much larger than for heavier particles.

Secondly, electrons themselves produce photons if they pass near an atom. This process is called *bremsstrahlung*, German for "braking radiation", because it happens whenever electrons are forced to slow down, as when they are attraced (or repelled, for positrons) by the nucleus of an atom. We won't describe this any further, only to say that the low mass of the electron means that it is the only particle for which bremsstrahlung matters. The emitted photon energy can be just about as large as the electron's incident energy.

Finally, positrons will eventually slow down and come close enough to an

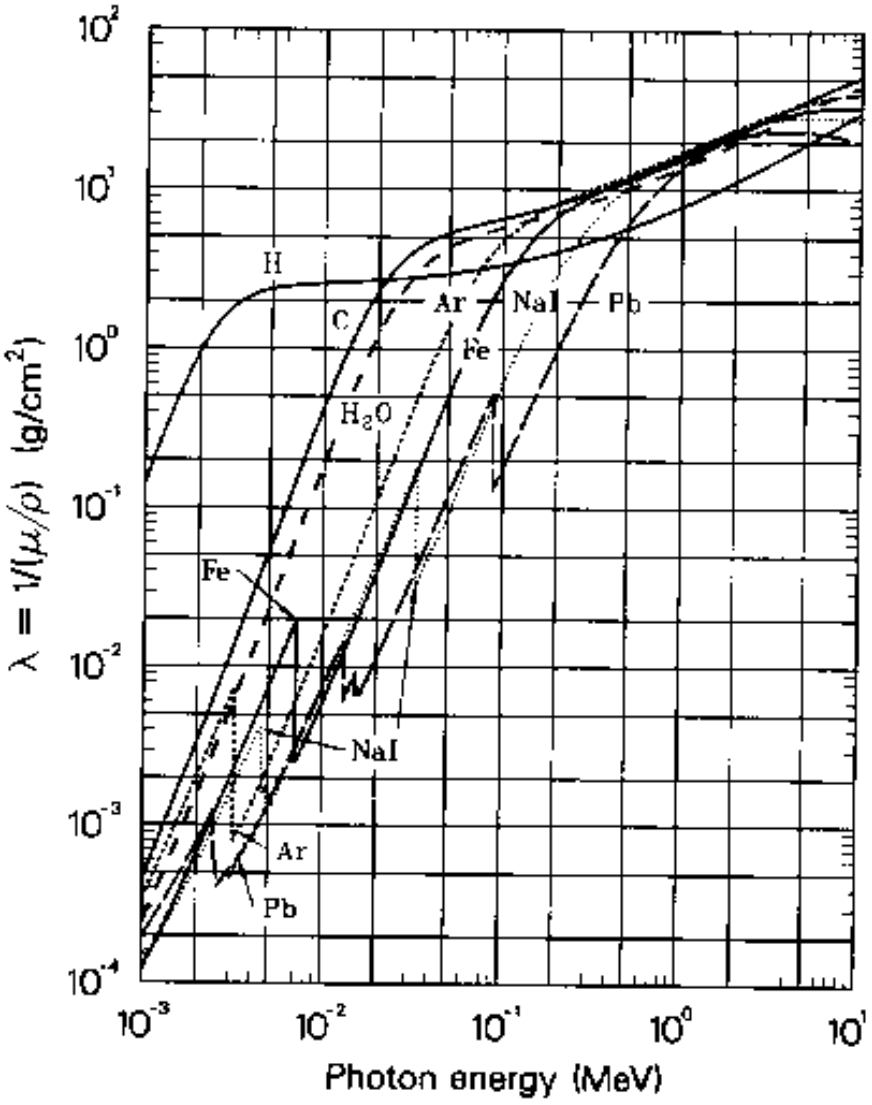


Figure 17.3: Photon mass attenuation coefficient μ/ρ plotted for various materials. Taken from *The Review of Particle Properties*.

atomic electron and annihilate[2], producing a pair of 511 keV photons.

Clearly, then, if a high energy (greater than several MeV) photon or electron impinges on a large block of material, the various processes produces photons from electrons and so forth, multiplying the number rapidly until the average energy falls below around an MeV. This "shower" is a very effective way to absorb and measure the energy of a high energy photon or electron. Such "shower counters" are very common in nuclear and particle physics for energies of $\sim$ 100 MeV or more, but we will not be using them in this course.

17.1.3 Neutrons

It turns out that neutrons are relatively easy to liberate in simple nuclear reactions, and they are commonly used in the laboratory. Frequently their use is limited to inducing some nuclear reaction and we study the reaction byproducts, but on occasion we want to detect the neutrons themselves. That can be quite difficult, because they have no charge and only interact significantly through interactions with the atomic nuclei in the detector. Those cross sections are small, so neutrons tend to go a long way before they interact.

For high energy neutrons, the reactions they induce give off fragments that, on the whole, are charged. These fragments are quite detectable, although a lot of energy escapes with the neutron, so it is hard to get a good measurement of the neutron total energy. If the material has a lot of free protons, i.e. it has a lot of hydrogen atoms as do hydrocarbon polymers, then the simple np elastic scattering reaction can be used to slow the neutrons down since the proton can take up a lot of the neutron's incident energy. As long as the neutron is slowed down, however, it will eventually wander close enough to some nucleus to be captured by it. When this happens, a γ-ray photon of several MeV is emitted, and this is typically detectable.

[2]See also Experiment 11.

17.1.4 Radiation Safety

Ionizing radiation is dangerous. Ionization does nasty things to living cells and that can affect people in various ways. If you get *a lot* of it all at once, it can kill you! You should take it seriously.

That said, there is nothing in our laboratory that can badly hurt you, at least if you take some simple precautions.

We'll first get some of the language straight, and then look at levels of radiation you are exposed to every day, and in this laboratory. Finally, we'll mention some safety rules. For more details, see Leo.

Units

People worried about the effects of ionizing radiation long before they understood it fully. That lead to some odd quantities and units which stay with us today. To make matters worse, there are now standard SI units for these things. Everyone is "supposed" to use the SI units, but almost nobody does.

Two quantities measure how much ionizing radiation you (or some other form of matter) receive:

- **Exposure:** This is the *ionization* produced per unit of matter. The standard unit is the *Roentgen* (R) which is the quantity of X-rays that make one electrostatic unit of ionized charge in 1 cm^3 of air. In more modern units, 1 R works out to be 2.58×10^{-4} Coul/kg in air at STP.

- **Absorbed Dose:** This is the *energy* deposited per unit mass. The old unit is the *rad* defined to be 100 erg/gram. The SI unit is the *Gray* (Gy) or 1 J/kg=100 rad.

To convert between dose and exposure, you need to know how much energy (in a particular material) produces so much ionization. For air, it takes (on the average for electrons) 33.7 eV of deposited energy to create an ion pair.

Therefore,

$$\begin{aligned} 1\text{ R} &\Rightarrow 2.54\times10^{-4}\frac{\text{Coul}}{\text{kg}}\times\frac{1}{1.6\times10^{-19}\frac{\text{Coul}}{\text{ion-pair}}}\times 33.7\frac{\text{ev}}{\text{ion}-\text{pair}} \\ &= 8.6\times10^{-3}\text{ Gy} = 0.86\text{ rad} \qquad \textit{in air} \end{aligned}$$

In *living tissue*, it turns out that 1 R is very close to 1 rad. For this reason, we tend to use Roentgen's and rad's interchangeably, so if you hear someone use "R", they probably don't care whether they're talking about exposure or dose.

Biological Effects

So is a 1 R exposure something to worry about? Well, it turns out that it can be, at least if you get it in a short time. However, you have to be rather careless to let that happen.

First, realize that neutrons, protons, and alpha particles are more dangerous than electrons or photons. This is because they drop a lot more energy in a very short distance. (See Fig. 17.1.) For that reason, we multiply the exposure by a factor between 5 and 20 for these particles. In the SI system, this (dimensionless) factor turns a Gray into a Sievert (Sv). (In the old system it turned a rad into a "rem", for "rad equivelent man".) The units have the same dimensions, but using Sieverts reminds us that this "quality factor" is taken into account.

So how much is too much? From natural sources (like cosmic rays and natural radioactivity), people in the US receive a dose of around 3 mSv/year. Based on statistics gained from incidents like Hiroshima and Nagasaki, and reactor accidents, we know that 2-3 Sv received in a short time can kill you. On the other hand, doses of 100 mSv/year over many years is probably safe. That is, such a dose is not likely to get you sick beyond your normal probability to get sick.

How does this compare to the radioactivity used in our laboratory? You can estimate exposure rate from standard sources using

$$\text{Exposure Rate} \approx \Gamma\times\frac{A}{d^2}$$

Table 17.2: Exposure rate constant for various radioactive sources.

Source	Γ (R·cm^2/hr·mCi)
^{137}Cs	3.3
^{22}Na	12.0
^{60}Co	13.2

where A is the source activity and d is the distance from the source. The "Exposure Rate Constant" Γ is different for different sources, because of the kinds of radiation they emit. Values of Γ are listed in Table 17.2.

Experiment 12 uses a (shielded) 10 mCi ^{137}Cs source which emits gamma rays in a narrow cone. In that region, at a distance of 1 m, the exposure would be

$$3.3 \text{ R/hr} \times \frac{10}{100^2} \approx 3.3 \text{ mrad/hr} = 3.3 \times 10^{-2} \text{ mSv/hr}$$

since we are talking about humans (i.e. 1 R is equivalent to 1 rad) and gamma rays (i.e. the quality factor is 1). So, if you stood 1 m in front of the source opening for four days, you'd get about the same dose as you would all year due to natural sources.

This is far less than what is considered a "safe" dose. The US regulation specifies 50 mSv (i.e. 5 rem) as the annual occupational dose limit. Below this limit, the risk of dying from cancer for a radiation worker is the same as that for a worker in a non-radiation environment.

Protecting yourself and others

The most important rule for radiation protection is

Don't be stupid.

For example, don't stand in front of the Compton Scattering source (the

hottest open source in the laboratory) for four days.

Formally, there are three things to keep in mind to minimize your exposure: time, distance, and shielding. Some simple things to do include:

- Don't leave unused sources open or lying around.
- Stay behind shielding blocks if they are present.
- Don't get unnecessarily close to any radioactive sources.
- Don't eat or drink in the laboratory.
- Wash your hands after working with sources.

If you'd like, you can obtain a radiation dosimetry badge from the Rensselaer Office of Radiation Safety so you can monitor your exposure.

17.2 Kinds of Particle Detectors

Now that we've learned that nuclear "radiation", or elementary particles, ionize atoms, how do we use that to detect particles and measure their properties? After working on this problem for 80 years or so, a lot of different techniques have been developed. Leo's book contains excellent discussions about almost all of these. We will concentrate on the two techniques we use in the typical undergraduate laboratory, namely gaseous ionization and scintillation detectors. First, however, we introduce a general physical concept called "solid angle".

17.2.1 Solid Angle

Solid angle is a three dimensional generalization of the planar angles you know so well. If you recall, a planar angle $\Delta\theta$ is just the length of a circular arc Δs, divided by the radius r of the circle, i.e. $\Delta\theta = \Delta s/r$. Solid angle $\Delta\Omega$

is just the area ΔA of a piece of a spherical surface, divided by the square of the radius, i.e. $\Delta\Omega = \Delta A/r^2$. Planar angles are measured in radians and solid angles are measured in sterradians. A circle subtends a planar angle of 2π and a sphere subtends a solid angle of 4π.

Solid angle is a useful concept whenever we are dealing with some sort of detector intercepting radiation which moves out in all directions from a source. Ionizing radiation and elementary particle detectors are just one example, but you would encounter the same thing in fields like optics or sonics.

Let's be a little more explicit with our definition. If $d\vec{A}$ is a vector whose magnitude is an area dA in some plane, and whose direction is normal to that plane, and $\hat{n}$ is a unit vector pointing towards the source which is a distance r away, then

$$d\Omega = \frac{\hat{n} \cdot d\vec{A}}{r^2} \equiv \frac{dA_\perp}{r^2} \tag{17.3}$$

where $dA_\perp$ is just the perpendicular component of the area. A spherical surface is most convenient since all surface elements are normal to the direction to the center. In spherical coordinates (r, θ, ϕ), where $0 \leq \theta \leq \pi$ is the polar angle and $0 \leq \phi \leq 2\pi$ is the azimuthal angle, a small rectangular piece of the surface has area

$$dA = \text{width} \times \text{height} = (r \sin\theta d\phi) \times (r d\theta) = r^2 \sin\theta d\theta d\phi$$

so the infinitesimal solid angle is just

$$d\Omega = \sin\theta d\theta d\phi \tag{17.4}$$

You will see Eq. 17.4 many times in physics.

Let's apply this to a specific case that is used a lot. See Fig. 17.4. This is a circular area with radius R located a distance d from a source. The face is normal to the direction to the source. There is perfect azimuthal symmetry, so we immediately integrate over ϕ to get

$$d\Omega = 2\pi \sin\theta d\theta$$

and integrate from $\theta = 0$ to $\theta_{MAX} = \tan^{-1}(R/d)$ to get

$$\frac{\Delta\Omega}{4\pi} = \frac{1}{2}\int_{\theta=0}^{\theta_{MAX}} \sin\theta d\theta$$

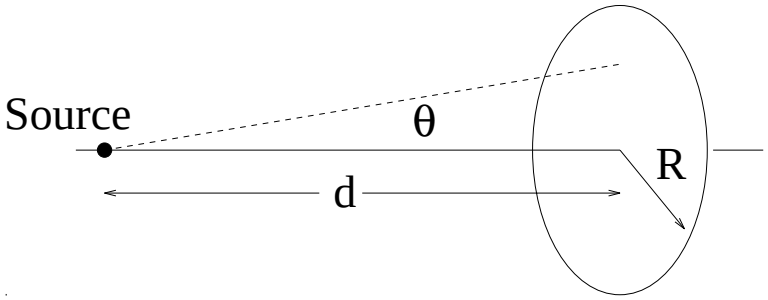


Figure 17.4: Calculating the solid angle of circular face.

where we have written the fraction of the total solid angle as $\Delta\Omega/4\pi$. This integral is done most easily by a change of variables to $\mu = \cos\theta$ with μ ranging from $\cos\theta_{MAX} = d/\sqrt{d^2 + R^2}$ to 1. Since $d\mu = -\sin\theta d\theta$,

$$\frac{\Delta\Omega}{4\pi} = \int_{\cos\theta_{MAX}}^{1} d\mu = \frac{1}{2}\left[1 - \frac{d}{(d^2 + R^2)^{1/2}}\right] \tag{17.5}$$

For $d = 0$, $\Delta\Omega/4\pi = 1/2$, that is, the surface covers one entire hemisphere. For $d \to \infty$, expand Eq. 17.5 to first order in R/d to find $\Delta\Omega/4\pi = R^2/4d^2$ or $\Delta\Omega = (\pi R^2)/d^2$ which is just what you expect from our basic definition of solid angle.

17.2.2 Gaseous Ionization Detectors

The simplest way to use ionization to detect particles is to put the detector material in an electric field and let the electrons and ions drift toward the anode and cathode respectively. Since the ionization is all over with in a very short time period (typically nanoseconds), you expect an electric "pulse" first at the anode from the lighter and faster moving electrons. Some time later, the ions would give a pulse at the cathode. These pulses would then be fed into electronic circuits, digitized, interfaced into computers, and so forth.

Among other things, this scenario assumes the electrons and ions move freely through the material. The easiest way to achieve this is to use a gaseous detector. These in fact were the first electronic elementary particle detectors. They are cheap to build and easy to use. They are almost never

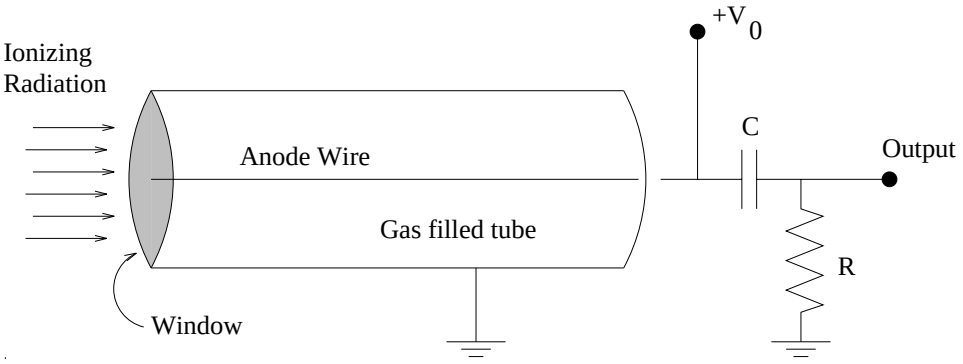


Figure 17.5: Schematic of a gaseous ionization detector.

used in their simplest form anymore, however, because of other advances in particle detectors. They survive today mainly as "radiation detectors" or "Geiger Counters", used for radiation protection monitors.

A Geiger counter is just one way of operating a simple gaseous ionization detector. Let's look at the basic design of such a thing and understand the different ways to operate it. For a more complete discussion see Leo or Dunlap. Leo in fact contains descriptions of the modern incarnations of gaseous detectors.

Figure 17.5 shows how you might construct a gaseous ionization particle detector. It is a gas-filled metal (i.e. conducting) cylindrical tube connected to ground with a thin wire along the axis. A thin window seals off one end and that is where the particles enter. The wire is insulated from the tube and held at some positive high voltage V_0. The wire is therefore the anode and will detect the pulse from the electrons. The cathode (tube) is grounded, and we will not bother with the pulse from the ions. A high pass filter is used to extract the signal pulse without exposing the downstream electronics to the possibly large DC V_0.

So what happens when an elementary particle ionizes some of the gas atoms inside the tube? Well, if $V_0 = 0$, the electrons and ions just hang out near each other and soon recombine. However, if V_0 is larger than 20 V or so, the electrons and ions separate before they can recombine, and you get a voltage pulse on the anode wire. You've detected an elementary particle!

Let's get an idea of how large the signal is. According to Eq. 17.1, an electron with an MeV of energy deposits around 2 MeV/(gm/cm^2) in material. A typical gas density is around 10^{-3} gm/cm^3 at room temperature and pressure. Therefore, if the tube in Fig. 17.5 is 5 cm long and the electron passes all along it, then it loses $\sim$ 0.02 MeV of energy along its path, or around 1% of its total energy. It costs something like 10 eV of energy to ionize an atom, so our very rough calculation predicts something like 2000 electron-ion pairs created by the electron in the detector. That may sound like a lot, but it is actually a very small signal because the electron charge $e = 1.6 \times 10^{-19}$ C is very small. Even if all those electrons are collected over 10 ns, it only amounts to an average current of 32 nA or a 1.6 μV voltage drop across 50Ω.

This is not enough voltage to work with if you are trying to detect single particles. On the other hand, this type of device is suitable for measuring a large flux of particles from a beam or intense source, for example. Operated in this way, the gaseous ionization detector is called an *ionization chamber*. It is still used today, albeit rather infrequently, at accelerator laboratories and other installations.

We can still use the device in Fig. 17.5 to detect individual particles. The trick is to amplify the signal in the detector by simply increasing the voltage V_0. When V_0 exceeds 100 V or so, the electrons gain enough energy on their journey to the anode wire that they ionize the atoms they collide with. This increases the size of the signal to the point where it can be amplified with external electronics without worrying about fundamental limitations from Johnson noise and so forth. What's more, up to some maximum value of V_0 the size of the signal that comes out is still proportional to the ionization caused by the incident particle. That is, the signal size still measures the amount of energy deposited. Operated in this mode, the detector is called a *gas proportional counter*.

If V_0 is increased still further, more than several hundred volts, things become quite different. The avalanche of electrons caused by successive collisions grows very large. There is so much charge near the anode wire that the electric field is severely distorted, and the chain reaction spreads charge out along the entire length of the wire. The response of the detector saturates, and the result is one large output pulse that is the same size, independent

of the number of ions that the elementary particle first created. This is a *Geiger counter*. It is a useful way to get a large pulse for detecting pretty much any kind of radiation, although you can't get any other information about the particle other than that it entered the detector. This is fine for monitoring and counting radioactivity.

Now suppose we want to measure the total kinetic energy carried by an incident particle. We can do that by stopping it in the detector, and making sure the output is proportional to the number of ions created. The latter can easily be achieved in proportional mode in a gas ionization detector, but stopping the particle is another matter. As we showed above, a 1 MeV electron only loses 1% of its energy in a 5 cm long counter. That means we would have to build a detector 5 m long to stop it, and 5 m in diameter to contain electrons going in different directions! This is no good, and a different idea is needed. It took many years before scientists figured out something other than gas detectors to make this possible.

17.2.3 Scintillation Detectors

Despite their importance, gaseous ionization detectors clearly have their limitations. The signals are rather small (if you want the output to be proportional to the deposited energy), it takes a relatively long time to collect the charge, and gases have very low density. If you want to stop a particle to measure its total energy, for example, you need a more dense medium, and that means going to a solid. The problem with solids, though, is that electron-ion pairs don't move freely, so you can't detect the ionization directly by collecting the electrons.[3] We would like to detect the primary ionization some way other than by collecting the charge.

Scintillation detectors are the most popular way to solve this problem. The idea is to convert the primary ionization into a pulse of light, and then use some optical technique to detect the light. There are, therefore, two essential components to a scintillation detector. One is some detector material, called

[3]One way around this is to use a semiconductor diode as the detector medium. In fact, silicon and germanium diodes have been made into high quality particle detectors. They are expensive, however, and tricky to operate, and we don't use them in our laboratory.

a "scintillator", which produces detectable light when ionized. The second is the light detector, almost always a photomultiplier tube (see Sec. 11.2.2) but sometimes a photodiode (Sec. 11.2.3).

When an ionized atom picks up an electron and deexcites, it emits light. You might think, therefore, that it is not hard to find materials that scintillate, but it is actually not so easy. For one thing, the material needs to be transparent to the light produced by ionization, otherwise the light won't make it to the photomutliplier tube. What's more, the light needs to be at wavelengths for which the photomultiplier is sensitive, and this is usually in a rather narrow region in the blue. (See Fig. 11.4.) In addition, since you want the detector to measure total energy, you need the number of optical photons emitted by the scintillator to be at least roughly proportional to the primary ionization. A variety of suitable scintillators have been identified over the years, but we will concentrate on two of them, namely NaI(Tl) and plastic scintillator.

NaI(Tl) Scintillation Detectors

For a good reference specifically on using NaI(Tl) detector for γ-ray spectroscopy, I recommend

- *Applied Gamma-Ray Spectrometry*, C.E. Crouthamel
  Pergamon Press (1960)

It is old, but the techniques remain sound and there are lots of useful tables. I leave it on reserve in the library.

Figure 17.6 diagrams your basic NaI(Tl) scintillation detector, several of which are found in our laboratory. Light is produced in a single crystal of NaI(Tl), that is sodium iodide doped with around 0.1% thallium, generally shaped into a cylinder measuring perhaps two to three inches in diameter and two to three inches thick. The crystal is housed in an aluminum can, lined on the inside with white reflective MgO, except on the face which couples to the photomultiplier tube where a glass window is used. The glass window is mounted to the can forming a hermetic seal which protects the crystal from

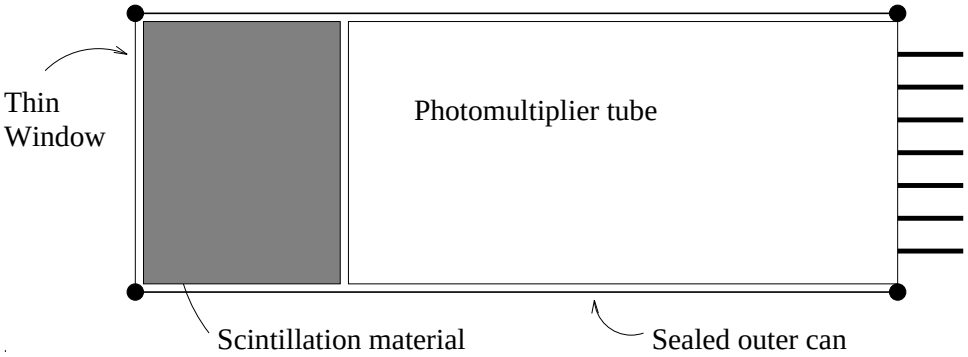


Figure 17.6: A typical NaI(Tl) scintillation detector. These detectors are usually referred to by the crystal size. A detector with a 3″ diameter and 3″ long crystal would be called a 3″×3″, for example.

the atmosphere. If any moisture or humidity were to come in contact with the crystal, it would absorb water and quickly become useless.

Sodium iodide makes a fine scintillation detector, particularly for photons because of it's high density and high average Z (from the iodine), both of which make it possible to absorb all the energy of a photon (or a charged particle) in a relatively small crystal. It is also very efficient, producing one optical photon for every 25 eV of deposited energy, and the wavelength spectrum is nicely matched to a photomultiplier tube, peaking at 413 nm and nearly all contained within $350 \leq \lambda \leq 500$ nm. The scintillation signal rises quickly, but decays with a relatively long time constant, around 230 ns.

