LABORATORY MANUAL

PHYSICS 133

SPRING 2000

LABORATORY MANUAL

PHYSICS 133

SPRING 2000

Table of Contents

1. INTRODUCTION

1.1 Motivation

The Intermediate Laboratory (Physics 133) is designed to give you a strong start on skills that physicists spend most of their time using every day: good writing, data and error analysis (even theorists!), a bit of computer programming, and, not least, working with typical laboratory equipment. This will be accomplished through three experiments chosen for their variety, level of difficulty, and historical importance. We encourage you to have fun with them and we will support as well as we can anything that you'd like to do with the experiments beyond the assigned tasks.

1.2 Procedures

1.2.1 General Description of Course Requirements

This section gives a general description of each of the course requirements. For more details, including due dates, see the course syllabus, which will be provided separately from this laboratory manual.

- Completion of the three experiments described in this manual.
- Keeping of a laboratory notebook, which is turned in at the end of the quarter.
- Completion of any exercises indicated as required by your instructor, either to be written in your laboratory notebook or handed in as homework, and any homework sets on statistics and data analysis assigned by your instructor.
- A written report on each experiment, handed in by the appropriate due date. The preparation of these written reports may consume half of your time in this course. A considerable part of that preparation will be working with the computer to do analysis, make graphs, and write your reports.
- Attendance at lectures about radiation safety, experimental technique, probability, statistics, error analysis, etc.

1.2.2 More Detailed Descriptions of Course Requirements

Completion of Experiments:

Proper Behavior in the Laboratory. It is important for many reasons to follow established procedures in the laboratory. These reasons are: your safety, efficiency of work, and courtesy to others in the sense of not making it difficult for others to do their work.

Working Procedures. You are expected to work in pairs or groups of three, but you will keep separate lab notebooks and turn in separate lab reports. You must not miss *any* of the lab time without informing your lab partners and the instructor well in advance. Grade penalties for doing this even once may be severe; see your instructor's syllabus. Do not take data without your lab partner(s) present. If someone is missing, work on other aspects of the course and make up the data acquisition later. If you and your partner(s) *all agree* that you have completed your experimental work for the day or for the entire experiment, you are free to skip out of the lab early, skip a day, or, if you have all agreed on it, start a session late.

If you are in a group of three, there is a risk that one person will be passive or sidelined, and not get much out of the lab. It is the responsibility of **all three of you** to make sure that each person spends some time operating the equipment, spends some time taking data, and always contributes actively to the discussions about what to do. The more talkative among you need to remember to solicit the ideas of the less talkative.

The quarter will be divided into three parts. You will sign up for experiments and (different) partners for each. You must finish each experiment on time, as the equipment you are using will be needed by the next group.

Reference Materials. This manual includes, in addition to written descriptions of the three experiments, materials on statistics and the analysis of error. Your instructor may also provide material on useful programming tools for data analysis.

In addition to this manual, there are, in the lab, technical manuals describing each piece of equipment. These documents may not be removed from the lab except for photocopying, in which case, they must be checked out with your lab instructor.

Equipment. Most of the experimental apparatus is stored in cabinets which are locked when the lab is not in use. At the close of each lab session, please return **all** equipment ---including cables, small parts and tools---to its proper place for the next group.

If Something is Wrong. If something isn't working the way you think it should, or if you can't find a piece of apparatus, tell your lab instructor or TA. At least one of us will be around at all times.

Most of the apparatus is electrical. Please be careful not to exceed rated voltages and currents. Recall that a sudden change of current through an inductor can produce VERY large voltages, apt to damage apparatus. THEREFORE: **Switch the power OFF before connecting or disconnecting cables or wires**. Oscilloscopes, multi-channel analyzers, and frequency counters are particularly susceptible. **See other tips on using electronics below**.

Your Laboratory Notebook:

As mentioned above, you are expected to work in pairs, but you must keep your own laboratory notebook, which must be complete in itself (no references to another student's notebook). It is

acceptable to make a photocopy of data tables to place in a laboratory notebook. You must use a notebook that is permanently bound (not spiral bound) and quad ruled (i.e. each page is graph paper). Please write in permanent blue or black ink. The idea is that it is impossible to tear out or add pages or erase results.

Your notebook will be turned in at quarter's end. **In the lab notebook you enter everything about your laboratory experience in real time:** date and time every day, sketches of your apparatus, serial number and model number for every device (so you know, a year later, whether you used "that bad 'scope", and that's the reason your data were lousy), descriptions of your procedures, raw data, problems encountered, and preliminary data analysis, including hand-drawn graphs. This latter item (data analysis with graphs) is extremely important and is too often neglected by students. As you take data, it is imperative that you make at least a crude analysis to verify that the data you are gathering are correct and relevant. Note both requirements. If your data are wrong, you have crashed at the beginning. If you are taking correct data that are in the wrong range of parameters, then you are wasting valuable time; with those data you will not be able to calculate the quantities of interest in the experiment.

The phrase "in real time" is essential -- write down what you are doing **as you do it**, not after you are finished for the day and not even after you have tried "just one more thing." If you are not really frustrated by having to stop and write all the time, you aren't doing it right!

Make your lab notebook readable! This requires more work to write down what might be obvious at the time but cannot be remembered a few days much less weeks later. It is worth it to take that extra time to write a bit more! The lab notebook is the permanent record of your work. It is the standard method all scientists use. Many scientists have made serendipitous discoveries because they carefully kept complete journals. (Mme. Lavoisier kept one worth emulating. If you are in Paris, see it along with the scientific apparatus she and Antoine built, in the Science Museum.) In industry, an intelligible notebook is essential for winning patent suits.

To reiterate, your lab notebook is a substantial part of your grade and will be graded on these characteristics:

1. Completeness -- if any data or results are described in your reports that do not appear in your lab notebook, we may assume that you falsified them or copied them from another student -- you probably didn't, but we *ought to* assume that you did.

2. The presence of good **descriptive** material saying what you did, how things were set up, etc. -- a lab notebook is not just a list of data tables;

3. The consistent use of dates for all your work, and the preservation of chronological sequence (don't skip around, don't tear out or cross out bad starts and abandoned work, etc. your notebook is everything you did, whether or not it turned out good enough to put in your report);

4. The presence of figures -- both schematics of your experimental setups and hand-drawn graphs of data as they arrived. These last are crucial because they show that you were checking that your data made sense as they came in; and

5. (If requested by your instructor): Solutions to "Exercises" that are described in the experiments.

Written Reports:
The style of the reports should be the style of articles in scholarly journals. Your instructors may assign you to read some and comment on them, or they may provide some examples. If not, find some articles of interest to you in *The Physical Review* online or in the library to look at the style of writing.

Your report must communicate clearly, concisely, and effectively. This is as important as the quality of your data! In any scientific writing, the most important thing is **to have your audience in mind** as you write every section -- every sentence -- and, in the case of jargon, even every word. When writing for a highly specialized journal, your typical audience member might be yourself-just-before-you-did-this-experiment. When writing for a general journal like *Science* or *Nature*, it is someone with strong quantitative skills and a good overall science background, but not necessarily a physicist. **When writing lab reports for this course, your audience is someone like yourself before you started taking this course.** Therefore:

- Do not assume your reader has read the lab manual.

- **Do** assume your reader has as much background as is implied by a reasonable retention of Physics 5ABCD.

Put into your report enough information that another researcher could reproduce what you've done in another laboratory. Imagine this is to be a published paper in the *Journal of Infinitely Repeated Undergraduate Lab Experiments*. You may *formally cite* the lab manual just a few times for equations, procedures, etc. that you think are of interest but not vital to your presentation, but do not assume that your reader would always go dig them up. Your report should be self-sufficient. Do **not** copy out portions of the manual in your report -- first, not all of it will be vital to your presentation and line of argument, and second, we want you to practice putting what you did (not what you were supposed to do!) in your own words.

Your reports are NOT repetitions of your lab notebooks. They contain introductory material and can stand alone without the lab manual (the notebooks don't have to). They should exclude dead ends and bad data (meaning data taken in the wrong conditions, not data you just don't understand -- **that** you have to keep and try to explain!).

What follows is an outline of the required format, meant to emulate most experimental journal articles, with comments about the content of each section.

1.2.3 Report Format Guide

Cover Page.

The cover *must* have a descriptive title, date, your name and the full name of your lab partner, and an abstract. Also, include the course title (Physics 133) and the instructor's name.

The abstract is a single paragraph that summarizes the work that was done and the primary results (including quantitative values if they can be briefly summarized). In a research paper, the abstract serves two purposes: 1) to convey the primary result to those who don't care enough to read the paper, and 2) to let the people who would care know that this is a paper they should read all the way through. Remember that **where ever** you report a quantitative result you must also include the value of uncertainty. Results without uncertainties are meaningless!

Introduction.

The introduction of a scientific paper usually starts with a brief summary of the status of previous work on the topic and the open question(s) that the work in the paper sheds new light on. If you like, it's nice to replace this, in our case, with a little historical context for the experiment (not necessarily given in the lab manual!) This is optional. You may skip it and proceed to:

A paragraph stating the purpose of the experiment. For example, "The purpose of this experiment was to measure quantitatively the effective circuit elements that represent the contents of a number of 'black boxes'…" (note there are more purposes than this one to that lab).

Next, it is good to put most of the theory and equations related to your experiment into this section.

Finally, include a brief outline of the rest of the report. For example, "Section 2 describes the experimental apparatus, Section 3 gives the raw data, Section 4 gives the data analysis and results, and section 5 provides the conclusions. The error analysis is described in the Appendix."

Notice the difference between the abstract and the introduction: the introduction gets the reader ready to see your procedure and results. The abstract is a quick summary of your most important results (with just enough background to understand their context and importance).

Apparatus and Procedure.

The purpose of this section in a research paper is to allow other interested groups to reproduce your result in their labs -- or to find any flaws your experiment might have had (which you didn't know were flaws, of course -- if you know them, it's better to point them out yourself than have someone else do it!). As stated before, you are not writing this section for the course staff -- you are writing it for a reasonably well-educated reader who has **not** seen the laboratory manual.

Do **NOT** put in a long list of all the pieces of equipment you used. Mention each device where it comes in naturally in the description of *how* the experiment was set up.

Since not every single detail in your notebook can or should be transferred to your paper, here you need to make some judgments. The more specific to this experiment a piece of apparatus is, the

more detail you should give in describing it. Your notebook should contain the serial number and model number for every device you use, including oscilloscopes, multimeters, etc., but in the writeup you should concentrate on mentioning the equipment which is unique to the experiment.

Schematic diagrams of the experimental setup are *required*.

Include details of the procedures you followed. Do *not* describe the calculations you will do in order to obtain final results. That is *not* procedure; that is data analysis. Write a description of what you (the experimenter) actually did, *not* in the form of instructions to the reader. Do not describe what you should have done. In other words, this is right:

> "We calibrated the fragistam by running seven chilled samples for three hours each."

These are wrong:

> "Next calibrate the fragistam by running at least seven chilled samples for three hours each." (you are not writing a manual and giving someone else instructions)
>
> "Next we were supposed to calibrate the fragistam by running at least seven chilled samples for three hours each." (what does "supposed" mean? You are writing as though this is your own experiment, not as if you were following someone else's instructions, even though that is sometimes true)
>
> "Next we calibrated the fragistam by running at least seven chilled samples for three hours each." (this is close, but you didn't run "at least seven", even if those are the words in the manual; you actually ran some specific number, and that's what you should say).

If you took some wrong turns in your work that you later did right, you should not include these in your report, even though they should be very clearly discussed in your notebook. In the report, just describe what you did that you consider best and valid.

You should not name what software you are using (e.g. EXCEL) unless it does something unique. Unique means specific to this kind of experiment, just as it does for hardware. We don't have such kinds of software in 133, but examples might be a code that simulates circuit behavior, or allows an unusual kind of automated operation of a spectrometer.

Results.

This section should include your data, in tables, the techniques you used to analyze it, and your final results. As in the procedures section above, this should be a complete, flowing narrative that tells an independent researcher everything they need to know to reproduce your work.

You must copy into your report all the data that directly enter into any determination of a final result. Typically this means that your report has a number of tables of the measurement results,

but don't forget single measurements that are necessary as well. For example, in the Atomic Spectra experiment, the angles of many lines are measured for each gas. But also measured is the angle of the undeflected light, and that single angle for each gas is extremely important.

All tables and figures that you include **must** be referred to and described by text in the body of your report.

The raw data are much more important than any intermediate steps used to calculate your final quantities, because with the raw data any calculation can be repeated, but if you only give the results of calculations, no check can be done for arithmetic or other errors.

Definition of raw data: One of the important aspects of experimental technique you will learn in this class is something you may think you know already but probably don't: the definition of raw data. This is the level at which you should record every measurement in your lab notebook. The best example we can give is reading an oscilloscope to find the voltage level of a signal. The raw data are the number of centimeters from the origin to the point you are measuring, plus the definition of the origin point, plus the scale of voltage/cm from one of the dials. Even the voltage itself is NOT raw data, because to get it you need to do a calculation (multiplying number of centimeters times the scale)! In all cases, write down EVERYTHING about the measurement: direct observation (centimeters, or the number on the LED screen of the instrument), scale, any connector used, such as an attenuator, in the wire between the circuit and the meter, brand of instrument and its serial number. For the optical experiment or the radioactivity experiment, similar considerations apply.

It is also important at the raw-data level to estimate uncertainties as best you can. You MUST estimate uncertainty quantitatively for every measurement you make.

For the data tables in your lab report, you don't need to present quite as "raw" a format; for example, you can record an oscilloscope reading of voltage in volts (with a suitable error derived from the truly raw data in your notebook). However, quantities that require a further level of derivation (for example the current in a circuit which you have deduced from a voltage reading) cannot stand alone; you should include in your table the voltage and resistance that you used as well as the current you derived. If a column in a table is just derived from other columns, that should be made VERY CLEAR in either the heading of the column or a footnote.

Note the following about the way data are presented in tables:

- They must have a reasonable number of significant figures. This includes calculated as well as raw data. You did **not** calculate that the current in your circuit was 4.12348376 A. Don't put that in your data table. It was probably 4.1 or 4.12 A (and you should have a good reason for knowing which).

- Some programs like EXCEL will drop zeros off the end of numbers even when those zeros are actually important significant figures. If three successive measurements are 4.17, 5.00, 6.32, and they appear in your table as "4.17, 5, 6.32," that is a real and major error in your result. You need to fix it, even if it means copying the entire table out of EXCEL

and into another program.

- Use scientific notation to avoid numbers like “0.000000000217” and “42300000000.00” in your tables. These should be 2.17×10^{-10} and 4.23×10^{10} respectively. You need not use scientific notation in tables if all your values are between 0.1 and 999. Note that “2.17E-10” is not scientific notation, it is programmer's “slang”. Use the proper format given above.

After presenting your data, you present any equations you use to transform the raw data into final quantities of interest, with uncertainties, and you present the final results themselves. For the Low-Frequency Impedance and Radioactivity experiments you will be presenting plots of results here. *Each plot must be numbered and mentioned in the text!* For example, you would say “Figure 7 shows impedance as a function of frequency for Box C.” You do not have to place the figures within the text; they can be gathered at the end of the report if you prefer. Note that you can have numbered Figures, Tables and Equations in your text (Figure 3, Equation 7, Table 4, *etc*.). You do not use any other terms (like “Graph 1”, “Image 3”, etc.) Graphs, diagrams, and every other imaginable type of illustration are numbered and referred to as “Figures.”

“Throwing out” data: When you take a datum in the lab, **you are stuck with it**. In the end, either: 1) it must end up in your published results, or 2) you may throw it out if you understand, and can explain in your lab notebook, exactly why your setup was bad when you took it. Note that again: why your **setup** was bad. Under no circumstances is it permissible to delete data simply because either 1) you don't understand it, or 2) it seems to disagree with the rest of your data. Furthermore, if you throw out one or more data points on the argument of a bad setup, you must also throw out **all** the data taken with that same setup, even if some of what you're throwing away looks like it might be good. To rephrase: 1) you must have a reason **unrelated to the *value*** of a data point for throwing it away, and 2) all data that are susceptible to the same reason must be thrown away **regardless of their value**. This principle is the heart of ethical -- and effective -- experimentation. Finally, of course, all records and all data, even those eventually discarded from publication, must be retained in your lab notebook.

Discussion and Conclusions.

Here is the place where, in a real journal article, you would discuss how your result relates to all the other work in the field. That's not so relevant to us.

Also, a long research paper which presents a variety of material often requires a summary at the end in order for the reader to recall what was covered and to pick what the author considers to be the primary results. You should be the judge as to whether this is necessary for your own report. If your writing has been organized, concise, and skillfully explanatory, it probably isn't. But if you think you might have written something of a tangle, this is one way to help make up for it (rewriting and untangling everything else is better, of course).

For this section in particular, if you write one, watch your tone. Your whole report should be very objective. Do *not* give any personal opinions about your “experience” in the laboratory. These

reports are meant to be comparable to research articles that you would submit to a journal; they are *not* personal memoirs.

Acknowledgements.

Acknowledge any significant help you received in doing the experiment, calculating results, or interpreting the results.

References.

Cite bibliographic references in a consistent way. In the text, you can use either a number in parentheses or in a superscript, or you can refer by author name and year. At the end, list the full information for all the references you cited. If you used numbers, the references are in the order you cited them. If you used author/year, then the references are in alphabetical order.

Tables and Figures.

All Tables and Figures must be numbered and referenced in the text of your report. You will state something like "Table 3 includes the angles of spectral lines observed for Helium gas." Or, "Figure 1 shows a schematic of the experimental apparatus." It is not necessary to interleaf figures and tables inside the body of the report; they may be collected at the end. Figures and tables should have sources cited if they are not original.

Some general notes about report formatting.

Double-space all text, so that there will be ample room for comments and corrections.

It is important to include lots of figures, because they clarify results and other points very quickly. **Label all axes** and *include units wherever appropriate!* Include an informative figure caption with all relevant information.

Whether you hand- or machine-graph, always label the axes (including the appropriate units), and use an appropriate scale. The latter means that the data fill the space well. Sometimes using a logarithmic scale on either axis, or both, illustrates the important features of the data better; you should know how to create these kinds of plots in whatever graphics package/programming language you use. But don't knock yourself out figuring out how to do labels, axes, etc. in your plotting program. You can hand-write things in if necessary. It's having the right content that's important.

Plot your data points with error bars. Your instructor will explain more about errors and error analysis. **Do not connect the dots:** your graphs can contain data points and they can contain a function which you have fit to the data, but they should NOT contain either line segments connecting the points or spline curves that pass through all the data points. If you **do** include a fit function, then either the caption or the text should contain the relevant information about the fit: e.g., parameters of the fitting function and chi-square values.

Do **not** include printouts of code, raw output, etc. from data analysis programs (unless of course these are your data tables and graphs). You'll never see such things in a published paper. You should know what is actually relevant and incorporate it into the text of your report.

The standard reference for scientific communication is the *American Institute of Physics Style Manual*. It provides much useful information. We have filed a copy in the laboratory reference-manual file cabinet.

1.2.4 Computer aspects of the course

You may use any software to write your lab reports as long as they conform to the requirements above. Your instructor or TA is likely to have used both Word and LaTeX, and so can readily help you with either of those. There are other experts around, including in some cases your fellow students!

Students need computer access and capability in this course for report writing, data analysis, and graphing. If you already use a programming language you are comfortable with for analyzing and plotting data, you may use it as long as the graphs you produce are suitable in format. Your instructor will also provide at least one package that the course staff will support you in learning to use -- this might be in Python or it might be a simple, widely available package like gnuplot. If you choose to use a spreadsheet package like EXCEL, do **not** use the packaged statistical programs, as they usually do not incorporate errors of individual points properly.

1.3 Radiation Safety Rules

One of our experiments uses low-level gamma-ray sources, and you will be given a lecture on radiation safety at the start of the course. Below are appropriate precautions and rules required by the UCSC Environmental Health and Safety Department in compliance with the State Department of Health Services.

1.3.1 Rules to Observe

1. Any foods, including drinks, are absolutely prohibited in any laboratory where there are radioactive sources. Leave them outside the lab. Do not dispose of empty food wrappers or drink containers in the lab wastebaskets.

2. Smoking is not permitted in any campus building, especially in laboratories.

3. Don't put sources close to your eyes. Corneal damage may result.

4. Return sources immediately to the instructor when you finish, and don't take them from the laboratory at any time!

5. Keep away from sources as much as practicable. Do not handle them unnecessarily, and do not put them in your pockets.

6. Handle lead with gloves if possible; if not, minimize handling and wash hands thoroughly with soap and hot water when finished handling lead. Lead toxicity is probably more of a health risk than radiation for the sources we use in this lab.

2. UNCERTAINTY IN MEASUREMENT

Note: Chapters 2, 3, and 4 parallel discussions in the textbook by Lyons in some places; in some places they move a bit beyond that text; and many things are not covered here. Statistics is difficult and is something you will need lifelong; so please read both sources and choose the explanations that work best for you.

"Noise" can be defined as any variation in a measurement that cannot be predicted, and scatters your measurement randomly within a region centered on the correct value. Sometimes these variations are interesting in themselves, but more often they are getting in the way of making a precise measurement. Because of noise, every measurement of any physical quantity is *uncertain*. For example, here is a recorder trace of the output voltage from an ohmmeter:

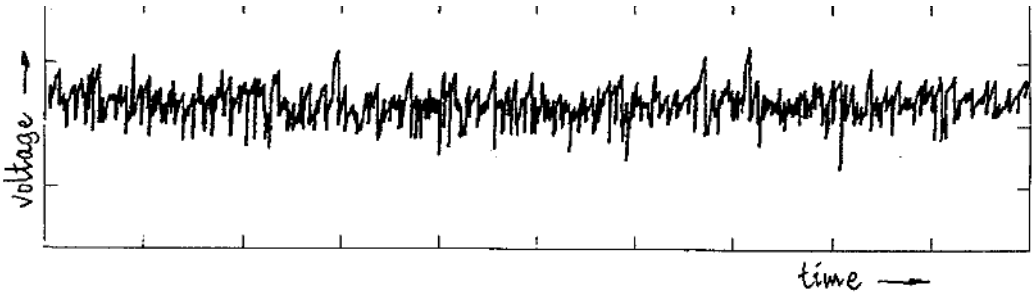

Figure 1. Noise in a meter reading. The voltage fluctuates because of noise.

In another example, taken directly from the Radioactivity experiment, the intensity of a radioactive source is monitored with a Geiger counter. The counter is used to count the number of pulses in each of a sequence of one-second intervals, producing this graph of counting rate vs time:

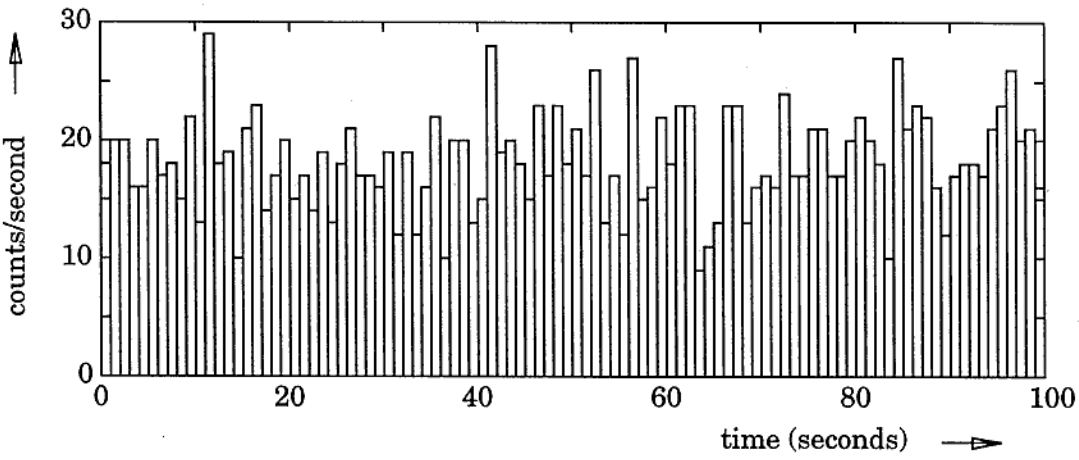

Figure 2. Noise in a pulse counter.

The number of counts recorded in each interval will *fluctuate* from one interval to the next. We use the term *noise* to describe such fluctuations. It is our aim in the following paragraphs to understand noise as a source of uncertainty, to describe techniques for quantifying it, and to give meaning to the concept of *precision*.

Noise is also called *random error*, or *statistical uncertainty*. It is to be distinguished from *systematic error*. Systematic error, which is an error in measurement arising from a defect, such as the mis-calibration of a meter or some physical effect not taken into account in the measurement, can in principle be checked and corrected for.[1] Noise, on the other hand, is more basic. It arises, as in the first example (Figure 1), from the thermal motion of individual atoms, or, as in the second example (Figure 2), from the quantum-mechanical uncertainty associated with the radioactive emission of particles.[2]

In this second example, the question arises: How accurately may we estimate the "true" intensity of the radioactive source, or the "true" counting rate, when we measure for only a finite number of time intervals? Such a finite number of measurements, which in the above example is 100 (in general we'll call it n) is called a "sample", or more precisely, a "random sample" of the *total* population of such measurements. In this example, the total population is infinite.[3] If we could make an infinite number of measurements, we could, in principle, reduce the statistical uncertainty to an infinitesimal value. Since we clearly cannot make an infinite number of measurements, we are stuck with a finite sample of n measurements, and hence with a finite statistical uncertainty in the determination of the counting rate.

For any such sample of n measurements, a few key statistical parameters may be calculated that serve the purpose of describing the measurement sample in the context of its associated noise. There are three parameters that are particularly useful:

1. The *sample mean* $\bar{x}$:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{1}$$

Here x_k is the k^{th} measurement.

2. The *sample variance* s^2:

$$s^2 = \frac{1}{n-1}\sum (x_k - \bar{x})^2 \tag{2}$$

The square root of the sample variance is s, and is called the sample *standard deviation*.

3. The *variance of the mean* $\sigma_{\bar{x}}^2$:

$$\sigma_{\bar{x}}{}^2 \approx \frac{1}{n} s^2 \tag{3}$$

---

[1] The event depicted on the cover of John Taylor's monograph, *An Introduction to Error Analysis* very likely arose from a *systematic* error in engineering design.

[2] Noise can also arise from what has recently been described as *deterministic chaos*---see, for example, James Gleick's recent book entitled *Chaos---Making a New Science* (Penguin Books, 1988). Connections may exist between such deterministic chaos and the thermal fluctuations or quantum fluctuations on the atomic scale; such connections are the object of recent research.

[3] We make the assumption that our source of radioactive particles is inexhaustible, which of course cannot be strictly true. This has no bearing on the point of our discussion, however. We'll just take "infinite" to mean "very very large."

Note the distinction between the *sample* variance and the variance of the *mean*. The square root of the variance of the mean is $\sigma_{\bar{x}}$, and is called the *standard deviation of the mean*. The meaning of the approximation sign in (3) is that the quantity s^2/n is an *estimate* of the variance of the mean.

An experimental result, *i.e.*, the best estimate we can make of the "true" value of *x*, is conveniently expressed in the form

$$\text{RESULT} = \bar{x} \pm \sigma_{\bar{x}} \approx \bar{x} \pm \frac{1}{\sqrt{n}} s \tag{4}$$

As we shall see in the discussion contained in the following paragraphs, the meaning of this statement is that we expect the "true" value of *x*, taking into account only the random effects of noise or random error, to have about a 68 percent chance, or *level of confidence*, of lying between $\bar{x} - \sigma_{\bar{x}}$ and $\bar{x} + \sigma_{\bar{x}}$.[4] These two values of *x* are the approximate *confidence limits*. They delimit a range of *x*-values called the *confidence interval*.

There is one further point that we shall discuss later in more detail. It frequently happens that we wish to determine the mean, and the variance of the mean, for a quantity *u* that is a *function* $f(x, y, \cdots)$ of a number of experimentally measured, *independent* quantities $x, y, \cdots$. That is, $u = f(x, y, \cdots)$.

The value of $\bar{u}$, the mean of *u*, and the best estimate for $\sigma_{\bar{u}}^2$, the variance of the mean of *u*, can be calculated using the following formulas:

$$\bar{u} = f(\bar{x}, \bar{y}, \cdots) \tag{5}$$

and

$$\sigma_{\bar{u}}{}^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_{\bar{x}}{}^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_{\bar{y}}{}^2 + \ldots \tag{6}$$

Each of the variances on the right side of Eq.(6) may be estimated using an expression like that of Eq. (3). Hence a result for the derived measurement of *u* should be expressed in the form

$$\text{RESULT} = \bar{u} \pm \sigma_{\bar{u}} \tag{7}$$

The process of doing the calculations described by Eqs. and (6) is called the *propagation of uncertainty through functional relationships*. These formulas, which are valid if $\sigma_{\bar{x}}, \sigma_{\bar{y}}, \cdots$ are not too large, are quite general.

In what follows, we discuss the details of each of these points. Further references are cited at the end of this chapter.

[4] Equation (4) is not quite correct. See footnote (8) regarding further discussion of Eqs. (4) and (7).

2.1 The sample mean $\bar{x}$

The sample mean $\bar{x}$ is simply the average of the n individual measurements:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{8}$$

Consider our second example shown graphically in Figure 2. The number of counts in each of the first 25 one-second intervals is 18, 20, 20, 16, 16, 20, 17, 18, 15, 22, 13, 29, 18, 19, 10, 21, 23, 14, 17, 20, 15, 17, 14, 19, 13; $n = 25$, and

$$\bar{x} = \frac{1}{25}\{18+20+20+16+16+20+17+\cdots\} = \frac{444}{25} = 17.76$$

For this particular sample, certain numbers appear more than once. 13, 14, 15, 16 and 19 each appear twice, 17 and 18 appear three times, and 20 appears four times. In general, the value x_k might appear $g(x_k)$ times; $g(x_k)$ is called the *frequency* of the value x_k. Thus, an expression equivalent to Eq. (8) may be written as

$$\bar{x} = \frac{1}{n}\sum_{x_k} x_k g(x_k) \tag{9}$$

Note that while the sum in Eq.(8) is over k (the interval number), the sum in Eq. (9) is over the values of x_k.

For our example, $g(18) = 3$, $g(20) = 4$, $g(16) = 2$, etc., and Eq. (9) looks like this:

$$\bar{x} = \frac{1}{25}\{10\cdot 1+13\cdot 2+14\cdot 2+\cdots+22\cdot 1+23\cdot 1+29\cdot 1\}$$

Now

$$\sum_{x_k} g(x_k) = n \qquad \text{so that} \qquad \bar{x} = \frac{\sum_{x_k} x_k g(x_k)}{\sum_{x_k} g(x_k)}$$

Furthermore, we expect that as n becomes very large, the quantity $g(x_k)/n$ will approach the *probability* $p(x_k)$ that the value x_k will appear. This defines $p(x_k)$:

$$p(x_k) \equiv \lim_{n\to\infty} \frac{1}{n} g(x_k) \tag{10}$$

The introduction of the probability $p(x_k)$ now leads us to a discussion of the total population from which our finite sample of measurements is taken.

2.1.1 Background: Properties of the total population

The probability $p(x_k)$ is descriptive of the total (in our case, infinite) population of all possible measurements. (The *total* population is also called the *parent* population). In general, we expect that $p(x_k)$ will be *normalized*:[i5]

$$\sum_{x_k} p(x_k) = 1$$

Although for infinitely large populations such as the one we are considering, $p(x_k)$ is not accessible to us (we can only *estimate* it through the measurement of large samples), it is conceptually well-defined, and with it we can define the *mean* μ and the *variance* σ of the total population:[6]

$$\mu = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} x_k = \sum_{x_k} x_k \, p(x_k) \tag{11}$$

and

$$\sigma^2 = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \left(x_k - \mu\right)^2 = \sum_{x_k} \left(x_k - \mu\right)^2 p(x_k) \tag{12}$$

Note that these definitions are similar to Eqs. (1) and (2) defining the mean and variance for a particular finite sample of measurements; the difference is that we are here considering the *total* population.

In general, the *mean* value, also called the *average* value, or *expectation* value of *any* function $f(x_k)$ is given by

$$E[f(x_k)] = \text{ave}[f(x_k)] = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(x_k) = \sum_{x_k} f(x_k)\, p(x_k) \tag{13}$$

where $E[f]$ stands for the *expectation* value of f.

Note that μ and σ^2 are the mean values, or average values of particular functions of x_k. Thus $\mu = E[x_k]$ and $\sigma^2 = E[(x_k - \mu)^2]$.

The square root of the population variance σ^2 is σ, the *standard deviation* for the total population. σ is a statistical parameter describing the *dispersion* of the (infinite) number of measured values about the population mean μ. It describes how closely the measured values are clustered about the mean, and thus gives a measure of the width of the distribution of measured values.

2.1.2 The Normal Distribution

---

[5] In the following discussion we assume that x is limited to only the discrete values indicated by x_k. If x is in fact a continuous variable, sums over x_k should be replaced by integrals over x. Thus, for example

$$\sum p(x_k) = 1 \qquad \text{becomes} \qquad \int p(x)dx = 1$$

[6] Greek letters are often used to denote parameters that are descriptive of the *parent* population.

The interpretation of the parameter σ is easily envisaged if the measured quantities are distributed according to a *normal*, or *Gaussian* distribution, which is a common occurrence. The probability distribution function for a normally distributed continuous random variable x is given by[7]

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(x-\mu)^2 / 2\sigma^2] \tag{14}$$

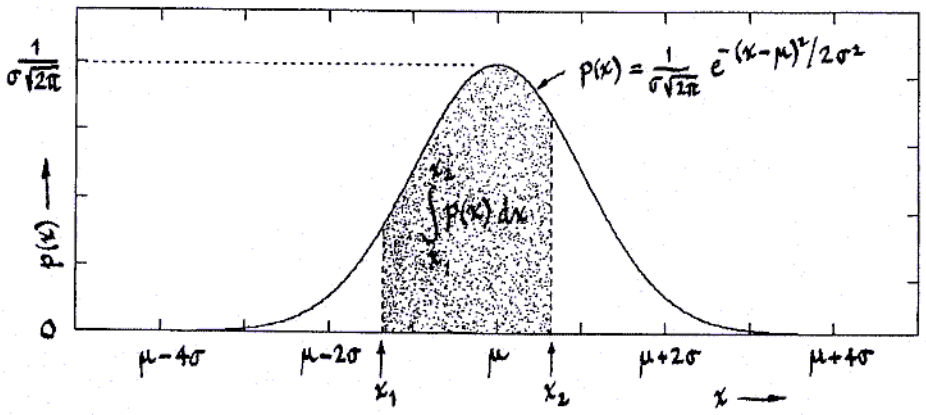

Figure 3. The Gaussian distribution---$p(x)$ vs. x.

$P(x)dx$ is the probability that any particular value of x falls between x and $x + dx$, and

$$\int_{x_1}^{x_2} p(x)dx$$

is the probability that any particular x falls between x_1 and x_2. This integral is represented by the shaded area in the graph above.

If $x_1 = -\infty$ and $x_2 = +\infty$, then it is certain that any particular x falls in this interval, and

$$\int_{x_1}^{x_2} p(x)dx = 1$$

The normalization factor $1/\sigma\sqrt{2\pi}$ ensures that this is the case. If σ is reduced, $p(x)$ becomes more sharply peaked.

If $x_1 = \mu - \sigma$ and $x_2 = \mu + \sigma$, the shaded area is approximately 0.6827. That is, for a normal distribution, there is approximately a 68 per cent chance that any particular x falls within one standard deviation of the mean. Furthermore, the chance that an x will fall within two standard deviations of the mean is approximately 0.9545, and within three standard deviations, approximately 0.9973. Equivalently, the probability that x falls outside 2σ is 4.55 percent; outside 3σ is 0.27 percent.

[7] See The Taylor Expansion Method in section 3.

From a random sample of *n* measurements one may form a frequency distribution that may be compared with any particular probability distribution function *p*(*x*). Figure 4 is a bar graph, or histogram, formed from the data shown in Figure 2.

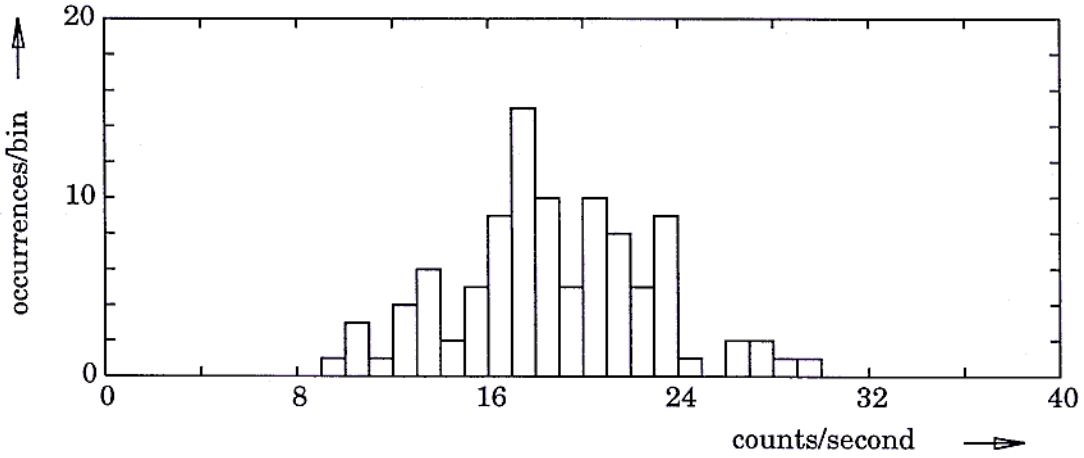

Figure 4. A sample distribution.

Note that it looks qualitatively similar to the Gaussian Distribution shown in **Figure 3**. A quantitative comparison may be made using *Pearson's Chi-square Test*, as described in Chapter 4 of this manual, and in practical terms, in the Radioactivity experiment.

2.1.3 Motivation for the Normal Distribution

We have mentioned that the fluctuations in measured quantities are commonly found to be approximately described by a Normal, or Gaussian distribution. Why? The answer is related to a powerful theorem, much beloved by physicists, called the *Central Limit Theorem*.

This theorem states that if we have a number of random variables, say $u, v, w, \cdots$, and that if we form a new variable *z* that is the *sum* of these ($z = u + v + w + \cdots$), then as the number of such variables becomes large, *z* will be distributed *normally*, *i.e.*, described by a Normal distribution, *regardless* of how the individual variables $u, v, w, \cdots$ are distributed.

While we won't prove the Central Limit Theorem here (it's not an easy proof), we can present a "physicist's proof"---an example that is easily tested: Let each of $u, v, w, \cdots$ be real numbers randomly and uniformly distributed between 0 and 1. That is, each is drawn from a *flat* distribution---clearly *not* a Normal distribution. Then let $z = u + v + w + \cdots$. For even as few as 4 or 5 such terms in the sum, *z* will be nearly normally distributed. (This may be easily verified using a simple numerical simulation requiring but a ten-line computer program.) In fact, if there are only 2 terms, we can already see the peaking near the center, with the result being a *triangular* distribution as shown in Figure 5.

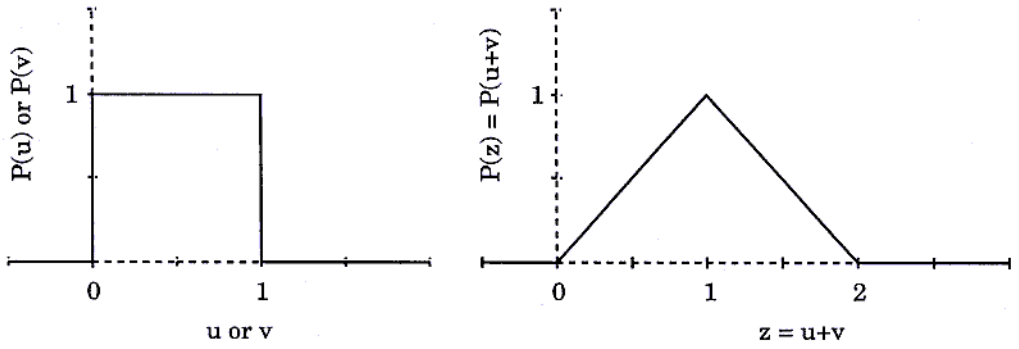

Figure 5. A triangular distribution

Now a typical quantity measured in a physics experiment results from the sum of a large number of random processes, and so is likely to be distributed normally. For example, the pressure of a gas results from summing the random motions of a very large number of molecules, so we expect measured fluctuations in gas pressure to be normally distributed.

Nevertheless, we must be careful to not put too much faith in the results of the Central Limit Theorem. One frequently sees measured values that are obviously non-normal---too far away from the mean----that could arise, say, from some voltage spike or a truck passing by. Not every data point can be expected to fall within this classic bell curve.

2.1.4 How is $\bar{x}$ related to μ ?

In general, our desire is to determine, from a finite sample of measurements, best estimates of parameters, such as μ and σ^2, that are descriptive of the total population. The simplest relationship is that between $\bar{x}$ and μ: $\bar{x}$ is the best estimate of μ. This is equivalent to saying that the expectation value of $\bar{x}$ is μ, or $E[\bar{x}] = \mu$. While this statement may seem intuitively obvious, here is a proof:

$$E[\bar{x}] = E[\frac{1}{n}\sum_k x_k] = \frac{1}{n}\sum_k E[x_k] = \frac{1}{n}\sum_k \mu = \frac{1}{n}\cdot n\mu = \mu$$

2.2 The sample variance s^2, and the sample standard deviation

The *sample variance* s^2 is defined by:

$$s^2 \equiv \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2 \tag{15}$$

By substituting $\bar{x} \equiv \frac{1}{n}\sum x_k$ we obtain

$$s^2 = \frac{n\sum x_k^2 - (\sum x_k)^2}{n(n-1)} \tag{16}$$

which is an expression useful for numerical calculation, in that it involves only

$$\sum x_k \ , \ \sum x_k^2 \ , \ n$$

which are easily computed, and involves only one final division. For the complete measurement sample shown in Figure 2,

$$\sum x_k = 1825, \ \sum x_k^2 = 35013, \ n = 100$$

which yields $s^2 = 17.240$ as the sample variance, and $s = (17.240)^{1/2} = 4.152$ as the sample standard deviation.

2.2.1 How is s^2 related to σ^2?

The sample variance s^2 is the best estimate of the variance σ^2 for the total population. This is equivalent to the statement that the expectation value of s^2 is equal to σ^2. The proof of this statement takes a few lines, but runs as follows. We start by taking expectation values of both sides of Eq. (15):

$$\begin{aligned} E[s^2] &= \frac{1}{n-1} E\left[\sum (x_k - \bar{x})^2\right] \\ &= \frac{1}{n-1} E\left[\sum x_k^2 - 2\bar{x}\sum x_k + \sum \bar{x}^2\right] \\ &= \frac{1}{n-1} E\left[\sum x_k^2 - 2n\bar{x}^2 + n\bar{x}^2\right] \\ &= \frac{1}{n-1} \left\{\sum E[x_k^2] - nE[\bar{x}^2]\right\} \\ &= \frac{n}{n-1} \left\{E[x_k^2] - E[\bar{x}^2]\right\} \end{aligned}$$

To evaluate $E[x_k^2]$ we note, from Eq. (12), that

$$\sigma^2 = E[(x_k - \mu)^2] = E[x_k^2] - 2\mu E[x_k] + E[\mu^2] = E[x_k^2] - \mu^2 \qquad (17)$$

so that

$$E[x_k^2] = \mu^2 + \sigma^2$$

To evaluate $E[\bar{x}^2]$, we expand to find

$$E[\bar{x}^2] = E[\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right)^2] = \frac{1}{n^2}\left\{E[\sum x_k^2] + E[\sum_{k \neq j} x_k x_j]\right\}$$

where the quantity

$$\sum_{k \neq j} x_k x_j$$

represents all cross-products of two different measurements of x. Since x_k and x_j are independent for $k \neq j$, we have

$$E[x_k x_j] = E[x_k]E[x_j] = \mu \cdot \mu = \mu^2$$

Since there are n terms of the form x_k^2 and $n(n-1)$ cross-product terms of the form $x_k x_j$, we have

$$E[\bar{x}^2] = \frac{1}{n^2}\left\{n(\sigma^2 + \mu^2) + n(n-1)\mu^2\right\} = \frac{\sigma^2}{n} + \mu^2 \tag{18}$$

Hence, we find (finally!)

$$E[s^2] = \frac{n}{n-1}\left\{(\sigma^2 + \mu^2) - (\frac{\sigma^2}{n} + \mu^2)\right\} = \sigma^2$$

and the assertion is proved.

This proof also provides the justification for dividing by $n-1$, rather than n, when we calculate the sample variance. Qualitatively, when we calculate the sample variance s^2 using Eq. (15), the use of $\bar{x}$ as an estimate of μ in that expression will tend to somewhat reduce the magnitude of

$$\sum (x_k - \bar{x})^2$$

That is,

$$\sum (x_k - \bar{x})^2 < \sum (x_k - \mu)^2$$

Division by $n-1$ rather than n serves to compensate for this slight reduction.

2.3 The variance of the sample mean $\sigma_{\bar{x}}^2$ and its associated $\sigma_{\bar{x}}$

For a sample of n measurements x_k we have seen that $\bar{x}$ is the best estimate of the population mean μ. If the x_k are *normally distributed*, an additional single measurement will fall within $\bar{x} \pm s$ at approximately the 68 per cent level of confidence. This is the interpretation of the standard deviation s for the sample of n measurements.

If we take additional samples, of n measurements each, we expect to gather a collection of sample means that will be clustered about the population mean μ, but with a distribution that is *narrower* than the distribution of the individual measurements x_k. That is, we expect the variance of the sample means to be *less* than the population variance. The desired result may be succinctly stated for a sample of n measurements:

$$\sigma_{\bar{x}}^2 = \frac{1}{n}\sigma^2 \tag{19}$$

or, since one may estimate the value of σ^2 by calculating s^2

$$\sigma_{\bar{x}}^2 \approx \frac{1}{n} s^2 \tag{20}$$

The quantity $s/\sqrt{n}$ thus provides us with an estimate of the *standard deviation of the mean*. An experimental result is conventionally stated in the form shown in Eq. (4), namely[8]

$$\text{RESULT} = \bar{x} \pm \frac{s}{\sqrt{n}} \tag{21}$$

As an example, we look once again at the data sample of 100 measurements of the counting rate shown in Figure 2. Since for that sample we have $\bar{x} = 18.25$ and $s = 4.152$, we may express our measured counting rate in the form

$$\text{Counting rate} = 18.25 \pm \frac{4.152}{\sqrt{100}} = 18.25 \pm 0.42 \text{ counts/second.} \tag{22}$$

This result implies that if one were to take an additional sample of 100 measurements, there would be about a 68 per cent chance that this new sample mean would lie between 17.83 and 18.67 counts/second.

Note that it is meaningless to include more than two significant figures in the uncertainty, or more significant figures in the result (18.25 above) than are implied by the uncertainty.

A result so expressed thus allows us to compare our own experimental result with those of others. If the result stated in the form of Eq. (21) brackets, or overlaps, a similar result obtained elsewhere, we say that the two experimental results are in agreement. We have ignored, of course, any *systematic errors* that may be present in either measurement.

Equation (19) may be easily proved:

$$\sigma_{\bar{x}}^2 = E[(\bar{x} - \mu)^2] = E[\bar{x}^2 - 2\mu\bar{x} + \mu^2] = E[\bar{x}^2] - E[\mu^2]$$

From Eq. (18) we have

---

[8] Equations (4), (7) and (21) are not quite correct. Because of the non-normal distribution of the sample variance, it should be written

$$\text{RESULT} = \bar{x} \pm t_{n-1} s/\sqrt{n}$$

Where t_{n-1} is a constant called the Student *t*-factor ("Student" was the pseudonym of William Sealy Gosset, the Guiness Chief Brewer, 1876-1937). In the general case, t_{n-1} depends on the level of confidence chosen and the sample size *n*. If, as usual, we chose a confidence level of 68.27 per cent, t_{n-1} approaches 1.0 for large *n*, and is not much larger than 1.0 even for small *n*. In this course, our interest in t_ν is largely academic, and frequently, as in Eq. (21), we omit it. With more conservative confidence intervals such as 95 or 99 per cent, its use becomes more meaningful. Its use also arises in the fitting of data to a mathematical model, where confidence intervals on the estimates of parameters are desired. The computer programs we use for such data modeling include Student *t*-factors in the estimation of confidence intervals. For a complete discussion of the Student *t*-factors, see the book by Bennett and Franklin.

$$E[\bar{x}^2] = \frac{\sigma^2}{n} + \mu^2,$$

from which it follows that $\sigma_{\bar{x}}{}^2 = \sigma^2 / n$.

Finally, there is one additional point worth discussing: Suppose we measure a quantity *x* several times, or by several different methods, and for each measurement x_i we estimate its uncertainty σ_i. The σ_i are not necessarily equal; some of the measurements will be better than others, because of larger sample sizes (more repetitions) or because of other factors---like better apparatus. How do we determine our best estimate of *x*, and how do we find the uncertainty in that estimate?

For example, suppose one group measures a length *x* n_1 times to arrive at a value

$$\bar{x}_1 = \frac{1}{n_1}\sum_k x_k \qquad \text{with} \qquad \sigma_1^2 \approx \frac{1}{n_1} s^2$$

while a second group measures the same length *x* n_2 times to find

$$\bar{x}_2 = \frac{1}{n_2}\sum_j x_j \qquad \text{with} \qquad \sigma_2^2 \approx \frac{1}{n_2} s^2$$

(Here σ_1 is shorthand for $\sigma_{\bar{x}_1}$, *etc.*). How should $\bar{x}_1$ and $\bar{x}_2$ be combined to yield an overall $\bar{x}$, and what is the uncertainty in this final $\bar{x}$? Clearly

$$\begin{aligned}
\bar{x} &= \left[\frac{1}{n_1 + n_2}\right]\left[\sum_k x_k + \sum_j x_j\right] \\
&= \left[\frac{1}{n_1 + n_2}\right](n_1\bar{x}_1 + n_2\bar{x}_2) \\
&= \left[\frac{1}{1/\sigma_1^2 + 1/\sigma_2^2}\right]\left[\frac{\bar{x}_1}{\sigma_1^2} + \frac{\bar{x}_2}{\sigma_2^2}\right]
\end{aligned}$$

with

$$\sigma_{\bar{x}}^2 = \frac{s^2}{n_1 + n_2} = \frac{1}{\dfrac{1}{\sigma_1^2} + \dfrac{1}{\sigma_2^2}}$$

If there are *n* values of *x*, here is the generalized result, in a form that depends *only* on each x_k and its uncertainty σ_k:

$$\bar{x} = \frac{\sum_{k=1}^{n} x_k / \sigma_k^2}{\sum_{k=1}^{n} 1/\sigma_k^2}; \qquad \sigma_{\bar{x}}^2 = \frac{1}{\sum_{k=1}^{n} 1/\sigma_k^2} \tag{23}$$

Note how more measurements, or more accurate measurements, *reduce* the uncertainty by *increasing* its reciprocal. The results expressed in Eq. (21) may also be derived from a principal of maximum likelihood or a principle of least squares. An explicit example appears in chapter 4 of this manual.

2.4 The propagation of uncertainty through functional relationships

It frequently occurs that one wishes to determine the uncertainty in a quantity that is a function of one or more (independent) random variables. As we have seen, if we measure a counting rate *x*, we may express our result as $\bar{x} \pm \sigma_{\bar{x}}$. Suppose, however, we are interested in a quantity *u* that is proportional to the square of *x*, that is, $u = ax^2$, where *a* is some constant. What is the resulting uncertainty in *u*?

Using the concepts of differential calculus, one expects that if *x* fluctuates by an amount *dx*, then *u* will fluctuate by an amount $du = (\partial u / \partial x)dx = 2axdx$. In statistical terms, where the sign of the fluctuation is irrelevant, and if the fluctuations are not too large, one expects that

$$\sigma_u = \left|\frac{\partial u}{\partial x}\right| \sigma_x$$

and also, for the standard deviation of the mean,

$$\sigma_{\bar{u}} = \left|\frac{\partial u}{\partial x}\right| \sigma_{\bar{x}}$$

In each case, the derivative should be evaluated at the point $x = \bar{x}$. We may generalize this idea to include situations where *u* depends on more than one random variable: Suppose $u = f(x, y, \cdots)$, where $x, y, \cdots$ are random independent variables. Then

$$\sigma_u^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + \cdots \tag{24}$$

and also, for the variance of the mean,

$$\sigma_{\bar{u}}{}^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_{\bar{x}}{}^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_{\bar{y}}{}^2 + \cdots \tag{25}$$

Note that we sum the squares of the individual terms; this is appropriate when the variables $x, y, \cdots$ are *statistically independent*.

We illustrate the idea with an example taken from the Radioactivity experiment, one of the required experiments in the intermediate lab. In that experiment, the counter dead time τ may be estimated from a measurement of three counting rates, x, y and z. Here x is the counting rate from one source, y is the counting rate from a second source, and z is the counting rate from both sources simultaneously. For this illustration, we use a simple (albeit inaccurate) formula for the dead time:

$$\tau \approx \frac{x+y-z}{2xy} \tag{26}$$

From this expression, we can calculate

$$\frac{\partial \tau}{\partial x} = \frac{z-y}{2x^2 y} \qquad \frac{\partial \tau}{\partial y} = \frac{z-x}{2y^2 x} \qquad \frac{\partial \tau}{\partial z} = \frac{-1}{2xy} \tag{27}$$

In a particular experiment, the number of counts in one minute were measured for each of the three configurations, yielding

$$x = 55319 \quad \sigma_x = 235$$
$$y = 54938 \quad \sigma_y = 234$$
$$z = 86365 \quad \sigma_z = 294$$

where the units of all quantities are counts per minute.