The "pulse height" signal from the photomultiplier (actually the integrated current out of the anode) is proportional to the amount of light detected, which is proportional to the amount of energy deposited. Therefore, if the elementary particle stops in the crystal, then you might expect the pulse height to be one fixed value. This is not the case because the energy resolution is not perfect, and has a random, statistical spread to it. This spread comes about partly because of the Poisson-statistical uncertainty in the number of detected photons, but it is mainly due to differences in how the light is collected and inhomogenieties in the response of the crystal. It is best to illustrate this with an example.

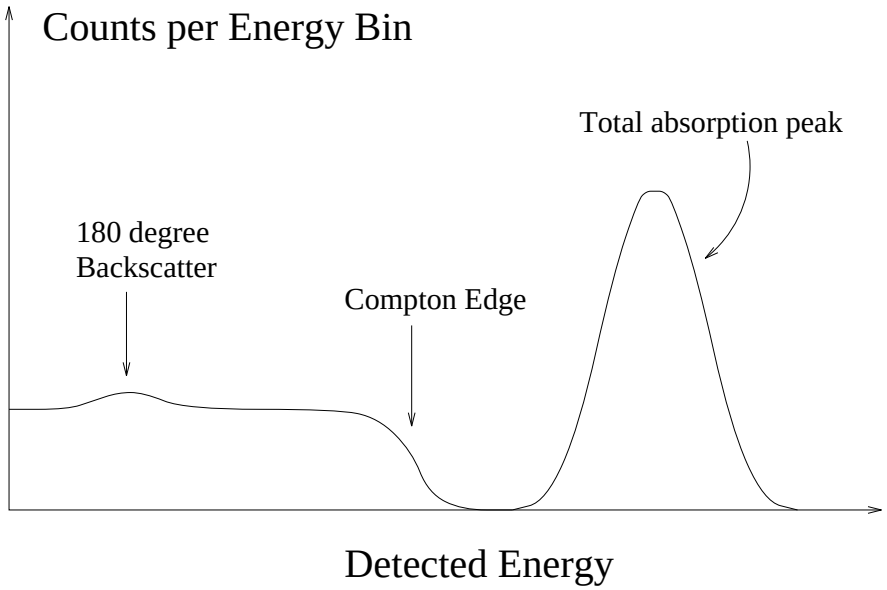


Figure 17.7: A typical γ-ray spectrum from a NaI(Tl) detector. The γ-ray energy is $\sim$ 1 MeV.

Figure 17.7 shows the distribution of pulse heights from a typical NaI(Tl) detector when exposed to 1 MeV γ-rays. The large peak corresponds to full absorption of the γ-ray (i.e. a deposited energy of 1 MeV), and because of the various resolution effects is $\sim$ 100 keV wide at half the peak height. Notice also the valley for pulse heights just below the peak, and the relatively flat distribution below the valley. This structure is typical of γ-ray spectra in scintillation detectors, and it corresponds to less than full absorption of all the γ-ray energy. It is easy to see where the valley comes from. At these energies, the γ-rays interact first mainly through Compton scattering. (See Fig. 17.2.) If the scattered γ-ray photon escapes the detector (remember that the mean free path $1/\mu$ is around 15 gm/cm^2=1.6 in. for sodium iodide for 1 MeV photons), then all the ionization is due to the initial recoil electron. This electron can have any energy up to that corresponding to 180° scattering of the photon, and this value is less than the full photon energy.

Detection Efficiency for NaI(Tl) Detectors. There will be lots of times you want to know the actual decay rate or reaction rate, where the NaI(Tl)

detector measures some γ-ray to let you know something happened. In that case, you need to know the *efficiency* ϵ with which the detector actually observes the photon. This is a complicated business, since ϵ depends on many things like the photon energy, the distance to the source, and the shape of the detector. In practice, it is best determined with calibrated radioactive sources, but *for a detector of a particular geometry* you can essentially reduce the efficiency to the product of two factors:

$$\epsilon(E_\gamma, d) = \epsilon_{\rm intr}(E_\gamma, d) \times P(E_\gamma, d)$$

where E_γ is the photon energy and d is the distance to the source. The first factor $\epsilon_{\rm intr}(E_\gamma, d)$ is called the *intrinsic* efficiency and measures the probability that the photon deposits a measurable amount of energy in the detector. The second factor $P(E_\gamma, d)$ is called the *photopeak* efficiency, and measures the probability that if the photon *does* deposit a measurable amount of energy, then it in fact deposits *all* of its energy. This separation is of course artificial since both factors depend on E_γ and d, but it is convenient since we will make different approximations to reduce $\epsilon_{\rm intr}$ and P to simpler forms.

The intrinsic efficiency is dominated by the fact that the detector only subtends a small portion of the solid angle into which the photon can radiate. It also depends to some extent on the γ-ray energy, since a high enough energy photon can pass through a detector without depositing any energy. We separate these two effects using the approximation

$$\epsilon_{\rm intr}(E_\gamma, d) \approx \frac{\Delta\Omega}{4\pi} \times \epsilon_{\rm dep}(E_\gamma) \tag{17.6}$$

The fractional solid angle can be taken from Eq. 17.5. In other words, we write the intrinsic efficiency as the product of the probability that the photon intercepts the detector times the probability that it deposits any energy at all in the detector if it intercepts it.

Note that $\epsilon_{\rm dep}$ still has some distance dependence in it because because the back of the detector covers a smaller solid angle than the front, if the detector is cylindrical. That is, should you use the front or the back to calculate $\Delta\Omega/4\pi$? (The right answer is "somewhere in between".) The intrinsic efficiency is best calculated using Monte Carlo techniques, and you can find tables of these efficiencies in various books. You can then use Eq. 17.6 to interpolate between different cases for the case closest to yours. A table taken from Crouthamel for a $3'' \times 3''$ detector is reproduced in Table 17.3.

Table 17.3: Intrinsic γ-Ray Efficiencies for a $3'' \times 3''$ NaI(Tl) Detector

Energy	Distance (cm)						
(MeV)	0	0.5	1.0	2.0	4.0	10.0	20.0
0.10	0.500	0.432	0.368	0.260	0.133	0.0318	8.70×10^{-3}
0.15	0.500	0.424	0.355	0.246	0.124	0.0303	8.46×10^{-3}
0.20	0.497	0.407	0.334	0.227	0.114	0.0286	8.17×10^{-3}
0.30	0.465	0.360	0.289	0.192	0.971	0.0255	7.58×10^{-3}
0.40	0.428	0.323	0.257	0.170	0.087	0.0234	7.09×10^{-3}
0.50	0.399	0.297	0.235	0.156	0.079	0.0218	6.68×10^{-3}
0.60	0.378	0.279	0.220	0.146	0.075	0.0207	6.39×10^{-3}
0.80	0.347	0.254	0.200	0.132	0.068	0.0191	5.95×10^{-3}
1.00	0.325	0.236	0.186	0.123	0.063	0.0179	5.61×10^{-3}

The photopeak efficiency P is a very strong function of energy, but a weak function of the distance between the source and the detector. For example, in the same $3''\times3''$ detector, the photopeak efficiency drops from 0.9 for ~ 0.2 MeV photons, down to ~ 0.4 for 1 MeV. In fact, the photopeak efficiency can be measured for a particular setup using a radioactive source which emits a line of one specific energy. A plot of the photopeak efficiency, or "total absorption factor" for a $3''\times3''$detector, is shown in Fig. 17.8. Note that P is very sensitive to the size and shape of the detector. For a smaller detector, for example, a greater fraction of the incident photon energy might leak out the sides or back, so the photopeak efficiency will be smaller.

Plastic Scintillation Detectors.

A different class of scintillator materials are created by dissolving various organic compounds in clear plastic. Relative to NaI(Tl), plastics are cheap, hard to break, and give signals which decay away in much less time (~ 3 ns). They do not absorb moisture, so they need not be handled with as much care. What's more, it is also possible to form them into a variety of shapes. On the other hand, plastic scintillators are much less dense than NaI(Tl) (1 gm/cm^3 as compared to 3.7 gm/cm^3) and being made mainly of carbon and hydrogen, they have a low average Z. This makes them less suitable for photon

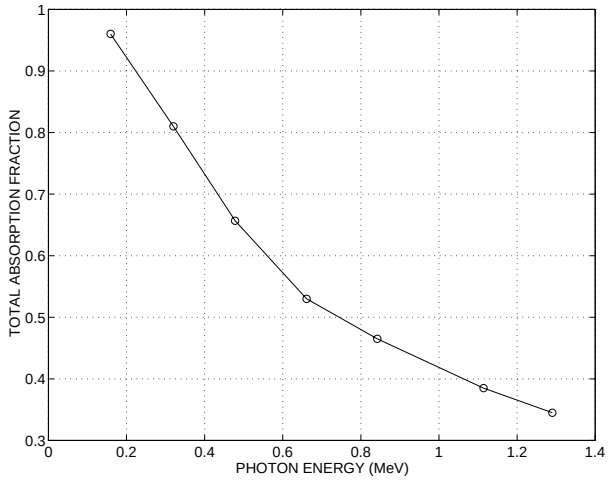


Figure 17.8: Photopeak efficiency for detecting γ-rays with a $3'' \times 3''$ NaI(Tl) scintillation detector.

detection, and they are mainly used to detect charged particles. Plastics are also less efficient than NaI(Tl), requiring around 100 eV of deposited energy to created a detectable optical photon. Just as for NaI(Tl), the emitted wavelength spectrum peaks at around 400 nm and spreads between 350 nm and 500 nm.

Depending on the geometry, a plastic scintillator can be mounted to a photomultiplier in a variety of ways. Probably the simplest is to glue the scintillator to the photomultiplier window, but more elaborate schemes can be used to "pipe" the light along some tortuous path using a lucite guide. This may be necessary if the scintillator sits in some enviornment that the photomultiplier cannot tolerate, such as cryogenic or in a large magnetic field. Depending on how efficiently the light can be collected, the energy resolution may be limited by the Poisson statistics on the average number of detected photons.

Since plastic scintillators are much less dense than NaI(Tl), as well as having lower Z, there are considerably less efficient as far as photopeak efficiency is concerned. In fact, plastic scintillation detectors are almost never used as total energy detectors for photons. On the other hand, they can be quite useful as electron detectors, if they are large enough for the energies

involved.

17.3 Pulse Processing Electronics

We conclude this chapter with a discussion of the types of electronics used in the nuclear physics laboratory. In particular, we are talking about the devices which accept the types of electrical pulses produced by particle detectors. Typical time scales for these pulses are 100 ns or less for scintillation counters, and considerably longer for Geiger counters.

Nearly all such electronics in our laboratory subscribe to the *Nuclear Instrument Module* or NIM standard. This dictates the physical size, voltage requirements, and input and output characteristics for all modules. The modules themselves are about 10" long and 8" high. Any NIM module will fit into a "NIM bin" which supplies the necessary voltages for that module. A NIM bin can accept up to 12 single width NIM modules. A single width NIM module is about $1\frac{1}{2}$" wide, but double width NIM modules are not uncommon. The power is supplied at the rear of the module through rectangular connector with some number of pins which plugs into the bin. Power can be drawn from taps which supply ± 6 V, ± 12 V, and ± 24 V, as well as the 110 V AC line.

Coaxial cable, usually with 50Ω characteristic impedance, is used to carry signals from the detector to the module, or from one module to another. BNC connectors are usually used on the cables and on the modules, but sometimes thin cable is used (but still 50Ω) with a smaller connector standard called LEMO.

17.3.1 Amplifiers

A pulse signal from a photomultiplier tube is a burst of electrons, corresponding to some ionization event in the detector, which varies over a period of time somewhere between 10 ns and 200 ns. The integral of this current signal over time gives the total number of electrons in that pulse, and this

is the number we care about, especially if we want a measurement of the deposited energy. An *amplifier* not only takes this signal and increases its magnitude to make it easier to use, it also gives the pulse a shape that is more convenient. For example, the *height* of the amplifier output pulse, as opposed to the integrated area of the input, corresponds to the magnitude of the signal.

The gain of the amplifier and pulse shaping parameters can be varied by turning various knobs on the front of the module. It is also possible, in most cases, to adjust the output of the amplifier so that the ambient level is close to ground potential.

In our laboratory, you will likely find, for example, the Canberra model 2012 or the Ortec model 570 pulse amplifiers. Lots more models are available from these and other vendors.

17.3.2 Discriminators and Single Channel Analyzers

You frequently need to do one simple thing with the analog output of a detector or an amplifier. This is to determine whether or not the analog pulse was greater or smaller than some particular value. The module which does this is called a *discriminator*.

The input to the discriminator is the output of the detector or amplifier. There is an adjustment of some sort with which you can vary the "threshold level" of the discriminator, that is, the voltage above which (or below which, if the analog signal is negative) the discriminator "fires" and gives you an output logic pulse. These logic pulses are then used further down the line in your electronics setup in some other NIM module.

Logic pulses are defined by one of three common standards in use in the modern laboratory. One of these, the ECL standard, is not used in our laboratory, so I won't discuss it any further. On the other hand, we use each of the other two, the NIM (actually, the "fast NIM") standard and the TTL standard. Both of these use zero (i.e. ground) voltage to correspond to "off", or a logical "false". In the NIM standard, an "on" or logical "true" is based on current requirements and specifies -16 mA (or within some range of this)

into 50Ω, or around −800 mV. A logical "true" in the TTL (for Transistor-Transistor Logic) standard is based on voltage and must be between +2 V and +5 V. Discriminators which accept the fast pulses from scintillators, for example, usually have only NIM outputs, while discriminators for shaped signals generally provide TTL or both TTL and NIM. Depending on which standard you are using, of course, you must be consistent down the line. Some NIM modules are available which convert NIM to TTL or the other way around.

A variation of the discriminator called the *single channel analyzer* is also quite useful. The single channel analyzer, or SCA, gives a logical true output if the input analog pulse is not only larger than some value E, but also smaller than some value $E + \Delta E$. The threshold level E and the "window" ΔE are both set by knobs on the front panel.

17.3.3 Processing Logic Signals

Logic signals are used in a semi-infinite number of ways to accomplish many different things. I will give you two examples of their use, but you will likely encounter several others.

Frequently, you will want to perform pulse height analysis or particle counting or something like that, only while some particular logic condition is satisfied based on your detector setup. For example, you might want to analyze the pulse height in a NaI(Tl) γ-ray detector only when a second scintillator indicated that a β-decay occurred. You would use a logic pulse, probably generated by the β-decay detector signal through a discriminator, to *gate* the multichannel analyzer. That is, you would use the logic signal as an input to the multichannel analyzer gate input which says "Don't perform pulse height analysis on your analog input signal unless I'm giving you a logical 'true' signal", assuming your multichannel analyzer is so equipped. We use the phrase "gate" in many similar ways, all of which are supposed to tell the module to do something only if the gate signal is "true". Some NIM modules, called *gate generators*, are designed to provide particularly flexible gate signals.

Another common use of logic signals is to know if two events happen at the same time. In this case, a NIM module called a *coincidence module* accepts two or more logic signals as input, and provides a "true" output only if the inputs are all "true" at the same time. Otherwise, the output is a logical "false". Practically, of course, it is important to define what "the same time" means. Different coincidence modules have different definitions. One of the most common definitions is that the leading edge of the two (or more) logic pulses arrive within some time window, called the "resolving time". On some modules, the resolving time can be adjusted using a front panel knob.

17.4 Exercises

1. A ^{22}Na radioactive source emits 0.511 MeV and 1.27 MeV γ-rays. You have a detector placed some distance away. You observe a rate of 0.511 MeV photons to be 2.5×10^3/sec, and of 1.27 MeV photons to be 10^3/sec, with just air between the source and the detector. Use Fig. 17.3 to calculate the rate you expect for each γ-ray if a $\frac{1}{2}$ in. thick piece of iron is placed between the source and the detector. Repeat the calculation for a 2-inch thick lead brick.

2. A radioactive source is situated near a particle detector. The detector counts at a rate of 10^4/second, completely dominated by the source. A 2 cm thick slab of aluminum (density 2.7 gm/cm^3) is then placed between the source and the detector. The radiation from the source must pass through the slab to be detected.

a. Assuming the source emits only 1 MeV photons, estimate the count rate after the slab is inserted.

b. Assuming the source emits only 1 MeV electrons, estimate the count rate after the slab is inserted.

3. Consider a small rectangular surface far away from a source. The surface is normal to the direction to the source, and subtends an angle α horizontally and β vertically. Show that the solid angle subtended is given by $\alpha\beta$.

4. A photomultiplier tube with a $2''$ active diameter photocathode is located 1 m away from a blue light source. The face of the PMT is normal to the direction of light. The light source isotropically emits 10^5 photons/sec. Assuming a quantum efficiency of 20%, what is the count rate observed by the photomultiplier?

5. Refer to Table 17.3.

a. Compare the tabulated intrinsic efficiency for detecting 100 keV γ-rays to those calculated using Eq. 17.6. Normalize the calculated values to the tabulated value at $d = 0$. A plot might be best, including a plot of the ratio of the two values.

b. Use the curve to estimate the intrinsic efficiency for detecting 100 keV γ-rays at a distances of 15 cm and 1 m.

c. Repeat (a) and (b) for 1 MeV gamma rays.

d. Do you expect the plots to look very different for the 100 keV and 1 MeV cases?

6. Two scintillation particle detectors are constructed as shown in Fig. 17.6. In one case, the scintillator is NaI(Tl). However, in the other case it is an ordinary form of plastic scintillator material. The output pulse height is digitized into a spectrum by the multichannel analyzer. Assume that the resolution is dominated by the random statistics of the number of optical photons. Plastic scintillators produce about 10 photons per KeV of deposited energy, while NaI(Tl) gives around 40 photons per KeV.

a. A monochromatic, well-collimated source of 1 MeV electrons impinges on the detector. On the same set of axes, sketch the spectrum determined by the MCA for each of the two scintillators and label the two curves. Label the horizontal axis in units of *detected* energy in MeV.

b. The electron source is replaced by a monochromatic, well collimated gamma source. As in (a), sketch and label the response of the two detectors.

7. Two scintillation detectors separated by 3 m can measure the "Time-of-Flight" for a particle crossing both of them to a precision of $\pm 0.20\ ns$. Each

detector can also measure the differential energy loss $dE/dx = constant/\beta^2$, $\beta = v/c$, to $\pm 10\%$. For a particle with velocity of 80% the speed of light (i.e. $\beta = 0.8$), how many individual detectors are needed along the particle path to determine the velocity v using dE/dx to the same precision as is possible with Time-of-Flight?

8. A Čerenkov detector is sensitive to particles which move faster than the speed of light in some medium, i.e. particles with $\beta > 1/n$ where n is the index of refraction of the medium. When a particle crosses such a detector, it produces an average number of detected photons given by

$$\mu = K\left(1 - \frac{1}{\beta^2 n^2}\right)$$

The actual number of detected photons for any particular event obeys a Poisson distribution, so the probability of detecting no photons when the mean is μ is given by $e^{-\mu}$. When 1 GeV electrons ($\beta = 1$) pass through the detector, no photons are observed for 31 out of 19761 events. When 523 MeV/c pions ($\beta = 0.9662$) pass through, no photons are observed for 646 out of 4944 events. What is the best value of the index of refraction n as determined from this data? What is peculiar about this value? (You might want to look up the indices of refraction of various solids, liquids, and gases.)

Ch 18

Experiment 10: Radioactivity

Around the year 1900 or so, scientists started to get a glimpse of what nature looked like at distances smaller than the atom. Rutherford discovered something deep inside the atom that had to be very small. The Curies discovered high energy radiation emerging from atoms, and in retrospect, it is clear the source had to have a small size to go with the high energy. These discoveries have led us to the modern fields of Nuclear Physics and Particle Physics.

In this experiment, you will study some of the same phenomena as the Curies, that is, radioactivity. This is the classic name given to radiations that come from the decay of the nucleus. Of course, we've learned a lot since then about the nucleus and about how to do experiments. Modern techniques will allow you to make a number of measurements of nuclear decay and some of the properties of nuclear radiation. Furthermore, you will watch several species of short-lived nuclei decay away. Some of these you will actually create yourself using a nuclear reaction technique called "neutron activation".

The material we've covered in Chapters 9 and 17 will be particularly important for this experiment. I suggest you review that material, as well as what is covered in this chapter. You may also want to consult other sources. For a good basic introduction to nuclear physics and radiation detection, see

- *Physics*, Robert Resnick, David Halliday, and Kenneth Krane, John Wiley and Sons, Fourth Edition (1992), Chapter 54

A more thorough treatment of the different types of radioactivity is given by

- *Introduction to the Structure of Matter*,
  John J. Brehm and William J. Mullin, John Wiley and Sons (1989),
  Chapters Fourteen and Fifteen.

- *Experimental Physics: Modern Methods*, by R. A. Dunlap,
  Oxford University Press (1988); Chapter 11

The rest you can get from these notes and from online documentaion for the data acquisition program.

18.1 Nuclear Decay

Rutherford discovered that the most of the size of the atom was filled with empty space, which we now understand to be the electrons. Most of the mass of the atom, though, was concentrated at a very small distance, around 10^5 times smaller than the atom itself. This is the atomic nucleus, and the study of it is called nuclear physics. All Rutherford could tell was that it was very small. We of course know a lot more about the nucleus today.

For the purposes of this course, you can consider the nucleus as a collection of protons and neutrons, bound together quite tightly. Protons and neutrons are almost identical types of elementary particles, having nearly the same mass and interacting with matter pretty much the same way. The biggest difference is that the proton has a charge $+e$ while the neutron is neutral. In fact, we tend to refer to protons and neutrons as different "states" of the same particle, called the nucleon.

A nucleus is characterized by the number of protons Z and the number of neutrons N that it contains. For a neutral atom, of course, Z also counts the number of electrons, and so it specifies the chemical element. In other words, nuclei with the same Z but different values of N, called *isotopes*, give rise to atoms with essentially identical chemical properties. On the other hand, the nuclear properties are more closely identified with the atomic mass number

$A = Z + N$. We designate the nucleus with the notation "$^A(Z)$" where (Z) is the one- or two-letter chemical symbol that designates the atom with the appropriate value of Z. For example, the nucleus of the carbon atom ($Z = 6$) with $N = 6$ ($A = 12$) is ^{12}C. Despite the way it is written, we still say "Carbon-12".

The protons and neutrons in the nucleus move around constantly interacting with each other in much the same way as electrons do in an atom. The biggest difference is the distance scale, and therefore the energies involved. For an atom, the sizes are on the order of 1Å$= 10^{-10}$m and energies are on the order of several eV, while for the nucleus size is on the order of several[1] fm$= 10^{-15}$m and the energies are in the MeV region. In other words, just as atoms have energy levels separated by some eV of energy, nuclei also have energy levels, but they are separated by some MeV. Excited states decay to lower states by emitting photons, but for nuclei these photons are obviously much higher in energy. We generally refer to these MeV photons as γ-rays.

Nuclear energy levels are specified in terms of their total angular momentum, called *nuclear spin*, and their *parity*, which is either $+1$ or -1. The parity of a nuclear level just tells whether or not the wave function for that state changes sign ($P = -1$) or not ($P = +1$) when you make the the substitution $\vec{r} \rightarrow -\vec{r}$ in the argument of the wavefunction. That is, $\psi(-\vec{r}) = \pm\psi(\vec{r})$. The most important thing for you to know about parity for now, is that parity, along with the nuclear spin, will determine which lower energy level a particular state would rather decay to, and how long that state will stick around before it is likely to decay. More on this in a little while, but for now, let's talk a little more quantitatively about this notion of *lifetime*.

If you look one particular nucleus, or an atom for that matter, in some excited state, then you expect it to decay at some point. Quantum mechanics cannot tell you *when* it will decay, only what the probability is for how long it is likely to live. This is the essence of why radioactive decay is randomly statistical in nature. You will measure some decay rate for a collection of (identical) nuclei, based on the probability of decay.

To measure this probability, we resort to a determination of the nuclear

[1]The unit "fm" technically stands for "femtometer", but just about everyone calles it the "fermi" in honor of Enrico Fermi.

lifetime. For some large sample of nuclei, the decay rate[2] R, the number of decays per second, will be proportional to the number N of nuclei in the sample at any particular time. That is

$$R = \frac{dN}{dt} = -\lambda N$$

where the proportionality constant is called $-\lambda$, the minus sign reflecting the fact that the decay causes the number of nuclei to decrease with time. This differential equation has a simple solution, namely

$$N(t) = N_0 e^{-\lambda t}$$

where N_0 is just the number of nuclei present at $t = 0$. Obviously, λ characterizes the lifetime. The larger λ is, the faster the sample decays, and the shorter the lifetime. There are two definitions we use for the lifetime. One is the *mean life*

$$\tau = \frac{1}{\lambda} \qquad \text{Mean Life}$$

The other is more practically minded, and measures the time it takes for the sample to decay to $\frac{1}{2}$ its original number. This is called the *half life*, and it is determined by solving $N(t) = N_0/2$ for t. You find

$$t_{1/2} = \frac{\ln 2}{\lambda} = 0.693\tau \qquad \text{Half Life}$$

References *usually* quote the half life, but not always. Always be sure when you look up a lifetime, whether you are getting the half life or mean life.

The lifetime of a nuclear state is directly related to the quantum mechanical matrix element that connects that state to the final state. This matrix element depends not only on the wavefunctions for the two states, but also on the type of interaction that causes the decay. If the interaction is "strong", then the decay is highly probable and the lifetime is short. Weaker decays generally have longer lifetimes, but the answers vary a lot, largely because of the dependence on the wavefunctions.

So far the discussion pertains as much to the decay of excited atomic states as to nuclear states. We just have a special name for the photons

[2]For historical reasons, the standard unit for decay rate is the Curie$\equiv 3.7 \times 10^{10}$ decays per second. This is the number of decays per second in one gram of radium.

emitted in the nuclear decay, i.e. γ-rays, and we call the process γ-decay. However, nuclei can do some things that atoms can't. The ground state, that is, the lowest energy level, can also decay in some cases, creating a new nucleus in place of the old one. There are two such types of decays, namely α-decay and β-decay and they are as different as night and day.

Alpha decay is the process by which a nucleus emits an α-particle, that is a ^{4}He nucleus, reducing its Z by two and A by four. It is very common for very heavy nuclei to decay this way. For example, the uranium isotope ^{238}U decays to ^{234}Th through

$$^{238}\mathrm{U} \rightarrow ^{234}\mathrm{Th} + \alpha$$

with a half life of 4.5×10^9 years. (The lucky coincidence that this is about the age of the earth allows geophysicists to determine the age of the solar system using radioactive dating.) Such long lifetimes are not uncommon, mainly because the α-particle must quantum mechanically tunnel through a "barrier" at the edge of the nucleus. This is not very probable, so the lifetimes for α-decay tend to be long.

Beta decay is a fundamentally important reaction, since it was the first known example of the *weak interaction*. The weak interaction is not the same "force" that gives rise to nuclear binding or to the decay of excited states by γ-emission. (The latter is the normal, old electromagnetic interaction, used quantum mechanically.) In fact, the weak interaction is much weaker than either of these.

The weak interaction changes protons into neutrons or neutrons into protons. In the process, an electron (e^-) or positron (e^+) is emitted. In the jargon of nuclear physics, these processes are called β^- or β^+ decay respectively. In one form of the weak interaction, an electron from one of the inner atomic shells is captured, instead of emitting a positron. This process is called *electron capture* instead of β-decay, but both are manifestations of the weak interaction.

Neutrinos (ν_e) or antineutrinos ($\bar{\nu}_e$) are also emitted in β-decay. They are nearly impossible to detect because they have no charge and could only be detected through the weak interaction. However, they do have an important effect on β-decay because they carry away some of the energy. Therefore, unlike the photons emitted in γ-decay, the $e^\mp$ in β-decay *are not* monoener-

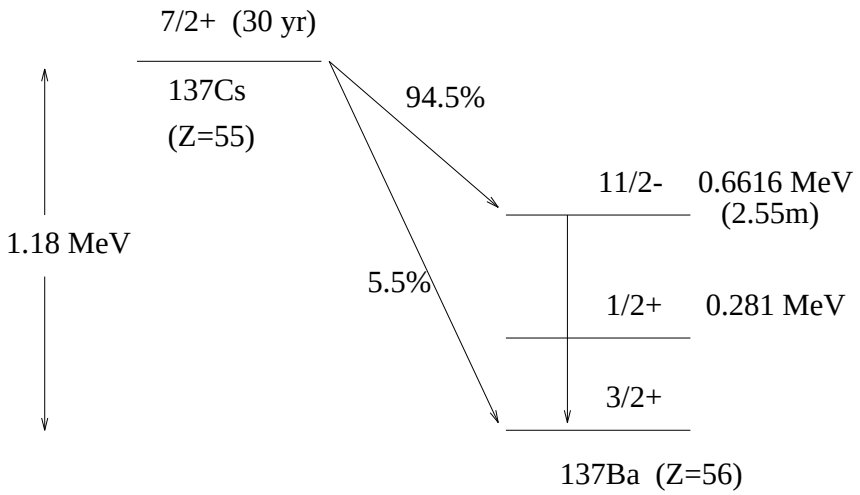


Figure 18.1: Decay scheme of ^{137}Cs. This is a β^- decay which proceeds through the reaction ^{137}Cs$\rightarrow^{137}$Ba$+e^- + \bar{\nu}_e$. You will detect both the e^- emitted in the β-decay, and the radiations from decay of the excited ^{137}Ba nucleus. The $\bar{\nu}_e$ goes undetected.

getic.

A β-decay reaction therefore takes the form

$$^A(Z) \xrightarrow{\beta^\mp} {}^A(Z \pm 1) + e^\mp + \{\bar{\nu}_e, \nu_e\}$$

whereas an electron capture reaction would be

$$^A(Z) + e^- \rightarrow {}^A(Z-1) + \nu_e$$

It is important to realize that the so-called daughter nucleus may in fact be in some excited state which consequently γ-decays. This is even the case more often than not.

Let's illustrate these points (with the exception of α-decay) by considering one particular nucleus that we will use a lot, namely ^{137}Cs. The decay scheme is shown in Fig. 18.1. The ground state of ^{137}Cs β^- decays with a half life of 30.0 years. The difference in the mass between the ^{137}Cs ground state and the ground state of the ^{137}Ba daughter is $\Delta Mc^2 = 1.18$ MeV, so this much energy ends up being divided between the e^- and neutrino kinetic energies, and any γ-rays from the daughter.

The ground state of the ^{137}Cs nucleus has spin 7/2 and positive parity. (This is written as $\frac{7}{2}^+$.) The rule of thumb is that β and γ decays prefer to go to states with larger energy release, and with as small a change in spin as possible. For β decay, nature would also prefer not to change the parity. In the case of ^{137}Cs, the change-in-spin rule dominates, and 94.5% of the decays go to the $\frac{11}{2}^-$ state in ^{137}Ba, which subsequently γ-decays to the ground state, and the rest proceed to the $\frac{3}{2}^+$ ground state. There are no β decays to the $\frac{1}{2}^+$ excited state.

Notice that the γ-ray transition in ^{137}Ba is from a $\frac{11}{2}^-$ excited state at 0.662 MeV to the $\frac{3}{2}^+$ ground state. That is a large change in angular momentum, and this decay proceeds very slowly for a γ-decay. The 0.662 MeV photon is emitted with a 2.55 min half life. You will be able to measure this half life, among other things, in this experiment.

18.2 Measurements

Instead of one main experiment, we will do a series of relatively simple measurements, all involving radioactivity and nuclear reactions. The last measurement outlined is a catch-all, but should give you ideas on more sophisticated measurements that can be tailored to the physics involved.

You will use a Geiger counter to detect the radiation in all the measurements, except for the last section where you might try other kinds of detectors. In any case, the particles of radioactivity are counted using a multi-channel scaling plug-in board, directly interfaced to a PC.

The Geiger counter comes in a neat little package with trays and holders that allow you to position the source several distances away from the detector window. You will also be able to place material between the source and the window, in order to study the attenuation of the radiations.

The plug-in board, manufactured by EG&G/ORTEC and called the MCS-plus system, comes complete with a set of software that runs under Microsoft Windows. The program, located in the "MCS" window and called MCS-

PLUS, is more or less self explanatory, but there is plenty of documentation on it. The data can be saved in a binary format and read back into the program for later use. The data can be printed out to the screen using the program PRINT MCS, located in the same window. A third program called MCS-CRICKET will convert the data to an ascii file suitable for reading into MATLAB or some other program.

18.2.1 Particle Counting Statistics

Before getting on with measurements of actual physical quantities, use the setup to verify the random statistical nature of radioactive decay. This will give you a chance to get used to operating the detector and the software as well.

We will study the mean μ and standard deviation σ for the number of counts N which you collect in some time interval. You can adjust the mean value of N by adjusting the time interval, or "dwell time" in the program, since the count rate is essentially fixed (for the ^{137}Cs source). If you prefer, you can also adjust N by changing either the absorbers or the position of the tray holding the source under the Geiger counter.

Set the ^{137}Cs source on the holder tray, near the detector window. Set the mean value of the number of counts $\mu = \bar{N}$ to be somewhere in the range between 2 and 5. Take on the order of 100 measurements of N, that is set the pass count to 100. Use MATLAB or some other program to calculate the mean μ, as well as the standard deviation σ. (You might review the various ways to use MATLAB to analyze data, mainly in Chapter 9.) As radioactive decay is supposed to be a good example of Poisson statistics, you should find that $\sigma = \sqrt{\mu}$. How good is this?

You can go farther and verify that the distribution of counts is in fact approximated quite well by the Poisson distribution. Make a histogram of the actual number of counts N you measure for each of the $M \sim 100$ measurements you make. (See Sec. 9.5.) Plot on top of this the predicted Poisson distribution, normalized to the number of measurements, that is

$$P(N) = M \times e^{-\mu} \frac{\mu^N}{N!}$$

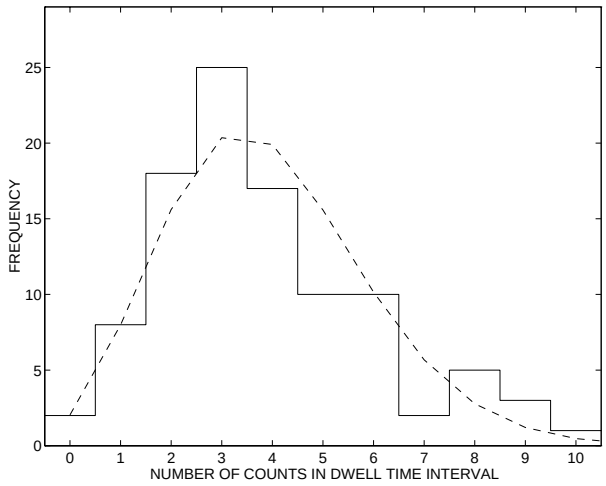


Figure 18.2: An example of counting statistics in radioactive decay. The histogram shows the number of times ("frequency") a certain number of counts appears in a specified time. The dashed line is the Poisson distribution predicted by the mean number of counts and the total number of measurements made.

as we did in Fig. 9.3. How good is this fit? You might go so far as to calculate the χ^2 for the comparison, taking the uncertainty for each bin of the histogram to be the square root of the number of entries in that bin. Recall that a "good fit" means $\chi^2/M \sim 1$.

An example[3] of this sort of data, analyzed with MATLAB, is shown in Fig. 18.2. A list of numbers N is read in from the MCS-plus program, and the mean μ is determined. The numbers are binned into a histogram, using for example the MATLAB function hist, and plotted as a stairs plot. The predicted poisson distribution is then overplotted. The procedure is very similar to that described in Sec. 9.5.

As the count rate or dwell time gets large, the random statistical uncertainty in the rate R will become a small fraction of the rate. That is, $\delta R/R = \sqrt{N}/N = 1/\sqrt{N}$, so as $\mu = \bar{N}$ gets very large, $\delta R/R$ gets very small. At some point, some systematic uncertainty will begin to dominate the uncertainty in R. This could be due to temperature or voltage fluctua-

[3]Data taken by Peter Thies and Dan Bentz, class of 1996.

tions, or many other things.

Try increasing the value of μ and again check the relation $\sigma = \sqrt{\mu}$. See if you can identify a point where there is clearly some other contribution to σ other than from simple counting statistics.

18.2.2 Detecting Radiation

Let's start with some very simple measurements using the ^{137}Cs source. As discussed in Sec. 18.1, this source emits a combination of β^- and γ radiation.

The first thing to do is get a good measure of the *background rate* in the detector. That is, with the ^{137}Cs source far away, measure the number of counts per unit time. This value will have to be subtracted from all other rates you measure, and it is probably a good idea to go back and remeasure it over the course of the experiments, just to make sure it doesn't change.

Don't forget to record the uncertainty in the background rate, as well as in all other rates you measure. This is simple to do, assuming that only random counting statistics contribute. If you determine a number of counts N during a time t, then the rate you measure is $R = N/t$ and the uncertainty in R, i.e. δR, is $\sqrt{N}/t$, and you report $R \pm \delta R$.

Dependence on Distance

The source radiates outward in all directions, and in principle has no preferred direction. Therefore, you expect the rate to vary pretty much like $1/r^2$ as you change the distance r between the source and the detector.

Test this hypothesis. Plot the data in any form you like, but include error bars on the rate to indicate the uncertainty. You might try plotting $r^2 R(r)$ and see if it is a constant, but if you prefer plotting $R(r)$, that is up to you.

Don't forget to subtract background. You can assume that the different rates you measure are random and uncorrelated, so the uncertainty in the

net rate is given by adding the uncertainties in quadrature, that is

$$R_{NET} = R_{SOURCE} - R_{BACKGROUND}$$

and

$$\delta R_{NET} = \sqrt{\delta^2 R_{SOURCE} + \delta^2 R_{BACKGROUND}}$$

Explain your results. Over what range of distances do you expect the $1/r^2$ rule to be valid.

Attenuation of Radiation

Now be a little more ambitious with the physics. Take the ^{137}Cs source and measure the count rate with each of the two sides closer to the Geiger counter window. Can you detect a difference in the count rate? Explain what you see. Remember that a Geiger counter is detecting particles that interact the gas, a very low density medium. Does that mean it is a better detector for β particles or for γ-rays?

Place different absorber materials between the source and the detector, and see if you can reduce the count rate. You can do this neatly by using the tray holders underneath the Geiger counter and the various aluminum absorbers that are provided. It is a good idea to measure the thickness of the aluminum absorbers carefully.

You should see a clear decrease in the count rate as you add thin aluminum absorbers. This is because you are detecting β^- which are attenuated rapidly. The range of β particles in matter is roughly given by the relation

$$R = 0.52 \times E_\beta - 0.09$$

where R is in gm/cm^2 and E_β is in MeV. You should be able to use this formula and your measurements to estimate the maximum β^- energy emerging from the source. Compare this to what you expect. Include some uncertainty estimate in E_β from an uncertainty you estimate in R. Is this "range method" a good way of measuring the energy of β particles?

At some absorber thickness, the attenuation should be only due to the loss of the 662 keV photons. Therefore, further attenuation should be governed

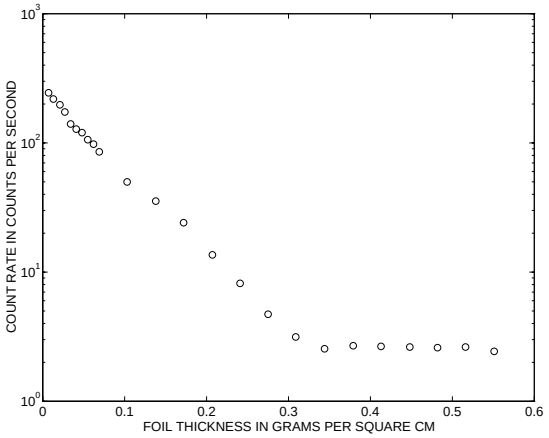


Figure 18.3: Sample data for β^- absorption in aluminum foils from a ^{137}Cs source.

by Eq. 17.2, where μ is the appropriate value for the material you are using and this photon energy.

A sample of this sort of data[4] is shown in Fig. 18.3. Notice how the rate falls, but abruptly becomes more or less constant. Where does this constant rate come from?

18.2.3 Half Life Measurements

Now let's measure the half lives of some short lived nuclear states. Obviously, you need to do some tricks to get short lived states that you can measure. One trick we will use is the chemical separation of barium from cesium. However, beyond that, we will in fact create new isotopes using a type of nuclear reaction called *neutron activiation*.

In neutron activation, reactions with neutrons are used to create radioactive isotopes from stable nuclei. Neutrons are produced using a Plutonium-Beryllium (PuBe) source, which is safely packaged away so you can't get near

[4]Data taken by Peter Thies and Dan Bentz, class of 1996.

it, but the neutrons can get out. Plutonium decays by α-emission, that is

$$^{239}\mathrm{Pu} \rightarrow ^{235}\mathrm{U} + \alpha$$

and the α particles react with the Beryllium

$$\alpha + ^{9}\mathrm{Be} \rightarrow ^{12}C + n$$

releasing neutrons. These neutrons are slowed down by collisions with protons (in all the paraffin, a hydrocarbon, surrounding the source), making them available for other reactions.

When you put an isotope in the neutron radiation "oven", make sure you "cook" it for at least a large fraction of one half life. Otherwise, you may not get enough rate for you to measure.

Production and Decay of ^{116}In

This is a good place to start. You will make ^{116}In using neutron capture on a piece of indium. Indium is a very common metal used for soldering compounds, and all of natural indium is the isotope ^{115}In. The decay scheme for ^{116}In is shown in Fig. 18.4. Notice that the ground state has a very short half life, only 14 sec. You will be detecting β^- decay of the *excited* state, 60 keV above the ground state. The decays proceed mainly to a couple of states at around 2.3 MeV and the available energy is 3.3 MeV, so the β^- typically have energies up to an MeV or so. These are easy to detect in your Geiger counter.

Irradiate the piece of indium for an hour or so. Remove it and place it on the Geiger counter platform, close to the counter window. Take data for an hour or so, setting the MCS program to count for intervals of something like a minute. Save the data, and convert it to ascii for further analysis.

It's probably a good idea to make a semilog plot of the data, and estimate the half life by hand, just to make sure the number looks about right. To do a better job, you can easily fit the data to a decaying exponential. Just use the MATLAB function polyfit to fit the logarithm of the number of counts versus channel to a straight line. In fact, this is a case where you can accurately

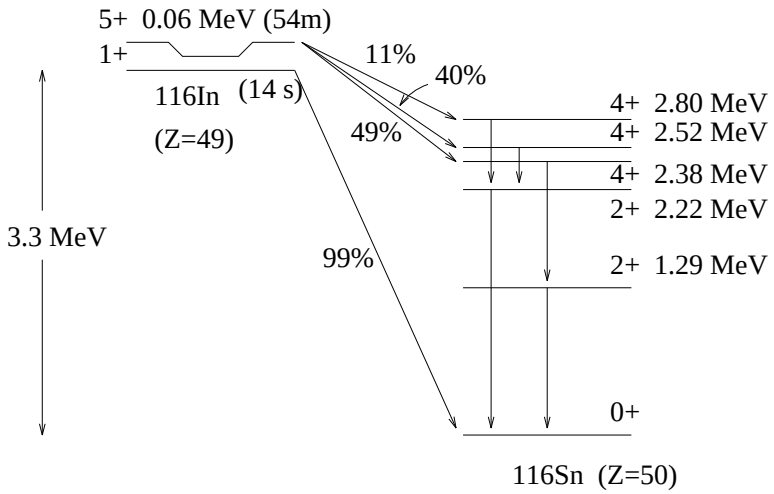


Figure 18.4: Decay scheme for ^{116}In.

write the random errors of the points, since they are governed by a Poisson distribution. That is, if there are N counts in any one channel, then the random uncertainty in N is $\delta N = \sqrt{N}$ and the random uncertainty in the logarithm of N is $\delta \log N = 1/\sqrt{N}$. Recall that the MATLAB routine linreg (Fig. 9.1) fits weighted data to a line.

A sample of data[5] on indium decay is shown in Fig. 18.5. Each channel represents 30 sec. The simple fit described above is shown in the *dashed* line. Notice that the fit isn't really very good. You can see that more clearly if you plot the difference between the fitted function and the data points.

In fact, it's not too surprising. When you put the chunk of indium in the neutron oven, it irradiates other things in the material as well, and you expect some background radiation. You can try subtracting a constant value (representing the background counts) from the data before you fit it, and see if it looks better. By calculating the χ^2 function, you can even optimize the background term by minimizing χ^2.

The MATLAB program shown in Fig. 18.6 was used to do exactly this. After reading in the values of channel and counts, the user is asked for a number of background counts. Then this value is subtracted from the data,

[5]Data taken by Peter Thies and Dan Bentz, class of 1996.

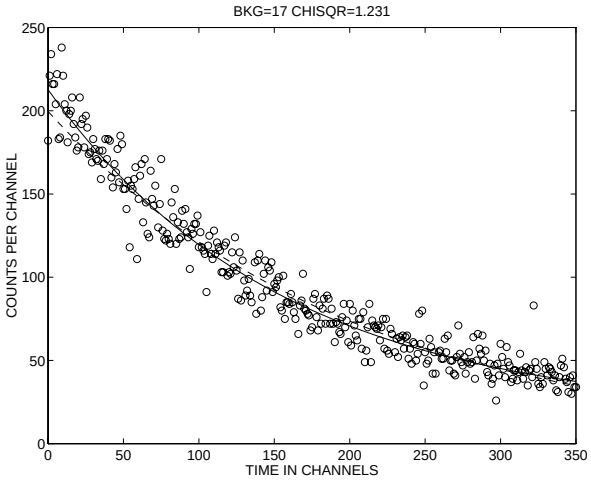


Figure 18.5: Data and fits for the decay of ^{116}In. The dashed line is fit to a decaying exponential, while the solid line includes a constant backgroud of 17 counts.

and care is taken to make sure the value is not less than one. (Remember, you are going to take a logarithm.) Two fits are done, one that is unweighted (using polyfit) and one that is weighted according to the Poisson uncertainty in the points (using linreg). The results, including the χ^2, are printed and plotted. By trying various backgrounds, you find that the lowest χ^2 (i.e. the "best fit") is found for 17 background counts. You can even estimate your *systematic* uncertainty by looking at how much the lifetime varies as you move around in χ^2 near the minimum. This can be large if the minimum in χ^2 is shallow. For this particular data set, we find

$$\tau = 160.7 \pm 2.0 \pm 10 \text{ channels}$$

where the first uncertainty is random and the second is systematic. Since each channel is 30 sec, we determine

$$t_{1/2} = \log 2 \times \tau \times \frac{1}{2}\text{min/channel} = 55.7 \pm 0.7 \pm 3.5 \text{ min}$$

which agrees well with the accepted value of 54 minutes. In fact, it seems we may have overestimated the systematic uncertainty.

Actually, this business of adjusting the background term to minimize χ^2

```
% LOAD AND EXTRACT DATA POINTS
load indium.dat
chan=indium(:,1);
data=indium(:,2);
%
% PREPARE DATA FOR FITTING LINE TO LOGARITHM
bkgd=input('Background counts ');
dnet=max(data-bkgd,1);
ndof=length(data)-2;
edata=sqrt(data);
ldata=log(dnet);
eldata=edata./dnet;
%
% UNWEIGHTED FIT
coefa=polyfit(chan,ldata,1);
fita=exp(polyval(coefa,chan));
chisqa=sum(((dnet-fita)./edata).^2);
fprintf('Unweighted fit:\n');
fprintf(' tau=%6.3e\n',-1.0/coefa(1));
fprintf(' chisquare/dof=%6.3f\n',chisqa/ndof);
%
% WEIGHTED FIT
[coefb,ecoefb,lfitb]=linreg(chan,ldata,eldata);
fitb=exp(lfitb);
chisqb=sum(((dnet-fitb)./edata).^2);
fprintf('Weighted fit:\n');
fprintf(' tau=%6.3e',-1.0/coefb(2));
fprintf(' uncert=%6.3e\n',ecoefb(2)/coefb(2)^2);
fprintf(' chisquare/dof=%6.3f\n',chisqb/ndof);
%
% PLOT RESULTS
plot(chan,data,'o',chan,fita+bkgd,'-');
title(['BKG=',num2str(bkgd),' CHISQR=',num2str(chisqa/ndof)]);
```

Figure 18.6: Program (i.e. m-file) used to fit indium data.

can be done automatically in MATLAB. That brings us into the world of nonlinear fitting (Sec. 9.2.3), and we'll do that next.

The Half Life of ^{137m}Ba

Now we'll measure the half life of another short lived isotope, ^{137m}Ba. The background is very clear in this case, and we'll use that to go a step further in our data analysis techniques. This isotope does not need to be produced in the neutron oven.

Recall the decay scheme of ^{137}Cs in Fig. 18.1. The daughter nucleus, ^{137}Ba, is produced in its ground state only 5.4% of the time. The rest of the time it is made in the excited state, called ^{137m}Ba for "metastable", which decays by γ-ray emission, but with a relatively large half-life (for γ decay) of around 2.5 minutes. Of course, ^{137m}Ba is produced all the time, as the very long-lived ^{137}Cs decays, so you can't isolate the ^{137m}Ba decay without somehow separating it from the ^{137}Cs.

You can make this separation because *chemically*, Cesium is very different from Barium. By passing a weak HCl solution through a ^{137}Cs source, Barium is captured and comes out in solution. Some Cesium comes through as well, but most of the radioactivity of the solution is from ^{137m}Ba. There are some small squeeze bottles of HCl in the lab, and specially prepared ^{137}Cs samples that allow you to force a few drops over the radioactive isotope. It is best if you squeeze the drops through *slowly*, enough to fill the small metal holder in about 30 seconds. Then, place the holder in the Geiger counter tray, and start the MCS program.

Realize that you are working with radioactivity and hydrochloric acid. **Don't be careless**. None of this is concentrated enough to be particularly dangerous, but you should take some simple precautions. Disposable gloves are located near the setup. It is also a good idea to wash your hands soon after you're finished.