At the point (x, y, z):

$$\frac{\partial \tau}{\partial x} = 9.35\times10^{-11} \quad \frac{\partial \tau}{\partial y} = 9.30\times10^{-11} \quad \frac{\partial \tau}{\partial z} = -16.5\times10^{-11}$$

with the units being min^2/count. Inserting these values into Eq. (24) yields

$$\sigma_t^2 = (9.35\times10^{-11})(235)^2 + (9.30\times10^{-11})(234)^2 + (-16.5\times10^{-11})(294)^2 = 3.31\times10^{-15}\ \text{min}^2$$

from which

$$\sigma_\tau = (3.31\times10^{-15})^{\frac{1}{2}} = 5.75\times10^{-8}\ \text{min} \approx 3.5\ \text{microseconds}$$

Using Eq. (26) to evaluate τ at the point (x, y, z) we find

$$\tau = \frac{x+y-z}{2xy} = 3.93\times10^{-6}\ \text{minutes} = 236\ \text{microseconds}$$

Hence, we may express the final result of this measurement of the dead time as

$$\tau = 236 \pm 3.5 \text{ microseconds}$$

It turns out that if we used the more accurate formula given in the experiment we would have obtained 300 microseconds instead of 236. These values may now be compared with the value obtained by measuring the dead time of the counter directly from the oscilloscope screen, which in this particular experiment was found to be about 220 microseconds, with an error of several tens of microseconds. The two results are thus found to be in rough agreement.

2.5 References

1. John R.Taylor, *An Introduction to Error Analysis* (University Science Books, 1982). This book is a good place to start. It includes almost all the material set forth in this chapter, but without some of the derivations and proofs.

2. Philip R.Bevington and D. Keith Robinson, *Data Reduction and Error Analysis for the Physical Sciences*, 2nd Ed. (McGraw-Hill, 1969). Bevington's book is comprehensive, and has long been a standard reference for physicists. It is, however, a little tedious.

3. Robley D. Evans, *The Atomic Nucleus* (McGraw-Hill, 1969). Here one may find clearly written sections relating to much of the preceding material. Moreover, it contains a good description of *Pearson's Chi-square Test*, which you may find useful for analyzing your data in the Radioactivity Experiment.

4. Carl A. Bennett and Norman L. Franklin, *Statistical Analysis in Chemistry and the Chemical Industry* (Wiley, 1954). Material for this chapter was gleaned from the first 60 or so pages of this humungous tome. The book is excellent, although it contains much more information than we need.

3. LEAST SQUARES: FITTING A CURVE TO DATA POINTS

3.1 An example to illustrate the motivation

We illustrate the method of the *least squares* fitting of a curve (here a straight line) to a set of data points by considering a classic experiment from introductory physics, in which a spring is hung from a rigid support, and a mass M is hung on the spring. If the mass is pulled down and released, the system will oscillate with a period T given by the expression

$$T = 2\pi(M/k)^{1/2}$$

where k is the spring constant.

The formula is correct if the spring has negligible mass. If the spring mass is not negligible, we can revise the simple formula to take the mass of the spring into account:

$$T = 2\pi[(M+m)/k]^{1/2} \tag{28}$$

where m is the effective mass of the spring. A somewhat questionable theoretical calculation shows that $m = M_s/3$, where M_s is the actual mass of the spring.[9]

We now think of doing an experiment to test this hypothesis, and so load the spring with a number (say N) of different masses, $M_1, M_2, \cdots, M_N$, and measure the associated periods $T_1, T_2, \cdots, T_N$.[10]We may use these N data points to test the theoretical relationship given by Eq. (28). We could plot our experimental points on a graph of T vs M, but if Eq.(28) holds, our experimental points would fall on a curved line, and it would be difficult to tell whether or not the functional form suggested by Eq. (28) actually describes the data. However if we plot our points on a graph of T^2 vs M, they should fall on a straight line if Eq. (28) is a good description of the system:

$$T^2 = \frac{4\pi^2}{k}M + \frac{4\pi^2 m}{k} = \alpha M + \beta \tag{29}$$

where $\alpha = 4\pi^2/k$ is the slope of the line and $\beta = 4\pi^2 m/k$ is its intercept with the T^2 axis. If the data do in fact fall on a straight line for this plot, we can estimate the slope and intercept of the line to provide estimates for the two parameters m and k.Note that for our example, the ratio of the

[9] This problem can be solved exactly. It turns out that a far better approximation is $m = (4/\pi^2)M_s$, which, incidentally, is exact as M goes to zero. This result also agrees well with the data presented here. (George Brown, private communication).

[10]We number our points from 1 to N, because it is traditional, natural, and notationally elegant. Such a vector is called a *unit-offset* vector. However in the C language, which we use to do the calculations we shall soon describe, a vector such as memory for a 9-dimensional vector it will have components $M[0]\cdots M[8]$, and you may well get garbage, or perhaps a "segmentation fault" if you ask for $M[9]$. While it is possible to twiddle the C code to handle unit-offset vectors and matrices, we do not do so. In referring to our programs you will just have to make the mental adjustment to shift the offset. We trust this will cause no confusion.

intercept to the slope of the line should equal m if the theory is correct. Figure 6 displays data from an actual experiment, in which a spring is suspended from a support and the oscillation period is measured for each of 9 different masses, ranging from 55 to 455 grams, which are hung on the spring. The mass of the spring itself was measured to be approximately 48.2 grams.

M(gm)	55	105	155	205	255	305	355	405	455
T(sec)	.496	.645	.761	.867	.957	1.037	1.113	1.194	1.254
T^2(sec)	.246	.416	.579	.752	.916	1.075	1.239	1.426	1.573

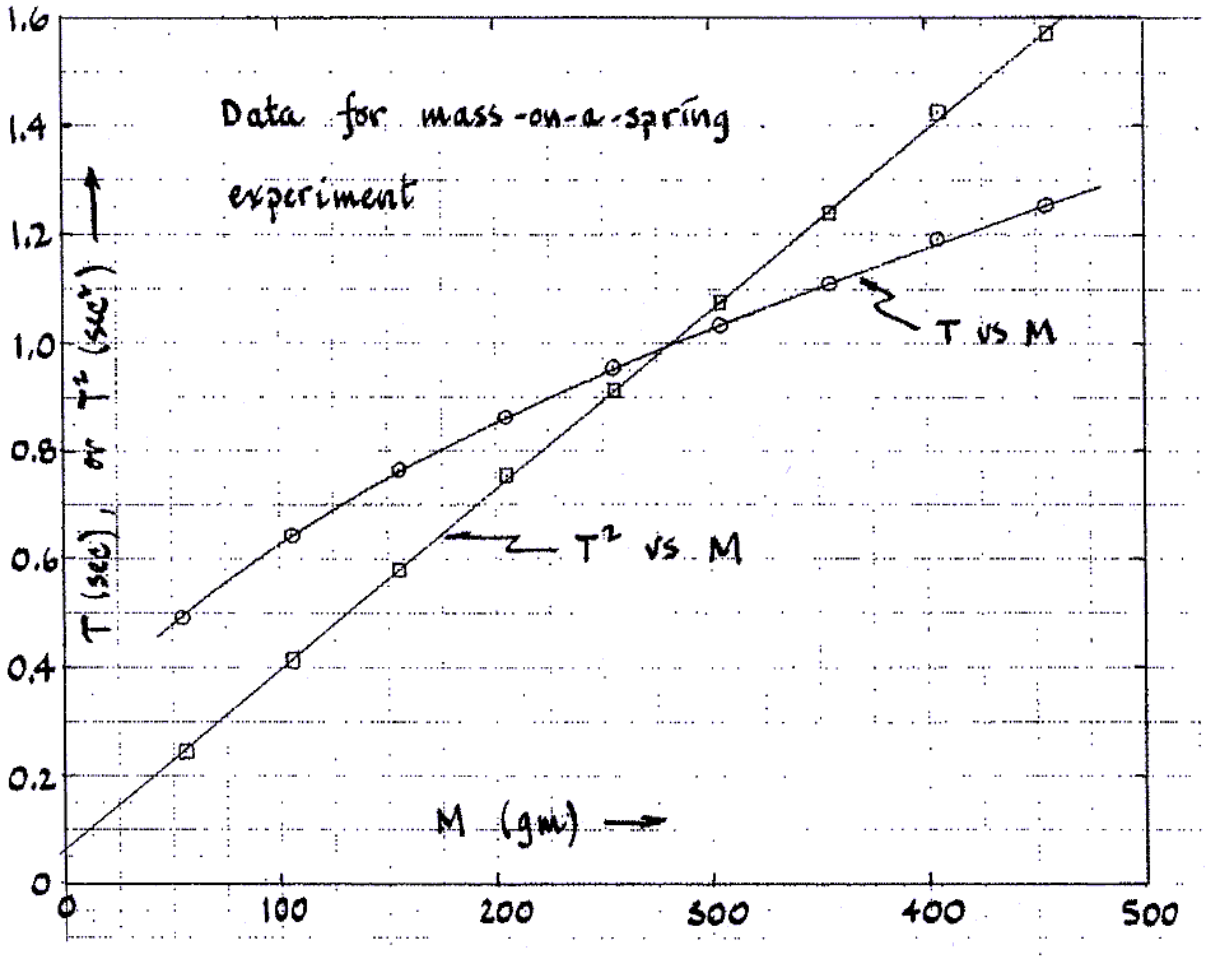

Figure 6. Data for mass on a spring.

Note that when we plot T_i vs M_i, the experimental points fall on a curve, but that when we plot T_i^2 vs M_i, the points appear to fall on a straight line, as we might expect from the theoretical hypothesis represented by Eq.(29). As we have mentioned, this straight line is much easier to analyze than the curve.

Now the question arises: What are the "best" values of m and k that can be determined from the data? Furthermore, once we have determined the "best" values for m and k, can we say that the data are consistent with the hypothesis that $m = M_s / 3$?

We can make some initial estimates by hand, just by drawing, by eye, a straight line through the data points (as shown on the graph of T^2 vs M), and reading off the slope and the intercept of this straight line. Thus, we can read an intercept of about $0.063\,\text{sec}^2$, and a slope of about $(1.390-0.063)/400 = 3.318\times10^{-3}\ \text{sec}^2/\text{gram}$, and hence a ratio of the intercept to the slope (this will be m) of about 19 grams. We can also make some rough estimates of how accurate these values are by shifting the straight edge on the graph while estimating (subjectively) whether the straight edge still goes through the points, and thus deduce that m is accurate to within a few grams.

We can, however, be still more quantitative, by making a *least squares* straight line fit to the data. Such a fit is also called a *linear regression* by the statisticians.

In the following section we discuss the general methods for fitting a straight line to a set of data points. Later on, we'll discuss the more general problem of fitting *any* hypothesized function to a set of data points. Although the going may get a little thick at times, we'll include all the details of the derivations, just to have all the relevant stuff right here in one place in our manual. For further reading, consult the references listed at the end of the chapter. Chapter 8 in the book by Taylor is a good place to start. The book of Numerical Recipes and the book by Bennett and Franklin provide more advanced and detailed discussion.

3.2 General discussion of the least squares method: Fitting a straight line

We consider the general problem of fitting a straight line, of the form $f(x) = \alpha x + \beta$, to a set of N data points (x_i, y_i), where i goes from 1 to N. Our hope is that y_i will be well approximated by $f(x_i)$. (In our example, $x_i = M_i$, and $y_i = T_i^2$, so we hope that T_i^2 will be well approximated by $\alpha M_i + \beta$).

We assume that the x_i are known exactly, and that for each x_i there is a normally distributed population of observations y_i, whose mean is $\alpha x_i + \beta$, and whose variance is σ_i^2.[11] If the x_i are not known exactly, but the y_i *are* known exactly, we can just reverse the roles of x_i and y_i. If neither x_i nor y_i are known exactly, a unique straight line fit is considerably more difficult.[12] In practice, we choose the independent variable x_i to be the most precisely measured quantity. Thus, in our example of the previous section, we shall take the masses M_i to be known exactly. The overall situation is shown graphically in Figure 7.

[11] In the ideal situation, independent estimates for σ_i^2 would be obtained from the data. For an experiment such as the one we describe here, this would require that several values of y_i be measured for each x_i, so that a sample variance in y may be calculated for each x_i. For some kinds of experiments, especially in particle physics, the y_i are numbers of counts, as registered, say, by a Geiger counter. If we make the reasonable assumption that such y_i obey Poisson statistics, an estimate of σ_i is available: $\sigma_i \approx \sqrt{y_i}$, since for a Poisson distribution the variance is equal to the mean. More commonly, an estimate of | is not available, in which case we must make some assumption about its value. Often it is assumed that all the σ_i are equal, so that each data point is considered to be weighted equally.

[12] See Reference 3 at the end of this chapter.

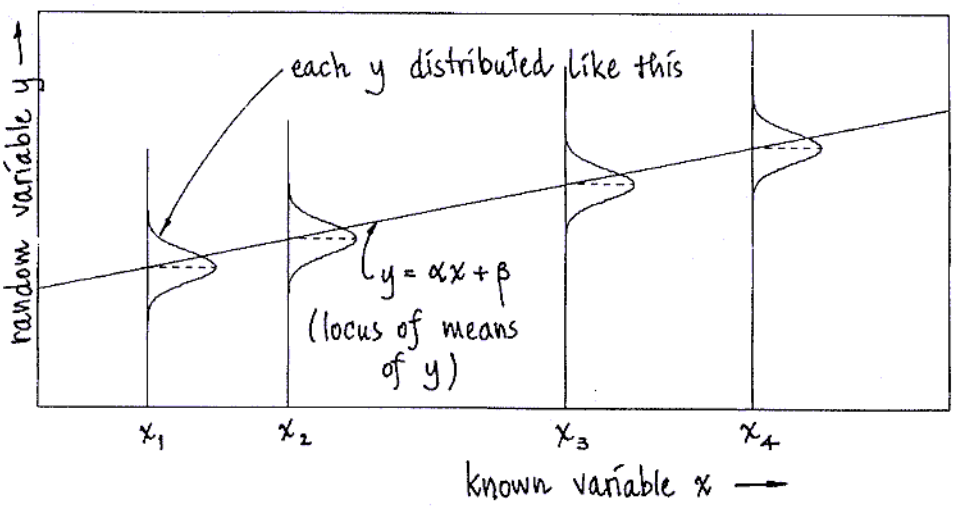

Figure 7. Graphical representation of the model for linear regression.

Now we hypothesize a straight line of the form

$$f(x) = ax + b \tag{30}$$

Our job is to determine values for a and b that are best estimates for the unknown population parameters α and β. To do this, we form the quantity Φ, the weighted sum of the squares of the residuals. It will be a function of a and b:

$$\Phi(a,b) = \sum_{i=1}^{N} w_i r_i^2 = \sum_{i=1}^{N} w_i [f(x_i) - y_i]^2 = \sum_{i=1}^{N} w_i (ax_i + b - y_i)^2 \tag{31}$$

Here $r_i \equiv f(x_i) - y_i$ is the *residual* for the i^{th} data point. The concept of a *residual* may be illustrated graphically as shown in Figure 8.

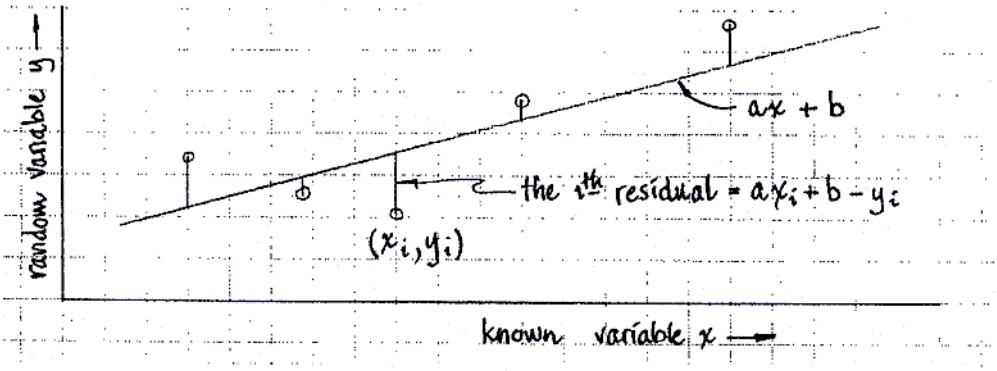

Figure 8. Graphical representation of residuals.

There are N such residuals.

w_i is the *weight* to be given to the i^{th} data point. It will be proportional to the inverse of the variance, $1/\sigma_i^2$, for that point. If w_i is taken to be *equal* to $1/\sigma_i^2$, where σ_i^2 is *independently* estimated for each point, the quantity Φ becomes equal to χ^2 (*chi-square*), a useful statistical quantity. For this reason a "least-squares fit" is sometimes called a "chi-square fit."

In practice, we shall often know only the *relative* weights, not having available independent estimates of the σ_i^2. If we know only that all σ_i are equal, we may take $w_i = 1$ for each point.

Note that Φ is a function of a and b, the parameters to be determined. The best values for these parameters will be those for which Φ is a minimum. This is the meaning of *least squares*. Moreover, the smaller Φ_{min} is, the better the fit.[13]

To minimize Φ, we differentiate Φ with respect to a and b, and set each of those derivatives equal to zero. That is, we set

$$\frac{\partial \Phi}{\partial a} = 0; \qquad \frac{\partial \Phi}{\partial b} = 0$$

Working out the derivatives, we find

$$\begin{aligned} X_2 a + X_1 b &= P \\ X_1 a + W b &= Y_1 \end{aligned} \tag{32}$$

where we represent relevant sums like this:

$$W \equiv \sum w_i \quad X_1 \equiv \sum w_i x_i \quad Y_1 \equiv \sum w_i y_i$$
$$X_2 \equiv \sum w_i x_i^2 \quad Y_2 \equiv \sum w_i y_i^2 \quad P \equiv \sum w_i x_i y_i$$

with all sums running from $i = 1$ to N. Equation (32) may be solved[14] for *a* and *b*:

$$a = \frac{1}{\Delta}(W \bullet P - X_1 Y_1 \quad b = \frac{1}{\Delta}(X_2 Y_1 - X_1 Y_2) \tag{33}$$

where Δ is the determinant of the coefficients

$$\Delta \equiv W \bullet X_2 - X_1^2$$

With these values of *a* and *b*, Φ phi assumes its minimum value:[15]

---

[13] The reasoning leading to the statement that the best fit is obtained by the method of "least squares" (*i.e.*, by minimizing chi-square) stems from the consideration of *maximum likelihood estimators*. Chapter 15 of *Numerical Recipes* and Appendix 5A of the book by Bennett and Franklin contain good discussions of this topic.

[14] As an exercise, try deriving Eq. (32) and solving to get Eq. (33).

[15] This compact expression for the minimum value of Φ can be deduced by writing Φ in the form

$$\Phi = \sum w_i (a x_i + b - y_i) r_i = a \sum w_i x_i r_i + b \sum w_i r_i - \sum w_i y_i r_i .$$

The first two sums on the right side of this equation will vanish by virtue of Eq. (32), leaving only the last sum, which is the right side of Eq. (34).

$$\Phi_{\min} = Y_2 - aP - bY_1 \tag{34}$$

We have thus determined values for *a* and *b*, the best estimate of the slope and intercept. We must now estimate the statistical uncertainties, or *confidence limits*, for each parameter. Without confidence limits our results are meaningless.

We have thus determined values for a and b, the best estimates of the slope and intercept. We must now estimate the statistical uncertainties, or *confidence limits* for each parameter. Without confidence limits our results are meaningless.

The idea is this: a and b are each functions of the y_i. Hence statistical fluctuations in the y_i will lead to statistical fluctuations in a and b. Later on we shall provide a concrete illustration of such fluctuations. For now, we simply note that if σ_i^2 is the variance of y_i, the variances of a and b will be, according to the theory described for the propagation of uncertainty in the preceding chapter,[16]

$$\sigma_a^2 = \sum\left(\frac{\partial a}{\partial y_i}\right)^2 \sigma_i^2 \qquad \text{and} \qquad \sigma_b^2 = \sum\left(\frac{\partial b}{\partial y_i}\right)^2 \sigma_i^2 \tag{35}$$

Now $\sigma_i^2 = C / w_i$, where C is some constant, and

$$\frac{\partial a}{\partial y_i} = \frac{w_i}{\Delta}(Wx_i - X_1); \qquad \frac{\partial b}{\partial y_i} = \frac{w_i}{\Delta}(X_2 - X_1 x_i)$$

This leads to

$$\sigma_a^2 = C\frac{W}{\Delta}; \qquad \sigma_b^2 = C\frac{X_2}{\Delta} \tag{36}$$

Now what is C? If we are fortunate enough to have available independent estimates s_i^2 of the y-variances σ_i^2, then $w_i \approx 1/s_i^2$, that is, $C \approx 1$. In that case our best estimates for σ_a^2 and σ_b^2 are just

$$s_a^2 = \frac{W}{\Delta}; \qquad s_b^2 = \frac{X_2}{\Delta} \tag{37}$$

The square roots of these quantities are estimates of the standard deviations of each of the parameters, providing confidence limits of approximately 68 per cent.

Furthermore, since it is known that $C \approx 1$, the value of $\Phi_{\min}$ is equal to a sample value of χ^2, a statistical quantity that may be used to test "goodness of fit", that is, whether the data may be

[16] Estimates of the variances σ_a^2 and σ_b^2 lead to "one-parameter" confidence limits---those discussed here. Such confidence limits ignore any correlation between a and b. Later on in this chapter we discuss the consequences of taking into account such correlations.

appropriately described by a straight line. For more details on the use of χ^2, see the following chapter.

More frequently, estimates of only the *relative* weights are available, in which case the best we can do is to estimate the value of C. We do this by *assuming* that our data are well-fit by the straight line, that is, we assume that χ^2 is equal to its mean value, the number of "degrees of freedom" or $N-2$.[17] Since $\Phi_{\min} = C\chi^2$, this means we simply replace C by $\Phi_{\min}/(N-2)$, so that our best estimates for σ_a^2 and σ_b^2 become

$$s_a^2 = \frac{W}{\Delta}\frac{\Phi_{\min}}{N-2} \qquad \text{and} \qquad s_b^2 = \frac{X_2}{\Delta}\frac{\Phi_{\min}}{N-2} \tag{38}$$

The square roots of the quantities given by either Eqns. (37) or (38) are estimates of the standard deviations of each of the parameters, which provide confidence limits of approximately 68 per cent. Thus our results may be stated succinctly:[18]

$$\text{SLOPE} = a \pm s_a; \qquad \text{INTERCEPT} = b \pm s_b \tag{39}$$

3.3 About the use of standard "Linear Regression" statistics routines

The usual "statistics" software routines available for use on personal computers do not usually calculate the standard deviations s_a and s_b of the slope and intercept of the straight line resulting from a "linear regression" calculation. Instead, they provide a calculation of either the *covariance* s_{xy} and/or the *correlation coefficient* $r = s_{xy}/s_x s_y$, where[19]

$$s_x^2 = \frac{1}{N-1}\sum (x_i - \bar{x})^2; \qquad s_y^2 = \frac{1}{N-1}\sum (y_i - \bar{y})^2$$

and

$$s_{xy} = \frac{1}{N-1}\sum (x_i - \bar{x})(y_i - \bar{y}).$$

Such a procedure is more general than that we have been considering, in that the *covariance* and the *correlation coefficient* may be calculated (and have meaning) even if the x_i are *not* known exactly.

However, if the x_i are assumed to be known exactly, s_a and s_b may be expressed in terms of the correlation coefficient r. Here are the appropriate equations:

---

[17] The number is N - 2 and not N because 2 degrees of freedom have been used up in the determination of a and b.

[18] The expressions given by Eq. (39) are not quite correct. To be correct, we should include the appropriate "Student" t-factor t_{N-2}, which depends on the number of data points and the confidence level chosen. Hence the correct forms for Eqns. (39) are: SLOPE = $a \pm t_{N\text{-}2}s_a$ and INTERCEPT = $b \pm t_{N\text{-}2}s_b$. The least-squares program fitline incorporates "Student" t-factors in this way.

[19] Weighting of points is implicit in these expressions.

$$s_a^2 = \frac{a^2}{N-2}\left(\frac{1}{r^2}-1\right); \qquad s_b^2 = \frac{X_2}{W}s_a^2 \tag{40}$$

with the correlation coefficient being

$$r \equiv \frac{s_{xy}}{s_x s_y} = \frac{WP - X_1 Y_1}{(WX_2 - X_1^2)^{1/2}(WY_2 - Y_1^2)^{1/2}}$$

Note that as $r \to \pm 1$, the correlation becomes exact, and s_a and s_b both $\to 0$. On the other hand, as $r \to 0$, y_i becomes uncorrelated with x_i, and both s_a / a and s_b / a increase without bound, as expected.

3.4 Continuing with our example, involving the mass on the spring

We start with the raw data, given in the table in Figure 6. We assume that σ_T^2, the variance in T, is independent of T, *i.e.*, the same for all the data points. Hence the variance in T^2 will *not* be the same for all data points, and if we want to perform a proper least-squares fit to the straight line of T^2 vs *M*, we should weight each point with its appropriate weight w_i.

(Actually, in the majority of cases any physicist will encounter, the fit obtained, assuming that all points are equally weighted, will not be very different from that in which correct weights are included. We do this example, however, including the correct weighting factors, to illustrate the method in detail).

To determine the w_i, let $y = T^2$. Then

$$\sigma_y^2 = \left(\frac{\partial y}{\partial T}\right)^2 \sigma_T^2$$

as seen from the preceding chapter. Now $\partial y / \partial T = 2T$, so that

$$\sigma_y^2 = 4T^2\sigma_T^2 \qquad \text{and} \qquad \frac{1}{\sigma_y^2} = \frac{1}{4T^2}\frac{1}{\sigma_T^2}$$

Now since w_i is proportional to $1/\sigma_y^2$, w_i will be proportional to $1/T_i^2$, and in fact may be taken equal to $1/T_i^2$, since for these data an estimate of σ_T^2 is not available. It is not difficult, using the computer program *fitline* to calculate the sums

$$W \equiv \sum w_i \quad X_1 \equiv \sum w_i x_i \quad Y_1 \equiv \sum w_i y_i$$
$$X_2 \equiv \sum w_i x_i^2 \quad Y_2 \equiv \sum w_i y_i^2 \quad P \equiv \sum w_i x_i y_i$$

and hence, using Eqs.(33) and (38), the values of a, b, s_a and s_b. The results, using the measured data from Figure 6 are

$$a = \text{SLOPE} = \frac{4\pi^2}{k} = (3.331 \pm 0.015) \times 10^{-3} \text{ sec}^2 / \text{gm}$$

and

$$b = \text{INTERCEPT} = \frac{4\pi^2 m}{k} = 0.0642 + -0.0032 \text{ sec}^2$$

These expressions for the slope and intercept, complete with the standard deviations s_a and s_b, imply that the "true" values of the slope and intercept have about a 68 per cent chance of falling within the specified limits, assuming that the experiments were carried out without any *systematic* error such as a mis-calibration of a meter. Now, what do these values of a and b imply for k and m? We note that $k = 4\pi^2 / a$. What is δk, the uncertainty in k, in terms of s_a?