You should choose a dwell time and pass length that allows you to get a relatively large number of points in each channel, but many channels over the expected decay time of a few minutes. You should be able to get several

hundred counts per bin in the first bin or two, and a background less than 20 counts per bin. (The background level will be clear after counting for a half hour.) You might need a few tries to get all of this where you want it. Save the data and convert it to ascii for later analysis.

You can use the program in Fig. 18.6 to fit the data and adjust the background counts, but that is tedious. In this case, since the background will be very clear, you can determine it precisely by averaging over the last many channels, and subtract that number from the data before fitting. However, MATLAB gives you the ability to fit things all at once.

What you need to do is minimize the χ^2 function numerically, and MATLAB gives you two numerical minimization functions. These are fmin, which minimizes a function of one variable. and fmins for many variables. You need to minimize χ^2 as a function of three variables, two for the exponential and one background value, so you need to use fmins.

First, write a simple m-file called expcon.m which calculates the function you are going to fit to the data:

```
function y=expcon(x,pars)
y=pars(1)*exp(-x/pars(2))+pars(3);
```

and then write another called chiexpcon.m which calculates χ^2:

```
function chisqr=chiexpcon(pars,xdata,ydata,edata)
chisqr=sum(((ydata-expcon(xdata,pars))./edata).^2);
```

Don't forget that for these data, the array of uncertainties edata is just the square root of the counts, i.e. edata=sqrt(edata).

Play around with some values of pars(1,2,3) so that you have a good starting point. (Just plot the data points, and then overplot the function expcon until it looks kind of close.) Then type the command

```
fmins('chiexpcon',pars,0,[],chan,data,edata)
```

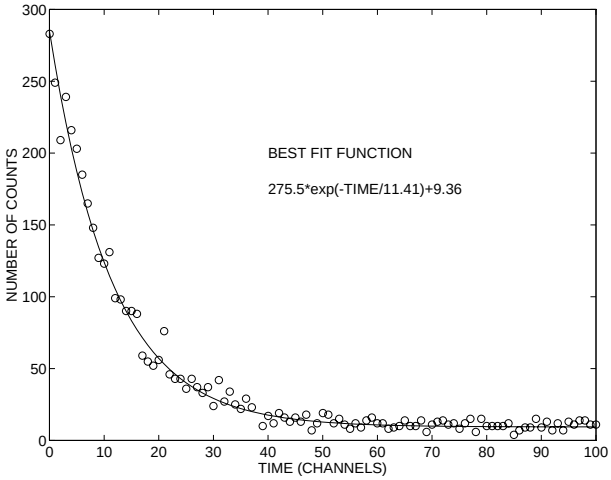


Figure 18.7: An example of a nonlinear fit. The data is from the decay of ^{137m}Ba, including some constant background. The MATLAB function fmins was used to make the fit.

and after a little chugging away, you will get the best fit values returned. (Check the manual for details of the arguments for fmins.) Exactly this procedure was followed to fit the data shown in Fig. 18.7.

You should probably check that $\chi^2/\text{dof} \approx 1$. Try to make some estimate of the uncertainty, and compare your result to the accepted value. You can estimate the uncertainty as we did for ^{116}In, by subtracting the background and using linreg to fit the log to a straight line. Be careful to propagate the errors correctly on the subtracted data.

Note that the radioactivity you detect from ^{137m}Ba decay is γ radiation, which is not detected very efficiently by a Geiger counter. You might try using a NaI(Tl) detector instead, keying in on the the particular γ-ray in question. This should greatly increase your counting statistics, as well as reduce the background. It might be easiest to do this after having worked on Experiment 11.

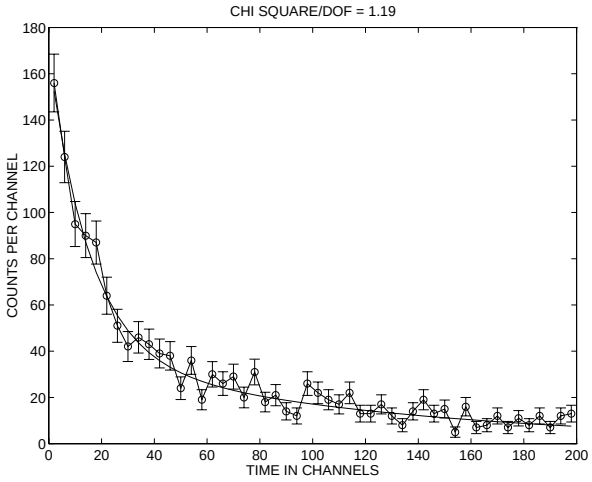


Figure 18.8: The decay of neutron activated natural silver, fit to two decaying exponentials. The plot was made using the MATLAB function errorbar.

Radioactive Silver Isotopes

Natural silver is pretty much evenly divided between two isotopes, ^{107}Ag and ^{109}Ag. Neutron activation captures a neutron equally well on these two isotopes, producing two radioactive isotopes ^{108}Ag and ^{110}Ag. Both of these decay with a pretty stiff β^- that is easy to detect, but one isotope has a half life of 24.4 sec and the other is 2.42 minutes. You might want to look up the decays to get more details.

Take a piece of the pure silver foil and cook it in the neutron oven for at least ten minutes. *Quickly* take it out, put it in the Geiger counter, and start the program. Don't forget that the lifetime of the shorter lived isotope is only half a minute. It should be clear from the MCS program that there are two lifetime components from the decay.

Representative data[6] is shown in Fig. 18.8, where each channel is 2.5 seconds long, but I've used MATLAB to add every four channels together to get better statistics in each channel. Errorbars are added to the data points using the errorbar function. The points are fit to a double exponential decay,

[6]Data taken by Peter Thies and Dan Bentz, class of 1996.

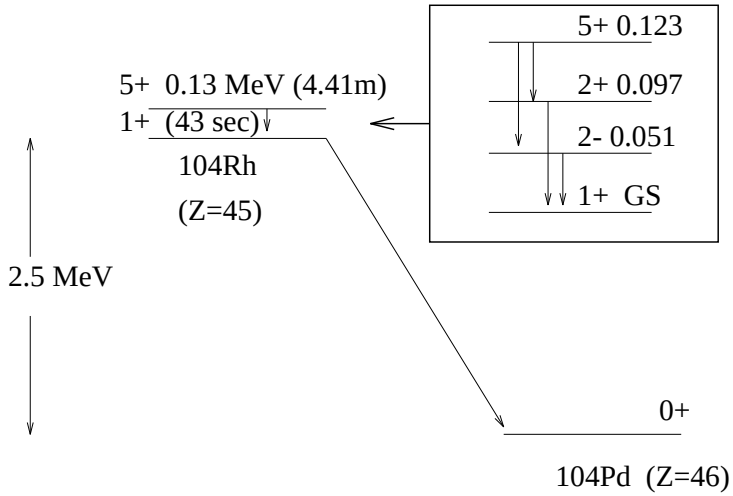


Figure 18.9: Decay scheme for ^{104}Rh.

completely analagous to the way we fit a constant plus and exponential to the ^{137m}Ba data. The only difference is that the m-files for the fit function and for the χ^2 are changed slightly.

It is difficult to determine the random uncertainty for these fits, but you can try the following. Subtract one of the exponentials from the data, take the logarithm of the remainder, and use the linreg to fit it (and its uncertainty) to a straight line. Does this over- or under-estimate the random uncertainty?

Production and Decay of ^{104}Rh

The decay scheme for ^{104}Rh is shown in Fig. 18.9. This is a peculiar isotope indeed. The 43 sec 0^+ ground state decays nearly all the time with a very stiff (up to 2.5 MeV) β^- to ^{104}Pd. (In fact, a small fraction of the time, the ground state decays in the other direction, via electron capture to ^{104}Ru, but the branching ratio is less than 1%.) The first excited state, on the other hand, lives a lot longer (4.41 min), but decays through a cascade of very soft γ-rays, with energies between 30 and 100 keV.

All of natural rhodium is ^{103}Rh, so neutron activation will make ^{104}Rh very nicely. Will it also make the 5^+ excited state? There doesn't seem to be

any difficulty with this in indium, but there the easy state to detect was the excited state. Here, the ground state decay is the easy one to see, at least with a Geiger counter.

Irradiate the piece of rhodium foil that is kept with the neutron oven. *Be careful with this foil. Rhodium is very expensive.* Place it in the Geiger counter, and try to identify the relatively short lived ground state decay. That should not be very hard to do.

The excited state decay will be harder. These low energy photons are not picked up very well in the Geiger counter. It would be much better to use a NaI(Tl) detector with a thin window that allows these low energy photons to penetrate into the crystal. If any of the γ transitions in Fig. 18.9 can be identified, then you can test to see if the excited state is indeed populated in neutron capture.

Ch 19

Experiment 11: Positron Annihilation

In this experiment, you will make measurements of different variables of radioactive decay. In particular, you will study the *coincidence* of two different γ-ray detectors. The two γ-rays come from the same radioactive decay, hence they should be detected at the same time. The setup is first used to study the particularly simple correlation of the two γ rays from positron annihilation, and you will have the opportunity to carefully measure properties of the NaI(Tl) detectors. Then you can measure the so-called $\gamma\gamma$ angular correlation from ^{60}Co decay, and explore the physics that goes along with it.

The physics and the technique are covered quite thoroughly in

- *Experiments in Modern Physics*,
  Adrian C. Melissinos, Academic Press (1966)
  Sections 9.3 and 9.4

Section 9.3 concentrates on the experiment, while section 9.4 goes into some detail on the theory of electromagnetic transitions in nuclei, i.e. γ-decay.

A lot of the detailed work you can do in this experiment specifically involves NaI(Tl) detectors for γ-rays. For these specific details, a good reference

is

- *Applied Gamma-Ray Spectrometry*, C.E. Crouthamel
  Pergamon Press (1960)

This book should be available on reserve at the library.

19.1 Correlated Pairs of γ-Rays

If two (or more) γ-rays emerge simultaneously from the same decay, then you expect them to be somehow correlated with each other. For example, their energies or relative angles would not be independent of each other. We will concentrate on two examples of this, one simple and one more sophisticated.

Let's consider the simple case first, namely the photons that emerge from positron annihilation with an atomic electron, where everything is at rest in the beginning. The reaction is

$$e^+e^- \rightarrow \gamma_1\gamma_2$$

We know there must be at least two photons emitted because there would be no way to conserve momentum if there were only one. If there is no net momentum in the beginning ("everything is at rest"), then there must be no net momentum in the end, and you can't do that with a single photon.

So, let's see if we can do it with two photons. We need to conserve not only momentum, but also energy. The energy of the two photons is just made from the mass of the e^+e^-, so

$$2mc^2 = E_1 + E_2$$

where m is the electron (and positron) mass and E_1 and E_2 are the energies of photons 1 and 2. As for momentum, the only way to make it zero in the end is if both photons are moving directly away from each other at 180°, and if the magnitude of their momenta are equal, that is

$$E_1/c = E_2/c$$

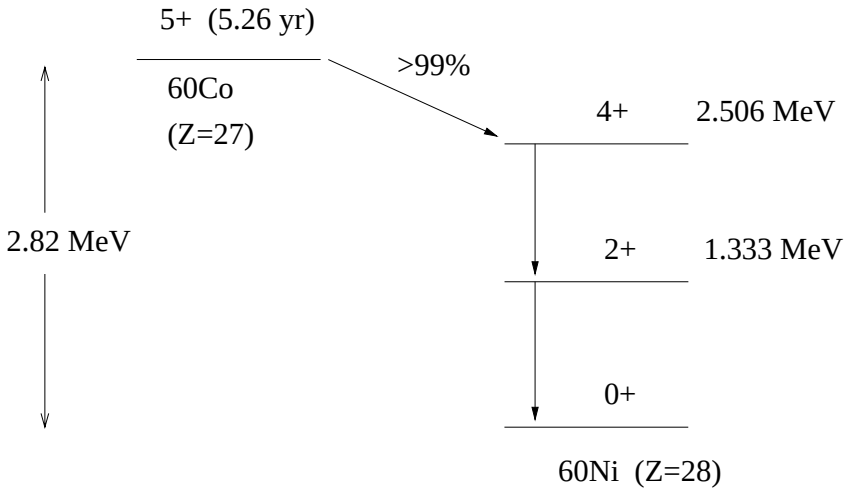


Figure 19.1: Decay scheme of ^{60}Co. Only the relevant states in ^{60}Ni are shown.

These equations can in fact be trivially solved for E_1 and E_2, and you get

$$E_1 = E_2 = mc^2 = 0.511 \text{ MeV}$$

You could probably have guessed that was what it had to be.

This is surely a simple correlation. The two photons must have equal energies, and they must be at 180° with respect to each other. Your job in this experiment will be to verify these assertions as carefully possible.

A different type of $\gamma\gamma$ angular correlations, which contains lots of neat physics, is provided by the β decay of ^{60}Co. The decay scheme of ^{60}Co is shown in Fig.19.1. This nucleus decays through β^- emission to ^{60}Ni with a 5.26 year half life. More than 99% of the decays go to the 4^+ 2.506 MeV excited state in ^{60}Ni. This state decays 100% of the time to the 2^+ 1.333 MeV excited state, which subsequently decays to the ground state. The lifetime of the 2^+ state is around 0.7 ps which is much shorter than anything you can be sensitive to in this laboratory. Therefore, ^{60}Co β decay is characterized by the emission of two simultaneous γ-rays, with energies of 1.173 MeV and 1.333 MeV.

Consider the physics of these emitted γ-rays. The first one, with energy 1.173 MeV, can come out in any direction it pleases. We are not going to

consider detecting the β^- along with the γ-rays, so there is absolutely no preferred direction in space.

But what about the second γ-ray, the one with 1.333 MeV energy? The first γ-ray obviously carried with it some angular momentum since the nucleus changed from a spin-4 state to a spin-2 state. That means that the spin vector of the spin-2 state has some orientation in space, relative to the direction of the emitted 1.173 MeV photon. This might therefore imply that the 1.333 MeV photon is not free to come out in any direction it pleases, and in fact it cannot.

Melissinos goes through the physics, which requires some understanding of the multipole expansion of the electromagnetic field, but the result is easy to express. It depends only on spin and electromagnetism, not on the particulars of the decaying nucleus. You find that the $\gamma\gamma$ angular correlation is given by

$$\alpha(\theta) \equiv \frac{R_{\gamma\gamma}(\theta)}{R_{\gamma\gamma}(90^\circ)} = 1 + a_1 \cos^2(\theta) + a_2 \cos^4(\theta) \tag{19.1}$$

where $a_1 = 1/8$ and $a_2 = 1/24$. In other words, it is around 16% more probable for the two photons to be emitted back-to-back than at 90° relative to each other.

19.2 Measurements

This experiment does not make heavy use of computerized data acqusition. Instead, your experience will be with the use of NaI(Tl) detectors and their use as γ-ray detectors. Taking the data is straightforward, but the interpretation will require more thought.

The setup is shown schematically in Fig.19.2. Photons are detected in two identical 2 in. diameter $\times$2 in. long NaI(Tl) scintillators. These detectors are mounted on a table that points both detectors at a center point, and lets one of them move around that center point through a large angular range. You can read the relative angle off the scale mounted on the table. You can also adjust the distance of the detectors to the center point by sliding it along the mounting rails.

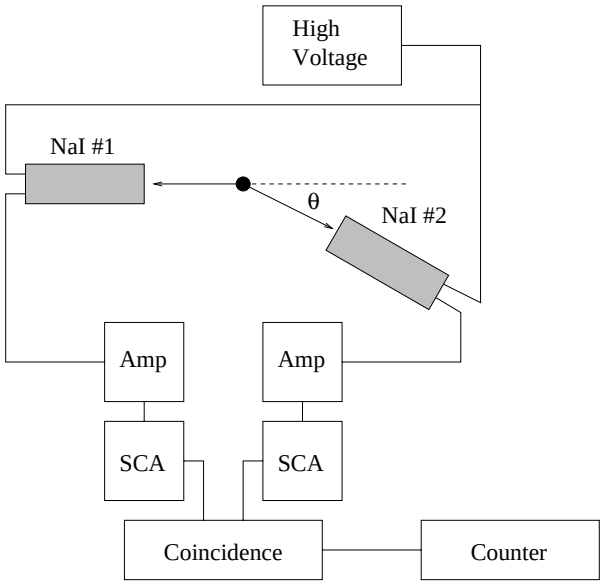


Figure 19.2: Setup for measuring $\gamma\gamma$ angular correlations.

The photomultipliers are powered with a high voltage DC supply, and the signal outputs pass into two identical Canberra 2012 pulse amplifiers. The amplifier outputs are fed into single channel analyzers, and the outputs of the SCA's are fed into a fast coincidence module. A Canberra 1772 visual counter lets you measure the count rate. You can measure "singles" (i.e. not-in-coincidence) counting rates either by changing the switches on the coincidence module, or by bypassing this module and putting the SCA output directly into the visual counter.

There are two kinds of SCA's you can use. One is a basic device like the Ortec 550A which delivers an output pulse as soon as the input crosses threshold. This is prone to "time slewing", however, since it leads to an output time that varies with pulse height. See Fig. 19.3. This effect is large, since the rise time of the amplified pulse is on the order of microseconds. Therefore, a large time window is necessary to allow for coincidences, and this increases your chance of getting an accidental coincidence, i.e. background.

Consequently a second type of SCA, called a "timing SCA", is also available. It uses a special circuit that is provides an output pulse at a time independent of the pulse height, regardless of when the pulse crosses thresh-

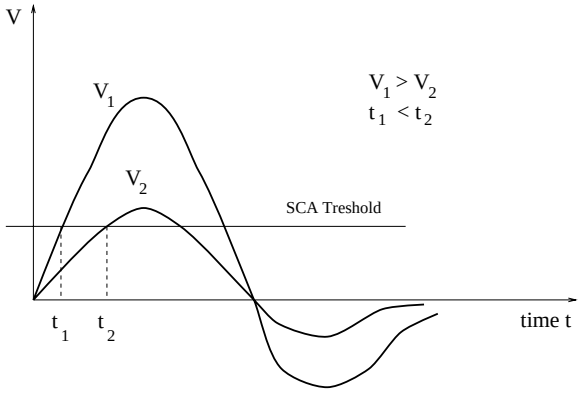


Figure 19.3: Time slewing with a basic Single Channel Analyzer. Rise times of the voltage pulse is on the order of a microsecond, so the output time can vary over hundreds of nanoseconds depending on the size of the pulse.

old. One example is the Ortec Model 420. You will want to use this device for careful coindicence timing.

Two kinds of coincidence modules are also available, the Ortec 414A and the Canberra 840. The 414A allows you to make more careful measurements as a function of the resolving time (more on this soon), but has limited dynamic range. The 840, on the other hand, has an order of magnitude more range, but cannot make such precise steps.

Positrons are produced using the ^{22}Na radioactive source. The decay scheme for ^{22}Na is shown in Fig. 19.4. More than 99.9% of the decays of this nucleus go to the first excited state of ^{22}Ne at 1.275 MeV. Most of these decays (90%) proceed through β^+ decays, which is where you get your positrons from, and the rest decay through electron capture. The maximum energy of the β^+ is rather small, around 0.55 MeV. You should estimate the thickness of material required to stop these positrons.

19.2.1 Procedure and Analysis

We'll go through several measurements you can make with this setup. They progress from first learning how to use NaI detectors to measure gamma ray

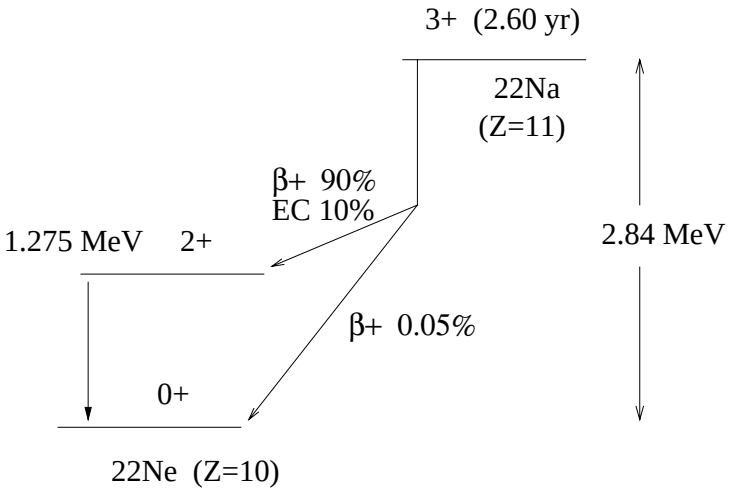


Figure 19.4: Decay sheme for ^{22}Na.

photons, to learning how to do coincidence photon detection, to measuring the $\gamma\gamma$ angular correlation in ^{60}Co. You should work on them more or less in the order listed, since you need skills from one to do a good job on the next.

Turn the high voltage on to around 1150 V. You should leave it on for a while (a half hour or so) to let the phototube bases warm up. In the meantime, check the connections and insert the radioactive source in the holder by removing the screws, and placing the source in the inside cup. Screw the holder into the center point of the rotating table.

Adjust the gain of the amplifiers so that each detector gives a ≈ 3 V signal for the 0.511 MeV annihilation γ-rays. (These pulses should be obvious to you when you look at the output of the amplifier on the oscilloscope.) Set the Ortec 550A SCA to "symmetric window" operation. In this mode, the "lower level" (E) is the center of the window, and the "window" (ΔE) knob controls the width of the window. That is, the SCA gives an output logic pulse if the input pulse height is between $E - \Delta E/2$ and $E + \Delta E/2$. The knobs set the value of E in volts (range 0-10 V) and ΔE in tenths of volts (range 0-1 V).

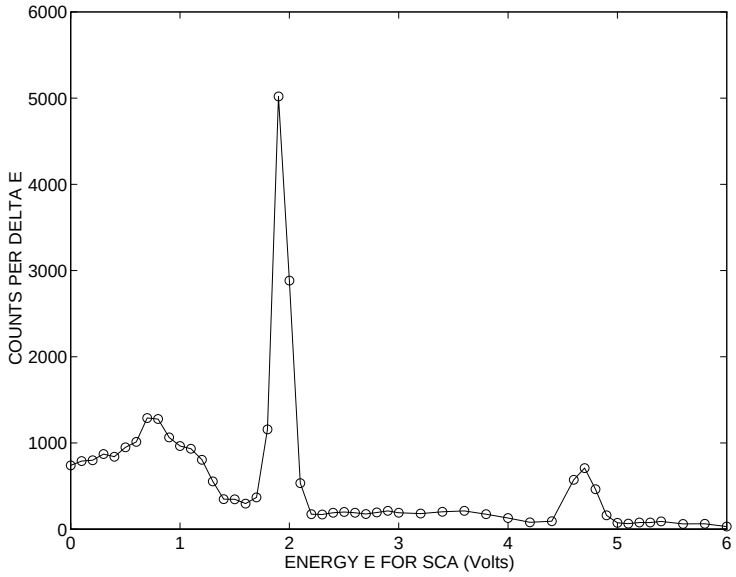


Figure 19.5: Singles data taken with the SCA and a ^{22}Na radioactive source.

Singles Measurements

Now measure the shape of the γ-ray spectrum for each of the two detectors. Slide the detectors so they are very close to the source, within a couple of cm. Set the window width on the SCA large enough so that you get a decent number of counts in a reasonable time, but small enough so that you do not wash out the peaks. Something like 0.1 V to 0.2 V should be alright. Keep an eye on the flashing red light on the front of the SCA. It flashes whenever it emits an output pulse, and it can be a handy way of checking whether the rate is large or small. Record the number of counts out of the SCA with the visual counter for some preset amount of time, as a function of E. Plot these numbers. The result should look like Fig. 17.7. Specific examples can be found in Crouthamel.

Typical data[1] is shown in Fig. 19.5. Notice the sharp peak in the count rate for $E \approx 3$ V, corresponding to the annihilation γ-rays. Identify the compton edge, and calculate the energy (in MeV) where it occurs. You should also be able to see the 1.27 MeV γ-ray. Is your data consistent with

[1]Data taken by Aaron Blow and Rick Vigil, Class of 1996

a linear energy response and zero energy corresponding to zero volts? You can put in other radioactive sources if you wish to check other γ-rays. You may want to adjust the offset level on the amplifier to make "zero energy" as close as possible to "zero volts".

Using your data, calculate the number of counts in the total absorption peak for the annihilation γ-rays. (A simple "triangle" approximation to the peak is good enough.) Realize that you are measuring $\Delta N/\Delta E$, so you must include the value of ΔE you used when you calculate the integral. Check this estimate by actually measuring the total number of counts under the peak. To do this, switch the SCA to "normal" mode, and adjust the "lower level" and "upper level" knobs (now both 0-10 V) so that the range just covers the peak. Now the counter gives you the total number of counts that come in the total absorption peak for the annihilation γ-rays. This should agree with your estimate based on your measurements of $\Delta N/\Delta E$.

Take the total number of counts in the absorption peak and calculate the activity (i.e. the total decay rate) of the ^{22}Na source. You will need to measure the distance from the source to the detector so that you can look up the intrinsic efficiency ε in Crouthamel. You can read the photopeak efficiency P from Fig. 17.8. Don't forget that there are *two* 0.511 MeV photons for each β^+ decay, and that only 90% of the decays are β^+. When you compare it to the value labeled on the capsule, don't forget to take into account the elapsed time between when the capsule was made and the day you make your measurements.

You might want to repeat some of these measurements as a function of the distance between the detector and the source. See if it is possible to observe any changes in the shape of the spectrum, possibly in the peak region. You might also want to put a collimator (i.e. a block of lead with a hole in it) in front of the detector and see what effect that has. Also, you expect the count rate to pretty much follow the intrinsic efficiency ε as a function of distance. Can you test that?

Try measuring the attenuation of photons in material. Some steel plates are provided which will absorb some but not all of the photons. See how the rate within the total absorption peak changes as a function of the steel thickness. Does it agree with what you expect from photon absorption curves?

What effect does the detector resolution have on this measurement?

You can learn a lot from these singles measurements, but the technique is obviously tedious. There are two ways, however, to get around the task of making many measurements by hand of the rate for different values of E. One way is to borrow the MCA-plus system used in Experiment 10. The input box allows you to analyze the amplifier output and use the system as an SCA, taking a full spectrum automatically. The other way is to borrow the multichannel analyzer and GRAPHWIN program setup from Compton scattering, Experiment 12.

Coincidence Measurements

Now let's move on to measuring *coincidences* between the two γ-rays. At this point, the SCA settings should be set to include the full absorption peak for both detectors. Check the SCA outputs on the oscilloscope to make sure you see the coincident logic signals. (You might want to go back and reread Sec. 3.5.1 and Sec. 3.5.3.) It is a good practice to use equal length cables for the two signals, but the time delay in cables is only ~ 1.5 ns/foot, which is smaller than you really need to be concerned with here. Make the time base short enough so that you can see the relative timing of the leading edge of the logic pulses. You will see a lot of "jitter" of one pulse relative to the other, mainly because of the slow response of the NaI detector.

Now put the two logic pulses into the coincidence module. Measure the coincidence count rate as a function of the resolving time, which you change using the knob on the front panel. Make a plot of the count rate versus resolving time. An example[2] is shown in Fig. 19.6 using the Ortec 420 timing SCA and the detectors separated by 8 cm. You should see the rate rise quickly until the resolving time covers the time jitter. You should be able to use your oscilloscope observations to estimate the resolving time at which this occurs. Past this point, the coincident rate should rise only very slowly, mainly due to accidental coincidences. The accidental coincidence count rate goes up with resolving time, of course, because there is a better chance of getting a random pulse into a longer time window.

[2]Data taken by Aaron Blow and Rick Vigil, Class of 1996

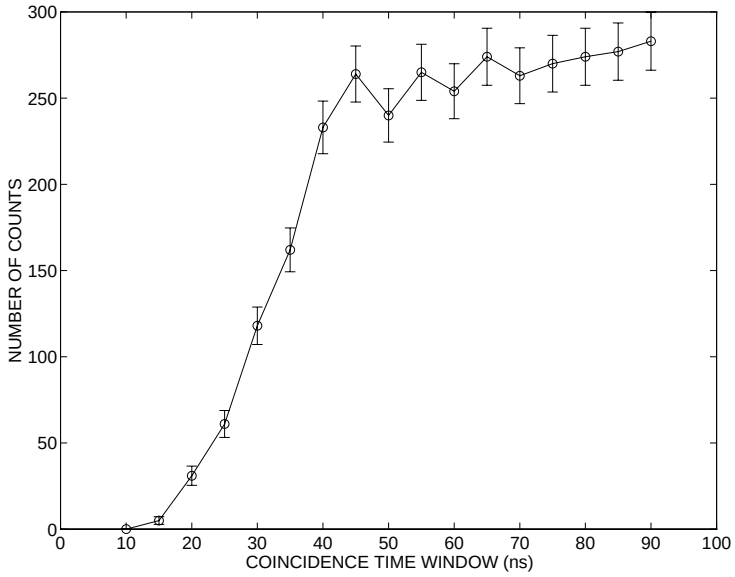


Figure 19.6: Coincident count rate as a function of resolving time using the timing SCA.

Measure the coincidence count rate as a function of the distance between the two detectors, or the distance each of them is set from the center point. Keep the two detectors back-to-back for the time being. Does the coincidence count rate vary as you expect?

Now change the angle between the two detectors, and record the coincident count rate. How well can you verify that the two γ-rays come out at 180°? What determines the width of the distribution for count rate as a function of angle? Estimate the width you expect from the geometry and compare it to what you measure. Test your ideas by changing whatever you think is reasonable, and see if the count rate varies the way you expect.

This measurement is relatively straightforward and you can learn a lot by carrying it out carefully. However, the correlation between the two γ rays in positron annihilation is very strong (pretty much a δ-function, in fact) and that makes it rather easy to do a clean job. If the correlation is weaker, then the measurement is harder, but the physics associated with the correlation is sweeter.

19.3 $\gamma\gamma$ Angular Correlation in ^{60}Co

This experiment is harder than positron annihilation for two reasons. First, the two γ-rays are not of the same energy, although they are close. (The energies are 1.17 MeV and 1.33 MeV. See Fig. 19.1.) That means that the SCA can be set for one γ-ray or the other, or for both. The other (and much more significant) reason the experiment is harder is because the angular correlation is not nearly as sharp. For positron annihilation, the two photons are really back-to-back, but for ^{60}Co, they are almost uncorrelated. This means that you have to be more careful about a lot of things, including details of your procedure and the accidental coincidence rate.

This measurement will be very sensitive to small changes in the photon detectors over the time of the measurement, because you will be trying to measure a small correlation. It is important to let everything come to thermal equilibrium, for example, so that temperature changes don't make for big gain drifts. You should turn everything on and leave it on for at least several hours before making your final measurements, but while things are warming up, it is a good time check things out and become familiar with the technique.

Since the two photons are almost completely *un*correlated, the coincidence rate will be very small. To make up for this, you should use a rather hot source. (This will make the issue of accidentals a problem, however, but we will return to that in a moment.) We have in the laboratory, in a large lead container, a sealed ^{60}Co source (TRACERLAB Catalog Number R-31, S/N B-405) that was 10.6 mCi when it was calibrated on February 27, 1961. Even with the 5.26 yr half-life of this isotope, there is still plenty of activity left. The source is actually at the very end of a long stainless steel rod, and you should hang it vertically, with the tip at the center of the setup, using the ring stand assembly. It is important that things stay stable, so make sure the ring stand is anchored and that you don't bump into it during the course of the measurements.

Now to make s ome measurements. First things first. Make sure you can see the two γ-rays clearly in the singles spectra. Just as you did for ^{22}Na, measure $\Delta N/\Delta E$ as a function of E. You should be able to easily see two nearby peaks, corresponding to the two γ-rays. An example is shown in

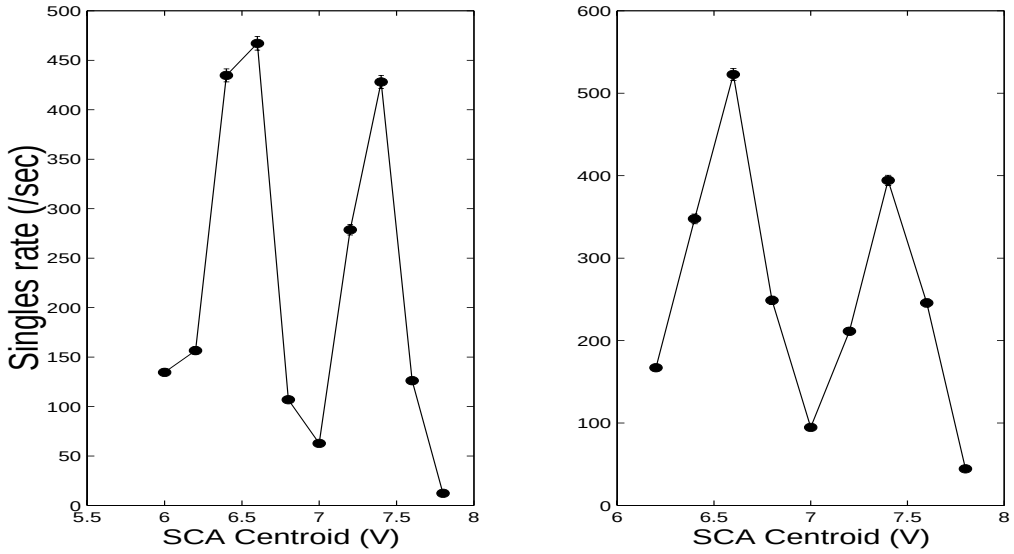


Figure 19.7: Single channel analyzer spectra for the two NaI detectors used in the angular correlation measurements, exposed to a ^{60}Co radioactive source. The two separate gamma ray energies are obvious in both detectors.

Fig. 19.7, for the two detectors each at 20 cm from the source. (For this data, the SCA is set to the "symmetric window" mode, with a 0.1 V window.) One detector seems to have slightly poorer resolution than the other, but these are fine for this job. Based on this data, the SCA's are switched to normal mode, and the lower and upper levels are set to 6.2 V (6.2 V) and 7.8 V (7.9 V) on detector one (two) respectively. The singles rates at these settings are 3.8 kHz for detector one and 4.0 kHz for detector two.

Now put the two detectors in coincidence using the 414A coincidence module. (It is actually a good idea to run even the singles measurements through the 414A, using the switches on the front panel to take out the appropriate detector. This way, no cables have to be changed and everything should be able to stay a lot more stable.) With the detectors at 20 cm from the source, you should get a coincidence rate of a few per second, assuming the coincidence window is large enough. Figure 19.6 suggests that 40 ns or so should be okay, but with a very small rate, we need to be extra careful of accidental coindicence background. Let's look at this a bit more carefully.

Accidental coincidences happen when two detectors are randomly firing at some rate, and there is a chance that two random pulses happen within the coincidence window. Consider the first detector, with rate R_1, as the one which "opens" the window. Then, the fraction of time that the window is open is $R_1\tau$ where τ is the width (in time) of the coincidence window. (We assume that this product is much less than one. If the rate is too high, the window is always open.) Then, the second detector, with rate R_2 comes along and will happen during the concidence time window with an accidental coincidence rate $R_A = R_2 \times (R_1\tau) = R_1 R_2 \tau$. If the two detectors are counting at pretty much the same rate $R = R_1 = R_2$, then we have $R_A = R^2\tau$.

Compare this to the *true* coincidence rate R_T. So long as the window accomodates the detectors, R_T will not depend on τ. It will, however, be proportional to the source activity, which is proportional to the rate R in the detectors. Therefore, the ratio of "accidentals" to "trues", R_A/R_T, will be proportional to $R\tau$, and we would like to keep this number as small as possible. Still, we need R to be relatively large so we can make our measurements in a reasonable amount of time, so keeping the coincidence window small is crucially important.

Let's look at the numbers for our case. As shown above, our detectors are counting at a rate $R \approx 4$ kHz. If $\tau = 100$ ns, then the accidental rate is $R_A = 1.6$/sec. This is a significant contribution to the observed coincidence rate of a few per second, and must be dealt with.

Figure 19.8 shows the coincidence rate for our setup as a function of resolving time. It is reminiscint of Fig. 19.6, but with a larger fraction of accidental to true coincidences. The "knee" rises rather slowly (it is hard to get a signal to turn on in a few ns), but by 80 ns it seems that we are accumulating all the true coincidence events. We will operate, therefore, with $\tau = 80$ ns.

With everything now set up, you can take angular correlation data. Being very careful not to disturb the setup between measurements, rotate the two detectors relative to each other and measure the coincidence rate. It is a very good idea to take a few measurements over again, after moving the detectors to one angle and then back again, to make sure you get numbers that are consistent to within error bars. You are trying to measure a small

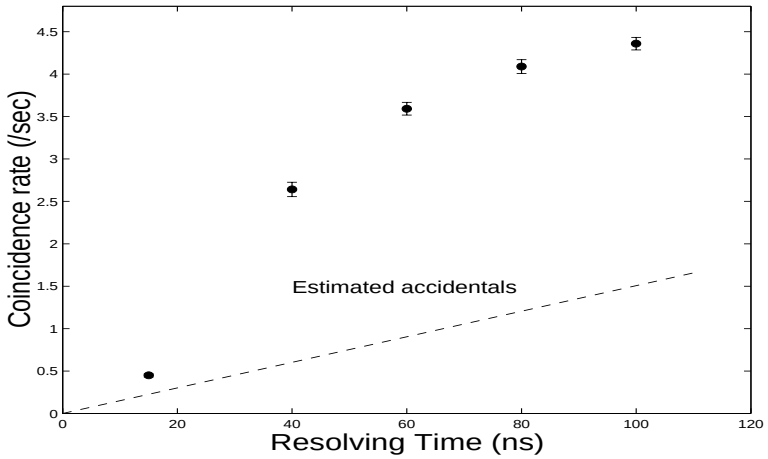


Figure 19.8: Coincidence rate as a function of resolving time for the $\gamma\gamma$ angular correlation apparatus, using the TRACERLAB ^{60}Co source. The estimated accidental rate, proportional to the resolving time, is indicated based on the measured singles rates in the two detectors. It is a significant fraction of the observed coincidences.

effect, about 16% difference between $\theta = 0^\circ$ and $\theta = 90^\circ$, so you should have error bars that are only a couple of percent. This requires a few thousand counts, so take data for about 1000 sec per point if the rate is a few per second. One such set of data is shown in Fig. 19.9. The data point at 90° (i.e. $\cos\theta = 1$) was taken twice to check consistency. *Note that the estimated accidental coincidence rate has been subtracted.* The curve drawn through the points is given by

$$N \times \left(1 + \frac{1}{8}\cos^2\theta + \frac{1}{24}\cos^4\theta\right)$$

where N is determined by minimizing the χ^2 with respect to the data points. The agreement is rather good, but the points at larger angles (smaller values of $\cos\theta$) seem to be below the curve by a small amount. You might consider writing a simple MATLAB program to vary the background level as well as the normalization to see what background is predicted, or to allow the coefficiencts of $\cos^2\theta$ and $\cos^4\theta$ to vary and compare them with the theory.

A more precise measurement of the accidental rates should be possible. You were likely able to get good coincidence timing simply by setting the

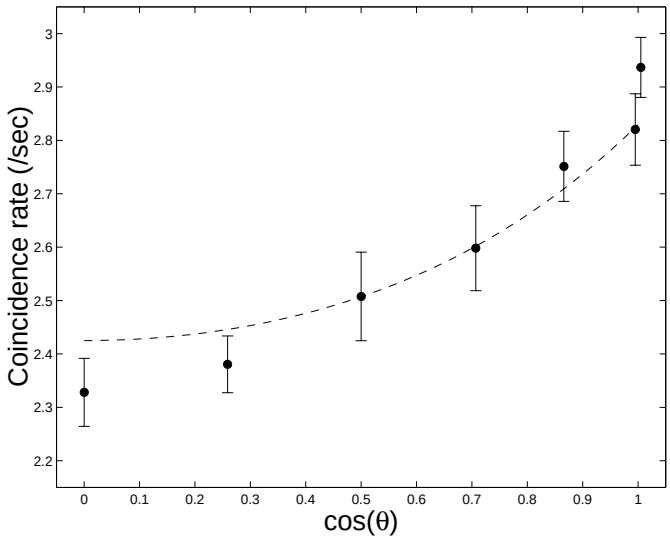


Figure 19.9: Accidental-subtracted rate for $\gamma\gamma$ coincidence events from a ^{60}Co source, as a function of the angle between the two detectors. The dotted line is the theoretical curve, normalized by the value that minimizes the χ^2.

two SCA's to the same delay time. (The two detectors are set up almost identically, and they are looking at the same energy gamma rays.) If you use the delay time switch on the front of one of the SCA's to set the pulses very much *out* of time with respect to each other, then all you will count is accidental background. In fact, you may want to experiment with the various SCA settings in order to optimize the shape of the resolving time dependence in Fig. 19.8.

Ch 20

Experiment 12: The Compton Effect

Why we do we believe that light behaves sometimes as a wave and sometimes as a particle? The answer, of course, is that experiments give us evidence of both types of behavior, and our picture of nature emerges that is consistent with experiment. Is there any experiment which is particularly compelling?

A particle is something of definite mass that can move with some velocity, and therefore can have momentum and (kinetic) energy. Consider an elastic collision of a moving particle with a stationary one. If you know the initial momentum of the incident particle, then measuring the angle through which it scatters tells you a lot of other things. In particular, conservation of momentum and energy tells you the energy of the scattered particle, and the recoil angle and energy of the target particle. What a fine way to demonstrate the particle nature of light to show that light behaves in exactly this way. This is called the Compton Effect. When it was discovered by A. H. Compton in 1922, it was the final and most convincing evidence that light is indeed "quantized".

The Compton Effect has long been a standard experiment in most undergraduate physics laboratories, and ours is modeled after them. It appears in some of the standard textbooks, such as

- *The Art of Experimental Physics*,
  Daryl W. Preson and Eric R. Dietz, John Wiley and Sons (1991)
  Experiment 19

- *Experiments in Modern Physics*,
  Adrian C. Melissinos, Academic Press (1966)
  Section 6.3

and there are several papers in the literature, such as

- *Verification of Compton Collision and Klein-Nishina Formulas - An Undergraduate Laboratory Experiment*,
  R. P. Singhal and A. J. Burns,
  American Journal of Physics **46**(1978)646

- *Compton Scattering Experiment*,
  Michael Stamatelatos, American Journal of Physics **40**(1972)1871

- *Compton Effect: Historical Background*,
  A. A. Bartlett, American Journal of Physics **32**(1964)120

- *Compton Effect: A Simple Laboratory Experiment*,
  A. A. Bartlett, American Journal of Physics **32**(1964)127

- *Compton Effect: An Experiment for the Advanced Laboratory*,
  A. A. Bartlett, American Journal of Physics **32**(1964)135

You will also find a lot of basic information on any introductory physics textbook that includes a discussion of modern physics. For example,

- *Introduction to the Structure of Matter*,
  John J. Brehm and William J. Mullin, John Wiley and Sons (1989),
  Chapter Two

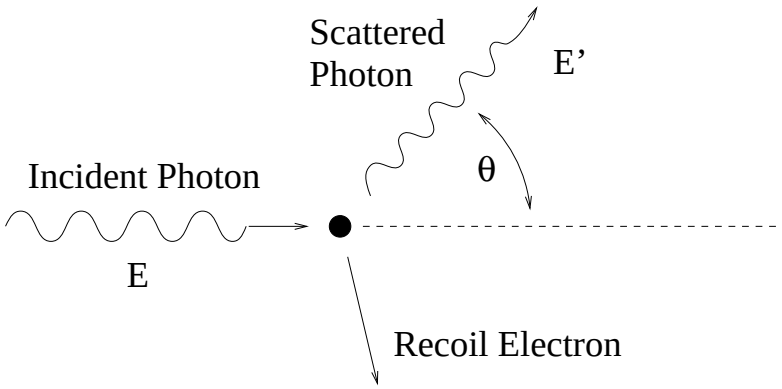


Figure 20.1: Kinematics of Compton scattering.

20.1 Scattering Light from Electrons

First, we will work out the kinematics of scattering a photon (that is, a particle with zero rest mass) from an electron. Then we will discuss the scattering probability, or cross section, for this reaction. We discuss the scattering cross section in the classical limit (Thompson scattering) and compare it to the quantum mechanical formula for Compton scattering.

20.1.1 Relativistic Kinematics

We consider the reaction

$$\gamma + e \rightarrow \gamma' + e' \tag{20.1}$$

for a photon of energy E and with the electron initially at rest. We want to calculate first the scattered photon energy E', when it scatters at an angle θ with respect to its incident direction. This is shown schematically in Fig. 20.1.

These kinematics are in fact all worked out in detail in Brehm and Mullin, in section 2.7. They use a traditional approach of solving the equations for conservation of momentum and energy to determine E' as a function of θ. Instead, I will show you a different way to do the calculation, using something called "four-vectors".

Four-vectors are simple extentions of the "three-vectors" you have been using since high school. They actually have a very profound importance which we won't go into here, but for now just think of them as a shorthand. If a particle has (three-)momentum $\vec{p} = (p_x, p_y, p_z)$ and total energy $E = K + mc^2$, then its four-vector momentum (or just "four-momentum" for short) is a combination of momentum and energy, namely

$$p = (E/c, p_x, p_y, p_z) = (E/c, \vec{p}) \tag{20.2}$$

A key to using four-vectors is the definition of the "dot product". It may look a little weird to you at first, but the dot product of two four vectors p_1 and p_2 is given by

$$\begin{aligned} p_1 \cdot p_2 &\equiv E_1 E_2/c^2 - p_{1x}p_{2x} - p_{1y}p_{2y} - p_{1z}p_{2z} \\ &= E_1 E_2/c^2 - \vec{p_1} \cdot \vec{p_2} \end{aligned} \tag{20.3}$$

The power of four-vectors, and the reason for the weird-looking dot product, starts to become clear when you consider the dot product of a four-vector with itself:

$$p^2 \equiv p \cdot p = E^2/c^2 - |\vec{p}|^2 = m^2c^2 \tag{20.4}$$

That is, the "square" of a four-vector is the square of the rest mass of the particle, after you've thrown in the appropriate factors of c.

So now let's return to reaction 20.1. Let k, p, k', and p_e be the four-momenta of the incident photon, target electron, scattered photon, and recoil electron respectively. We can write down the equations for the conservation of (total) energy and (each component of) momentum all just by writing

$$k + p = k' + p_e \tag{20.5}$$

Now rearrange this equation and square both sides:

$$\begin{aligned} p_e^2 &= (k + p - k')^2 \\ &= k^2 + p^2 + k'^2 + 2k \cdot p - 2k \cdot k' - 2p \cdot k' \end{aligned} \tag{20.6}$$

Now make use of Eq. 20.4. That is, $p_e^2 = p^2 = m^2c^2$ (where m is the mass of the electron), and $k^2 = k'^2 = 0$. We also know that $p = (mc, \vec{0})$ since the initial electron is at rest. We can then take Eq. 20.6 and write it as

$$\begin{aligned} 0 &= k \cdot p - k \cdot k' - p \cdot k' \\ &= Em - k \cdot k' - mE' \end{aligned} \tag{20.7}$$

Finally, we have $k \cdot k' = EE'/c^2 - \vec{k} \cdot \vec{k'} = [EE' - EE' \cos\theta]/c^2$ since $|\vec{k}| = E/c$ for the massless photon. This allows us to solve for E' in Eq. 20.7. We find that

$$E' = \frac{E}{1 + \frac{E}{mc^2}(1 - \cos\theta)} \tag{20.8}$$

This is what we were after. It tells the scattered photon energy E' in terms of the photon scattering angle θ.

Note that Eq. 20.8 says that for low photon energies, that is $E \ll mc^2$, $E' \approx E$. This is in fact the classical limit, that is, the result you expect from *classical* scattering of electromagnetic radiation from electrons. Let's discuss that now.

20.1.2 Classical and Quantum Mechanical Scattering

Scattering is a fundamental concept in physics. We encounter it now, as we discuss Compton scattering, but it shows up all over the place and you will certainly see it again sometime. Nevertheless, we will be using light-electron, or photon-electron, scattering to develop the principles.

As the electromagnetic wave passes by an atom, it makes the electrons oscillate. Oscillating electrons, since they are charged, generate their own electromagnetic radiation, which travels out from the atom in all directions. Thus, the electron "scattters" the incident wave. The frequency ν of this wave is the same as the frequency at which the electrons oscillate, which is the same as the frequency of the incoming wave. That is, the scattered wave has the same frequency and wavelength as the incident wave. In terms of photons, the energy of the scattered photon, $E' = h\nu$, is the same as the energy E of the incident photon. This is just what Eq. 20.8 predicts in the limit where the photon energy is much smaller than the rest energy mc^2 of the electron.

However, scattering is a lot more than just kinematics. In fact, the probability that a particular scattering event occurs, and the angles at which things prefer to scatter, tell us an enormous amount about the forces between the incident particle and the scatterer. To do this right, we have to first introduce the idea of the "differential cross section".

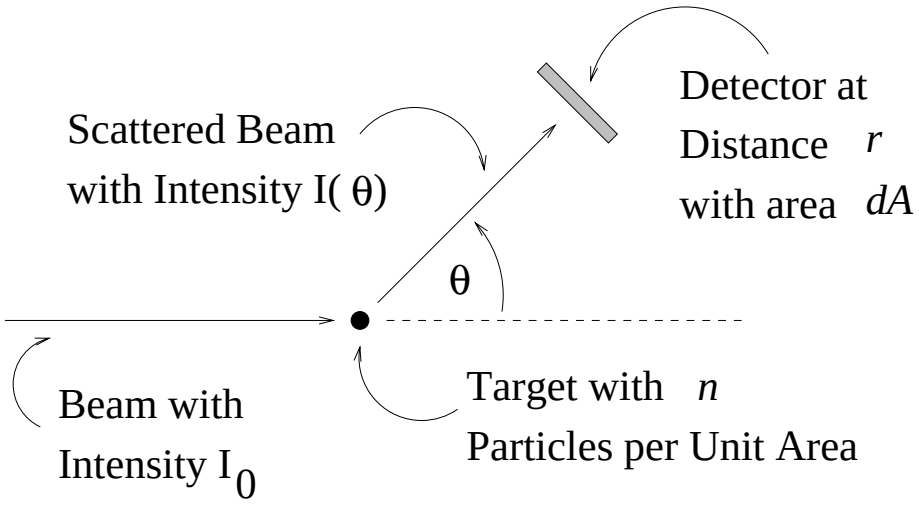


Figure 20.2: Schematic of a typical scattering measurement.