Again we propagate uncertainty through the functional relationship to find

$$(\delta k)^2 = \left(\frac{\partial k}{\partial a} \right)^2 s_a^2,$$

so

$$\delta k = \frac{4\pi^2}{a^2} s_a$$

Hence we get the result:

$$k \pm \delta k = (1.1852 \pm 0.0053) \times 10^4 \text{ dynes/cm}$$

Similarly, since $m = b / a$,

$$(\delta m)^2 = \left(\frac{\partial m}{\partial a} \right)^2 s_a^2 + \left(\frac{\partial m}{\partial b} \right)^2 s_b^2 = \frac{b^2}{a^4} s_a^2 + \frac{1}{a^2} s_b^2$$

and hence

$$m \pm \delta m = 19.27 \pm 0.97 \text{grams}$$

Finally, we ask, are the data consistent with the hypothesis that $m = M_s / 3$? Since the measured $M_s \approx 48.2$ grams, $M_s / 3 \approx 16.07$ grams. This value is not bracketed by the uncertainty in our experimentally deduced value of m, and in fact is less than our deduced value of m by over three standard deviations. Therefore, one must conclude that the value of m derived from the data is inconsistent with the hypothesis that $m = M_s / 3$.

Either (a) the hypothesis is incorrect, or (b) there is a systematic error in the measurements. We might want to check the accuracy of our apparatus, or try the experiment again, or rethink the theoretical analysis.[20]

[20] This problem can be solved exactly. It turns out that a far better approximation is $m = (4/\pi^2)M_s$, which, incidentally, is exact as M goes to zero. This result also agrees well with the data presented here. (George Brown, private communication).

NOTE TO 133 STUDENTS: The following section, which deals with non-linear curve fitting, is not required for any data analysis in the Intermediate Laboratory, and therefore is optional reading. We will use it, however, in the Advanced Laboratory (Physics 134).

3.5 Fitting curves nonlinear in the parameters: the Marquardt algorithm

The *least-squares* method is not limited to fitting a straight line to a set of data points. The method may be generalized to fit either (a) a linear combination of *any* K specified functions of x, or (b) *any* function of x and a specified number of parameters, even one nonlinear in the parameters. The former may be accomplished in one step through the simple inversion of a $K \times K$ square matrix, while the latter requires an iterative technique, one that we describe below.

In practice, an iterative method may be used for both (a) and (b), and the particular method we shall describe---the Marquardt algorithm---is widely used as a technique for modeling data in a broad range of situations.

For example, when silver is irradiated with neutrons, the resulting intensity of radioactivity $I(t)$ decreases with time according to the expression

$$I(t) = \beta_1 + \beta_2 e^{-\beta_3 t} + \beta_4 e^{-\beta_5 t} \tag{41}$$

Although $I(t)$ is linear in β_1, β_2 and β_4, it is nonlinear in both β_3 and β_5, and there is no way to transform this function into any two-parameter straight line. It is possible, however, to use the *least squares* technique to provide estimates for each of the five parameters, along with estimates for the uncertainty in each.

Other examples encountered in both the intermediate and advanced laboratory include the transient oscillatory decay that arises in the Cavendish experiment, and the determination of resonance curves or line shapes that may be encountered in other experiments.

The method we describe is used in a computer program called textfit that we wrote a few years ago. The fit program, written in the C language, contains functions (or subroutines) that can be used elsewhere. For example, there is a nice function that inverts a matrix, and another that can be used to make a graph of any specified function.

3.5.1 The general idea

Suppose we have a set of N data points with coordinates (x_i, y_i), that we wish to describe using a function

$$y = f(x, \beta_1, \beta_2, \cdots, \beta_K),$$

an expression we'll often write as

$$y = f(x, \boldsymbol{\beta}) \tag{42}$$

Here x is the independent variable whose values are presumed precisely known, and the β_j are the K parameters whose values we desire. $\boldsymbol{\beta}$ can be thought of as a K-dimensional vector.

Our goal is to adjust the values of these parameters so that y_i is well-approximated by $f(x_i, \mathbf{b})$, where $\mathbf{b}$ is our best estimate of the desired parameter vector $\boldsymbol{\beta}$.

To do this we minimize the quantity Φ ---the weighted sum of the squares of the residuals---just as we did in fitting a straight line to a set of data points. Φ will be a function of the b_j:

$$\Phi(\mathbf{b}) = \sum_{i=1}^{N} w_i r_i^2 = \sum_{i=1}^{N} w_i [f(x_i, \mathbf{b}) - y_i]^2 = \sum_{i=1}^{N} w_i (f_i - y_i)^2 \tag{43}$$

Here f_i is an abbreviation for $f(x_i, \mathbf{b})$. As with our fitting of a straight line, $r_i \equiv f_i - y_i$ is the i^{th} residual, and w_i is the weight of the i^{th} point, a number that is ideally equal to, or at least proportional to $1/\sigma_i^2$, the inverse of the observed y --variance of that data point.

The best values of the b_j will be obtained when we have found the minimum value for $\Phi(\mathbf{b})$.[21]This will happen when

$$\frac{\partial \Phi}{\partial b_j} = 0 \tag{44}$$

for each b_j. If f is a linear function of the parameters b_j, the problem of minimizing Φ is straightforward. For example, in fitting a straight line to a set of data points there are two parameters, the intercept and the slope (b_1 and b_2). That is, $f(x, \mathbf{b}) = b_1 + b_2 x$. As described earlier, Φ may be minimized to find the best straight line by solving the two simultaneous equations represented by Eq. (41). For this case, Φ, plotted as a function of b_1 and b_2, will have the shape of a paraboloid, and contours of constant Φ will be ellipses. If we take ϵ_1 and ϵ_2 to be the excursions of b_1 and b_2 from their best values, Φ has this form:

$$\Phi = \Phi_{\min} + (\sum w_i)\epsilon_1^{\,2} + 2(\sum w_i x_i)\epsilon_1\epsilon_2 + (\sum w_i x_i^2)\epsilon_2^{\,2}$$

In the general case, where $f(x, \mathbf{b})$ is a nonlinear function of the b_j, we can expect that near the minimum, Φ will also be at least approximately parabolic, that is, that Φ will have the approximate form

$$\Phi = \Phi_{\min} + \sum_j \sum_k A_{jk} \epsilon_j \epsilon_k \tag{45}$$

where the A_{jk} are constants:

$$A_{jk} = \frac{1}{2} \frac{\partial^2 \Phi}{\partial b_j \partial b_k} \qquad (\text{evaluated at } \Phi_{\min}) \tag{46}$$

[21] There is no guarantee that Φ will have only one minimum. Our solution may not be unique.

The main problem is to find the values of the b_j that will minimize Φ. There are a number of methods for doing this; we'll describe three. They are called (a) the Taylor expansion, or Newton method, (b) the gradient, or steepest descent method, and (c) the Marquardt method, which combines the Taylor expansion and gradient methods, making use of the best features of each.

Useful references include a Bell Labs memo regarding the *nllsq* program by Kornblit[22], Marquardt's original paper[23], and the more general references listed at the end of the chapter.

3.5.2 The Taylor expansion method

The Taylor expansion, or Newton method is similar to the Newton method for finding the roots of a function. If we expand the function $f(x_i)$ in the vicinity of an arbitrary point $\mathbf{b}$ in parameter space, we obtain a linear approximation to f :

$$f(x_i, \mathbf{b}+\boldsymbol{\delta}) \approx f(x_i, \mathbf{b}) + \sum_{j=1}^{K} \frac{\partial f_i}{\partial b_j}\delta_j = f(x_i, \mathbf{b}) + \sum_{j=1}^{K} p_{ij}\delta_j \tag{47}$$

Here δ_j is a Taylor expansion type increment in the parameter vector. The derivatives $\partial f_i / \partial b_j$, which we abbreviate by p_{ij}, are evaluated at the point $\mathbf{b}$. If the chosen point $\mathbf{b}$ is sufficiently close to the desired final $\mathbf{b}$ vector, the right side of Eq.(44) will be a reasonable approximation to f near the minimum of Φ.

Using this linear approximation, we can proceed to minimize Φ using straightforward methods:

$$\frac{\partial \Phi}{\partial \delta_j} = \frac{\partial}{\partial \delta_j}\sum w_i r_i^2 = 2\sum w_i r_i \frac{\partial r_i}{\partial \delta_j} = 2\sum w_i r_i p_{ij} \tag{48}$$

since $\partial r_i / \partial \delta_j = \partial f_i / \partial \delta_j = \partial f_i / \partial b_j = p_{ij}$. Hence, we have, since we want $\partial \Phi / \partial \delta_j = 0$,

$$\sum_i w_i r_i p_{ij} = \sum_i w_i (f_i + \sum_k p_{ik}\delta_k - y_i) p_{ij} = 0$$

or

$$\sum_i \sum_k w_i p_{ij} p_{ik} \delta_k = -\sum_i w_i (f_i - y_i) p_{ij} \tag{49}$$

Equation (49) represents a set of K linear equations that may be solved for the increments δ_j. The left side of this equation may be put in simple form by writing

---

[22] A. Kornblit, *nllsq -- Non Linear Least Square Fit in C*, Bell Laboratories Technical Memorandum (1979). This is not officially published, but a copy is available in the lab.

[23] D. W. Marquardt, *An Algorithm for Least-Squares Estimation of Nonlinear Parameters*, J. Soc. Indust. Appl. Math. **11**, 431 (1963).

$$A_{jk} = \sum_i w_i p_{ij} p_{ik}$$

Note that

$$A_{jk} = \frac{1}{2} \frac{\partial^2 \Phi}{\partial \delta_k \partial \delta_j}$$

as can be seen by differentiating Eq. (48) with respect to δ_k.[24] A_{jk} is sometimes called the curvature matrix, useful for thinking about the shape of Φ, particularly near $\Phi_{\min}$. See Eqs. (45) and (46).

The right side of Eq. (49) is $-(1/2)(\partial\Phi / \partial b_j)$, that is, half of the negative gradient vector component of the Φ surface, calculated at $\boldsymbol{\delta} = 0$. Denoting this quantity by $-g_j$, we see that Eq. (49) may be written

$$\sum_k A_{jk} \delta_k = -g_j \tag{50}$$

or in truly shorthand notation,

$$\mathbf{A}\boldsymbol{\delta} = -\mathbf{g}$$

Here $\mathbf{A}$ (the curvature matrix) is a K by K square symmetric matrix, while $\boldsymbol{\delta}$ and $\mathbf{g}$ are K-dimensional vectors. If $\mathbf{A}$ is nonsingular we can solve for $\boldsymbol{\delta}$ by inverting $\mathbf{A}$:

$$\boldsymbol{\delta} = -\mathbf{A}^{-1}\mathbf{g} \tag{51}$$

With luck, the δ_j, when added to the initial values of the parameters b_j, will produce a new set of parameters that lie closer to the point in b-space where Φ is a minimum. If this is the case, the process can be repeated (iterated) until a desired accuracy is achieved. If our starting point in parameter space is close to the point where Φ takes on its minimum value, this process will converge rapidly to the desired solution. On the other hand, if our initial guesses for the parameters are too far from the minimum of Φ, we may be led off to a never-never land, and the process will fail, leading to new points that actually increase the value of Φ. An alternative method, one that ensures that we find a new point for which Φ decreases, is the gradient method, which we describe next.

3.5.3 The gradient method

In the gradient method, we simply use a correction vector whose components are proportional to the components of the negative gradient vector:

---

[24] We neglect the term

$$2\sum_i w_i r_i \ (\partial^2 r_i / \partial \delta_k \partial \delta_j).$$

It vanishes when f_i is linear in the δ_j, as it surely will be near $\Phi_{\min}$.

$$\delta_k = -\alpha_k g_k$$

where α_k is a constant of proportionality. Note that α_k will depend on which parameter is being corrected, that is, on k. How large should α_k be? If it's too large, the correction will overshoot the minimum of Φ. If it's too small, the approach to $\Phi_{\min}$ is slow. Here's one way to think about what "too large" or "too small" mean: Note that δ_k has dimensions of b_k whereas g_k has dimensions of $1/b_k$. Therefore α_k has dimensions of b_k^2. The dimensions of $1/A_{kk}$ are also b_k^2, so we are led to write

$$\delta_k = -\frac{1}{\lambda A_{kk}} g_k \tag{52}$$

Here λ is a dimensionless factor, independent of k. (The reason for putting λ in the denominator will become clear shortly.)

Now A_{kk} is the curvature of the Φ surface in the k direction. If A_{kk} is small, we take a big step, whereas if it's big, we take a small step, which is what we want. The dimensionless parameter λ can be adjusted to achieve the optimum step length, that is, optimum convergence.

The gradient method will work well if we are so far from $\Phi_{\min}$ that the Taylor expansion method fails. Near $\Phi_{\min}$, however, the gradient method generally converges much more slowly than the Taylor expansion method since $\mathbf{g}$, and hence the correction vector that's proportional to it, becomes vanishingly small there.

3.5.4 The Marquardt method

The Marquardt method combines the best features of each of the above methods, emphasizing the gradient method at first if necessary, then switching to the Taylor expansion method as the minimum of Φ is approached. To achieve this, Marquardt combines the two, in a manner that we will explain. As we shall see, the factor λ will play a key role.

Equation (52) may be written

$$\lambda A_{jj}\delta_j = -g_k \qquad \text{or} \qquad \lambda \sum_k A_{jk}\delta_{jk}\delta_k = -g_j \tag{53}$$

where $\delta_{jk} = 1$ if $j = k$ and 0 otherwise. Marquardt combines Eq. (50) with Eq. (53) by adding the two together and ignoring the factor of 2:

$$\sum_k (1 + \lambda\delta_{jk}) A_{jk}\delta_k = -g_j \tag{54}$$

Given a value of λ, Eq. (54) may be solved for a correction vector $\boldsymbol{\delta}$.

As λ is decreased to 0 from a value much larger than 1, this $\boldsymbol{\delta}$ changes smoothly from a gradient-type to a Taylor expansion-type correction vector, since for $\lambda \gg 1$, Eq. (54) reduces to Eq. (53), while for $\lambda = 0$, Eq. (54) reduces to Eq. (50). Thus we may use a large value of λ to start if we are far from $\Phi_{\min}$, then reduce λ with each iteration, expecting Φ to decrease as we approach $\Phi_{\min}$. That is, with each iteration we solve for a correction vector $\boldsymbol{\delta}$, add it to $\mathbf{b}$ to produce a new vector $\mathbf{b}$ that is closer (we hope) to the desired parameter vector, then decrease λ and repeat the process. Eventually we hope that λ will be decreased to such a small value that we are in the Taylor expansion regime, so that convergence will be rapid. We continue this iterative process until the application of a suitable convergence test shows that we have achieved the optimum set of parameters.

In the following sections we describe the finer details of the Marquardt algorithm---scaling, testing for convergence, determining confidence limits for the parameters, and determining to what extent the parameters are correlated with each other.

3.5.5 The finer details

Scaling. In performing the computation it is useful to scale Eq. (54) so as to eliminate the dimensions. As we noted above in our discussion of the gradient method, the diagonal elements of the curvature matrix provide a natural scale for this problem. Thus, we are led to write for a scaled matrix element A^*_{jk},

$$A^*_{jk} \equiv \frac{A_{jk}}{\sqrt{A_{jj}}\sqrt{A_{kk}}} \tag{55}$$

Note that the diagonal elements of the scaled curvature matrix $\mathbf{A}^*$ are each equal to 1.

Substituting Eq. (55) into Eq. we obtain

$$\sum_k \sqrt{A_{jj}}\sqrt{A_{kk}}(1+\lambda\delta_{jk})A^*_{jk}\delta_k = -g_j$$

or, dividing both sides by $\sqrt{A_{jj}}$, and setting $A^*_{jj} = 1$ in the second term:

$$\sum_k (1+\lambda\delta_{jk})A^*_{jk}\delta^*_k = \sum_k (A^*_{jk} + \lambda\delta_{jk})\delta^*_k = -g^*_j \tag{56}$$

Here the $\delta^*_k = \sqrt{A_{kk}}\delta_k$ form the scaled correction vector, while $g^*_j = g_j / \sqrt{A_{jj}}$ is the scaled gradient vector. Equation (56), which may also be written in compact form as

$$(\mathbf{A}^* + \lambda\mathbf{I})\boldsymbol{\delta}^* = -\mathbf{g}^* \tag{57}$$

may be solved for the scaled correction vector $\boldsymbol{\delta}^*$ by inverting the matrix $\mathbf{A}^* + \lambda\mathbf{I}$. Then each component of the correction vector is *unscaled*:

$$\delta_k = \frac{\delta_k^*}{\sqrt{A_{kk}}} \tag{58}$$

δ_k is then added to the appropriate component of the parameter vector at the n^{th} iteration to obtain a new value for this component:

$$b_k^{(n+1)} = b_k^n + \delta_k \tag{59}$$

We expect this new set of parameters to be one for which the fit is improved, *i.e.*, one yielding a smaller value for Φ.

Given an initial guess for the parameter vector $\mathbf{b}$, we may summarize the steps of the algorithm so far as follows:

1. Compute the initial $\Phi(\mathbf{b})$.

2. Pick a modest value for λ, say $\lambda = 10^{-1}$.

3. Solve Eq. (57) and use Eq. (58) to obtain the correction vector $\boldsymbol{\delta}$ and compute $\Phi(\mathbf{b}+\boldsymbol{\delta})$.

4. If $\Phi(\mathbf{b}+\boldsymbol{\delta}) >= \Phi(\mathbf{b})$, *increase* λ by a factor of 10 (or any other substantial factor) and return to step 3.

5. If $\Phi(\mathbf{b}+\boldsymbol{\delta}) < \Phi(\mathbf{b})$, test for convergence (see below). If convergence is not yet achieved, *decrease* λ by a factor of 10, calculate a new (scaled) $\mathbf{A}$ matrix, and return to step 3.

6. If convergence is achieved, set $\lambda = 0$ and recalculate the (scaled) $\mathbf{A}$ matrix, which will be useful in determining the uncertainties in the final parameter estimates, along with correlations among the parameters, as described below.

Testing for convergence. At each iteration, it is necessary to test whether convergence has been achieved. In general, it is inadvisable to iterate until the machine round off limit is reached, since this would lead to results containing many more significant figures than are implied by ordinary data and is therefore inelegant.

Furthermore, it is fairly common, as one nears the minimum of Φ, to find parameters wandering around in small steps, searching for an ill-defined minimum in a flat valley of complicated topology. This is particularly likely when there are large correlations among the parameters.

Thus, it is necessary to establish appropriate criteria for stopping. The simplest (and crudest) is just to stop after some preset number of iterations. Of course, then we won't know, except by monitoring how the parameters change at each iteration, whether convergence has been achieved. Nevertheless it is advisable to set some maximum number of iterations, just in case no ordinary

convergence is reached. An appropriate number will depend on the particular problem at hand, but a number on the order of 20 or 30 is usually suitable.

In normal situations, we should stop when the changes in the parameters (the δ_j) are smaller than some specified value. Marquardt suggests stopping whenever

$$\left|\delta_j^*\right| < \epsilon\left(\tau + \left|b_j^*\right|\right) \qquad \text{for all } j \tag{60}$$

where ϵ and τ are constants. Suitable values might be $\epsilon = 10^{-5}$ and $\tau = 1.0$. The constant τ is there to take care of situations where a final parameter value might be close to zero. Kornblit calls this the "epsilon test" Scaled values of δ_j and b_j are used because they are dimensionless and likely to be of comparable magnitude.

Sometimes, particularly when the parameters are highly correlated so that the Φ surface in the vicinity of the minimum is quite flat, it may be found that λ will increase to values larger than would seem necessary to ensure a decreasing Φ. It is possible then that the correction vector is too large. In such situations it is helpful to monitor the angle γ between the scaled correction vector $\boldsymbol{\delta}^*$ and the scaled gradient vector $\mathbf{g}^*$. This angle can be calculated:

$$\gamma = \cos^{-1}\left(\frac{\mathbf{g}^* \cdot \boldsymbol{\delta}^*}{\left|\mathbf{g}^*\right|\left|\boldsymbol{\delta}^*\right|}\right)$$

The technique is to increase λ until γ is less than some chosen value (typically 45 degrees). This can always be achieved since γ will decrease monotonically toward zero as λ increases. Then we do not increase λ further, but halve the correction vector, replacing Eq. (59) by

$$\mathbf{b}^{(n+1)} = \mathbf{b}^{(n)} + \frac{1}{2}\boldsymbol{\delta}^{(n)} \tag{61}$$

We continue to halve the correction vector until either the epsilon test is passed or Φ decreases. Kornblit calls this the "gamma-epsilon" test.

Finally, there are two tests for non-convergence. The first is an attempt to deal with a singular Marquardt matrix $\mathbf{A}^* + \lambda\mathbf{I}$. If λ is too small and the parameters are too highly correlated this matrix may be judged singular by the "fit" program. If this is found, λ is automatically increased by a factor of 10 and the matrix recalculated. This will happen up to five times (resulting in a possible increase in λ by a factor of up to 10^5) before the program gives up in disgust.

The second is called the "gamma-lambda" test, which causes the program to stop if γ cannot be reduced to less than 90 degrees even with a λ of 10. The completion of this test is also an indication that the parameters are too highly correlated for a solution to be found.

It is not uncommon to find that the parameter values just grow larger without bound, as Φ decreases (usually slowly) toward an imagined minimum in outer space. If this is the case the searching will probably continue until the maximum number of iterations is reached, but with no meaningful results. The most likely cause for such behavior is that poor initial guesses have been made for the parameters.

Confidence Limits. The simplest and most often used method for estimating confidence limits for the parameters---the *one-parameter* confidence limits---is identical to that outlined earlier for the fitting of a straight line. Thus to estimate the variance in the j^{th} parameter we start with an expression analogous to Eq. (35):

$$\sigma_{b_j}^2 = \sum_i \left(\frac{\partial b_j}{\partial y_i} \right)^2 \sigma_i^2 \tag{62}$$

where b_j is the optimum value of the j^{th} parameter, and σ_i^2 is, as before, the y -variance at the i^{th} point, with σ_i^2 being inversely proportional to w_i: $\sigma_i^2 = C / w_i$, where C is a constant.

We need to calculate $\partial b_j / \partial y_i$ at the point where Φ is a minimum, that is, where $\mathbf{g} = 0$, or where for each m

$$g_m = \sum_k w_k (f_k - y_k) p_{km} = 0$$

To find $\partial b_j / \partial y_i$ we differentiate this equation with respect to y_i to get

$$\sum_k w_k p_{km} \sum_n \frac{\partial f_k}{\partial b_n} \frac{\partial b_n}{\partial y_i} - w_i p_{im} = 0$$

Thus

$$\sum_n \left(\sum_k w_k p_{kn} p_{km} \right) \frac{\partial b_n}{\partial y_i} = w_i p_{im}$$

or

$$\sum_n A_{nm} \frac{\partial b_n}{\partial y_i} = w_i p_{im}$$

which may be solved for $\partial b_j / \partial y_i$ by inverting (once again) the curvature matrix $\mathbf{A}$:

$$\frac{\partial b_j}{\partial y_i} = w_i \sum_m A_{jm}^{-1} p_{im}$$

where A_{jm}^{-1} is the jm^{th} element of $\mathbf{A}^{-1}$. Squaring this expression, multiplying by C / w_i, and summing over i yields

$$\sigma_{b_j}^2 = CA_{jj}^{-1} \tag{63}$$

If independent estimates s_i^2 of the y --variances σ_i^2 are available,[25] then $w_i \approx 1/s_i^2$, $C \approx 1$, so $s_{b_j}^2 = A_{jj}^{-1}$ is a good estimate for the variance of the j^{th} parameter b_j, and $s_{b_j} = (A_{jj}^{-1})^{\frac{1}{2}}$ is a good estimate of the standard deviation of the j^{th} parameter.

In addition, if estimates of the y -variances are available, the obtained value of $\Phi_{\min}$ becomes a sample value of χ^2 and may be used to test "goodness of fit."

If estimates of the σ_i^2 are not available (a frequent situation), then we are reduced, as described earlier for the fitting of a straight line to a set of points, to assuming that the fit is ideal, so that χ^2 is assumed to equal its mean value of $N-K$, the number of degrees of freedom. (In this case, any desire to test "goodness of fit" must be abandoned, since a sample χ^2 is *not* available.) Thus, we use $\Phi_{\min}/(N-K)$ as an estimate for C, so that our estimate of the standard deviation of the j^{th} parameter becomes

$$s_{b_j} = \left(\frac{\Phi_{\min}}{N-K} A_{jj}^{-1} \right)^{\frac{1}{2}} \tag{64}$$

To estimate the uncertainty in b_j, we multiply s_{b_j} by the appropriate "Student" t-factor, as explained earlier for the fitting of a straight line. Thus, if we have b_j as an estimate of the parameter whose "true" value is β_j, we expect that with a probability of the chosen confidence level,

$$b_j - t_{N-K} s_{b_j} \le \beta_j \le b_j + t_{N-K} s_{b_j} \tag{65}$$

where t_{N-K} is the appropriate "Student" t-factor for $N-K$ degrees of freedom and the chosen confidence level. For a 68.3 per cent confidence level, this factor is approximately 1, and so is often omitted.[26]

A different kind of confidence limit is one that defines the *joint confidence region*, that is, the region in parameter space within which *all* of the parameters will lie with some probability, say 68.3 per cent. Such a region will have a shape that is approximated by a K -dimensional ellipsoid, a surface of constant $\Phi > \Phi_{\min}$.

To envision why this is appropriate, imagine that the experiment yielding our data sample is repeated many, many times, so that we have a very large number of data samples. Each sample will produce a distinct parameter set, and hence a distinct point in parameter space. Thus, the collection

---

[25] For many measurement situations such estimates are not available. However if the y_i consist of *counts* drawn from a Poisson distribution, such as might be obtained using a Geiger counter, y_i itself is an estimate of σ_i^2, so that $1/y_i$ is an absolute estimate of w_i. For the least-squares theory described in this chapter, which assumes the y_i are drawn from a *normal*, or *Gaussian* distribution, y_i must be large enough so that the Poisson distribution may be assumed Gaussian. A common rule-of-thumb is to ensure that y_i is greater than or equal to 10.