Imagine some kind of beam incident on a target which may or may not cause the beam to scatter. Let I_0 be the intensity of the beam, either measured as area per unit area per second (as you would for a wave) or as particles per unit area per second. If the target scatters the beam, you would detect some intensity emanating from the target at some distance r away and at some angle θ with respect to the incident beam. This is shown in Fig. 20.2. There is some probability that the target scatters intensity out of the beam, and you measure a scattered intensity $I(\theta)$. You would certainly expect that the scattered intensity is proportional to the incident intensity, and that it decreases like $1/r^2$ as you move farther away. If you let $n = t\rho$ be the thickness of the target in particles per unit area (as viewed by the beam), where t is the linear thickness and ρ is the density of particles per unit volume, then you also expect $I(\theta)$ to be proportional to n. Finally, if the detector has some small area dA, then the scattered intensity you measure would also be proportional to dA. That is $I(\theta) \propto I_0 \times n \times dA/r^2$.

The proportionality constant only concerns the physics of the interaction between the beam and the particles that make up the target. We've taken out the dependence on everything else. We call the proportionality constant the *differential cross section*. We use the words "cross section" because it has the dimensions of an area, *not* because it has anything to do with a physical

area. We write the differential cross section as $d\sigma/d\Omega$,

$$I(\theta) = \frac{d\sigma}{d\Omega} \times I_0 \times n \times \frac{dA}{r^2} \tag{20.9}$$

and it may be a function of the energy E of the beam, the scattering angle θ, and any number of other things depending on the type of beam and target particle.

The notation we use, i.e. $d\sigma/d\Omega$, actually means something. The factor dA/r^2 in Eq. 20.9 is called the *solid angle* $d\Omega = dA/r^2$. (See the discussion in Sec. 17.2.1.) If you were to integrate the cross section over all solid angle, you would calculate the *total cross section* σ, that is

$$\sigma = \int \frac{d\sigma}{d\Omega} d\Omega$$

which represents the probability of any kind of scattering at all. (Recall the discussion in Expt. 2.)

Okay, now that we've got the machinery, we can talk about the differential cross section for scattering electromagnetic radiation, or photons, from electrons. Before we deal with photon scattering by electrons, we will review classical scattering of electromagnetic radiation by electrons. We will only worry about photons with energies significantly larger than the binding energies of electrons in atoms, that is, wavelengths much shorter than a few hundred nm. This process is called *Thomson scattering* and is covered in most upper level undergraduate texts on electricity and magnetism[1] The differential cross section for Thomson scattering is calculated from the radiation pattern made by the oscillating electron which was set in motion by the incident wave. You find that

$$\frac{d\sigma}{d\Omega} = r_0^2 \frac{1 + \cos^2\theta}{2} \qquad \text{Thomson Scattering} \tag{20.10}$$

where $r_0 = e^2/4\pi\varepsilon_0 mc^2 = 2.82 \times 10^{-13}$ cm is called the "classical electron radius".

[1]See, for example, Reitz, Milford, and Christy, Foundations of Electromagnetic Theory, 4th Edition (1992), Section 20-5.

The differential cross section for Compton scattering can be calculated using a combination of relativity and quantum mechanics. The result is

$$\begin{aligned} \frac{d\sigma}{d\Omega} &= r_0^2 \frac{1+\cos^2\theta}{2} \frac{1}{\left[1+\gamma\left(1-\cos\theta\right)\right]^2} \\ &\times \left[1+\frac{\gamma^2\left(1-\cos\theta\right)^2}{\left(1+\cos^2\theta\right)\left[1+\gamma\left(1-\cos\theta\right)\right]}\right] \end{aligned} \qquad (20.11)$$

Compton Scattering

where we use the shorthand $\gamma = E/mc^2$. This is called the Klein-Nishina formula after the physicists who first calculated it. Note that the Klein-Nishina formula reduces to the equation for Thomson scattering in the limit where $\gamma \to 0$, i.e. as the energy of the incident photon gets very small.

You will have the opportunity to measure the differential cross section in this experiment. As you will see, however, it is very difficult to get the systematic uncertainty to a level that allows you to actually determine a value for the cross section to a decent level of precision. Instead, you can measure the angular distribution of the cross section, after making some relatively straightforward corrections. Note a striking difference in the cross sections as a function of angle, between Thomson and Compton scattering. Thomson scattering is symmetric about $\theta = 90°$, that is, it predicts that the ratio of the differential cross sections at $\theta = 90° + \alpha$ and $\theta = 90° - \alpha$ should be unity. This is not the same as for Compton scattering, where this ratio is less than one, if α is positive.

So, Compton scattering, the scattering of photons from electrons, departs clearly from the classical case as the energy of the photon increases and becomes a significant fraction of the electron rest energy. This departure shows up *both* in the kinematics (by a shift in the energy of the scattered radiation) and by a change in the differential cross section. You will be able to measure both of these in this experiment.

20.2 Measurements

The basic Compton scattering setup is shown in Fig. 20.3. There are three essential components to a scattering experiment, namely the incident beam,

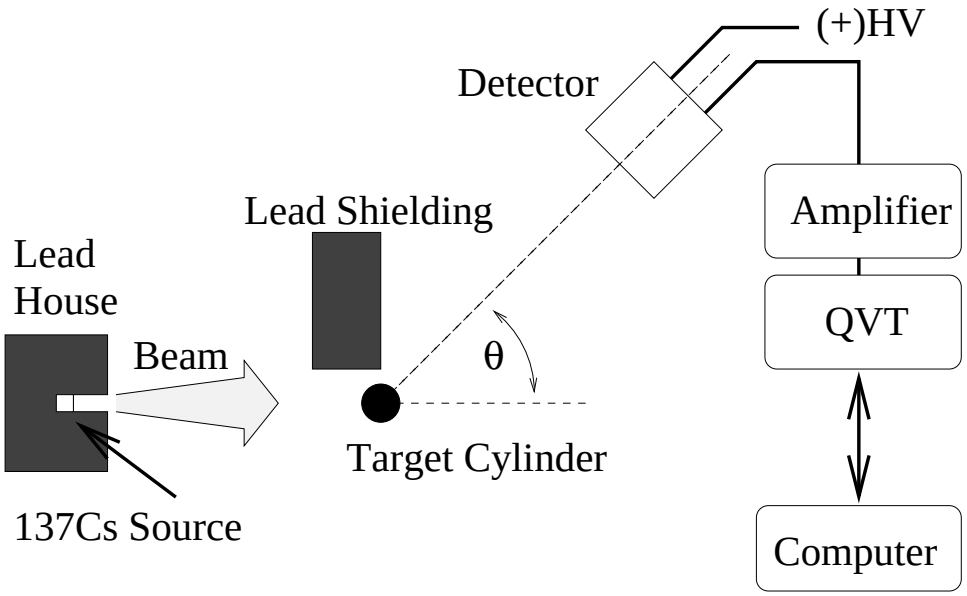


Figure 20.3: Setup used for Compton scattering singles measurements.

the target, and a detector for the radiation or particles scattered by the target. In our setup, the beam is provided by a ^{137}Cs γ-ray source giving monochromatic photons with $E = 0.662$ MeV. (See Expt. 10.) The detector is a 3 in. diameter by 3 in. long NaI(Tl) detector. A few different targets will be used, each of which is a long metal rod. You may consider each target as a collection of electrons, but when you determine the cross section the diameter and composition of the target will make an important difference in the analysis.

The ^{137}Cs source is from a commercial vendor, and they calibrated the activity level as 9.14 mCi on January 1, 1993. Their quoted error on this activity is $\pm 3.2\%$ at the 99% confidence level, i.e. $\pm 2.58\sigma$. To calculate the photon rate produced this source when you make your measurements, you need to know that only 94% of ^{137}Cs decays produce 0.662 MeV photons, and that the half life of ^{137}Cs is 30.0 years. The source capsule is captured inside a brass plug which in turn is inside the 4" $\times$ 4" $\times$ 8" lead house. The brass plug has a conical hole init which allows a small fraction of the emitted photons to emerge unimpeded. The angular profile of the beam[2] is shown in

[2]Data taken by Rick Hullinger, Class of 1996

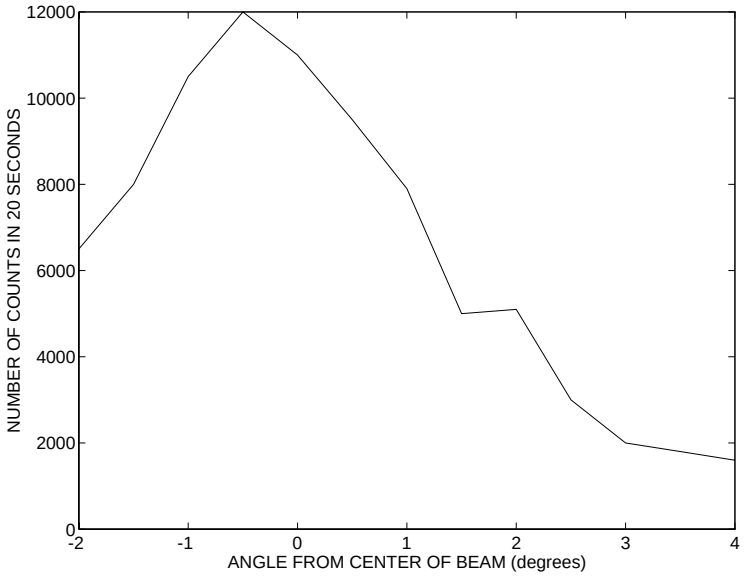


Figure 20.4: Angular profile of the beam from the ^{137}Cs source. Data was taken at a distance of 3.2 m with a $2'' \times 2''$ NaI detector.

Fig. 20.4.

You might want to review the discussion on radiation safety in Sec. 17.1.4. **THIS IS A VERY HIGH INTENSITY γ-RAY SOURCE. DO NOT ATTEMPT TO REMOVE THE BRASS PLUG OR THE SOURCE CAPSULE FROM THE LEAD HOUSE.** The front brick of the house has a handle making it easy to remove when data taking begins.

The NaI(Tl) detector is powered with a positive HV DC power supply, which should be set to $\sim 1300V$. It is a good idea to turn it on and let the photomultiplier tube base warm up for a half hour or so before you take actual measurements.

The detector signal is amplified using a standard γ-ray spectroscopy pulse amplifier, and the pulse height is processed with a multichannel analyzer. You will use the Lecroy Model 3001 qVt Multichannel Analyzer. The qVt is connected to a Model 3031 controller, which in turn communicates with the IBM/PC through a Model 1691A general purpose PC interface. The "V" input option should be used to analyze the amplifier output (i.e. you

are analyzing the Voltage level of the amplified pulse), and you should use the "internal" gate option. The BNC connector labeled "gate view" can be used to compare the timing of the internal gate with the amplified signal, using a dual trace oscilloscope. An XY oscilloscope should be connected to the horizontal and vertical outputs of the qVt, and this will serve as a live display.

You can control the qVt using the program GRAPHWIN which runs on the PC, including reading spectra from the qVt and manipulating them. The program is menu driven, and its use should be pretty much self explanatory. In any case, more detailed instructions are available if you need them.

20.2.1 Procedure

The first thing you need to do is calibrate the energy scale of the qVt spectra. You can do this using some or all of the standard energy calibration sources provided in the wooden box, or with some of the other sources available in the laboratory. If you arrange the gain of the amplifier so that the 0.662 MeV peak from ^{137}Cs is near the upper end of the range, then the sources you would find most useful are (low intensity) ^{137}Cs (662 KeV), annihilation photons from ^{22}Na β^+-decay (511 KeV), and ^{133}Ba (356 KeV, 302 KeV, and 80 KeV). If you want to have a larger energy range, reduce the gain of the amplifier, and you can also calibrate with the ^{60}Co source (1.17 MeV and 1.33 MeV) as well as with the 1.28 MeV photons from ^{22}Na.

Place each of the calibration sources near the point where the photon beam from the lead house would intersect the target. Then take a spectrum for some period of time and make sure you can clearly pick out the peaks corresponding to the calibration γ-rays. Read the spectra and store them. You should probably use GRAPHWIN to write down the peak positions as you take the data, and plot them versus energy in your log book, to determine the energy calibration. That is, get at least a rough idea of the constants A and B where

$$\text{Energy} = A \times \text{Channel} + B$$

so that you can check your Compton scattering data as it is taken.

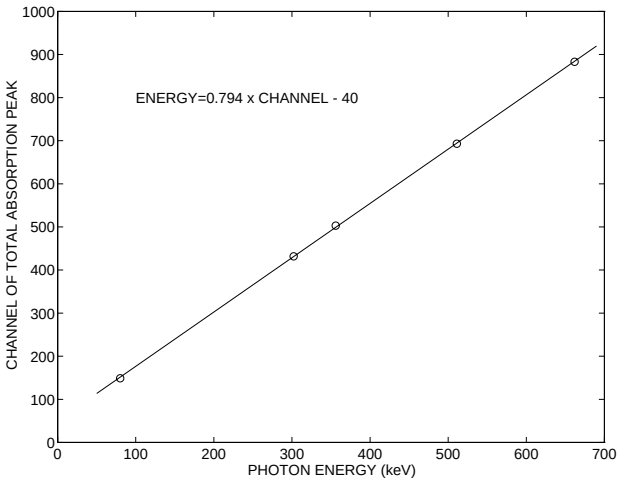


Figure 20.5: Energy calibration of the NaI detector using various radioactive sources.

An example[3] of calibration data, fit to a straight line, is shown in Fig. 20.5. If you are going to use your Compton scattering data to determine the differential cross section (Sec. 20.3.2), then you should also record the intensity calibration of each of the sources, as well as the amount of time you took data with them.

For any given target or detector angle, you must take two sets of data to get clean data on Compton scattering. This is because there is a good deal of background in the NaI(Tl) detector, i.e. signals coming from γ-rays which are scattered from the walls or perhaps leak out of the lead shield and have nothing to do with your scattering target. To subtract this background, take a set of data (for a fixed amount of time) with the target in place. Then remove the target and take another set of data for the same amount of time. GRAPHWIN allows you to subtract the two spectra, and the result should be a clear peak from Compton scattering. It is probably best, however, to save the individual spectra and do the subtraction in MATLAB.

Figure 20.6 plots data taken with the target in and target out. The Compton scattering peak is clear, and the two spectra fall on top of each other for energies greater than the peak. Below the peak, there is an excess

[3]Data taken by Ed Barnat, Class of 1996.

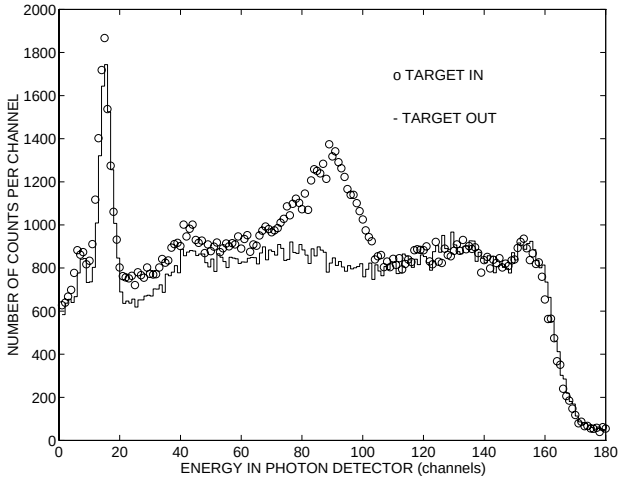


Figure 20.6: Sample data taken with the target in and the target out.

of counts for target in, just as you would expect because of the detector response to photons.

It is probably best to start at an angle near $\theta = 90^\circ$ because the background turns out to be relatively small there. Then change the angle in either direction, and keep track of the background rate. You might want to arrange various lead bricks to try and minimize the background, but leave the source house intact.

In principle, the answer you get should not depend on which target you select. However, you will get a better count rate (relative to the background) if the target presents a greater number of electrons to the photon beam. You should take a couple of angles with more than one target, and consider the results later. It should not be necessary to take data at all angles with more than one target.

20.2.2 Analysis

There are lots of things you can do with this data. We'll describe some of them here, starting from the more straightforward ones.

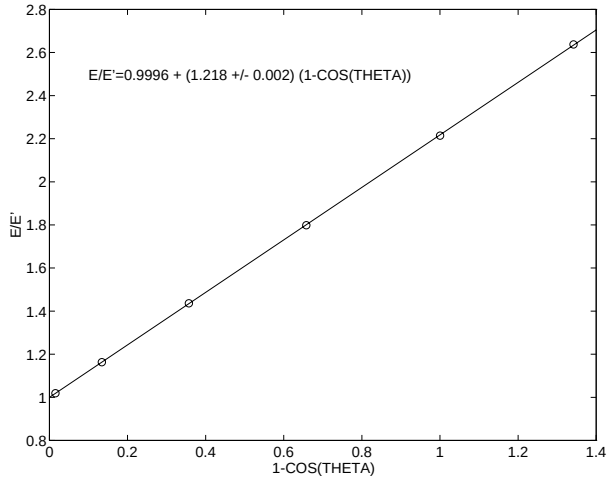


Figure 20.7: Test of Compton scattering kinematics. The quantity E/E' is plotted versus $1 - \cos\theta$, and the result is fit to a straight line. The intercept is 0.9996 and the slope, with an uncertainty given only by the straight line fit, is 1.218±0.002.

Verification of Compton Scattering Kinematics

The first thing to do is verify Eq. 20.8. The peak position of the Compton scattered γ-ray should be evident from your subtracted spectra. Use your energy calibration to turn this peak position into a γ-ray energy E'. Plot E' as a function of θ, and graph the result of Eq. 20.8 on top. Use a reasonable estimate for the uncertainty in the peak position so you can plot the measured values of E' with an error bar.

You may also choose to plot your data as $E/E' = 0.662$ MeV$/E'$ versus $(1 - \cos\theta)$. The result should look like a straight line. You can fit this line to get the best value for the slope and interecept, and compare these fitted values and uncertainties to the values you expect. An example[4] is shown in Fig. 20.7. The data are fit to a straight line, and the (random) uncertainties are determined using the formulas in Sec. 9.2.1. The result is

$$\frac{E}{mc^2} = 1.218 \pm 0.002$$

[4]Data taken by Ed Barnat, Class of 1996

Is this what you expect?

Consider the possible reasons that your may not agree within uncertainties with the expected values. You might be able to identify the 0.662 MeV peak from ^{137}Cs source in your spectra, from $\gamma-$rays leaking out of the lead house. Does the peak position stay constant over time? Can you use this peak position to make a correction to your data?

Determining the Angular Distribution

You should be able to test whether or not the angular distribution is more consistent with Compton scattering, or with Thomson scattering. That is, see if the angular distribution is consistent with a higher rate at angles less than 90° relative to angles greater than 90°, or instead symmetric about 90°. You have to be careful of corrections introduced because of the detector efficiency, as we discuss below.

Don't forget that the scattering "rate" is the number of counts under the full absorption peak with background subtracted, divided by the running time.

The simplest way to check the angular distribution is to tabulate the ratio of rates for angles around 90°, that is $R(90^\circ+\alpha)/R(90^\circ-\alpha)$ for various values of α. A different way is to plot the scattering rate as a function of angle. An example is shown in Fig. 20.8. (You may want to normalize your plot to the rate at one particular angle, and $\theta = 90^\circ$ is a good choice.) This count rate, or intensity, should change with angle because of the θ dependence of the cross section $d\sigma/d\Omega$. Does the result seem to be consistent with Eq. 20.10 or with Eq. 20.12?

Even though Eq. 20.9 contains no explicit angular dependence (except for the cross section), you have to be careful of corrections. One thing you should check is that this angular distribution is independent of the type of target you used. Use the angles at which there is data from more than one target to check the relative dependence on rate. In Fig. 20.8, how important is it to include the energy dependent efficiencies?

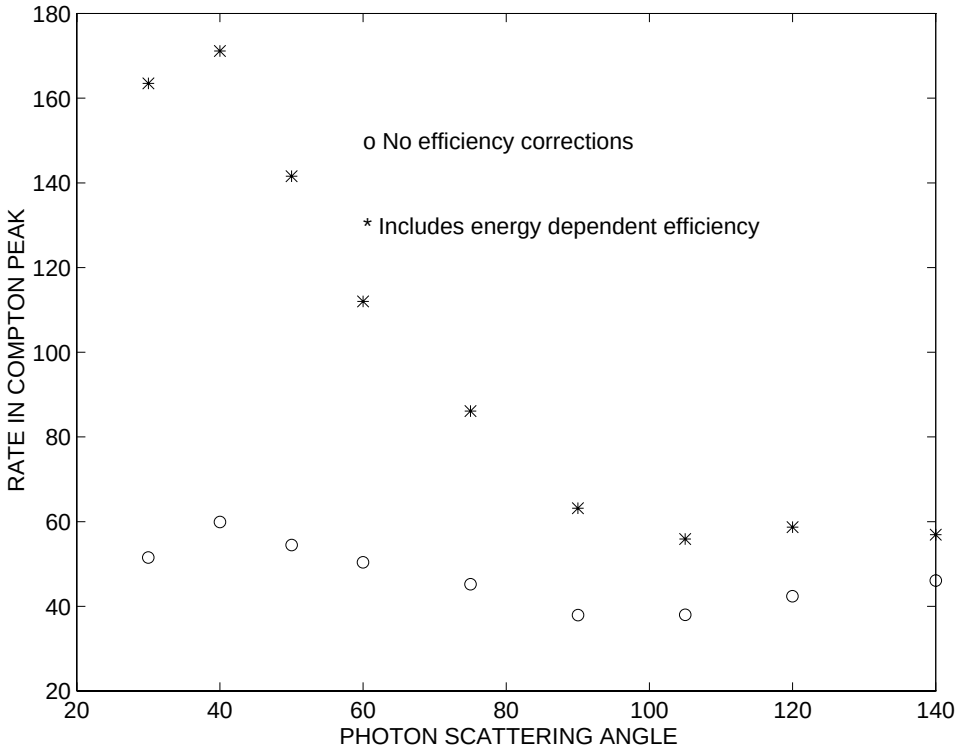


Figure 20.8: The rates under the Compton scattering peak after background subtraction, as a function of scattering angle θ. Energy dependent efficiencies include the intrinsic and photopeak efficiencies, both of which increase as the photon energy decreases. For this analysis, we used efficiencies as plotted in the paper by Singhal and Burns.

The biggest correction you likely have to make is from the detector efficiency. The efficiency is a function of angle because the Compton scattered energy changes as a function of angle. It is important to take into account both the intrinsic and photopeak efficiencies. These are tabulated and shown in Tab. 17.3 and Fig. 17.8. Estimate those efficiencies, and the uncertainty you get from reading the table or graph, and recalculate the angular distribution. Again, it might be wise to normalize to 90°. Make another plot including these corrections. Reconsider whether or not the result agrees with Thomson or Compton scattering, or both, or neither.

20.3 Advanced Topics

You can actually do quite a lot more with this detector setup. You can discuss some of the options with the teaching assistants or with me, but following are a couple of suggestions.

20.3.1 Recoil Electron Detection

We haven't really talked about another important piece of the Compton scattering picture. The photon scatters off of the electron and we detect the photon, but what about the electron? It carries off a considerable amount of energy, namely $E_e = E - E'$. We should be able to detect it.

Our setup can be extended to include these "coincidence" measurements, as opposed to the "singles" measurements where all we detect for any one event is the scattered photon itself. The extended setup used to measure Compton scattering is shown in Fig. 20.9. In this case, the "target" is actually a plastic scintillator detector. This detector consists of a 1/8" thick piece of plastic, mounted on a photomultiplier tube. The photomultiplier operates at *negative* high voltage, and you should set it to around −2000 V. This scintillator is thick enough so that the electrons which recoil in Compton scattering with 662 keV photons will stop in the detector, assuming they are produced not too far from the rear surface of the detector. (You should confirm this by estimating the range of the electrons involved, as a function

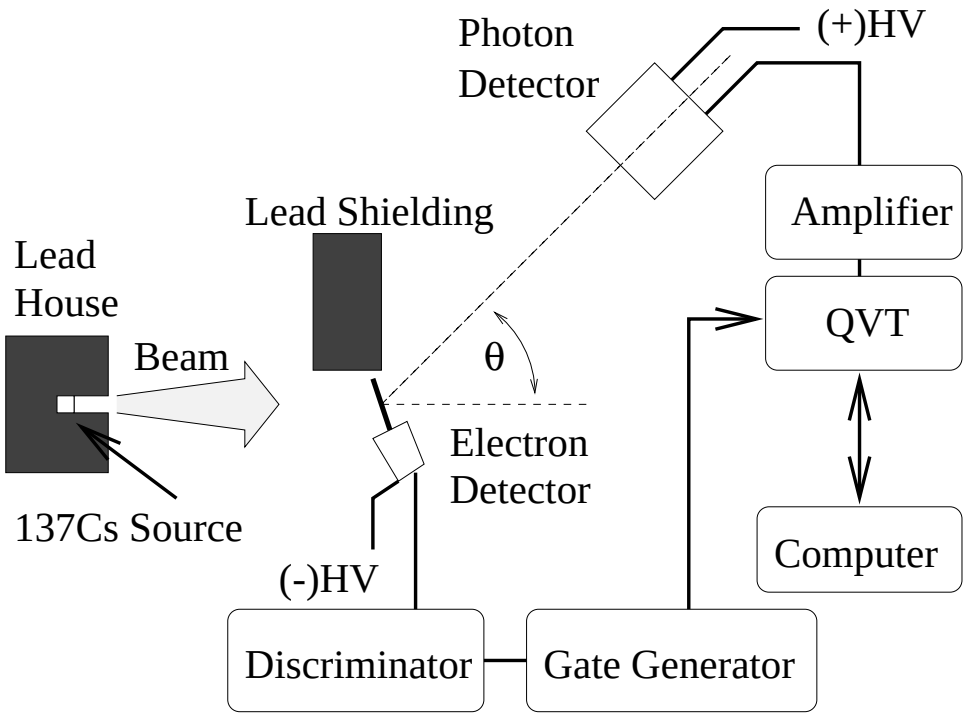


Figure 20.9: Setup used for Compton scattering coincidence measurements, where the scattered photon is detected in coincidence with the recoil electron.

of θ, using the Bethe-Bloch formula, Eq. 17.1).

The simplest way to observe the recoil electron is to demonstrate a coincidence between the plastic scintillator and NaI(Tl) detectors. You can do this by taking the output of the electron detector, passing it through a discriminator and gate generator, and using the output of the gate generator to control the qVt. *Make sure that you switch to "external gate" mode when running this way.* See your teaching assistant or me for details of the setup. You might also consult the paper by Stamatelatos.

First try to detect the Compton scattered γ-ray signal as before, but using the plastic scintillator as a target. Do not turn the plastic detector on, and leave the qVt mode in "internal gate". This makes a very thin target, so the peak will not be as clear as it was with the aluminum and brass cylinders, but it should still be visible if you count for a longer time and subtract background by removing the scintillator.

Next, set it up with the plastic scintillator turned on and the qVt in the "external gate" mode. This means that you only will observe γ-rays that are in coincidence with signals in the plastic scintillator detector. Most of the plastic detector's signals will come from photons which do not scatter into the NaI(Tl) detector, so there will be a lot of points at "zero" in the qVt spectrum. However, the Compton scattered γ-ray peak should appear clearly above the noise. Is it in the same position as it was before? Try this at a few different angles.

You can also demonstrate the change in the electron recoil energy, as you change the angle of the scattered photon. Measure the coincidence rate as a function of the discriminator threshold, for a few specific angles. How do you expect the electron's energy to change with angle? What does that imply for the threshold value at which the coincidence rate starts to fall off? Can you think of other ways to measure the electron energy?

20.3.2 Extracting the Differential Cross Section

It is possible, in principle, to determine the differential cross section absolutely using this setup. It is important to know the source activity and

effective target size to do this correctly, as well as to include the various efficiency corrections. A rather detailed discussion of this procedure is given in Melissinos, and in the paper by Singhal and Burns.

Because your ultimate experimental uncertainty will be dominated by the systematic uncertainties in the detection efficiencies, there is no point in reducing your relative background by using the plastic scintillator detector as a target. This will only give you a more difficult problem (the plastic scintillator efficiency must also be included), and the statistical (random) uncertainty due to the background subtraction with an aluminum target will make a negligible contribution.

Borrowing notation from Singhal and Burns, you determine the cross section experimentally by turning Eq. 20.9 around and including the various efficiency factors, thus

$$\frac{d\sigma}{d\Omega} = \frac{Rate(\theta)}{\Phi_\gamma \times N_e} \times \frac{1}{\varepsilon(E_\gamma) \times P(E_\gamma) \times \eta} \tag{20.12}$$

where *Rate* is the net number of counts per second, after subtracting the background; Φ_γ is the photon flux, in photons per cm^2 per sec incident on the target, and N_e is the number of electrons viewed by both the beam and the NaI(Tl) detector. You will have to estimate the geometry of the photon beam (which is actually a cone whose half-angle is determined by Fig. 20.4) and how much of the target illuminated by that cone that is seen by the detector.

The factors $\varepsilon(E_\gamma)$ and $P(E_\gamma)$ are the intrinsic and photopeak efficiencies, respectively, for photons of energy $E_\gamma = E'$. (See Sec. 17.2.3.) The solid angle factor $d\Omega$ is contained in the definition of intrinsic efficiency. The factor η, which is actually a rather weak function of E_γ, takes into account the fact that the photon beam (and the scattered photons) are attenuated by the target, and so the beam intensity is reduced. One finds $\eta \approx 0.8$ using a sophisticated numerical calculation (see Singhal and Burns), but you have the opportunity to make a rough estimate of the effect and confirm this number.

Photons get absorbed both on their way into and out of the target, and the amount of absorption depends on the incident energy and scattered energy

(and so the scattering angle). It also depends on the exact place where the incident photon scattered, because this determines the thickness of material through which the incident and scattered photons travel. We will sweep most of this under the rug with a simple model, namely that η is just given by the exponential attenuation formula for photons. (See Sec. 17.1.2.) That is,

$$\eta = e^{-\mu x} \tag{20.13}$$

Of course, the "true" values for both μ and x are different for every scattering event, but like I said, this is a simple model. Let μ be a number representative of the photons involved, perhaps $\mu^{-1} = (15\ \mathrm{gm/cm^2})/\rho$ where ρ is the density of the target. You can then use your data to determine a representative value for x, by measuring the scattered photon rate with, say, brass and aluminum targets with the same dimensions. From Eq. 20.12,

$$\begin{aligned} \frac{\text{Rate with Brass Target}}{\text{Rate with Aluminum Target}} &= \frac{N_e^{\mathrm{Cu}}}{N_e^{\mathrm{Al}}} \times \frac{\eta^{\mathrm{Cu}}}{\eta^{\mathrm{Al}}} \\ &= \frac{N_e^{\mathrm{Cu}}}{N_e^{\mathrm{Al}}} \times e^{-x/(15\ \mathrm{gm/cm^2})\left[\rho^{\mathrm{Cu}} - \rho^{\mathrm{Al}}\right]} \end{aligned}$$

(It is safe to assume that brass is the same as copper for these measurements.)

Your value for x should be about the same as the target thickness. Is it? Use this value of x to get a model value for η in Eq. 20.13. Does it agree with the value determined by Singhal and Burns? Do you get the same value for x and η using data at different scattering angles? If you could make another target out of a different material to further test your model, what materials might you pick?

Appendix A

Principles of Quantum Physics

Most of the experiments done in this course involve quantum mechanical phenomena. You will learn a lot about these phenomena and about quantum mechanics in general from doing the experiments. There are, however, a few general points which you will always come up against, and I've tried to collect them here.

Quantum mechanics implies that nature is "quantized", based on its postulates. All this means is that matter and energy cannot existed with any arbitrary value, but it instead must take on discrete values of one sort or another.

Don't try to figure out "why" quantum mechanics is the right way to describe the world. Nobody knows why. The principles of quantum mechanics just seem to work, so we believe them, or at least work with them. It makes no more sense to understand why quantum mechanics works than to wonder why Newton's laws of motion work.

There have been some fine experiments which looked for violations of basic quantum principles, but there had never been any substantiated that quantum mechanics is wrong.

A.1 Photons

Electromagnetic radiation, including everything from radio waves to light to gamma radiation, comes in packets of energy called *photons*. The energy of a photon is $E = h\nu$ where $\nu = c/\lambda$ and λ are the frequency and wavelength of the electromagnetic radiation. (Remember that Maxwell's equations insist that $\nu\lambda = c$, at least in a vacuum.) The fundamental constant h is called Planck's constant, and we can only determine it from experiment.

Factors of 2π are ubiquitous in physics because we are always integrating over a circle somewhere. For this reason, we have the definitions $\hbar = h/2\pi$, $\omega = 2\pi\nu$, and $k = 2\pi/\lambda$, and you might see the energy of a photon written as $E = \hbar\omega$ or $E = \hbar kc$.

Of course, a photon has momentum as well as energy. It has no rest mass, however, so according to special relativity we have

$$E = \sqrt{p^2c^2 + m^2c^4} = pc$$

for a photon. The photon's momentum can be written as $p = \hbar k = h/\lambda$.

It is tempting to forget about all those factors of $\hbar$ and c that show up in quantum physics. I have tried to be consistent and keep them throughout this book, but most people on the "outside" don't bother. Most people do the conversions when they need it by remembering that $c \approx 3 \times 10^8$ m/sec ≈ 1 ft/ns and that $\hbar c \approx 200$ MeV fm.

Of course, the reason we believe that light is made of photons is because experiments strongly suggest it. Two examples of groundbreaking experiments are blackbody, or cavity, radiation (Sec. 11.1.1) and Compton Scattering (Experiment 12).

A.2 Wavelength of a Particle

Just as light can behave as a particle, i.e. the photon, particles can also behave as waves. The wavelength of the particle was postulated by DeBroglie

to be completely analagous to the photon, that is $\lambda = h/p$. Of course, however, the relationship $\nu\lambda = c$ *is not* valid for particles because the waves are not governed by Maxwell's equations. Instead, there is a different wave equation that describes the motion of particles.

Before getting too technical, however, you can already see some important consequences of DeBroglie's hypothesis. If a particle is a wave with a definite wavelength, then if it moves in some confined area it is important to make sure that the wave doesn't interfere with itself and cancel away the particle's existence! (How can there be a particle if there is no wave, i.e., if the wave has no amplitude?) This leads, for example, to the Bohr Quantization Condition for the energy of a hydrogen atom. (Sec. 12.1.)

A simpler example is just a particle confined to move in one dimension, but constrained to be inside a "box" with walls at $x = \pm a$. We take this to mean that the wave corresponding to the particle must have zero amplitude outside the box. Unless the waves corresponding to the particle have nodes at $x = \pm a$, i.e. they are standing waves, then the continued reflections at the walls will destructively interfere the particle into oblivion. This means that an integral number of half-wavelengths must fit in the box. Therefore the allowed wavelengths must satisfy the condition $(2a)/(\lambda/2) = n$ or $\lambda = 4a/n$ where n is an integer. The (nonrelativistic) kinetic energy for this particle in a box can therefore take on the values

$$\begin{aligned} E_k &= \frac{p^2}{2m} = \frac{h^2}{2m}\frac{1}{\lambda^2} \\ &= \frac{h^2}{32ma^2}n^2 \end{aligned} \tag{A.1}$$

The allowed energies of the particle are *quantized*. They must have the discrete values $E_k = E_0$, $4E_0$, $9E_0$,/ldots, where $E_0 = h^2/32ma^2$. These energy levels are drawn in Fig. A.1. A particle is said to be in a definite "state" if its energy corresponds exactly to one of these energy levels. Figure A.1 also plots the wave function $\psi(x)$ which corresponds to the particle in that particular energy state.

This kind of energy quantization happens whenever we restrict the motion of a particle based on some kine of potential well. In particular, the well does not have to have infinitely high walls. In this case, the energy will be

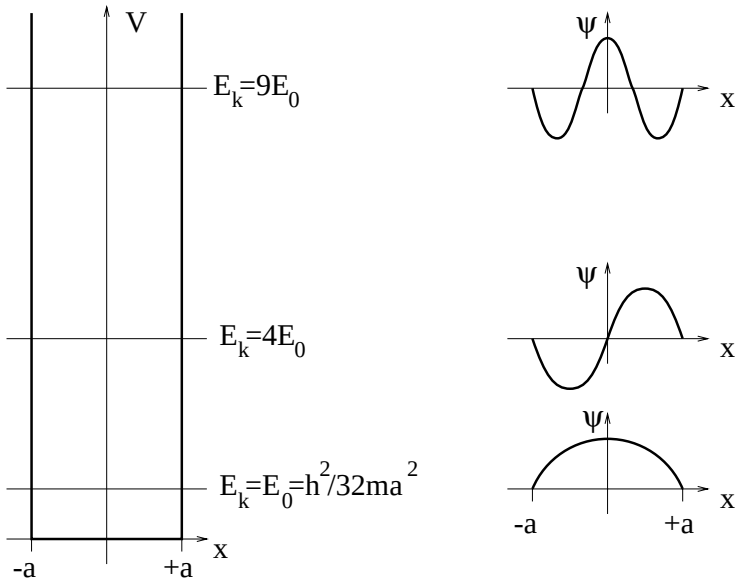


Figure A.1: Energy levels for a particle confined in a one-dimensional box with infinitely high walls.

quantized only if it lies below the highest value of the potential in the well. These are called *bound states*.

Schrödinger formulated a wave equation that you can use to determine the states and energy levels of any particle moving in any potential. It is based on conservation of energy, and although it is simple to write down, it can be hard to solve in practice. We discuss this in a bit more detail with regard to the Ramsauer Effect (Experiment 2), but for more information you need to go to a textbook on Quantum Mechanics.

A.3 Transitions between Bound States

Okay, so we have two manifestations of quantum mechanics. One is that light (i.e. electromagnetic radiation) is bundled up into photons of discrete energy. The other is that (massive) particles have wavelengths, and this leads to their having discrete energies if confined in some kind of potential.

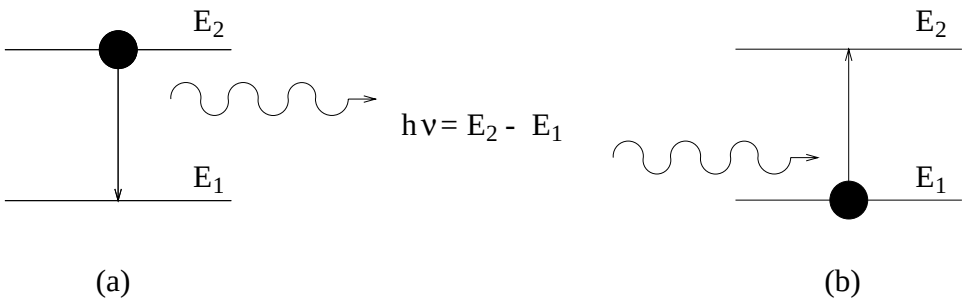


Figure A.2: Transitions between energy levels.

These two things are connected. If a particle is in a specific state with a specific energy, it can jump to another state with a different energy if it makes up the energy difference with a photon. The two cases are shown schematically in Fig. A.2. In case (a), a particle is initially in an upper energy state with energy E_2, when it spontaneously decides to drop down to a lower state with energy E_1. When this happens, a photon of energy $E = E_2 - E_1$ is emitted. Something like this has to happen, of course, so that energy is conserved. This is the principle used to produce the optical photons detected in Experiment 6.

Case (b) shows the opposite. Here the particle is initially in the lower energy state E_1 when a photon comes along and knocks it up to the state E_2. *This can only happen if the photon has precisely the right energy, namely* $E = E_2 - E_1$. One example of this is the precise resonance condition observed in Nuclear Magnetic Resonance (Experiment 9). In this case, the photons (from the radio frequency transition coils) induce transitions from E_2 to E_1 and back again.

In principle, transitions can occur between any two states, with the release or absorption of the right amount of energy. Sometimes, though, this energy cannot be in the form of photons, and some other kind of interaction (besides electromagnetism) must be invoked. One example is β-decay, studied in Experiment 10, where the transition between two different nuclei occurs and the energy release is *shared* between an electron (or positron) and a neutrino.

Appendix B

Principles of Statistical Mechanics

The physical interpretation of heat and temperature is the basis of Statistical Mechanics. It is hard to appreciate many of the experiments in this course without at least a cursory understanding of these principles. Some of the key ideas are collected here.

If you've studied Thermodynamics without a good connection to Statistical Mechanics, I suggest you put aside what you've learned about "temperature" and "heat" so far. Thermodynamics takes these quantities and makes them rather mysterious. I will approach them from the simple point of view of classical mechanics.

B.1 The Ideal Gas

Consider a collection of atomic or molecular sized particles which move around pretty freely inside some volume. Assume there is a very large number of such particles. We also assume that although the collide with each other frequently, all the collisions are elastic. That is, particles never stick to each other or excite each other in any way.

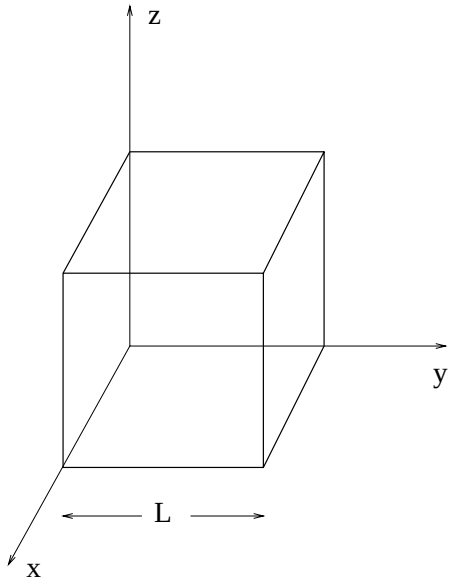


Figure B.1: A cubical box of side L for calculating the ideal gas law.

This collection of particles is obviously a *gas*. In fact, since the particles alway collide elastically, it turns out to be something called an *ideal gas*. Most common gases at or near room temperature and atmospheric pressure are pretty good approximations to the ideal gas.

Let's use classical mechanics to derive something called the *ideal gas law*. Put our collection of particles into a cubical box where the sides all have length L. This is shown in Fig. B.1. The particles also collide elastically with the sides of the box, so every time one of them bounces off of the wall at $x = L$, it imparts an impulse $\Delta(mv_x) = 2mv_x$ to the wall, where m is the mass of the particle and v_x is the x-component of its velocity. This particle will bounce around and come back again in a time $2L/v_x$, so the mean force F_x imparted by this particle is the impulse divided by the mean time for a bounce, i.e.

$$F_x = \Delta(mv_x)/(2L/v_x) = mv_x^2/L$$

This is the force from *one* particle. The mean force from all the particles is obtained by summing over all particles. In fact, we want the *pressure* p

which is the force per unit area, so

$$\begin{aligned} p &= \frac{1}{L^2}\left[F_{x_1} + F_{x_2} + \cdots\right] \\ &= \frac{m}{L^3}\left[v_{x_1}^2 + v_{x_2}^2 + \cdots\right] \\ &= nm\left[\frac{v_{x_1}^2 + v_{x_2}^2 + \cdots}{N}\right] \\ &= nm\langle v_x^2\rangle \end{aligned}$$

where N is the total number of particles and $n = N/L^3$ is the number density of the gas particles. (The mass density would just be $\rho = nm$.) It should also be clear that we can write the average of the $x-$components of the velocities-squared, i.e. $\langle v_x^2\rangle$, as $\frac{1}{3}$ of the average velocities-squared, since on the average, the particles are moving one-third in each of the x, y, and z directions. We can therefore write that

$$p = \frac{1}{3}nm\langle v^2\rangle = \frac{1}{3}nmv_{RMS}^2 \tag{B.1}$$

where $v_{RMS} = \sqrt{\langle v^2\rangle}$.

Now we can make the connection to heat and temperature. *Heat* is just the energy contained internally in the gas. For our ideal gas, the energy is strictly in the kinetic energy of the particles, since we've assumed there are no internal excitations like vibrations and rotations of the molecules. Therefore,

$$\begin{aligned} \text{Heat Energy in Ideal Gas} &= E_{K_1} + E_{K_2} + \cdots \\ &= N\langle E_K\rangle \\ &= N\frac{1}{2}m\langle v^2\rangle \\ &= N\frac{1}{2}mv_{RMS}^2 \end{aligned}$$

Temperature is a measure of how much heat energy is in the gas. In fact, we define temperature so that the mean kinetic energy of any particle in the gas is proportional to the temperature. The way we write the proportionality constant is a little weird, though. We define temperature with the relation

$$\langle E_K\rangle = \frac{3}{2}kT$$

(If the particles can rotate or vibrate or so on, then we include these motions in the energy and the factor in front of the kT changes. Don't worry about this for now.) The constant k is called Boltzmann's constant, which we measure in a roundabout way in Experiment 7.

So now that we know what we mean by temperature, we can connect pressure and temperature for the ideal gas. From Eq. B.1,

$$\begin{aligned} p &= \frac{2}{3}n\left[\frac{1}{2}mv_{RMS}^2\right] \\ &= \frac{2}{3}n\frac{3}{2}kT \\ \text{or} \quad p &= nkT \end{aligned} \tag{B.2}$$

This is the *Ideal Gas law*. It is a relation between the pressure, temperature, and density of an ideal gas. An alternative way to write the ideal gas law is

$$pV = NkT$$

where $V = L^3$ is the volume of the container. A chemist may prefer to express the number of particles in terms of the number of *moles* $n_m \equiv N/N_A$ where N_A is Avogadro's number. In this case, the ideal gas law becomes

$$pV = n_m RT$$

where $R = N_A k$ is called the "gas constant".

B.2 The Maxwell Distribution

We've learned a lot from this little exercise on the ideal gas law. It shows how we can take our basic physics principles and apply them to atomic sized particles to obtain properties of some big object like a gas-filled container.

Without making a big deal of it, though, we made a very important assumption. That is, we assumed that the particles are moving randomly through the entire volume, and that the number of particles is so large that we could ignore the random statistical fluctuations. This is what gives meaning to concepts like pressure, temperature, and heat energy.

Before leaving our brief discussion of statistical mechanics, let's take one look at the details of this assumption. We will look at the *distribution* of energies of the particles in the gas. That is, even though we know what the *mean* or *average* energy or velocity is, how broad is the distribution? For example, how probable is it to find a particle with energy much higher or lower than the mean energy? You may want to review the material on "distributions" in Section 9.4.

The concept of thermal energy distributions is very important. This can be particularly true in the case of quantum mechanical systems. Remember that to excite a particular quantum energy state, we need "precisely" the right amount of energy or it won't happen. Thermal distributions are a very common way to get to that precise energy value, since a broad range of energies are continuously covered.

It is not easy to derive the thermal distribution law, and we won't do it here. Be thankful that James Clerk Maxwell was so smart, and he derived it for us back in the 19th century. The *Maxwell Distribution* says that the particle speeds v are distributed according to the relation

$$\frac{dn}{dv} = 4\pi N \left(\frac{m}{2\pi kT}\right)^{\frac{3}{2}} v^2 e^{-mv^2/2kT} \tag{B.3}$$

where $dn = (dn/dv)dv$ is the number of particles with speeds between v and $v + dv$. Examples of the Maxwell distribution are shown in Fig. B.2 for one mole ($N = 6.02 \times 10^{23}$) of helium atoms ($m = 6.65 \times 10^{-27}$ kg) at various temperatures.

The Maxwell distribution has some important properties, which you can prove, namely

$$\int_0^\infty \frac{dn}{dv} dv = N$$

which just says that if you add up all the particles you get the total number of particles, and

$$\frac{1}{2}mv_{RMS}^2 = \frac{1}{2}m\int_0^\infty v^2 \frac{dn}{dv} dv = \frac{3}{2}kT$$

which is simply our definition of temperature. Note that the Maxwell distribution is asymmetric. The most probable value of the distribution, i.e. the velocity at which dn/dv peaks, *is not* the same as the average velocity.

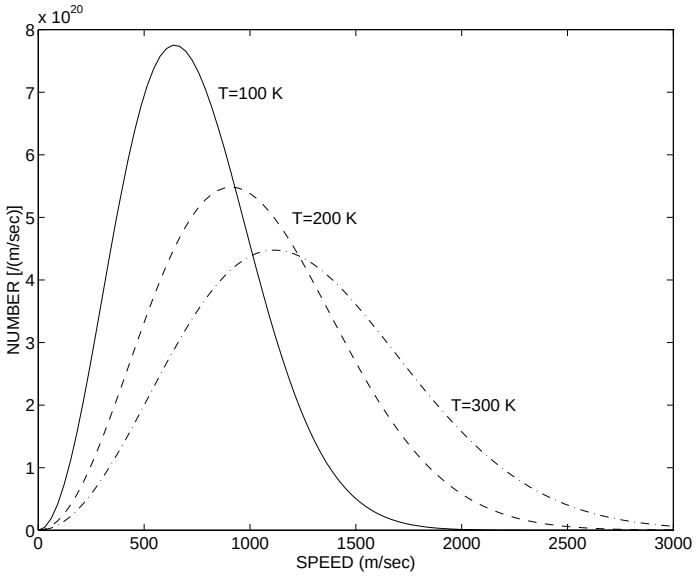


Figure B.2: Maxwell distribution of particle velocities in an ideal gas at different temperatures. As the temperature increases, the average shifts to higher velocities, but there is still plenty of overlap.