[26] Appropriate "Student" t-factors are calculated and used in our computer programs fit and fitline.

of repeated experiments will produce a cluster of points, one that we expect will be roughly ellipsoidal in shape, and we can arrange an ellipsoidal surface that will contain some specified fraction of the points. Its size will depend on the number of degrees of freedom and the desired confidence level. The interior of the ellipsoid so chosen is the *joint confidence region*. Statistics wizard G. E. P. Box has shown[27] that if the fit is assumed ideal (so that χ^2 is assumed equal to its mean value) the best estimate of the boundary of the joint confidence region is given by

$$\Phi = \Phi_{\min}[1 + \frac{K}{N-K} F_p(K, N-K)] \tag{66}$$

where $F_p(K, N-K)$ is the upper p per cent point of the F-distribution with K and $N-K$ degrees of freedom. p is the desired confidence level, say 68.3 per cent or 95 percent. Thus, if there are 2 parameters and 9 data points and we choose a 68.3 per cent confidence level, the boundary in parameter space of the joint confidence region is obtained by increasing Φ to $[1+(2/7)F_{0.683}(2,7)] \approx 1.39$ times its minimum value.

If the chosen function is linear in the parameters this boundary will have the shape of a K-dimensional ellipsoid. If the chosen function is nonlinear in the parameters the boundary will be only approximately ellipsoidal in shape; its actual form may be computed if desired.

The projection of the joint confidence region onto each of the parameter axes defines what is called the *support plane*. For an ellipsoidal region defined by Eq. (66), such a projection is given by

$$\Delta b_j = [K \cdot F_p(K, N-K)]^{\frac{1}{2}} s_{b_j} \tag{67}$$

and is called the *support plane error* for the j^{th} parameter.

The diagram shown in Fig. Figure 9 illustrates our discussion. To create this figure we have simulated 200 repetitions of the experiment described at the beginning of this chapter,[28] using our best (non-independently determined) estimates for the σ_i^2, the variance in $y_i = T_i^2$ for the i^{th} data point. Such a simulation is called a *Monte Carlo* simulation. Each simulated repetition produces an intercept-slope pair (b_1, b_2), a point in the 2-dimensional parameter space. Note that the cluster of points has an elliptical shape, as expected.[29] The degree to which the ellipse is skewed depends on the correlation between b_1 and b_2. If there were no correlation, the axes of the ellipse would coincide with the b_1 and b_2 axes. In our example this would happen only if $X_1 = \sum w_i x_i$ were zero, that is, if the curvature matrix $\mathbf{A}$ were diagonal.

---

[27] G. E. P. Box, in the two-volume set, *The Collected Works of George E. P. Box*, edited by George C. Tiao (Wadsworth, Inc., Belmont, Calif., 1985). See especially *The Experimental Study of Physical Mechanisms*, Vol. I, pp. 137-156, and *Application of Digital Computers in the Exploration of Functional Relationships*, Vol. II, pp. 381-388.

[28] Of course, the methods used for fitting nonlinear functions can also be used for fitting straight lines. No generality is lost by using this simple example to illustrate confidence limits.

[29] If there were 3 parameters, the cluster of points would have a 3-dimensional ellipsoidal shape (easily visualized in 3-space), while more parameters would result in an ellipsoidal cluster of still higher dimensions, not so easy to visualize.

Figure 10 shows a three-dimensional plot of Φ vs b_1 and b_2 in the vicinity of $\Phi_{\min}$, for the same example as that shown in Figure 9. Regularly spaced contours of constant Φ are shown projected onto the $b_1 - b_2$ plane; the innermost ellipse is approximately that shown in Figure 9, giving the boundary of the joint confidence region.

It is helpful to realize that contours of constant χ^2 play a definitive role in our discussion. The 68.3 per cent one-parameter confidence limits are roughly determined by projecting the contour that results from increasing χ^2 by 1.0 above $\chi^2_{\min}$ onto the parameter axes, whereas the joint confidence region is roughly determined by the contour resulting from an increase of χ^2 by an amount depending on the number of parameters. For two parameters this is approximately 2.70. Further discussion of this approach is contained in the Book of Numerical Recipes cited at the end of this chapter.

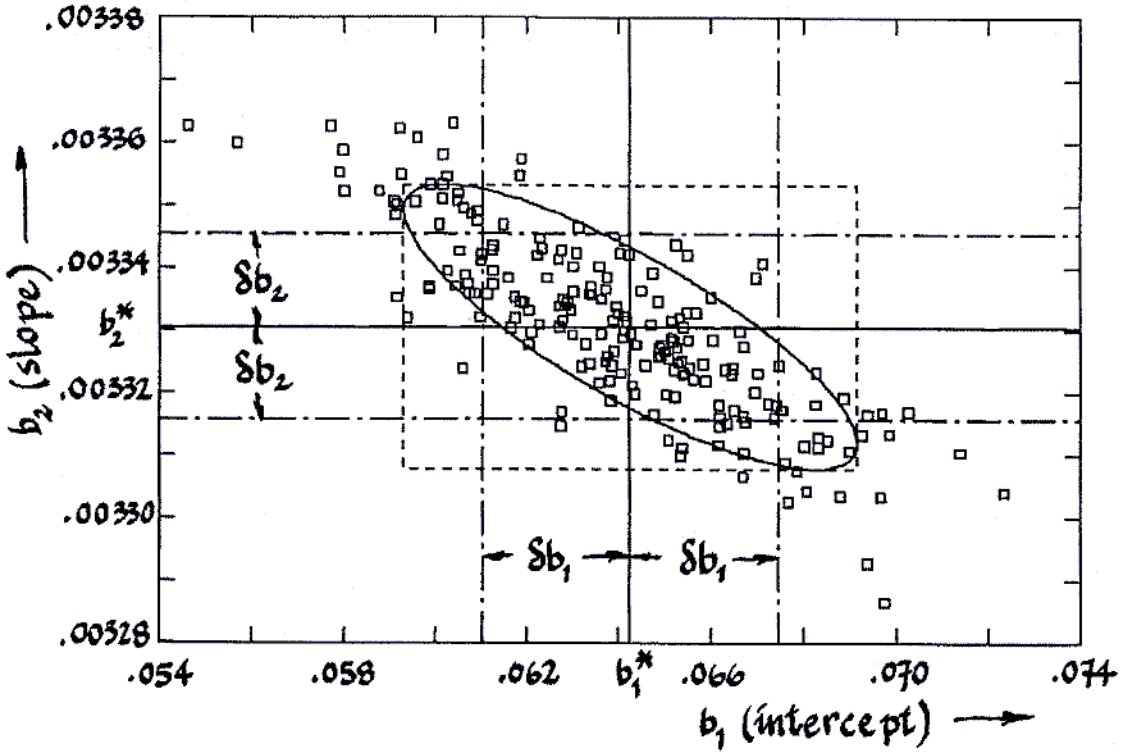

Figure 9. Confidence limits in parameter space, using the example of the straight line fit described at the start of this chapter. The best estimates of b_1 and b_2 are denoted here by b_1^* and b_2^*. The 200 points represent a Monte Carlo simulation of the experiment. δb_1 and δb_2 are estimates of the one-parameter confidence limits (see Eq. (65)); approximately 68.3 per cent of the points will fall within either the horizontal band or the vertical band delineated by these limits. The joint confidence region (see Eq.40), which contains about 68.3 per cent of the points, is delineated by the ellipse; the projection of this ellipse onto the parameter axes, indicated by the dashed rectangle, defines the *support plane*. The support plane will always extend beyond the one-parameter limits.

The correlation matrix. It is useful, when fitting functions nonlinear in the parameters to data, to know the degree to which parameters are correlated with each other. Strong correlations are the rule rather than the exception, and if the correlations are too strong the method will fail because the curvature matrix $\mathbf{A}$ will be judged singular. Excess correlations frequently occur when data are attempted to be fit by a function containing too many parameters, so that the attempted

fit is over-determined. In this case it is useful to know which parameters might fruitfully be abandoned. If two parameters are completely correlated then one of them may be eliminated with impunity.

Just as the *variance* of the j^{th} parameter is proportional to A^{-1}_{jj} (Eq. (63)), the *covariance* $\sigma_{b_j b_k}$ relating the j^{th} and k^{th} parameters is proportional to A^{-1}_{jk}.

Thus, it is not surprising that a matrix of correlation coefficients results from a scaling of the $\mathbf{A}^{-1}$ matrix. The diagonal elements of this scaled matrix will be 1.0 (each parameter will be completely correlated with itself), while an off-diagonal element such as $(A^{-1}_{jk})^*$ will indicate the degree of correlation between the j^{th} and k^{th} parameter. Off-diagonal elements close to ± 1 indicate a high degree of correlation, while those close to 0 indicate hardly any correlation at all.

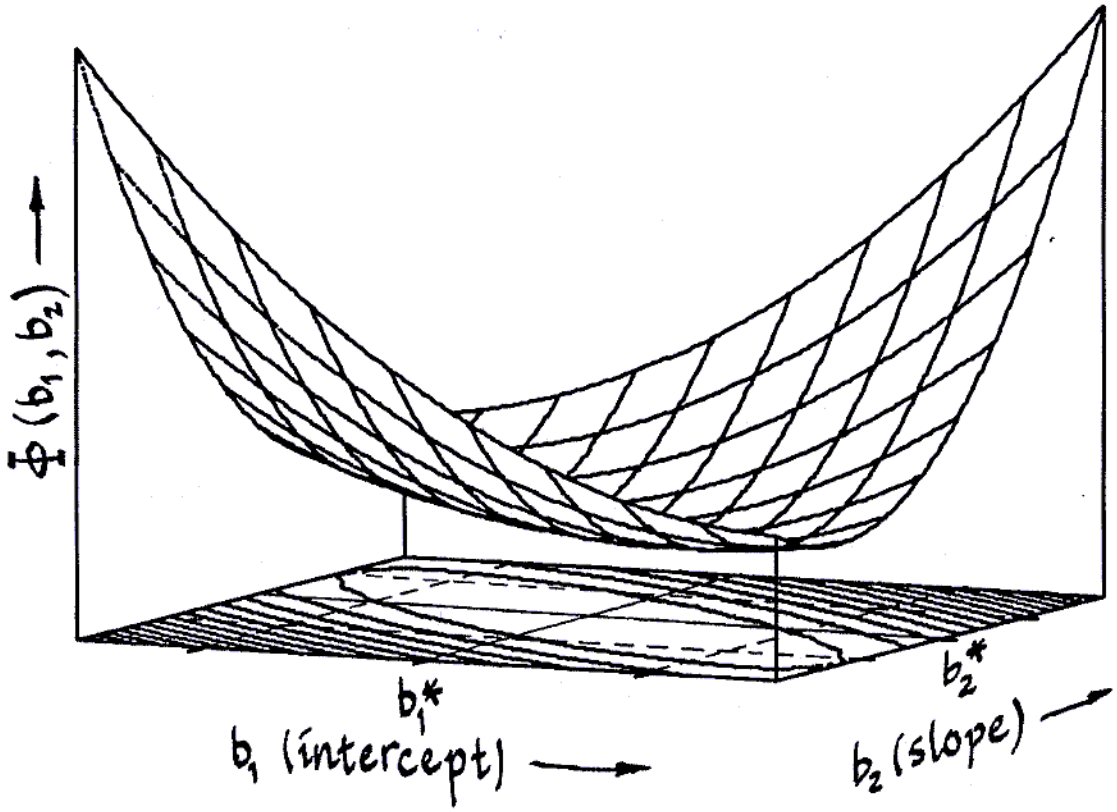

Figure 10. A plot of Φ vs b_1 and b_2 for the same example as that shown in Figure 9. Equally spaced contours of Φ are shown projected onto the $b_1 - b_2$ plane. A gradient correction vector $-\mathbf{g}$ would be perpendicular to these contours, whereas a Taylor correction vector $\boldsymbol{\delta}$ would normally point toward the minimum in Φ. Note the hammock-shaped minimum, a commonly encountered situation.

3.6 General references

1. Taylor, John R., *An Introduction to Error Analysis* (University Science Books, 1982). Taylor is somewhat misleading in his discussion of the chi-square statistic (which he mistakenly calls "chi-squared"), but for most of the basic concepts this is a good place to start.

2. Bevington, Philip R., and Robinson, D. Keith, *Data Reduction and Error Analysis for the Physical Sciences,* 2nd Ed. (McGraw-Hill, 1992). Bevington's comprehensive book is commonly found

on physicists' shelves. Unlike the *Numerical Recipes* book (see the following citation), it is, well, a little tedious. Think seriously about coffee if you want to delve into it.

3. Press, William H. et. al., *Numerical Recipes in C --- The Art of Scientific Computing*, 2nd Ed. (Cambridge University Press, New York, 1992). Chapter 15 of this useful volume contains extensive discussion of methods for fitting data by a "model."

4. Bennett, Carl A., and Franklin, Norman L., *Statistical Analysis in Chemistry and the Chemical Industry* (Wiley, 1954). These authors are remarkably thorough in their treatment.

4. CHI-SQUARE: TESTING FOR GOODNESS OF FIT

In the previous chapter we discussed procedures for fitting a hypothesized function to a set of experimental data points. Such procedures involve minimizing a quantity we called Φ in order to determine best estimates for certain function parameters, such as (for a straight line) a slope and an intercept. Φ is proportional to (or in some cases equal to) a statistical measure called χ^2, or *chi-square*, a quantity commonly used to test whether any given data are well described by some hypothesized function. Such a determination is called a *chi-square test for goodness of fit*. In the following, we discuss χ^2 and its statistical distribution, and show how it can be used as a test for goodness of fit.[30]

4.1 The definition of χ^2

If ν independent variables x_i are each normally distributed with mean μ_i and variance σ_i^2, then the quantity known as *chi-square* [31] is defined by

$$\chi^2 \equiv \frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}+\cdots+\frac{(x_\nu-\mu_\nu)^2}{\sigma_\nu^2}=\sum_{i=1}^{\nu}\frac{(x_i-\mu_i)^2}{\sigma_i^2} \tag{68}$$

Note that ideally, given the random fluctuations of the values of x_i about their mean values μ_i, each term in the sum will be of order unity. Hence, if we have chosen the μ_i and the σ_i correctly, we may expect that a calculated value of χ^2 will be approximately equal to ν. If it is, then we may conclude that the data are well described by the values we have chosen for the μ_i, that is, by the hypothesized function.

If a calculated value of χ^2 turns out to be much larger than ν, and we have correctly estimated the values for the σ_i, we may possibly conclude that our data are not well-described by our hypothesized set of the μ_i.

This is the general idea of the χ^2 test. In what follows we spell out the details of the procedure.

4.2 The χ^2 distribution

The quantity χ^2 defined in Eq.(68) has the probability distribution given by

$$f(\chi^2)=\frac{1}{2^{\nu/2}\Gamma(\nu/2)}e^{-\chi^2/2}(\chi^2)^{(\nu/2)-1} \tag{69}$$

[30] The chi-square distribution and some examples of its use as a statistical test are also described in the references listed at the end of this chapter.

[31] The notation of χ^2 is traditional and possibly misleading. It is a single statistical variable, and not the square of some quantity χ. It is therefore not *chi-squared*, but *chi-square*. The notation is merely suggestive of its construction as the sum of squares of terms.

This is known as the χ^2 - *distribution with* ν *degrees of freedom.* ν is a positive integer.[32] Sometimes we write it as $f(\chi_\nu^2)$ when we wish to specify the value of ν. $f(\chi^2)d(\chi^2)$ is the probability that a particular value of χ^2 falls between χ^2 and $\chi^2 + d(\chi^2)$.

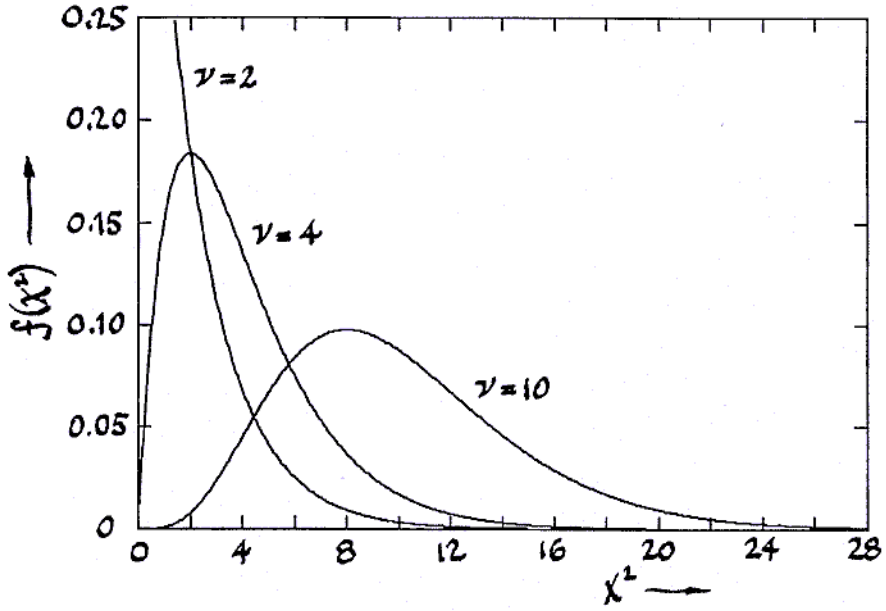

Figure 11. The chi-square distribution for $\nu = 2$, 4, and 10.

Note that χ^2 ranges only over positive values: $0 < \chi^2 < \infty$.

The mean value of χ_ν^2 is equal to ν, and the variance of χ_ν^2 is equal to 2ν. The distribution is highly skewed for small values of ν, and becomes more symmetric as ν increases, approaching a Gaussian distribution for large ν, just as predicted by the Central Limit Theorem.

4.3 How to use χ^2 to test for goodness of fit

Suppose we have a set of *N* experimentally measured quantities x_i. We want to test whether they are well-described by some set of hypothesized values μ_i. We form a sum like that shown in Eq.(69). It will contain *N* terms, constituting a sample value for χ^2. In forming the sum, we must use estimates for the σ_i that are *independently* obtained for each x_i.[33]

---

[32] $\Gamma(p)$ is the "Gamma function." defined by

$$\Gamma(p+1) \equiv \int_0^\infty x^p e^{-x} dx .$$

It is a generalization of the factorial function to non-integer values of *p*. If *p* is an integer, $\Gamma(p+1) = p!$. In general, $\Gamma(p+1) = p\Gamma(p)$, and $\Gamma(1/2) = \sqrt{\pi}$.

[33] In the previous chapter, we showed how a hypothesized function may be fit to a set of data points. There we noted that it may be either impossible or inconvenient to make independent estimates of the σ_i, in which case estimates of the σ_i can be made only by *assuming* an ideal fit of the function to the data. That is, we assumed χ^2 to be equal to its mean value, and from that, estimated uncertainties, or confidence intervals, for the values of the determined parameters. Such a procedure *precludes* the use of the χ^2 test.

Now imagine, for a moment, that we could repeat our experiment many times. Each time, we would obtain a data sample, and each time, a sample value for χ^2. If our data were well-described by our hypothesis, we would expect our sample values of χ^2 to be distributed according to Eq. (69) , and illustrated by example in Figure 11. However, we must be a little careful. The expected distribution of our samples of χ^2 will *not* be one of *N* degrees of freedom, even though there are *N* terms in the sum, because our sample variables x_i will invariably *not* constitute a set of *N* independent variables. There will, typically, be at least one, and often as many as three or four, relations connecting the x_i. Such relations are needed in order to make estimates of hypothesized parameters such as the μ_i, and their presence will *reduce* the number of degrees of freedom. With *r* such relations, or *constraints*, the number of degrees of freedom becomes $\nu = N - r$, and the resulting χ^2 sample will be one having ν (rather than *N*) degrees of freedom.

As we repeat our experiment and collect values of χ^2, we expect, if our model is a valid one, that they will be clustered about the median value of χ^2_ν, with about half of these collected values being greater than the median value, and about half being less than the median value. This median value, which we call $\chi^2_{\nu,0.5}$, is determined by

$$\int_{\chi^2_{\nu,0.5}}^{\infty} f(\chi^2) d\chi^2 = 0.5$$

Note that because of the skewness of the distribution function, the median value of χ^2_ν will be somewhat less than ν, which is its mean, or average value. For example, for $\nu = 10$ degrees of freedom, $\chi^2_{10,0.5} \approx 9.34$, a number slightly less than 10.

Put another way, we expect that a single measured value of χ^2 will have a probability of 0.5 of being greater than $\chi^2_{\nu,0.5}$.

We can generalize from the above discussion, to say that we expect a single measured value of χ^2 will have a probability of α of being greater than $\chi^2_{\nu,\alpha}$, defined by

$$\int_{\chi^2_{\nu,\alpha}}^{\infty} f(\chi^2) d\chi^2 = \alpha$$

This definition is illustrated by the inset in Figure 14.

Here is how the χ^2 test works:

1. We hypothesize that our data are appropriately described by our chosen function, or set of μ_i. This is the hypothesis we are going to test.

2. From our data sample we calculate a sample value of χ^2 (chi-square), along with ν (the number of degrees of freedom), and so determine χ^2 / ν (the normalized chi-square, or the chi-square per degree of freedom) for our data sample.

3. We choose a value of the significance level α (a common value is.05, or 5 per cent), and from an appropriate table or graph (*e.g.*, Figure 14), determine the corresponding value of $\chi^2_{\nu,\alpha}/\nu$. We then compare this with our sample value of χ^2/ν.

4. If we find that $\chi^2/\nu > \chi^2_{\nu,\alpha}/\nu$, we may conclude that either (1) the model represented by the μ_i is a valid one but that a statistically improbable excursion of χ^2 has occurred, or (2) that our model is so poorly chosen that an unacceptably large value of χ^2 has resulted. (1) will happen with a probability α, so if we are satisfied that (1) and (2) are the only possibilities, (2) will happen with a probability $1-\alpha$. Thus, if we find that $\chi^2/\nu > \chi^2_{\nu,\alpha}/\nu$, we are $100\cdot(1-\alpha)$. per cent confident in *rejecting* our model. Note that this reasoning breaks down if there is a possibility (3), for example if our data are *not* normally distributed. The theory of the chi-square test relies on the assumption that chi-square is the sum of the squares of ν *random normal deviates*, that is, that each x_i is normally distributed about its mean value μ_i. However, for some experiments, there may be occasional non-normal data points that are too far from the mean to be real. A truck passing by, or a glitch in the electrical power could be the cause. Such points, sometimes called *outliers*, can unexpectedly increase the sample value of chi-square.

5. If we find that χ^2 is too small, that is, if $\chi^2/\nu < \chi^2_{\nu,1-\alpha}/\nu$, we may conclude only that either (i) our model is valid but that a statistically improbable excursion of χ^2 has occurred, or (ii) we have, too conservatively, over-estimated the values of σ_i, or (iii) someone has given us fraudulent data, that is, data "too good to be true." A too-small value of χ^2 cannot be indicative of a poor model. A poor model can only increase χ^2.

Generally speaking, we should be pleased to find a sample value of χ^2/ν that is near 1, its mean value for a good fit.

In the final analysis, we must be guided by our own intuition and judgment. The test, being of a statistical nature, serves only as an indicator, and cannot be iron clad.

4.4 An example

The field of particle physics provides numerous situations where the χ^2 test can be applied. A particularly simple example[34] involves measurements of the mass M_Z of the Z^0 boson by experimental groups at CERN. The results of measurements of M_Z made by four different detectors (L3, OPAL, Aleph and Delphi) are as shown in Figure 12.

Detector	Mass in GeV/c^2
L3	91.161 ± 0.013
OPAL	91.174 ± 0.011
Aleph	91.186 ± 0.013
Delphi	91.188 ± 0.013

Figure 12. Measurements of M_Z made by four different detectors.

[34] This example is provided by Pat Burchat (private communication).

The listed uncertainties are estimates of the σ_i, the standard deviations for each of the measurements. The figure below shows these measurements plotted on a horizontal mass scale (vertically displaced for clarity).

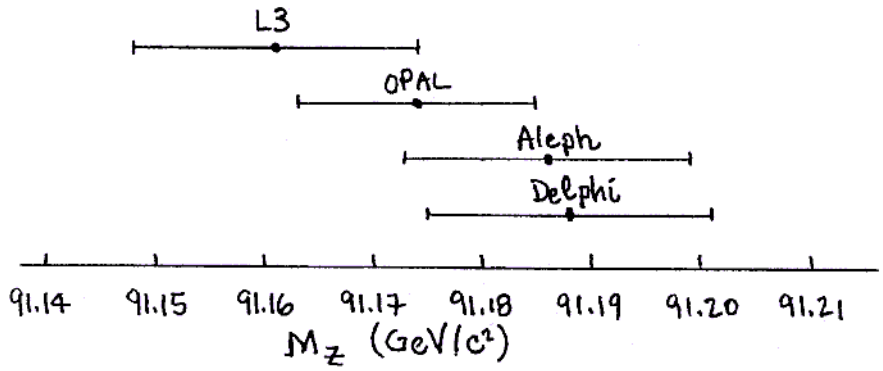

Figure 13. Measurements of the Z^0 boson.

The question arises: Are these measurements consistent with an estimate of M_Z made by determining the weighted mean of the four measurements? We find the weighted mean $\overline{M_Z}$ and its standard deviation like this:

$$\overline{M_Z} = \frac{\sum M_i / \sigma_i^2}{\sum 1/\sigma_i^2} \qquad \text{and} \qquad \sigma^2_{\overline{M_Z}} = \frac{1}{\sum 1/\sigma_i^2}$$

to find

$$\bar{M}_Z \pm \sigma_{\overline{M_Z}} = 91.177 \pm 0.006$$

Then we form χ^2:

$$\chi^2 = \sum_{i=1}^{4} \frac{(M_i - \overline{M_Z})^2}{\sigma_i^2} \approx 2.78$$

We expect this value of χ^2 to be drawn from a chi-square distribution with 3 degrees of freedom. The number is 3 and not 4 because we have used the mean of the four measurements to estimate the value of μ, the true mass of the Z^0 boson, and this uses up one degree of freedom. Hence $\chi^2/\nu = 2.78/3 \approx 0.93$. Now from the graph of χ^2 vs χ^2/ν shown in Figure 14, we find that for 3 degrees of freedom, α is about 0.42, meaning that if we were to repeat the experiments we would have about a 42 per cent chance of finding a χ^2 for the new measurement set larger than 2.78, assuming our hypothesis is correct. We have therefore no good reason to reject the hypothesis, and conclude that the four measurements of the Z^0 boson mass are consistent with each other. We would have had to have found χ^2 in the vicinity of 8.0 (leading to an α of about 0.05) to have been justified in suspecting the consistency of the measurements. The fact that our sample value of $\chi^2/3$ is close to 1 is reassuring.

4.5 Using χ^2 to test hypotheses regarding statistical distributions

The χ^2 test is used most commonly to test the nature of a statistical distribution from which some random sample is drawn. It is this kind of application that is described by Evans in his text, and is the kind of application for which the χ^2 test was first formulated.

Situations frequently arise where data can be classified into one of *k* classes, with probabilities $p_1, p_2, \cdots, p_k$ of falling into each class. We suppose that $\sum p_i = 1$. Now suppose we take data by *classifying* it: We *count* the number of observations falling into each of the *k* classes. We'll have n_1 in the first class, n_2 in the second, and so on, up to n_k in the k^{th} class. Suppose that there is a total of *N* observations, so $\sum n_i = N$.

It can be shown by non-trivial methods that the quantity

$$\frac{(n_1 - Np_1)^2}{Np_1} + \frac{(n_2 - Np_2)^2}{Np_2} + \cdots + \frac{(n_k - Np_k)^2}{Np_k} = \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i} \tag{70}$$

has approximately the χ^2 distribution with $k - r$ degrees of freedom, where *r* is the number of constraints, or relations used to estimate the p_i from the data. *r* will always be at least 1, since $\sum n_i = \sum N p_i = N \sum p_i = N$.

Since Np_i is the mean, or expected value of n_i, the form of χ^2 given by Eq. (70) corresponds to summing, over all classes, the squares of the deviations of the observed n_i from their mean values divided by their mean values.

Note that this special form looks different from that shown in Eq. (68), in that the variance for each point is replaced by the mean value of n_i for each point. Such an estimate for the variance commonly arises in situations involving counting. For a Poisson distribution, for example, the mean is precisely equal to the variance.

Equation (70) forms the basis of what is sometimes called *Pearson's Chi-square Test*. Unfortunately, this equation is also sometimes used to *define* χ^2, but this form is not the most general form, in that it applies only to situations involving counting, where the data variables are dimensionless.

4.6 Another example

Here is an example, in which the χ^2 test is used to test whether a data sample consisting of the heights of 66 women can be assumed to be drawn from a Gaussian distribution.

We first display the data in the form of a *frequency distribution*, listing for each height *h*, the value of *n*(*h*) (see Table 1). Here *n*(*h*) is the number of women in the sample whose height is *h* where *h* is in inches.

Table 1. The sample distribution

h	58	59	60	61	62	63	64	65
n(*h*)	1	0	1	4	6	7	13	8

h	66	67	68	69	70	71	72	73
n(*h*)	11	2	7	4	1	0	0	1

We make the hypothesis that the heights are distributed according to the Gaussian distribution, namely that the probability $p(h)dh$ that a height falls between h and $h+dh$ is given by

$$p(h)dh = \frac{1}{\sigma\sqrt{2\pi}} e^{-(h-\mu)^2/2\sigma^2} dh$$

This expression, if multiplied by *N*, will give, for a sample of *N* women, the number of women $n_{\text{th}}(h)dh$ theoretically expected to have a height between h and $h+dh$:

$$n_{\text{th}}(h)dh = \frac{N}{\sigma\sqrt{2\pi}} e^{-(h-\mu)^2/2\sigma^2} dh \qquad (71)$$

In our example, $N = 66$. Note that we have, in our table of data above, grouped the data into *bins* (we'll label them with the index *j*), each of size 1 inch. A useful approximation to Eq. (71), in which *dh* is taken to be 1 inch, gives the expected number of women $n_{\text{th}}(j)$ having a height h_j:

$$n_{\text{th}}(j) = \frac{N}{\sigma\sqrt{2\pi}} e^{-(h_j-\mu)^2/2\sigma^2} \qquad (72)$$

Now the sample mean $\bar{h}$ and the sample standard deviation *s* are our best estimates of μ and σ. We find, calculating from the data:

$$\bar{h} = 64.9 \text{ inches} \qquad \text{and} \qquad s = 2.7 \text{ inches}$$

Using these values we may calculate, from Eq. (72), the number expected in each bin (shown in Table 2).

Table 2. The expected distribution

h	58	59	60	61	62	63	64	65
$n_{th}(j)$	0.3	0.9	1.8	3.4	5.5	7.6	9.3	9.9

h	66	67	68	69	70	71	72	73
$n_{th}(j)$	9.1	7.3	5.0	3.1	1.6	0.7	0.3	0.1

Table 3. The sample and expected data grouped into classes

h	≤ 61	62	63	64	65	66	67	68	≥ 69
$n(h)$	6	6	6	13	8	11	2	7	6
$n_{th}(j)$	6.5	5.5	7.6	9.3	9.9	9.1	7.3	5.0	5.8

In applying the χ^2 test to a situation of this type, it is advisable to group the data into classes (or bins) such that the expected number occurring in each bin is greater than 4 or 5; otherwise the theoretical distributions within each bin become too highly skewed for meaningful results. Thus, in this situation we shall put all the heights of 61 inches or less into a single bin, and all the heights of 69 inches or more into a single bin. This groups the data into a total of 9 bins (or classes), with actual numbers and expected numbers in each bin being given as in Table 3. (note the bin sizes need not be equal).

Now we calculate the value of χ^2 using these data, finding

$$\chi^2 = \frac{(6-6.5)^2}{6.5} + \frac{(6-5.5)^2}{5.5} + \cdots + \frac{(6-5.8)^2}{5.8} = 6.96$$

Since we have grouped our data into 9 classes, and since we have used up three degrees of freedom by demanding (a) that the sum of the n_j be equal to N, (b) that the mean of the hypothesized distribution be equal to the sample mean, and (c) that the variance of the hypothesized distribution be equal to the sample variance, there are 6 degrees of freedom left.[35] Hence $\chi^2/\nu = 6.96/6 \approx 1.16$, leading to an α of about 0.33. Hence, we have no good reason to reject our hypothesis that our data are drawn from a Gaussian distribution function.

[35] Note that if we were hypothesizing a Poisson distribution (as in a counting experiment), there would be 7 degrees of freedom (only 2 less than the number of classes). For a Poisson distribution the variance is *equal* to the mean, so there is only 1 parameter to be determined, not 2.

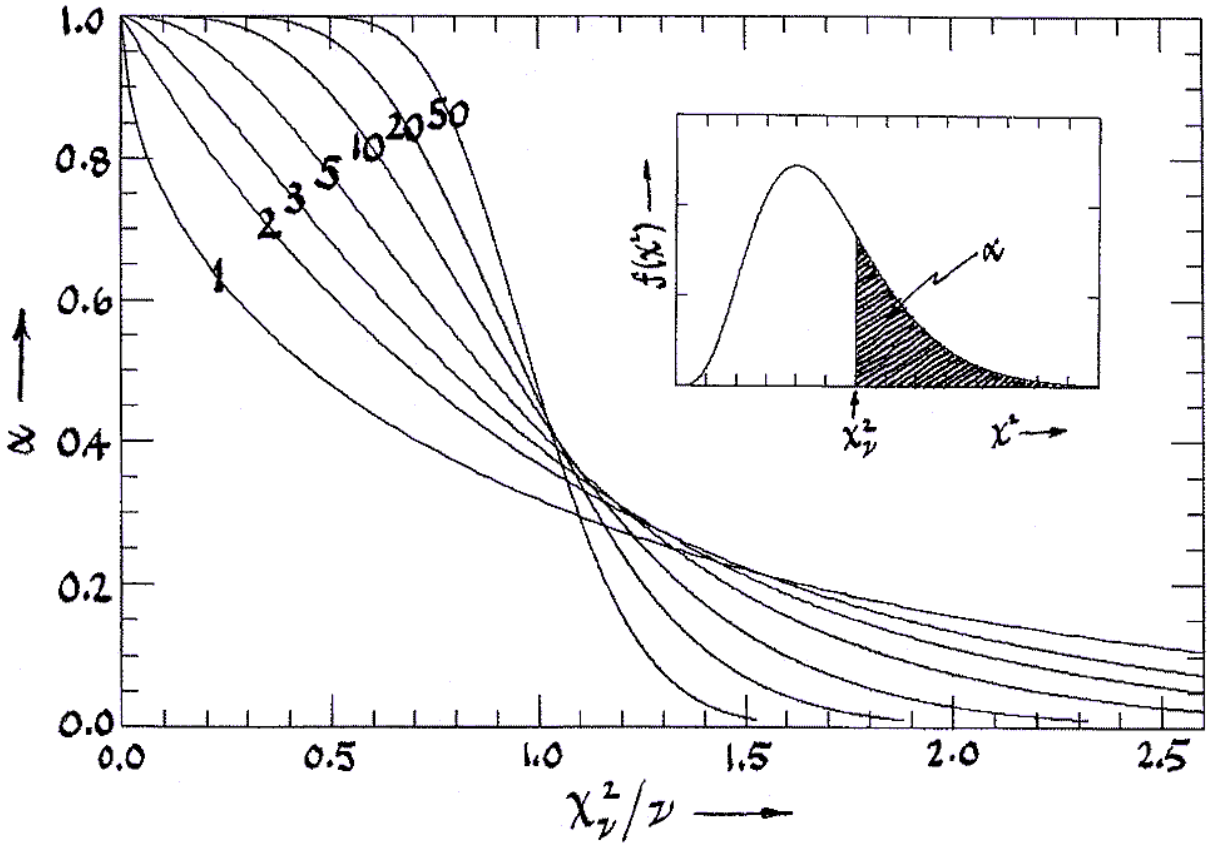

Figure 14. α *vs*. the normalized quantity χ^2_ν / ν. α is the probability that a sample chi-square will be larger than χ^2_ν, as shown in the inset. Each curve is labeled by ν, the number of degrees of freedom.

4.7 General references

1. William H. Press *et. al*., *Numerical Recipes in C --- The Art of Scientific Computing*, 2nd Ed. (Cambridge University Press, New York, 1992). Press devotes considerable discussion to the subject of chi-square. Noteworthy is his wry remark on page 657, about those who deem a fit acceptable if a graph of data and model "looks good." He calls this approach "*chi-by-eye*", and notes that its practitioners get what they deserve.

2. Carl A.Bennett, and Norman L.Franklin, *Statistical Analysis in Chemistry and the Chemical Industry* (Wiley, 1954). An excellent discussion of the chi-square distribution function, with good examples illustrating its use, may be found on pages 96 and 620.

3. Robley D.Evans, *The Atomic Nucleus* (McGraw-Hill, 1969). Chapter 27 of this advanced text contains a description of *Pearson's Chi-square Test*.

5. ATOMIC SPECTRA

In this experiment, your main project will use a high-precision (if somewhat old-fashioned) visible light spectrometer to confirm the expected energy level pattern of the hydrogen atom and derive a value for the Rydberg constant (which determines the separation of the energy levels of hydrogen). This will be the main subject of your report; your instructor may tell you to include other elements, so be sure you know what you're expected to include in your report, in your lab notebook, and in any associated homework your instructors may assign.

The experiment was designed with the following learning goals in mind:

- To demonstrate the power and sensitivity of spectroscopy, possibly the most important experimental tool in physics and astrophysics;
- To give you some experience in the careful and consistent calibration and use of sensitive equipment; and
- To familiarize you with particular aspects of error propagation and statistical analysis. This includes the estimation of error when reading dials; correctly combining several measurements of the same quantity to determine the best value and its error; and correctly propagating errors so that correlated errors are not erroneously treated as uncorrelated (i.e. when the same raw measurement has contributed to several derived measurements, whose errors are then combined).

5.1 Historical Background

In 1666, when he was only 23, Isaac Newton bought a glass prism "to try therewith the phenomena of colours." Here's how he describes what he did with it:[36]

"In a very dark Chamber, at a round Hole, about one third Part of an Inch broad, made in the Shut of a Window, I placed a Glass Prism, whereby the Beam of the Sun's Light, which came in at that Hole, might be refracted upwards toward the opposite Wall of the Chamber, and there form a colour'd Image of the Sun. This Image was Oblong and not Oval, but terminated with two Rectilinear and Parallel Sides, and two Semicircular Ends. On the Sides it was bounded pretty distinctly, but on it Ends very confusedly and indistinctly, the Light there decaying and vanishing by degrees."

For these "phenomena of colours" he later coined the word *spectrum*---like a specter, a ghost, or an apparition. These were the wonderful colors, spread out from red to violet, just as in that most exquisite of natural phenomena---the rainbow. Newton's *spectrum* resulted, however, not from the chance placement of millions of water droplets, but from his own placement of a simple piece of glass.

[36] The quoted passage is from Book One, Part I of Newton's *Opticks*, first published in 1704. A modern reprinting (Dover 1952) is available at the UCSC Science Library.

Nearly a century later, in 1752, a 26-year-old Scotsman named Thomas Melvill made the first recorded observations of what we now call spectral lines, looking through a similar prism at burning spirits into which he had introduced various salts. He observed for the first time what nearly every high school chemistry student now sees in the candle flame: the yellow lines of sodium. Unfortunately, Melvill died the following year.

The birth of modern spectroscopy is usually attributed to Joseph Fraunhofer, born in 1787, the son of a Bavarian glazier. Assisting his father, he became skilled in the melting and grinding of glass. With his rare combination of theoretical and practical skill he became a maker of fine lenses and telescopes. He used the sodium yellow lines (formed with a narrow slit) in making careful measurements of the refractive index of certain glasses, and in 1814 he analyzed, in the spectrum of the sun, hundreds of dark absorption lines, among which are those we now call the Fraunhofer lines. Indeed, you will find in the lab room a poster with the most prominent Fraunhofer lines identified. You will also find a Project Star hand-held reflection grating spectrometer that you may use for identifying prominent spectral lines from your spectrum tubes, from street lights, from sunlight reflected from clouds, etc.

The field of optical spectroscopy subsequently flowered. In *The Outline of Science*, first published in 1922, J. Arthur Thompson writes: "That the spectroscope will detect the millionth of a milligram of matter, and on that account has discovered new elements, commands our admiration; but when we find in addition that it will detect the nature of forms of matter and that it will measure the velocities with which these forms of matter are moving with an absurdly small per cent of possible error, we can easily acquiesce in the statement that it is the greatest instrument ever devised by the brain and hand of man."

The optical spectrometer is still at the forefront, a fine instrument, extending our senses, allowing us to understand and appreciate not only the stuff in our immediate vicinity, but also the stuff in the most remote regions of the universe. In this experiment you will have the chance to use an optical spectrometer to observe and measure the wavelengths of radiation emitted from excited atoms. The instruments we use are of a traditional design, some might say a little old-fashioned (the instruction manual for the older instrument was first copyrighted in 1938), but they are well-made, and when properly aligned and calibrated, can yield remarkably accurate measurements for the wavelengths.

5.2 Goals and Expected Results

You will use not a prism but rather a grating to disperse light into its component colors. The grating works by interference (see below); a prism works by having an index of refraction that varies with frequency, of course.

In the first part of the lab, you will take a helium lamp as a *known* wavelength standard and use the data from that lamp to calculate the spacing of the lines in your grating ("grating constant"). With this value, you will take data for hydrogen, treating the hydrogen lines as unknown. The hydrogen data will then be used to calculate the value of the Rydberg constant.

You should expect to measure the grating constant of your grating to a few parts in ten thousand! Furthermore, you will determine Rydberg's constant to a similar level of accuracy. Please keep these numbers in mind as you take measurements and do calculations. Do not make approximations that are incompatible with your expected accuracy.

5.3 Operation of the spectrometer

IMPORTANT: The angle to which you have rotated the spectrometer's telescope is measured with a Vernier scale. It utilizes two sets of tick marks, one having a slightly different spacing than the other. The accuracy to which you can measure an angle is *many times* better than the smallest tick mark on either scale. If you do not already know how to read a Vernier, type "using a Vernier scale" into your favorite Internet search engine and you will find a number of useful tutorials. We can also help. Do *not* assume you know what you're doing if you haven't used one before.

There are two Verniers on the protractor circle; it is well to record the readings of both. Be sure to make an estimate of the precision with which an angle may be measured.

LOVING CARE NEEDED: The apparatus used in this experiment is delicate, and should be treated with loving care. Be especially careful with glass parts, such as lenses and mirrors, as they are apt to break if dropped. Also, do not touch the surface of any diffraction grating with your fingers; handle them (and mirrors too) by their edges.

By measuring the angle through which the telescope is rotated away from the central (straight-through) image, the wavelengths of the diffracted light may be calculated if the grating constant is known. In order to obtain a sharply focused spectrum that will yield accurate measurements, it is necessary to align the instrument, using a procedure we shall outline in a moment.

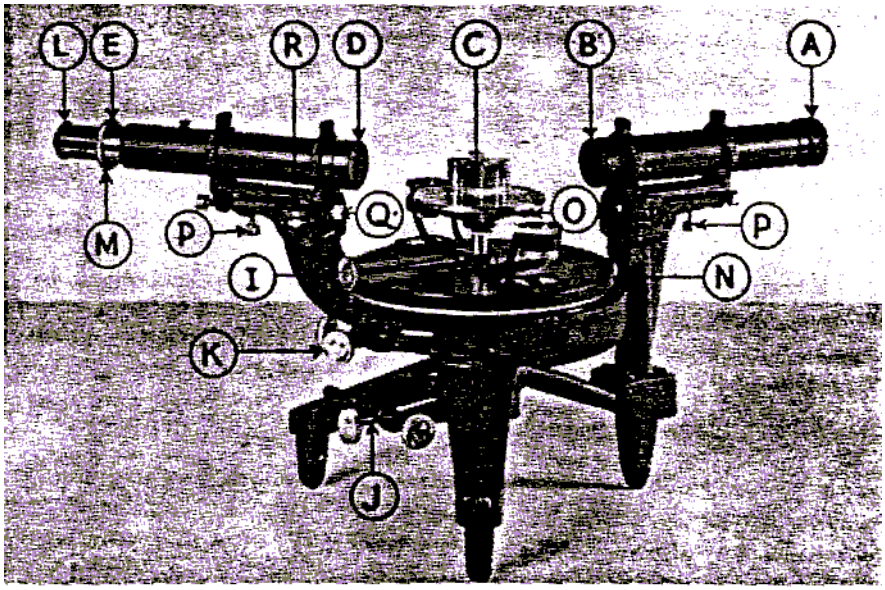

Figure 15. The Spencer Spectrometer. **A**: Entrance slit; **B**: Collimator objective; **C**: Diffraction grating; **D**: Telescope objective; **E**: Plane of cross-hairs; **I**: Grating table clamping screw; **J**: Grating table clamping and tangent screws; K: Telescope arm clamping and tangent screws; **L**: Eyepiece ring; **M**: Telescope focusing ring; **O**: Grating table leveling screws; **P**: Collimator and Telescope leveling screws.

Figure 15 shows the Spencer spectrometer and its parts. The light to be analyzed enters the instrument through an adjustable slit at **A**. This slit should be set in the focal plane of the collimator objective lens at **B**, so that light emerging from the collimator forms a parallel beam which then impinges on the diffraction grating at **C**. The diffracted light (still parallel) then enters the telescope objective at **D**, a lens which then brings the light to a focus at **E**, where a set of cross-hairs is located. In this way an image of the entrance slit **A** is formed at the cross-hairs, an image that may be viewed with the aid of the eyepiece **L**.

> **Exercise:** At this point, you should draw for yourself a diagram of the optical rays that go through the spectrometer from the source to the eyepiece, and you should *understand* your diagram! This does *not* mean drawing only on-axis rays. It means drawing the aperture-limited rays as well. That is, Figure 16 in the write-up of this experiment is *not* the requested drawing.

Before starting the alignment procedure, note that the telescope and the grating table may be rotated about the central axis using various clamping and adjustment thumbscrews. *Start by becoming familiar with these thumbscrews*. They should all be treated with a gentle touch. There are three clamping screws that may be loosened to allow for coarse adjustment. The first is denoted by **I** in Figure 15,[37] a screw that when loosened, will allow the grating table itself to be rotated. Normally it will be adjusted and tightened at the start, and then left alone during a series of observations.

The second is denoted by **J**, which also allows rotation of the grating table when loosened. Try loosening it, and note that the inner protractor rotates with the grating table. If screw **J** is tightened (gently), fine adjustments to the rotation of the grating table and protractor may be accomplished using the associated "tangent'" screw (also denoted by **J** in Figure 15).

The third is denoted by **K**, which is another pair of screws. They allow the rotation (again coarse and fine adjustment) of the telescope about the central axis, along with its protractor circle (the outer one). Try loosening this clamping screw. Note that the telescope can be moved freely to the left or right.

5.4 Alignment procedure

Once you have become familiar with the general principles of operation and the various adjustment screws, you will need to align the spectrometer. A well-aligned instrument will have the entrance slit in the focal plane of the collimator objective and the cross-hairs in the focal plane of the telescope objective, with the image of the entrance slit superposed on the cross-hairs. The cross-hairs should also be in sharp focus through the eyepiece. The grating surface should lie in the plane of the rotation axis, with the grating rulings parallel to that axis. Finally, both the telescope and collimator axes should be perpendicular to, and intersect with, the rotation axis of the instrument.

You will find at your bench a detailed procedure for accomplishing all of these. The alignment steps, if you have performed them carefully, will leave you with a well-aligned instrument capable

[37] Figure 15 shows the Spencer instrument. There are similar screws on the instruments that we will use.

of accurate measurements. After you have completed the alignment procedure you should be careful not to change the focus of the telescope. Any tweaking of the focus during your experimenting should be done by adjusting the focus of the collimator, *not the telescope*. Do you see why?

5.5 Measure the angles of the spectral lines from He and H

You will take full sets of data for two lamps: helium and hydrogen. Helium will be your calibration source -- you will *assume* its lines are known and use them to determine the grating spacing. Hydrogen will be your object of scientific study: knowing the grating spacing, you will make your best measurements of the wavelengths of the hydrogen lines. In each case you must choose which lines to record. In general, you will choose the brightest lines available, and the number of lines will depend on which lamp you are using. You should record at least eight lines for helium. Be sure to record the color of each line as well.

When you observe the spectrum from the hydrogen lamp, you may notice a few spectral lines that do not correspond to lines of the Balmer series. These could be radiations that are emitted by molecular hydrogen, which produces a much more complex spectrum than that from atomic hydrogen. It is difficult to produce a hydrogen lamp that does not emit some of the lines in the molecular spectrum, because it is energetically favorable for hydrogen atoms to combine to form molecules. The molecular spectrum is particularly evident in older tubes. You must find *at least* (and probably no more than) three lines of atomic hydrogen in the presence of these other confusing lines! Therefore, you need to have at least a crude idea of where the atomic-hydrogen (Balmer) lines are expected.

In order to use the spectrometer to determine spectral wavelengths, we must first determine the value for *d*, the average spacing between the rulings on the diffraction grating. *d* is also called the *grating constant*. Although the nominal grating constant is printed on the grating, the exact constant varies from grating to grating, on account of manufacturing variability. To calibrate the grating, we shall assume that some set of spectral wavelengths are already known: those produced by helium. Such a set of wavelengths then becomes a *calibration spectrum*, and using a theoretical model for the optical behavior of our grating, we may deduce a value for *d*. The fact that you are going to use helium lines to determine *d* is irrelevant to how you take data. Just use the following procedure for all lamps.

For each lamp, you must record the angle of the straight-through, undeflected light, as well as the angles of the diffracted lines. This undeflected angle is crucial for checking that your instrument is aligned properly.

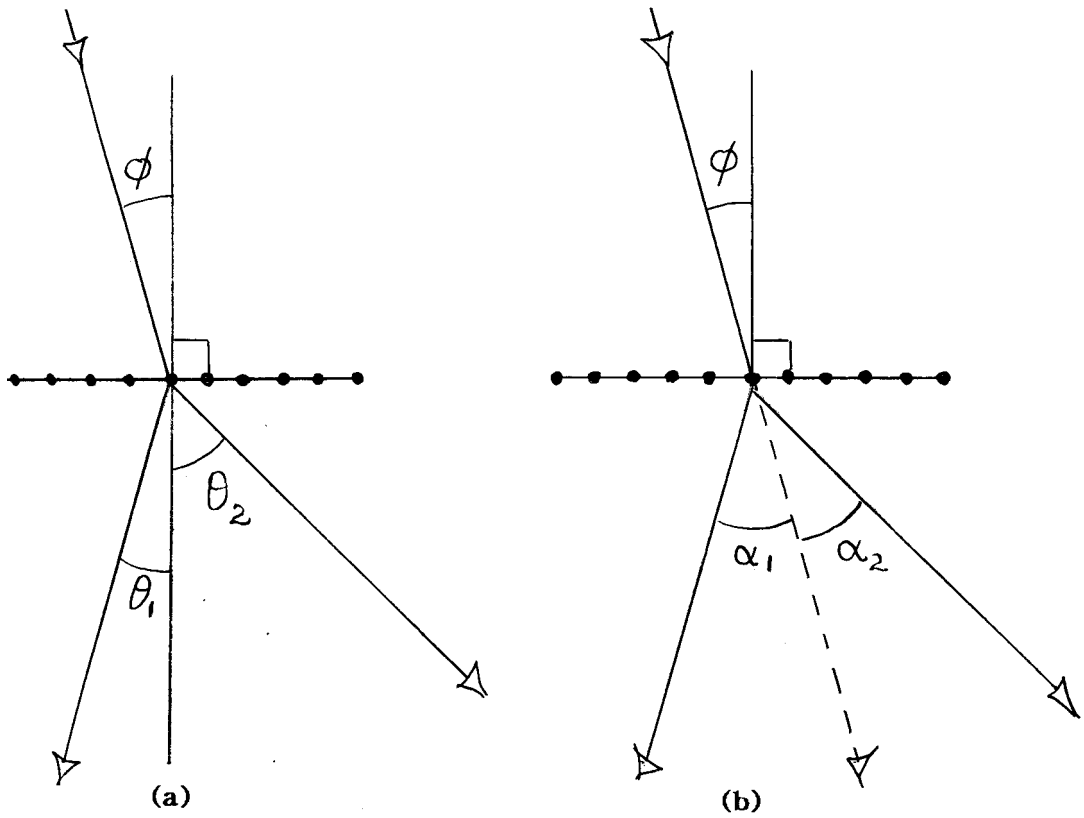

Figure 16. (a) Geometry with *diffraction* angles displayed. (b) Geometry with *measured* angles displayed.

Referring to Figure 16, we will assume for the moment that the incident beam makes a nonzero angle ϕ with respect to the *grating normal*. For a given *m*, there will be two diffracted beams, at angles θ_1 and θ_2 with respect to the grating normal.[38] These angles satisfy the grating diffraction equation, where *m*, the order of the spectrum, is a positive integer.

$$\begin{aligned} m\lambda / d &= \sin\theta_1 + \sin\phi \\ m\lambda / d &= \sin\theta_2 + \sin\phi \end{aligned} \tag{73}$$

With our spectrometer, one is able to measure the angles α_1 and α_2, which are the angles of the diffracted beams with respect to the *direct* (un-diffracted) beam. With a perfect instrument that is perfectly aligned to $\phi = 0$, the diffracted beams would be symmetric, and $\theta_1 = \theta_2 = \alpha_1 = \alpha_2$ However, if $\phi \neq 0$, then $\alpha_1 \neq \alpha_2$. In this latter case, we need to relate the measured angles α_1 and α_2 to the diffraction angles θ_1 and θ_2. If we define $\theta \equiv \frac{1}{2}(\alpha_1 + \alpha_2)$ and $\Delta \equiv \frac{1}{2}(\alpha_1 - \alpha_2)$, one can show that

[38] The two beams may have unequal intensities. This is because the grating grooves are assymetric, to intentionally "throw" more light into one beam at the expense of the other. The grating asymmetry is called the grating "blaze."

$$m\lambda / d = \left[1 + \left(\frac{\sin\Delta}{\cos\Delta - \cos\theta} \right)^2 \right]^{-\frac{1}{2}} \sin\theta \tag{74}$$

One can also show that if ϕ is small (less than a few arc minutes), then Δ is extremely small, of order ϕ^2. In this case, this equation reduces to the familiar diffraction equation:

$$m\lambda / d = \sin\theta \tag{75}$$

You will also notice that there are two engraved dials on the spectrometer, enabling one to make redundant measurements of α_1 and α_2. Such measurements will give you an estimate of the systematic instrumental uncertainty.