.

Finally, we take a moment to rewrite the Maxwell distribution in terms of the kinetic energy of the particles, $E_K = \frac{1}{2}mv^2$. You can do this yourself and find that

$$\frac{dN}{dE_K} = \frac{2N}{\sqrt{\pi}} \frac{1}{(kT)^{3/2}} E_K^{\frac{1}{2}} e^{-E_K/kT}$$

This is in fact called the Maxwell-Boltzmann energy distribution. It tells you how the (kinetic) energy is distributed for an ideal gas.

If we are working with systems other than ideal gases, the form of the distribution changes somewhat, but the factor

$$e^{-E/kT}$$

is always present and dominates the behavior of the distribution at low and high temperatures.

Appendix C

Principles of Mathematics

We take some time to review the basic principles of mathematics. Some of this is aimed at making the kinds of approximations physicists use all the time.

C.1 Derivatives and Integrals

Don't forget your basics. The definition of the derivative of a function $y = f(x)$ is just

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

If you need to approximate a derivative, just pick a small interval Δx, and estimate the value of Δy, perhaps from a graph of y vrs. x, and take $dy/dx \approx \Delta y/\Delta x$. Obviously, if Δx is small enough, then

$$\Delta y \approx \frac{dy}{dx} \Delta x$$

is essentially a perfect approximation. If Δx and Δy are indeed small enough, we call them dx and dy, that is

$$dy = \frac{dy}{dx} dx$$

and the "differential" notation seems pretty obvious.

Deal with partial derivatives like $\partial z/\partial x$ and $\partial z/\partial y$ where $z = f(x, y)$ in exactly the same way. If you're deailing with x then make believe y is a constant, and vice versa.

To approximate integrals, again make use of the fundamental definition, which just boils down to calculating the area under a curve. Depending on the situation, you can use a more or less simple approximations to the area. For example, if you use a simple rectangular approximation with a manageable number of intervals, depending on how you draw your boxes you might be either overestimating or underestimating the integral. Do both, take the average as your best value, and one-half of the difference as your uncertainty.

More sophisticated techniques are possible for estimating integrals (trapezoidal rule, Simpson's rule, gaussian integration,...) and they come with ways of estimating the uncertainty in the technique, but remember to not spend more time beating one level of uncertainty into the ground when somewhere else in your experiment, some source of uncertainty is sticking out like a sore thumb.

C.2 Taylor Series

Any curve can be approximated by a straight line. In fact, that is just what the derivative tells you. In other words, near any point x_0, the derivative of a function $f(x)$ is

$$\frac{df}{dx} \approx \frac{f(x) - f(x_0)}{x - x_0}$$

and therefore

$$f(x) \approx f(x_0) + \left.\frac{df}{dx}\right|_{x_0} (x - x_0)$$

Taylor's theorem simply points out that you can go further and approximate any curve by a polynomial, not just a straight line. In fact, if you go to infinite order in the polynomial, then you've got the exact function.

Each successive term in the polynomial comes with the next higher order derivative, divided by the factorial of the order. That is,

$$f(x) = f(x_0) + \sum_{k=1}^{\infty} \frac{1}{k!} \left. \frac{d^k f}{dx^k} \right|_{x_0} (x - x_0)^k$$

It is usually common to arrange things so that you expand about $x_0 = 0$. In this case, some of the more used Taylor expansions are

$$\begin{aligned}
e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \\
\sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots \\
\cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots \\
\ln(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} + \cdots \\
(1+x)^\alpha &= 1 + \alpha x + \alpha(\alpha - 1)\frac{x^2}{2!} + \cdots
\end{aligned}$$

Note how the last example truncates the infinite expansion and turns into the simple polynomial if α is a positive integer.

The importance of Taylor expansions in science cannot be overstated, mainly as they are used in approximations. They typically converge rapidly, and a limited number of terms (usually only one) is good enough to make the pertinent physics point. An important place this is used in physics is estimating the vibration frequency for a particle placed near the bottom of some arbitrary potential well. If you Taylor expand the potential function $V(x)$ around the minimum of the potential, then the first derivative is zero and the expansion is quadratic in the displacement - a harmonic oscillator! Even the simple first order approximation is often used, and the second order term becomes our estimate of the uncertainty.

C.3 Natural Logarithms

The idea of natural logarithms and the value of e is one of the sweetest pieces of mathematics, I can't resist reviewing it. The reason I like it so much is because it makes you realize why e is such an important number.

One of the first things we realize when studying calculus is that

$$\frac{d}{dx}x^n = nx^{n-1}$$

and therefore

$$\int x^n dx = \frac{1}{n+1}x^{n+1}$$

where n can be anything. But how can you do the integral when $n = -1$? That is, what is $\int \frac{1}{x}dx$, since the formula obviously gives you nonsense? Let's try to answer the question scientifically. We'll use guesswork!

Define the function

$$f(x) \equiv \int_1^x \frac{1}{t}dt$$

and let's examine some of its properties. Obviously, we have

$$f(1) = 0$$

By using a change of variables $u = 1/t$ it is pretty easy to see that

$$f\left(\frac{1}{a}\right) = -f(a)$$

By doing some other simple tricks, you can also see that

$$f(ab) = f(a) + f(b)$$

These are all properties of the logarithm function $\log_b(x)$*, where* b *is the base of the logarithm and can have any value.* So, maybe $f(x) = \log_b(x)$?

Let's try it. We know that the derivative of $f(x)$ must be $1/x$ since that is where it came from in the first place. Go back to the definition of the

derivative and apply it to the function $\log_b(x)$:

$$\begin{aligned}
\frac{d}{dx}\log_b(x) &= \lim_{\Delta x\to 0}\frac{\log_b(x+\Delta x)-\log_b(x)}{\Delta x} \\
&= \lim_{\Delta x\to 0}\frac{\log_b\left(1+\frac{\Delta x}{x}\right)}{\Delta x} \\
&= \frac{1}{x}\lim_{\Delta x\to 0}\frac{x}{\Delta x}\log_b\left(1+\frac{\Delta x}{x}\right) \\
&= \frac{1}{x}\lim_{\Delta x\to 0}\log_b\left(1+\frac{\Delta x}{x}\right)^{\frac{x}{\Delta x}}
\end{aligned}$$

Well that's a mouthful, but all it boils down to is that

$$\frac{d}{dx}\log_b(x) = \frac{1}{x}$$

if and only if

$$b = \lim_{\epsilon\to 0}(1+\epsilon)^{\frac{1}{\epsilon}} \equiv e$$

in which case $\log_b b = 1$. We call this special base the number e since it obviously will be used a lot. It is the base of the "natural" logarithm, that is, the logarithm that has the simplest derivative.

This of course points up a simple way to estimate e. Just take "one plus a small number" raised to "one over that small number":

$$\begin{aligned}
(1+1)^1 &= 2 \\
(1+0.5)^2 &= 2.25 \\
(1+0.1)^{10} &= 2.5937 \\
(1+0.01)^{100} &= 2.7048 \\
(1+0.001)^{1000} &= 2.7169
\end{aligned}$$

The correct value for e, quoted to several decimal places, is $e = 2.71828$.

Notation is a bit of bugaboo. Most people use $\ln x$ to mean $\log_e x$. I try to stick with that, but it is more *natural* to just use $\log x$ to mean the natural log, and explicitly put in the base if it is something else.

C.4 Complex Variables

Complex numbers are based on the "imaginary" number $\imath \equiv \sqrt{-1}$. A complex number is the sum of a "real" number (whose square is positive) and an imaginary number (whose square is negative, i.e., proportional to $\imath$). Complex numbers can be explicitly written as

$$z = x + \imath y$$

where x and y are both real. We say that z has both real and imaginary "parts" and write

$$\begin{aligned} Re(z) &= x \\ \text{and} \quad Im(z) &= y \end{aligned}$$

which are both real numbers. The "complex conjugate" number z^* is given by

$$z^* = x - \imath y$$

The "magnitude" of z, written as $|z|$, is given by

$$|z| = \sqrt{z^* z} = \sqrt{x^2 + y^2}$$

and is obviously a real number. It is no accident that the same symbol is used for the magnitude of a complex number, as for the absolute value of a real number. If z has no imaginary part, then these are the same thing.

It is natural to say that a complex number lies somewhere in the "complex plane". The horizontal (i.e. x) axis in this plane represents the real numbers and the vertical (i.e. y) axis represents the purely imaginary numbers. This is shown in Fig. C.1. You can get a simple pictoral representation of complex numbers this way. For example, the magnitude is the length of the line from the origin to the point (x, y) and the complex conjugate is the reflection in the real axis. We will take this further in a moment.

It is simple to do operations with complex numbers. Just remember that $\imath^2 = -1$ and you will have no problem. For example,

$$\begin{aligned} z_1 z_2 &= (x_1 x_2 - y_1 y_2) + \imath\,(x_1 y_2 + x_2 y_1) \\ \text{and} \quad \frac{1}{z} &= \frac{x}{|z|^2} - \imath \frac{y}{|z|^2} \end{aligned}$$

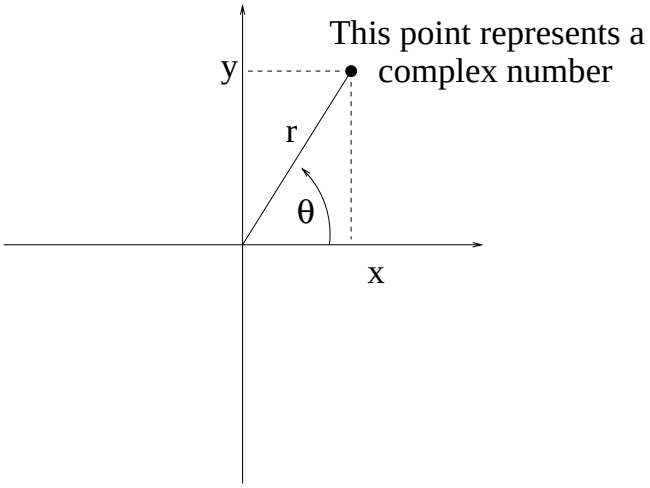


Figure C.1: A number $z = x + \imath y = re^{\imath\theta}$ in the complex plane.

and in particular, $1/\imath = -\imath$.

The real power of complex numbers starts to become clear when you use "Euler's Formula" which states that

$$e^{\imath\theta} = \cos\theta + \imath \sin\theta \tag{C.1}$$

You can think of this as a change from the cartesian coordinates (x, y) to polar coordinates (r, θ) (see Fig. C.1), where $r = 1$ in Eq. C.1 and $r = |z|$ in general.

The formal theory of complex numbers is lovely and plenty of important applications, but they are not really necessary for this course so I won't go into them. Just realize that the mathematical basis for doing some "natural" things is in fact on quite solid footing. For example, you can convince yourself that Eq. C.1 is valid by expanding $e^{\imath\theta}$ in a Taylor series (Appendix C.2). You know that for any *real* number x, you can write

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

so you assume that for any complex number z you can do the same, i.e.

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots$$

Now if you put $z = \imath\theta$, you get

$$\begin{aligned} e^{\imath\theta} &= 1 + \imath\theta + \frac{(\imath\theta)^2}{2!} + \frac{(\imath\theta)^3}{3!} + \cdots \\ &= \left[1 - \frac{\theta^2}{2!} + \cdots\right] + \imath\left[\theta - \frac{\theta^3}{3!} + \cdots\right] \\ &= \cos\theta + \imath\sin\theta \end{aligned}$$

which is just Eq. C.1.

So, we can now write any complex number z as

$$z = |z|e^{\imath\phi} \qquad \text{and} \qquad z^* = |z|e^{-\imath\phi} \tag{C.2}$$

where

$$\phi = \tan^{-1}\left[\frac{Im(z)}{Re(z)}\right] \tag{C.3}$$

is called the "phase" of z. The phase is critically important in many areas of physics, not the least of which is in electronics. For example, we might specify a sinusoidally varying voltage $V_{IN} = V_0 e^{\imath\omega t}$ as an "input" to some kind of device. The output voltage will have the same angular frequency ω (assuming that the device behaves linearly), so we can express the output voltage in terms of the input voltage and a "gain" $g = |V_{OUT}/V_{IN}|$ and a *relative phase* ϕ which tells where the output sinusoid "starts" relative to the input.

Appendix D

A Short Guide to MATLAB

This appendix collects some information that should help you navigate your way through MATLAB. The MATLAB User's Guide is a very useful reference, but there is much more in there than you will need in this course. Also remember that you can get help online from the world wide web at

http://www.mathworks.com

This site includes a long, searchable list of frequently asked questions, and it's a good bet that yours is among them.

D.1 A MATLAB Review

This review was prepared by Prof. Peter Persans.

The following is a brief summary of key MATLAB commands and procedures gleaned from the body of these notes, and from *The Student Edition of* MATLAB *Version 4 User's Guide*. The on-line introductory tutorial is also worthwhile.

A lot of the stuff mentioned here is used in the example in section 9.5 in these notes.

Input Modes. Commands can be executed one by one in the command-line mode in MATLAB or you can write a program consisting of the appropriate command lines in a convenient word processor such as notes in Windows or nedit on RCS and store it as a file with the ".m" extension such as *program-name*.m.

Data input. See pages 15 and 16 in these notes. Lists of data points are usually input as one-dimensional matrices (vectors). You can do this in a command line within MATLAB:

```
x=[1 2 3 4 5 6];
y=[0.1 0.2 0.3 0.4];
```

(The semicolon at the end of the line is *not* necessary, but if you do not include it, then MATLAB will echo values.) You can also store data in ASCII columns in a file with the ".dat" extension, such as *mydata.dat*. If the x data is in the first column and the y data is in the second column of your ASCII file, then you would use the following commands to load it into your MATLAB session:

```
load mydata.dat
x=mydata(:,1);
y=mydata(:,2);
```

Simple arithmetic. To get an on-line list of simple functions, type *help elfun*. Formatting for simple calculations with numbers is straightforward: Addition is a+b, subtraction is a-b, multiplication is a*b, division is a/b, and raising to a power is a^b. Scientific functions include

- abs(x) for absolute value

- round(x) to round to the nearest integer
- real(x) to take the real part of a complex number
- sign(x) to find the sign (it returns $+1$, -1, or 0)
- log(x) for the natural logarithm
- log10(x) for the logarithm to base 10
- sqrt(x) to find the square root

as well as the familiar trigonometric and hyperbolic functions and their inverses, sin(x), cos(x), tan(x), asin(x), acos(x), atan(x), sinh(x), cosh(x), tanh(x), and so on.

Vector construction. The easiest way to create a vector with regularly spaced elements is with the command

x=(*start*:*increment*:*last*)

where *start* is the first element of a vector, *last* is the last element, and *increment* is the step size between the elements. For example, x=(0:0.1:1) creates the vector

x = [0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0]

(The parenetheses "()" are optional, or they could be replaced with brackets "[]".) This is also equivalent to using the function linspace(*start*,*last*,*number*), where *number* is the number of entries in the vector. If you'd like to define a vector where the increments are logarithmic, i.e. separated by a constant factor instead of a constant difference, use logspace(*start*,*last*,*number*).

Array arithmetic. To get an on-line list of matrix functions, type *help elmat*. For operations between a *scalar and an array*, addition, subtraction, multiplication, and division of an array by a scalar looks just like simple arithmetic and the operation applies to every member of the array.

For operations between *two arrays of the same length*, addition, subtraction, multiplication, and division apply on an element-by-element basis, *but* the syntax for multiplication and division is different than for simple arithmetic. Multiplication is written `a.*b` and division is `a./b`, where `a` and `b` are vectors of the same length. (Multiplication and division without the dot correspond to normal matrix multiplication and division.)

Data analysis. See section 6.2.3 (page 94) of these notes. There are some simple MATLAB functions for calculating often-used quantities for analyzing a vector `x` of data values:

- `length(x)` returns the number of elements in the vector
- `sum(x)` adds all the elements in the vector
- `mean(x)` averages all the elements in the vector
- `std(x)` finds the standard deviation of the elements

Note that `std(x)` is equivalent to `sqrt(sum((x-mean(x)).^2)/(length(x)-1))`.

The command `[n,x]=hist(y,nb)` takes a vector `y` of data values, calculates a histogram with `nb` equally spaced bins, and returns vectors `n` and `x` which give the frequencies and midpoints, respectively, of the binned data.

Least squares fitting. This is discussed in some detail in the notes, on pages 134–138. When the data points are equally weighted, all of the operations necessary to fit a polynomial to a set of `(x,y)` data points are included in the command `p=polyfit(x,y,m)` where `m` is the order of the polynomial. A fit to a straight line is therefore `p=polyfit(x,y,1)`. The vector `p` holds the best fit values in order of decreasing polynomial order. For example, if `m=2`, then you are fitting to a quadratic function $ax^2 + bx + c$ and `polyfit` returns `p=`$[a,b,c]$.

The values of the fitted function can be computed for a set of x values x1 using the command y1=polyval(p,x1). (If you want to compute the fitted function at the data points, just use something like yfit=polyval(p,x).)

If the data points are not equally weighted, then you can use Garcia's function linreg (Table 9.1, page 136) to fit to a line. Note that you can retrieve this code (and lots more!) from the MATLAB web site.

Nonlinear least squares fitting. If you can't express the function you want to fit as a polynomial, then you can't use polyfit or linreg. If the function is still linear in the fitting parameters, though, you can use matrix techniques to solve the equations. However, it may be simpler just to resort to numerical techniques to minimize χ^2 directly. You are forced into this situation if the function is nonlinear in the fitting parameters anyway. For example, if you want to fit some decay data to $y = Ae^{-x/\lambda}$ then you can instead fit a straight line to $\log y = \log A - x/\lambda$, but if there is a background term, as in $y = Ae^{-x/\lambda} + B$, then you have to use numerical techniques.

Defining the χ^2 function in MATLAB is quite straightforward, and there is a MATLAB function called fmins which will do all the hard work of finding the values of the parameters which minimize the χ^2 function. This is outlined in some detail for the case of radioactive decay in these notes, pages 354–355.

Simple plots. See page 15 in these notes. There are several simple variations on the plot command which will give you everything you need for this course. If you really want to do more, see the next section of this appendix.

- plot(y) plots the column values of y versus index. It autoscales the axes. Points are connected by solid lines.
- plot(x,y) plots vector y y (vertical) versus vector x (horizontal) on an autoscaled plot. Points are connected by solid lines.
- plot(x,y,'*linetype*') allows you to specify the type of line which connects the points of the type of symbol which is printed on a data point. For "linetype" use "-", ":", "- -", or "-." for solid, dotted, dashed, or

dot-dash lines, respectively; or use ".", "o", "x", "+", or "*" for the corresponding plot symbol.

- bar(y) draws a bar graph of the elements of y versus index.
- bar(x,y) draws a bar graph of y at the locations specified by vector x.
- stairs(y) and stairs(x,y) draw "stairstep" histogram plots.

You can plot more than one set of data, or data and a fit, by specifying more than one set of vectors in plot. For example, plot(x,y,'o',x,yfit,'-') plots "data" vector y versus x as little circles, and then overplots the "fit" vector yfit as a solid line through the points. Another way to overlay plots is to hold a plot and then just repeat the plot command with new vectors. When you are finished collecting overlays, use the command hold off.

Simple labels are put on the graph using the commands

- xlabel('*label on the x-axis*')
- ylabel('*label on the y-axis*')
- title('*title for your plot*')
- text(x,y,'*some text*') puts *some text* at point (x,y)

To print your plot on the default printer, use print. Printing to files or to other printers will depend on which system you are using to run MATLAB. Consult the online help or the User's Manual for details.

D.2 Making Fancy Plots in MATLAB

It is simple to make MATLAB plots with the default characteristics. Sometimes, however, that isn't quite what you want, especially if you are preparing a formal lab report.

You can also, of course, consult the Mathworks web page help directly for some hints. For example, if you want to know how to add Greek characters

to your plot, click "Tech Support Solution Search" on the web page, and search for keywords "Greek AND plot". You will find

"492 How can I place Greek characters in my plot?"

in the search results list. Clicking on this solution tells you not only how to do it, but also tells you how to get an m-file which will make a chart for you that shows the mappings for all the various Greek letters and symbols.

You can dress up plots quite a bit in MATLAB using what is called "handle graphics". You can read about it in the manuals, but following is a primer written by Drea Thomas of The Mathworks that gets you through the basics very quickly. This primer is from "Drea's Desk", which is featured in old issues of the Mathworks digest. If you want to subscribe to the digest, consult the web page.

D.2.1 Drea's Handle Graphics Primer

Whenever I go to a conference or a tradeshow, I'm struck by the number of questions I get that are fairly basic Handle Graphics ones. It seems there are a lot of people that have been using MATLAB for quite a while but haven't taken the leap into learning Handle Graphics. The reason for this is probably that it is a little difficult to get started unless you read the manual or someone points out the two or three things you need to know. Since reading the manual is clearly out of the question (I never read manuals unless it is absolutely necessary), this Desk will be a brief introduction to Handle Graphics.

There are only two commands you need to know to master handle graphics. They are,

- GET – Get the properties of an object.
- SET – Set a property (or properties) of an object.

Every MATLAB graphics object has a unique handle and a set of properties (hence, "handle" graphics). The handle is just a number used to identify an object.

Handle graphics objects have a hierarchy that looks like,

```
                    Root
                   Figure
          Axes                Uicontrol  Uimenu
 Image Patch Line Surface Text
```

In handle graphics parlance, we'd say that Axes are children of Figures, Lines are children of Axes. An Image's parent is an Axes.

Let's look at a simple case,

```
figure
h = plot(1:10)
```

h is the handle to the line on the plot.

```
get(h)
```

You will see a list of properties of the line, most of which are settable. For instance, let's change the width of the line,

```
set(h,'LineWidth',10)
```

One property that is not settable is the Parent. This contains a handle to the axes that "owns" the line. To look at some axes properties,

```
ha = get(h,'Parent');
get(ha)
```

In this case, ha is the current axis so you could use GCA (get current axis).

```
get(gca)
```

Similarly, GCF gets the current figure and GCO gets the current object.

If you want to see what the possible values are of each property (and you don't want to open the manual), just,

```
set(h)
```

The root (handle 0) has some very interesting properties, including:

- Diary: You can turn diary on and off and set the diary filenames with this property.
- PointerLocation: You can tell where the pointer is on the screen and even modify it (*shudder*). For some fun, take a look at an M-file developed by someone with *way* too much free time,

  ftp://ftp.mathworks.com/pub/contrib/games/dropcurs.m

  Figure window objects also have a number of interesting properties as well including,
- PaperPosition: Along with PaperOrientation, PaperUnits, and PaperType, this property specifies the size and orientation of hardcopy.
- KeyPressFcn: Want to capture keyboard input from a figure window without using an editable text uicontrol? This contains a string that gets evaluated each time someone presses a key while focus is in the figure window. The key pressed is stored in CurrentCharacter. For instance,

  ```
  set(gcf,'keypressfcn','get(gcf,''CurrentCharacter'')')
  ```

  echos keystrokes to the command window.

- WindowButtonMotionFcn: This contains a string that gets executed as often as possible when the pointer is in motion over the figure window. Here is an example of how to use it.

```
load clown;image(X);colormap(map)
set(gcf,'windowbuttonmotionfcn', ...
'map=colormap;colormap([map(2:length(map),:);
                        map(1,:)]);')
```

Axes objects have a very large number of settable properties that allow you to customize virtually all aspects of a plot. Particularly useful ones include:

- X/Y/ZTick: These properties allow you to control where tick marks are placed and with X/Y/ZTickLabels, you can control what strings are used to label them.

- X/Y/Zdir: You are an oceanographer and are used to seeing plots with depth going down rather than up. By setting YDir to "reverse", you can flip the direction of the Y axis.

- ButtonDownFcn: This is a string that is executed whenever you click on an axis (the actual axes, not the line in the axis). For instance, if you want to animate your plots try,

```
subplot(2,2,1);plot(1:10);
set(gca,'buttondownfcn','for i=1:36,view(i*10,90);
    drawnow;end')
subplot(2,2,4);plot(magic(10));
set(gca,'buttondownfcn','for i=1:36,view(i*10,90);
    drawnow;end')
```

Click on each plot and see what happens.

- ColorOrder: Don't like the default colors we choose for lines on a plot? Use this property to set your own. The easiest way to use this property is to set the default axes colororder of the figure window.

```
set(gcf,'defaultAxesColorOrder',hot(16))
```

hot(16) returns a 16x3 matrix that represents the red, green, and blue values of 16 colors. Let's contour peaks now and see what happens.

```
contour(peaks,16)
```

Virtually all the settable properties of objects can be given default values. See Drea's Desk in the April 1994 edition of the digest for details,

ftp://ftp.mathworks.com/pub/doc/tmw-digest/apr94

I could go on forever about interesting object properties (but I won't) but it is more enjoyable to experiment yourself. The reference manual has verbose descriptions of each object and their properties so if you can't figure out what a property does from the name, that is your best bet. If you want to put buttons and sliders on your figure windows, there is a manual called "Building a Graphical User Interface" that talks about it in detail.

Happy matlabing.

-Drea-