Experiment: Using the helium spectra tube, measure the position angle of the telescope for the central (un-deviated) image, and for each of the spectral lines you choose. Be sure to make measurements on both sides of the central image, and also measure second-order lines if at all possible. (You should be able to do this for most of the helium lines. You should also be able to measure a few third-order lines). If the slit width is narrow, the angles may be measured with good precision. *Record the color of each line***.**

5.6 Interpreting your spectra

5.6.1 Determination of *d*.

Equation (75) may be used to determine a value for *d* for each α_1, α_2 pair observed in your calibration spectrum. Use your helium data and the table of known wavelengths provided in the Appendix to this experiment. (It is of interest that different editions of the *CRC Handbook of Chemistry and Physics* have different values for some of the wavelengths, and some lines do not appear in some editions. Other tables are available as well, such as the *MIT Wavelength Tables* (QC 453 M35). In this class, we will officially use the wavelengths from the 75th Edition of the *CRC Handbook*, from which the tables in the Appendix have been taken.) In general, you may also see lines that do not fit the pure spectrum. This is because of contamination in the gas of other molecules or elements. These lines will usually be fairly faint.

To make correct assignments of the wavelengths, you may find it helpful to plot out the spectral line wavelengths along an axis so as to be able to look for and recognize unique patterns, such as closely spaced doublets. Make sure that your various values for *d* agree with each other to within their errors (more or less). Otherwise you might not be identifying lines correctly!

5.6.2 Determination of the Rydberg.

When an atom is given sufficient energy, say by colliding with another atom or by absorbing radiation of sufficiently short wavelength, it may emit radiation whose wavelengths are characteristic of the atom. An atomic spectrum is therefore a kind of signature by means of which the atom may be identified. From the spectrum it is also possible to deduce information about the physical structure of the atom. The central idea relating the spectrum to the structure is the existence of discrete energy levels in the atom. An atom in a state with energy E_1 can make a transition to a state with a lower energy E_2 by emitting a photon whose energy is E_1 - E_2. The energy of the photon is in turn related to its frequency υ and wavelength λ by the familiar Planck relation

$$E_1 - E_2 = h\nu = \frac{hc}{\lambda} \tag{76}$$

where c is the speed of light and h is Planck's constant. An atom can also be raised, or excited, from a lower energy state to a higher energy state by the absorption of a photon whose energy equals the difference in energy between the two states.

The hydrogen atom is the simplest of all atoms, consisting of a single electron and a single proton, and its spectrum displays a corresponding simplicity. The quantum theory of the hydrogen atom describes the energy levels E_n for the hydrogen atom:

$$E_n = -\frac{\mu e^4}{8\varepsilon_0^2 h^2}\frac{1}{n^2} \tag{77}$$

Here e is the charge on the electron, ε_0 is the permittivity of free space, and n is a positive integer called the *principal quantum number*. The lowest energy state or *ground state* is the state with n = 1. As n increases, the proton and electron become increasingly separated (the separation growing like n^2) until, at $n = \infty$, the energy $E = 0$ is reached, and the atom is *ionized*, with the proton and electron completely separated. In Eq. (78), μ is called the *reduced mass*:

$$\frac{1}{\mu} \equiv \frac{1}{m} + \frac{1}{M} \tag{78}$$

where m is the mass of the electron and M is the mass of the proton. The proton mass is much larger than the electron mass. Sometimes, as an approximation, the proton mass is assumed infinitely large. In this case μ becomes equal to m, and Eq. (77) is written with the electron mass rather than the reduced mass.

The wavelengths in the emission spectrum of hydrogen correspond to all possible transitions between energy levels. These wavelengths may be deduced from Eqs. (76) and (77), yielding an expression for λ, the wavelength of any particular line in the spectrum:

$$\frac{1}{\lambda} = \frac{E_{n_1} - E_{n_2}}{hc} = \frac{\mu e^4}{8\varepsilon_0^2 h^3 c}\left(\frac{1}{n_2^2} - \frac{1}{n_1^2}\right) \tag{79}$$

where n_1 and n_2 are the principal quantum numbers of the initial and final states, respectively. The combination of constants outside the brackets in Eq.(79) is called R_H, the *Rydberg* constant.[39] The subscript H (for hydrogen) distinguishes it from the corresponding constant containing m rather than μ, which is denoted by R_∞. The accepted experimental value for R_H is $10967758\ \mathrm{m}^{-1}$, with an error of only $1\ \mathrm{m}^{-1}$! Thus, the wavelengths of the hydrogen spectrum are given by

$$\frac{1}{\lambda} = R_H\left(\frac{1}{n_2^2} - \frac{1}{n_1^2}\right) \tag{80}$$

Of the many groups of lines comprising the emission spectrum of hydrogen, only one group contains wavelengths visible to the eye, all of which terminate on the state with $n = 2$. This particular series of wavelengths is called the *Balmer* series, named for Johann Balmer, a Swiss mathematician and teacher at a girls' school in Basel in the nineteenth century. In 1884, when he was 60 years old, he presented a paper in Basel in which he displayed a formula equivalent to Eq. (80). It was his first research effort, carried out long before any theory of atomic structure was developed---the result of his attempts to make sense of spectra that had been recorded by Angstrom and others. Thirteen years later he published a paper that includes a remarkable geometrical interpretation of this formula. His diagram is reproduced in Figure 17.[40]

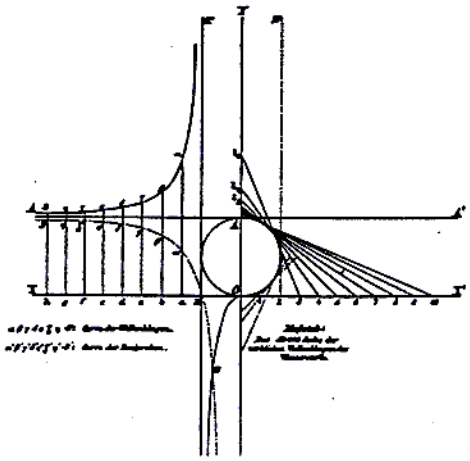

Figure 17. Balmer's Geometric interpretation.

In this diagram, the integers on the *X*'-axis are the values of n_1. The tangents to the circle intersect the *Y*-axis at l_3, l_4, l_5, etc., with distances from the origin *O* being proportional to the wavelengths.

[39] Johannes Rydberg, who worked at the University of Lund in Sweden, suggested the use of this constant in 1890, only a few years after Johann Balmer had worked out his formula describing the spectral series in hydrogen, although Rydberg and Balmer worked completely independently.

[40] This figure appears in Balmer's paper: *Annalen der Physik und Chemie* **60**, 380-391 (1897). Thanks are due to former Physics 133 student Jack Schonbrun for guiding us to the literature on Balmer.

The shortest wavelength---the series limit---is equal to the diameter of the circle. It's a nice problem to show that this geometrical interpretation does indeed duplicate Eq. (80).

You have measured three lines of the hydrogen spectrum. You have used your helium measurements to determine a value for the grating spacing *d*. For each of the hydrogen lines you can now determine an experimental value for R_H, the Rydberg constant, assuming that the theory of Eq. (80) is correct. You may then find the mean of these three values to determine your result for R_H complete with an experimental uncertainty.

If you have time, determine a value for the Rydberg constant using the $m = 2$ and, if possible, the $m = 3$ spectra. Do the resulting values of the Rydberg constant have a better or worse experimental uncertainty?

5.7 The spectra of neon, and of a helium-neon laser.

The highly coherent red light commonly emitted by a helium-neon laser results from a spectral transition between two highly excited energy levels in the neon atom. In the ordinary spectral emissions that we have observed so far, each atom radiates independently, with a typical lifetime that is short, on the order of 10^{-8} seconds. Such radiation is called *spontaneous*. In many of the more complex atoms there are excited energy levels that have a much longer lifetime. Such levels are called *metastable*. Transitions from such levels have a small likelihood and these are called *forbidden* transitions. If, however, radiation is already present that has the frequency corresponding to such a forbidden transition, it can stimulate the transition to occur. This is called *stimulated emission* of radiation. (The word LASER is an acronym, standing for Light Amplification by Stimulated Emission of Radiation.) In this process, the emitted photon is coherent, or in phase, with the stimulating photon. Figure 18 illustrates the process.

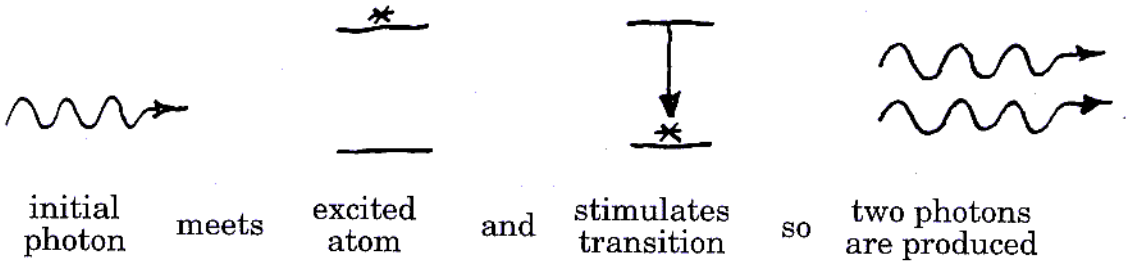

Figure 18. Illustration of the Laser process.

In a laser, the working substance, which may be either a gas, a liquid or a solid, is placed between two parallel mirrors, so that light is reflected back and forth between the mirrors, forming a standing wave pattern. It is this light that stimulates the emission from atoms in the working substance. By making one of the mirrors only partially silvered, a fraction of the energy can be extracted as an external beam. The energy in the standing wave is constantly replenished by emission from the atoms, provided there is a continuous and sufficient supply of atoms in the excited state.

In a helium-neon laser there is a mixture of helium and neon gases, typically about 1 mm pressure of helium, and about 0.1 mm pressure of neon. When an electrical discharge is caused to occur in the tube it excites mainly helium atoms, many of which are raised to excited metastable states. By

a fortuitous coincidence, the energy of one of these excited helium states is nearly equal to the energy of one of the excited states of neon, which allows radiationless transfer of energy from the helium to the neon, resulting in a population inversion for particular pairs of states of the neon. That is, many of the neon atoms will be preferentially in an upper state. These neon atoms can be stimulated to emit radiation, resulting in light amplification by stimulated emission of radiation, or LASER action. The emitted light will form a coherent beam that will be continuous if the end mirrors are properly aligned. The laser emission occurs at three distinct wavelengths, two of which are in the infrared part of the spectrum, at 3.39 microns and 1.15 microns, while the third occurs in the visible, at 0.6328 microns, or 6328 Angstroms. This is the characteristic red light emitted by the laser.

As the last required experiment in this lab, set up the helium-neon laser so that you can simultaneously observe the laser light and the light from a gas lamp through the spectrometer. This can be accomplished by scattering some of the laser light into the entrance slit of the spectrometer from a piece of paper attached to the neon lamp, such that the edge of the paper is about half-way up the slit. If you arrange the setup carefully you should be able to see just the gas-lamp spectrum in the upper half of the field of view, and just the laser light (a single line) in the lower half. Does the laser radiation coincide with any of the lines observed in the neon spectrum? How about the helium spectrum? Answer this question in two ways: by eye (does the laser line look like it overlaps another line *exactly*), and by measuring the wavelength of all relevant lines by your usual method.

Whether you find that the laser line matches another line or not, clearly state in your report both your conclusion from the data and your expectation from the theory. They may agree or disagree. You may understand the theory well enough from this discussion, or you may have to look it up in outside sources or get help from the instructors.

Table 4 (Below): Hydrogen, Helium, mercury, and neon emission spectra. I signifies the neutral atom and II signifies the singly ionized atom.

HYDROGEN (H)
Z = 1

H I
Ref. 214 — W.C.M.

Intensity	Wavelength (Air)	
5	3835.384	I
6	3889.049	I
8	3970.072	I
15	4101.74	I
30	4340.47	I
80	4861.33	I
120	6562.72	I
180	6562.852	I
5	9545.97	I
7	10049.4	I
12	10938.1	I
20	12818.1	I
40	18751.0	I
5	21655.3	I
8	26251.5	I
15	40511.6	I
4	46525.1	I
6	74578	I
3	123685	I

HELIUM (He)
Z = 2

He I and II
Ref. 16, 94, 173, 183, 317
W.C.M.

Intensity	Wavelength (Air)	
7	2385.40	II
9	2511.20	II
50	2577.6	I
1	2723.19	I
12	2733.30	II
2	2763.80	I
10	2818.2	I
4	2829.08	I
10	2945.11	I
40	3013.7	I
20	3187.74	I
3	3202.96	II
15	3203.10	II
1	3354.55	I
2	3447.59	I
1	3587.27	I
3	3613.64	I
2	3634.23	I
3	3705.00	I
1	3732.86	I
10	3819.607	I
1	3819.76	I
500	3888.65	I
20	3964.729	I
1	4009.27	I
50	4026.191	I
5	4026.36	I
12	4120.82	I
2	4120.99	I
3	4143.76	I
10	4387.929	I
3	4437.55	I
200	4471.479	I
25	4471.68	I
6	4685.4	II
30	4685.7	II
30	4713.146	I
4	4713.38	I
20	4921.931	I
100	5015.678	I
10	5047.74	I
5	5411.52	II
500	5875.62	I
100	5875.97	I
8	6560.10	II
100	6678.15	I
3	6867.48	I
200	7065.19	I
30	7065.71	I
50	7281.35	I
1	7816.15	I

MERCURY (NATURAL) (Hg)
Z = 80

Hg I and II (nat.)
Ref. 34, 45, 90, 117, 133, 189, 235, 304, 327, 328 — R.W.S.

Intensity	Wavelength (Air)	
90	2026.97	II
90	2052.93	II
70	2148.00	II
5	2247.55	I
60	2262.23	II
20	2302.06	I
15	2323.20	I
5	2340.57	I
20	2345.43	I
20	2352.48	I
100	2378.32	I
20	2380.00	I
40	2399.38	I
20	2399.73	I
10	2400.49	I
60	2407.35	II
50	2414.13	II
5	2441.06	I
20	2446.90	I
15	2464.06	I
40	2482.00	I
30	2482.72	I
40	2483.82	I
90	2534.77	I
15000	2536.52	I
25	2563.86	I
25	2576.29	I
5	2578.91	I
15	2625.19	I
5	2639.78	I
250	2652.04	I
400	2653.69	I
100	2655.13	I
5	2674.91	I
50	2698.83	I
50	2699.38	I
80	2705.36	II
80	2752.78	I
20	2759.71	I
40	2803.46	I
30	2804.43	I
2	2805.34	I
2	2806.77	I
150	2814.93	II
750	2847.68	II
50	2856.94	I
150	2893.60	I
150	2916.27	II
60	2925.41	I
150	2935.94	II
400	2947.08	II
1200	2967.28	I
300	3021.50	I
120	3023.47	I
30	3025.61	I
50	3027.49	I
400	3125.67	I
320	3131.55	I
320	3131.84	I
400	3208.20	II
400	3264.06	II
80	3341.48	I
100	3385.25	II
400	3451.69	II
200	3549.42	II
2800	3650.15	I
300	3654.84	I
80	3662.88	I
240	3663.28	I
30	3701.44	I
35	3704.17	I
30	3801.66	I
100	3806.38	II
20	3901.87	I
60	3906.37	I
100	3918.92	II
200	3983.96	II
1800	4046.56	I
150	4077.83	I
40	4108.05	I
250	4339.22	I
400	4347.49	I
4000	4358.33	I
100	4398.62	II
90	4660.28	II
80	4855.72	II
5	4883.00	I
5	4889.91	I
80	4916.07	I
5	4970.37	I
5	4980.64	I
20	5102.70	I
40	5120.64	I
100	5128.45	II
20	5137.94	I
20	5290.74	I
5	5316.78	I
60	5354.05	I
30	5384.63	I
1100	5460.74	I
30	5549.63	I
160	5675.86	I
240	5769.60	I
100	5789.66	I
280	5790.66	I
140	5803.78	I
60	5859.25	I
60	5871.73	II
20	5871.98	I
20	6072.72	I
1000	6149.50	II
30	6234.40	I
80	6521.13	II
160	6716.43	I
250	6907.52	I
250	7081.90	I
200	7091.86	I
40	7346.37	II
100	7485.87	II
20	7728.82	I
100	7944.66	II
2000	10139.75	I

NEON (Ne)
Z = 10

Ne I and II
Ref. 56, 58, 118, 150, 230 —
S.P.D.

Intensity		Wavelength AIR								
80		2007.01	II	150		3297.73	II	120	4457.05	II
80		2025.56	II	150		3309.74	II	100	4522.72	II
150		2085.47	II	300		3319.72	II	10	4537.754	I
180		2096.11	II	1000		3323.74	II	10	4540.380	I
120		2096.25	II	150		3327.15	II	100	4569.06	II
80	p	2562.12	II	100		3329.16	II	15	4704.395	I
90	w	2567.12	II	200		3334.84	II	12	4708.862	I
80		2623.11	II	150		3344.40	II	10	4710.067	I
80		2629.89	II	300		3345.45	II	10	4712.066	I
90	w	2636.07	II	150		3345.83	II	15	4715.347	I
80		2638.29	II	200		3355.02	II	10	4752.732	I
80		2644.10	II	120		3357.82	II	12	4788.927	I
80		2762.92	II	200		3360.60	II	10	4790.22	I
90		2792.02	II	120		3362.16	II	10	4827.344	I
80		2794.22	II	100		3362.71	II	10	4884.917	I
100		2809.48	II	120		3367.22	II	4	5005.159	I
80		2906.59	II	12		3369.808	I	10	5037.751	I
80		2906.82	II	40		3369.908	I	10	5144.938	I
90		2910.06	II	100		3371.80	II	25	5330.778	I
90		2910.41	II	500		3378.22	II	20	5341.094	I
80		2911.14	II	150		3388.42	II	8	5343.283	I
80		2915.12	II	120		3388.94	II	60	5400.562	I
80		2925.62	II	300		3392.80	II	5	5562.766	I
80	w	2932.10	II	100		3404.82	II	10	5656.659	I
80		2940.65	II	120		3406.95	II	5	5719.225	I
90		2946.04	II	100		3413.15	II	12	5748.298	I
150		2955.72	II	120		3416.91	II	80	5764.419	I
150		2963.24	II	120		3417.69	II	12	5804.450	I
150		2967.18	II	50		3417.904	I	40	5820.156	I
100		2973.10	II	15		3418.006	I	500	5852.488	I
15		2974.72	I	120		3428.69	II	100	5872.828	I
100		2979.46	II	60		3447.703	I	100	5881.895	I
12		2982.67	I	50		3454.195	I	60	5902.462	I
150		3001.67	II	100		3456.61	II	60	5906.429	I
120	p	3017.31	II	100		3459.32	II	100	5944.834	I
300		3027.02	II	25		3460.524	I	100	5965.471	I
300		3028.86	II	30		3464.339	I	100	5974.627	I
100		3030.79	II	30		3466.579	I	120	5975.534	I
120		3034.46	II	60		3472.571	I	80	5987.907	I
100		3035.92	II	150		3479.52	II	100	6029.997	I
100		3037.72	II	200		3480.72	II	100	6074.338	I
100		3039.59	II	200		3481.93	II	80	6096.163	I
100		3044.09	II	25		3498.064	I	60	6128.450	I
100		3045.56	II	30		3501.216	I	100	6143.063	I
120		3047.56	II	25		3515.191	I	120	6163.594	I
100		3054.34	II	150		3520.472	I	250	6182.146	I
100		3054.68	II	120		3542.85	II	150	6217.281	I
100		3059.11	II	120		3557.80	II	150	6266.495	I
100		3062.49	II	100		3561.20	II	60	6304.789	I
100		3063.30	II	250		3568.50	II	7	6328.165	I
100		3070.89	II	100		3574.18	II	100	6334.428	I
100		3071.53	II	200		3574.61	II	120	6382.992	I
100		3075.73	II	50		3593.526	I	200	6402.246	I
120		3088.17	II	30		3593.640	I	150	6506.528	I
100		3092.09	II	15		3600.169	I	60	6532.882	I
120		3092.90	II	20		3633.665	I	150	6598.953	I
100		3094.01	II	150		3643.93	II	70	6652.093	I
100		3095.10	II	200		3664.07	II	90	6678.276	I
100		3097.13	II	20		3682.243	I	20	6717.043	I
100		3117.98	II	12		3685.736	I	100	6929.467	I
120		3118.16	II	200		3694.21	II	90	7024.050	I
10		3126.199	I	10		3701.225	I	100	7032.413	I
300		3141.33	II	150		3709.62	II	50	7051.292	I
100		3143.72	II	250		3713.08	II	80	7059.107	I
100	p	3148.68	II	250		3727.11	II	100	7173.938	I
100		3164.43	II	800		3766.26	II	150	7213.20	II
100		3165.65	II	1000		3777.13	II	150	7235.19	II
100		3188.74	II	100		3818.43	II	100	7245.167	I
120		3194.58	II	120		3829.75	II	150	7343.94	II
500		3198.59	II	150		4219.74	II	40	7472.439	I
60		3208.96	II	100		4233.85	II	90	7488.871	I
120		3209.36	II	120		4250.65	II	100	7492.10	II
120		3213.74	II	120		4369.86	II	150	7522.82	II
150		3214.33	II	70		4379.40	II	80	7535.774	I
150		3218.19	II	150		4379.55	II	60	7544.044	I
120		3224.82	II	100		4385.06	II	100	7724.628	I
120		3229.57	II	200		4391.99	II	120	7740.74	II
200		3230.07	II	150		4397.99	II	300	7839.055	I
120		3230.42	II	150		4409.30	II	120	7926.20	II
120		3232.02	II	100		4413.22	II	400	7927.118	I
150		3232.37	II	100		4421.39	II	700	7936.996	I
100		3243.40	II	100	p	4428.52	II	2000	7943.181	I
100		3244.10	II	100	p	4428.63	II	2000	8082.458	I
100		3248.34	II	150	p	4430.90	II	100	8084.34	II
100		3250.36	II	150	p	4430.94	II	1000	8118.549	I

6. LOW FREQUENCY IMPEDANCES

In this experiment, your main project will be to analyze the impedance versus frequency of several "mystery boxes" containing combinations of linear components (resistors, capacitors, and inductors), and deduce what combination is in the box and what the values of the components are. This will be the main subject of your report; your instructor may tell you to include other elements, so be sure you know what you're expected to include in your report, in your lab notebook, and in any associated homework set your instructor may assign.

The experiment was designed with the following learning goals in mind:

- Refreshing and deepening the understanding of simple DC and AC circuits that you began developing in Physics 5C;

- Giving you practice in operating a digital oscilloscope in a number of display modes, and giving you a better understanding of how it works, what it can do, and its limitations; and

- Making you comfortable in a data-analysis situation where measurement error is *not* the main cause of disagreement between your data and your model. Instead, it will be the fact that the models of linear circuits themselves are approximate that causes disagreement. In this situation, what kind of statements can be rigorously made, and how can the uncertainty in parameters (like the values of resistances, capacitances, and inductances) be estimated?

6.1 Background

Here are some things you should understand about the measurements you will be taking in this laboratory; if you're not clear about anything on this list, be sure to work with your instructor or TA:

- **Reading a dial that you set, or the label on a component, is not a measurement.** Measure resistances, voltages, and currents with a multimeter; measure frequencies with an oscilloscope. Do not trust labels.

- **Understand what AC and DC coupling in an oscilloscope mean,** and be sure you're using the right one for your purpose.

- **Every circuit has some inductance and some capacitance.** Understand why. When not produced by discrete components on purpose, these are called “stray”. One will give you trouble at very high frequencies, and one at very low. Which is which? Before you take data for the “black boxes,” estimate in advance what range of frequencies you might want to take to avoid these problems. You will need to know that the inductances and capacitances that we're trying to measure are in the range of mH and μF to nF, respectively.

- **It is easy to destroy components in the resistor substitution boxes by putting too much power through them!** They are rated to only 1/2 watt. Calculate *in advance* what voltage/current/power you're likely to get for a given resistor box setting. If the power exceeds 1/2 Watt, use one of the higher power resistors in the parts box. It's safe to start at very high resistance settings to get an idea of where you're headed, and stop before you get to settings that are too low. It is also a good idea to check the resistance in the box with an ohm meter. Some of the resistors may have been damaged.

6.2 Preliminary Experiment: Output Impedance (Resistance)

Here we learn about the common behavior of power supplies, which do not behave in the simple way we assumed in Physics 5A, particularly when trying to put current through a low impedance.

If current is drawn from a "9 volt" flashlight battery, the voltage at the terminals of the battery will be less than 9 volts. This is because the battery itself has resistance, across which there will be a voltage drop, which has the effect of reducing the voltage available at the battery terminals.

Although the innards of a flashlight battery (or indeed any supplier of electrical power) may be quite complex, we may think of those innards in a simple way.

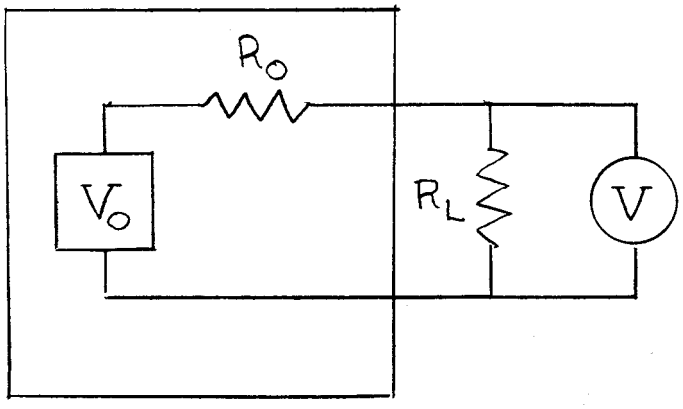

Figure 19. Thevenin's model for a power source.

Thevenin's theorem of linear circuit analysis states that any two-terminal network of resistors and voltage sources is equivalent to a *single* resistor R_0 in series with a *single* voltage source V_0. A flashlight battery or other power supply, or even (with some generalization) an oscillator, is such a two-terminal network, with the two terminals being the output terminals of the device. Thus, we may represent any such device as in Figure 19.

If no load resistance is connected, no current flows, and the output voltage is just V_0. However, when a load resistor R_L is connected to the power supply, a current I will flow through it and the output voltage will be reduced to V, where

$$V = V_0 - IR_0 \tag{81}$$

The relation between output voltage and output current for any power source is called the *output characteristic*. For Figure 19, the graph of the output characteristic is a straight line whose voltage intercept is V_0 and whose slope is R_0. It looks like Figure 20. Note that the current *I* is simply the output voltage divided by R_L, as it must be regardless of the source.

Actually, many circuits and sources are nonlinear and follow a characteristic of this sort only for small output currents. At high currents heat may be generated that changes the internal resistance, or a regulating network with complex feedback loops may be driven out of its range, as in a regulated power supply, and this may cause the output characteristic to be nonlinear. Nevertheless, the concept of output impedance is still useful in describing the linear, small-current region in which these devices are normally operated.[41]

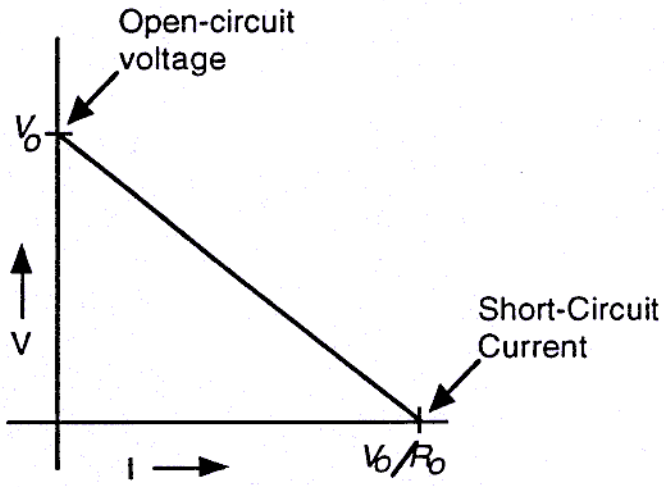

Figure 20. The linear output characteristic of a simple power source.

In some situations we may want to maximize the energy per second (the power) available from a source. It is not hard to prove that the maximum power will be transferred from the source to the load if the load resistor R_L is made equal to the output resistance R_0. This is called *matching the impedance* of the load to the source.[42] If the impedances are matched, the voltage drop across the output resistance will be equal to the voltage drop across the load resistance, and the output voltage will then be equal to half the open circuit voltage, *i. e.*, $V = V_0 / 2$.

[41] In some situations involving ac sources, the output impedance, as the generic word "impedance" implies, will not be totally resistive, but will contain a reactive (capacitative or inductive) component as well. In such cases the situation may not be quite as simple as the case we are considering, but the small-current output characteristic will still be linear, with a slope equal to the output resistance if the load is resistive.

[42] The concept of *matching the impedance* pops up in a variety of situations, not necessarily of an electrical nature. When impedances are matched between two systems, the maximum energy, or the maximum power, is transferred from one system to the other. Thus, for example, ears are shaped so as to couple sound waves to the ear drum with optimum efficiency, or a well-designed turbine will provide the maximum transfer of energy from a flowing stream of fluid to a useful machine. The electronic example we consider here provides us with the clearest example of impedance matching.

On the other hand we may be concerned, not with transferring the maximum power to a load, but with minimizing any perturbation of the source by the load. For example, if we use a meter to measure the output voltage of some device, or the voltage between any two terminals in a complex circuit, we are interested in knowing what the voltage would have been in the absence of the meter. In this case we want $R_L \gg R_0$, where the load resistance R_L is now the input resistance of the meter. If the circuit's output resistance R_0 is large, we must take care to ensure that the input resistance of the meter is much greater than the output resistance of the source.

Now we can start some experimenting. We can find the output resistance R_0 and the open circuit voltage V_0 of a source by measuring its output voltage at five or so values of load resistance (a range of 500 Ohms to 3,000 Ohms is a good start), and making a graph of V vs. I. From the slope and intercept (assuming a straight line results), we may determine R_0 and V_0.

To test these ideas, please determine the output resistance and the open circuit voltage for the source enclosed in a numbered aluminum box. Referring to Figure 19, measure the voltage in parallel with the load resistor. Please record the box number, but don't peek inside! Finally, please disconnect the load resistor when you have completed your measurements.

6.3 Preview of the main experiment: Introduction to Impedances

The six two-terminal black boxes A through F contain linear passive elements consisting of resistors, capacitors and inductors, either as individual components or as two or three components connected in a series or parallel configuration. The object of this section is to determine what components are inside each box, how they are connected, and what their values are, by passing an alternating current through the box, measuring the voltage across it, and examining the relationship between the current and the voltage as a function of frequency. In general, the voltage will not oscillate in phase with the current. The relationship between the voltage and the current may be described theoretically by using the idea of *complex impedance*. Some possible configurations can (and should) be quickly ruled out by measuring the dc resistance of the box when first encountering it.

Before examining boxes A-F it will help to review the basic theory of complex impedance. Although we shall focus on these simple electronic systems, the use of complex variables to describe the response of a linear system to a sinusoidally varying stimulus turns out to be quite general. Read carefully through the following background material, including the appendix on resonant circuits before starting on the experiments involving boxes A-F.

You need to take at least 20 data points. Cover the range from a few tens of Hz to a bit over 100 kHz, but be aware that stray capacitances and inductances may have effects on each extreme end. If you see behavior that is not monotonic (*i. e.* the impedance does not just continuously increase or decrease with frequency), that's a sign of a resonance. Take extra data points around the resonance; by reading the background materials in this chapter carefully, you should be able to figure out more than one way to use them!

6.4 Preview of experiment on Nonlinear Elements

You will be looking at some devices where impedance may or may not vary with frequency, but where it *will* vary with applied voltage. These are called nonlinear devices. This section is qualitative rather than quantitative but it is just as important to observe carefully and record honestly. This is a place where students often say that their observations agree with the theory or their expectations, when in fact they don't. If your observations and expectations don't agree, we're still quite happy as long as you say so clearly. But if you don't catch the disagreement, and you say everything is fine, and we do catch it, prepare for lots of angry red ink.

6.5 How to Use Complex Variables to Analyze AC Circuits

The use of complex variables can greatly simplify the description of the relationship between the voltage and the current flowing in a circuit containing *reactances*, *i. e.*, inductors and capacitors. We discuss some basic ideas here. Further discussion involving applications to resonant circuits appears in an appendix. The basic difference between reactance and resistance is that reactance will take energy from the power source, store it in fields (electric or magnetic) and return it to the circuit at a different phase of the oscillation, while resistance will permanently dissipate that power into thermal energy.

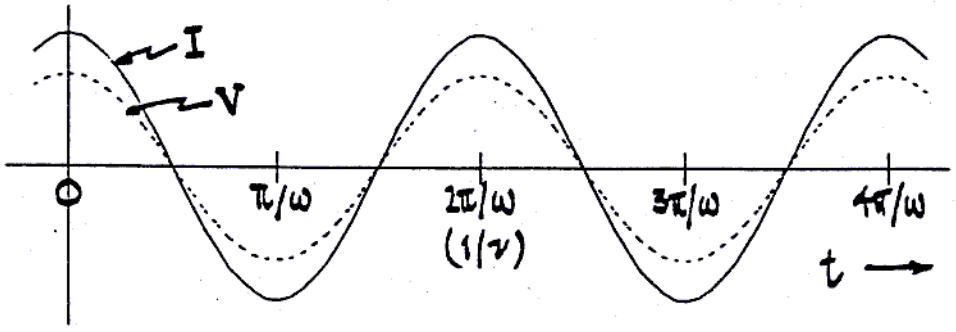

Figure 21. The relationship of voltage to current in a resistor.

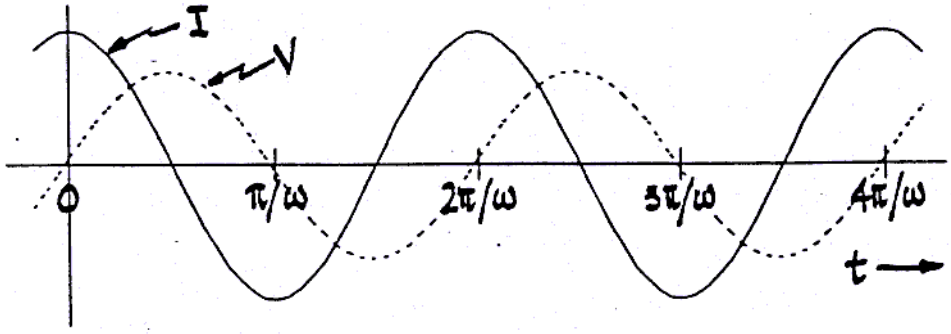

Figure 22. The relationship of voltage to current in an inductor.

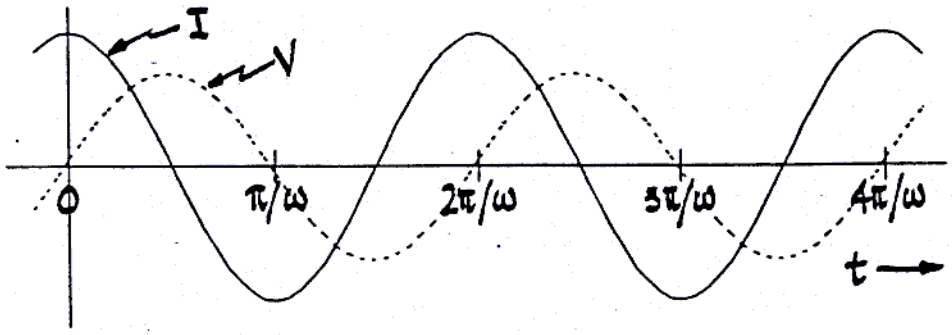

Figure 23. The relationship of voltage to current in a capacitor.

Consider a resistor *R* through which a current $I = I_0 \cos(\omega t)$ flows. Here ω is the angular frequency, equal to $2\pi\nu$, where ν is the actual frequency in Hertz. The voltage across the resistor is simply $V = RI = RI_0 \cos(\omega t)$. We say that the voltage *V* is "in phase" with the current I (as in Figure 21).

Now consider a pure inductance *L* with a current $I = I_0 \cos(\omega t)$ flowing through it. The voltage across it will be

$$V = L\frac{dI}{dt} = -\omega L I_0 \sin(\omega t) = \omega L I_0 \cos(\omega t + \frac{\pi}{2}) \tag{82}$$

We say that the voltage *leads* the current by a phase angle of $\pi/2$, or 90 degrees. We could also say that the current *lags* the voltage by 90 degrees. Graphically it's like Figure 22**.**

For a capacitance C, again with a current $I = I_0 \cos(\omega t)$ flowing in the wires attached to the capacitor, the charge Q appearing on the capacitor plates will be

$$Q = \int I dt = \int I_0 \cos(\omega t) dt = \frac{I_0}{\omega} \sin(\omega t) \tag{83}$$

and the voltage across the capacitor becomes

$$V = \frac{Q}{C} = \frac{I_0}{\omega C} \sin(\omega t) = \frac{I_0}{\omega C} \cos(\omega t - \frac{\pi}{2}). \tag{84}$$

We can set the constant of integration to zero. Why? What would it imply if it weren't zero? We say that the voltage *lags* the current by a phase angle of 90 degrees, so the graphical picture looks like Figure 23.

Note that while the case of a resistor is simple, with $V = IR$, the cases involving inductors and capacitors are not. In general, the voltage will not be in phase with the current, and therefore will *not* be simply proportional to it. We can simplify the analysis, however, by using complex variables.

To see how this works, consider a pure inductance L through which an actual current I_r flows. The voltage across the inductor is then

$$V_r = L\frac{dI_r}{dt} \tag{85}$$

If the actual current is I_i instead of I_r, the voltage will be

$$V_i = L\frac{dI_i}{dt} \tag{86}$$

The purpose of the subscripts *r* and i will become clear in a moment. Since d/dt is a linear operator, a linear combination of I_r and I_i will produce the *same* linear combination of V_r and V_i. Hence it must be true that

$$V_r + jV_i = L\frac{d}{dt}(I_r + jI_i) \tag{87}$$

where $j = \sqrt{-1}$.[43] Here I_r and I_i can be thought of as the real and imaginary parts of a *complex* current $I = I_r + jI_i$ producing a *complex* voltage $V = V_r + jV_i$. Of course, being complex, I can't be a physical current, nor V a physical voltage; they are just mathematical constructions.

The constructions are useful, however, whenever the current is sinusoidal in time, for if $I_r = I_0\cos(\omega t)$, we can *choose* $I_i = I_0\sin(\omega t)$, so that[44]

$$I = I_0[\cos(\omega t) + j\sin(\omega t)] = I_0 e^{j\omega t} \tag{88}$$

Then

$$V = L\frac{dI}{dt} = (j\omega L)I_0 e^{j\omega t} = (j\omega L)I \tag{89}$$

and V is now *proportional* to I. We call this constant of proportionality the *complex impedance*, and denote it by Z. In this case, $Z = j\omega L$, the complex impedance for an inductor.

The quantity Z is useful because it allows us, when a given sinusoidally varying current flows through any given combination of resistors, capacitors and inductors, to calculate the voltage across that combination using algebra alone. The recipe (or the algorithm) is always as follows:

1. Take the current to be $I = I_0 e^{j\omega t}$. The physical current is hence the real part of this quantity, or $I_0\cos(\omega t)$.[45]

---

[43] We use j rather than i to avoid possible confusion with i for current.

[44] One often sees the words "Let the current be $I = I_0\exp(j\omega t)$." What is meant is "Let the current be the *real* part of $I_0\exp(j\omega t)$."

[45] If we want the physical current to have a phase δ different from zero, that is, of the form $I_0\cos(\omega t + \delta)$, we can take the corresponding complex current to be $I_0\exp(j(\omega t + \delta))$. It is usually simplest, however, to take $\delta = 0$.

2. Find Z for the particular combination of resistors, capacitors and inductors being considered. For a resistor, $Z = R$. For an inductor, $Z = j\omega L$, while for a capacitor, $Z = 1/(j\omega C)$. For any series or parallel combination of these components, the total Z may be found using the usual rules for adding resistors in series or parallel. Thus, for an inductor and a resistor in series, $Z = R + j\omega L$, while for an inductor and a resistor in parallel, $1/Z = 1/R + 1/j\omega L = 1/R - j/\omega L$, etc.
3. Find V simply by multiplying I by Z: $V = IZ$.
4. Finally find the real part of V to get the physical voltage across the combination as a function of time.

It is sometimes useful to use the *admittance* $Y = 1/Z$ in doing such calculations. We illustrate this process with the following two examples.

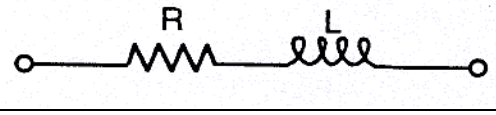

Figure 24. A resistor and inductor in series.

6.5.1 Resistor and inductor in series.

Symbolically a resistor and inductor in series looks like Figure 24; the complex impedance is $Z = R + j\omega L$.

As with any complex quantity, we can write

$$Z = |Z| e^{j\theta} \tag{90}$$

where $|Z| = (Z \cdot Z^*)^{1/2}$ is the *amplitude* of the complex impedance, and θ is its *phase*. Thus,

$$|Z| = (R^2 + \omega^2 L^2)^{\frac{1}{2}} \qquad \text{and} \qquad \tan\theta = \frac{\omega L}{R} \tag{91}$$

Hence

$$V = IZ = I_0 |Z| e^{j(\omega t + \theta)} \tag{92}$$

and

$$V_r = I_0 |Z| \cos(\omega t + \theta) \tag{93}$$

The voltage will lead the current by the phase angle θ.

We check that the limiting cases make sense: At low frequency, $\omega L \ll R$, so

$$\theta \to 0, \qquad |Z| \to R, \qquad \text{and} \qquad V_r \to I_0 R \cos(\omega t) \tag{94}$$

Here the impedance of the inductor is negligible, and the voltage is in phase with the current, as expected. At high frequency, $\omega L \gg R$, so

$$\theta \to \frac{\pi}{2}, \qquad |Z| \to \omega L, \qquad \text{and} \qquad V_r \to I_0 \omega L \cos(\omega t + \frac{\pi}{2}) \tag{95}$$

Here the impedance of the resistor is negligible, and the voltage leads the current by 90 degrees.

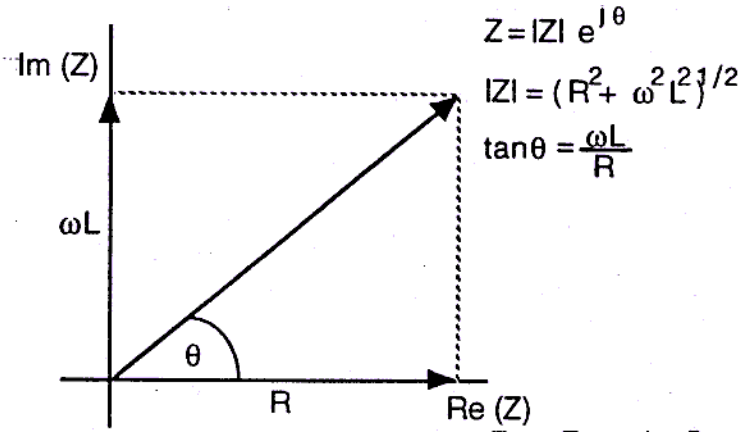

Figure 25. The phasor representing $Z = R + j\omega L$

6.5.2 A side note about phasors.

There is a graphical representation of the complex impedance that is useful for interpreting the current-voltage relationships in circuits. Any complex quantity can be represented by a two-component vector, sometimes called a *phasor*, in the complex plane. Thus, the quantity $Z = R + j\omega L$, for example, can be represented like Figure 25.

Phasor diagrams are also useful for representing complex admittances, as well as complex voltages and currents. Note how easily the limiting cases may be analyzed.

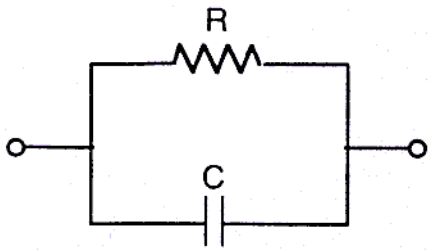

Figure 26. A resistor and capacitor in parallel.

6.5.3 Resistor and capacitor in parallel.

For parallel circuits it is more convenient to work with the complex *admittance*, $Y = 1/Z$, instead of the complex impedance *Z*. This familiar from adding resistors in parallel, where the quantities $1/R$ are what add linearly. The complex admittance for a resistor and capacitor in parallel (Figure 26) is $Y = 1/Z = 1/R + j\omega C = |Y|e^{j\phi}$,where

$$|Y| = \left(\frac{1}{R^2} + \omega^2 C^2\right)^{\frac{1}{2}} \qquad \text{and} \qquad \tan\phi = \omega CR \tag{96}$$

Hence

$$V = \frac{I}{Y} = \frac{I_0}{|Y|}e^{j(\omega t - \phi)} \tag{97}$$

and

$$V_r = \frac{I_0}{|Y|}\cos(\omega t - \phi) \tag{98}$$

Note that the voltage now *lags* the current by the angle ϕ. In this example, where components are connected in parallel, it is convenient to use the complex admittance rather than the impedance in analyzing the circuit.

Again we check that the limiting cases make sense:

At low frequency:

$$\omega C \ll \frac{1}{R}, \qquad \text{so} \qquad \phi \to 0, \qquad |Z| \to R \qquad \text{and} \qquad V_r \to I_0 R\cos(\omega t) \tag{99}$$

Since the impedance of the capacitor is very high we can neglect its presence, and the voltage is nearly in phase with the current.

At high frequency:

$$\omega C \gg \frac{1}{R}, \qquad \text{so} \qquad \phi \to \frac{\pi}{2}, \qquad |Z| \to \frac{1}{\omega C} \qquad \text{and} \qquad V_r \to \frac{I_0}{\omega C}\cos(\omega t - \frac{\pi}{2}) \tag{100}$$

Here the capacitor dominates, and the voltage lags the current by nearly 90 degrees.

6.6 Exercises

Work these out in your lab notebook; they don't go in your report. Do them before you start examining boxes A-F.

1. If a total current equal to $I_0\cos(\omega t)$ flows through a circuit consisting of a resistor R and an inductor L connected in parallel, show that the voltage across the circuit is given by

$$V_r = \frac{I_0}{|Y|}\cos(\omega t - \phi) \tag{101}$$

where

$$|Y| = \left(\frac{1}{R^2} + \frac{1}{(\omega L)^2} \right)^{\frac{1}{2}} \qquad \text{and} \qquad \tan\phi = -\frac{R}{\omega L} \tag{102}$$

2. If a voltage equal to $V_0 \cos(\omega t)$ is applied to a circuit consisting of a resistor R and a capacitor C connected in series, show that the current through the circuit is given by

$$I_r = \frac{V_0}{|Z|}\cos(\omega t - \theta) \tag{103}$$

where

$$|Z| = \left(R^2 + \frac{1}{(\omega C)^2} \right)^{\frac{1}{2}} \qquad \text{and} \qquad \tan\theta = -\frac{1}{R\omega C} \tag{104}$$

In this exercise, you are asked to find the current through a network with a given voltage across it rather than the other way around. Such is often the case. In step 1, take the *voltage* to be $V = V_0 e^{j\omega t}$, and in step 3, find *I* simply by dividing V by Z. Then in step 4, find the real part of I to get the real total current through the network. Note that the results of first exercise above can be expressed as

$$|Y|^2 = \left(\frac{1}{R^2} + \frac{1}{L^2}\left(\frac{1}{\omega^2} \right) \right). \tag{105}$$

Therefore, a plot of $|Y|^2$ vs. $(1/\omega^2)$ will be a *straight line*.

In the second exercise,

$$|Z|^2 = \left(R^2 + \frac{1}{C^2}\left(\frac{1}{\omega^2} \right) \right)$$

so that a plot of $|Z|^2$ vs. $(1/\omega^2)$ will be a straight line.

In most of the following linear-element experiments, it will be possible (and highly recommended) to cast your measurements in a form that allows you to fit to a straight line whose intercept and slope will yield the quantities of interest. That is, you will make a hypothesis that you have a particular circuit arrangement, figure out what *should* be linear for that configuration ($|Z|^2$ vs. $(1/\omega^2)$? $|Z|^2$ vs. (ω^2)? $|Z|$ vs. (ω)? $|Y|^2$ vs. $(1/\omega^2)$? *etc.*). For example, in exercise 2 above, the intercept is R^2, and the slope is $1/C^2$. Note that the intercept is the value of $|Z|^2$ when $(1/\omega^2)$ is zero: that is, at infinite frequency.

6.7 Units

With electrical impedance, it is crucial to keep your units straight. If you are told you have an inductor of 20 millihenries at a frequency of 60 kiloHertz, what is its impedance in ohms?

$$Z = \omega L = 2\pi fL = 2\pi \times 6 \times 10^4 \times 20 \times 10^{-3} = 7540\,\Omega \tag{106}$$

That is, don't forget the 2π ! Likewise, a capacitor of 10 nanofarads (nF) and frequency of 30 kHz has an impedance of

$$Z = \frac{1}{\omega C} = \frac{1}{2\pi \times 30 \times 10^3 \times 10 \times 10^{-9}} = 530\,\Omega \tag{107}$$

6.8 Finally the Main Experiment: Linear Elements

As noted in the preceding discussion, one way to determine the contents of an unknown linear two-terminal network is to pass a current $I_0 \cos(\omega t)$ through it and monitor the voltage across it while varying the frequency. It is usually helpful first to measure the impedance of the unknown at zero frequency with an ohm meter. Inductors have a small dc resistance, while capacitors have a nearly infinite resistance.

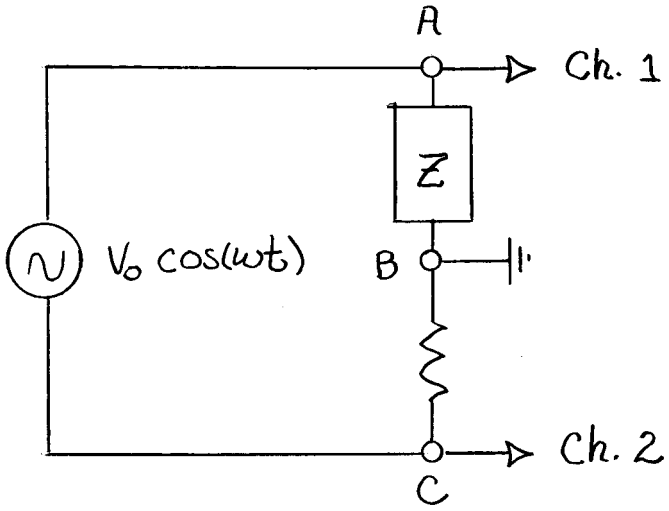

Figure 27. A circuit useful for analyzing the nature of *Z*.

To observe Z as a function of frequency, it is effective to use the circuit shown in Figure 27. You can simultaneously measure the voltage V_{CB} across the sensing resistor and the voltage V_{AB} across the unknown impedance, since they have a common ground. Note that the voltage across the sensing resistor is proportional to the *negative* of the current through the unknown impedance. To minimize stray pickup, at each frequency you should choose a sensing resistance (with the resistance decade box) so that the amplitudes of the two respective signals are coomparable. The phase and magnitude of the impedance can be determined from the phase and amplitude of the voltage across the unknown impedance, relative to the phase and amplitude across the sensing resistor. In fact, if the sensing resistance is chosen to give identical amplitudes to both, then the

magnitude of *Z* is just the sensing resistance! Using the MEASURE function on your oscilloscope, you can accurately measure amplitudes, frequencies, and phases of sinusoidal signals on each oscilloscope channel. Also note that for convenience you can invert the signal on one or more oscilloscope channels. This is useful because the voltage across the sensing resistor is proportional to the *negative* of the current through the unknown impedance.

Please note that the impedance of the oscilloscope is not infinite. In fact, it is 1 megohm, in parallel with a capacitance of 20 pF. Thus, the oscilloscope will "load" any circuit whose impedance is comparable to one megohm or more. **As a consequence, this technique will not be accurate for sensing resistors greater than about 100 kΩ.**

An excellent method to use in analyzing your data is to see whether a straight-line plot can be made, involving functions of $|Z|$ and ω, from which component values might be deduced from the slope and intercept of the straight line. While it is convenient to prepare such graphs using our computing facilities, you'll understand the process better if you do at least one or two graphs by hand.

Let us consider an example in which using straight lines is not so obvious. Suppose our box contains a parallel resonant circuit as shown in Figure 32, with a resulting equation:

$$|Y|^2 = \frac{1}{R^2} + \left(\omega C - \frac{1}{\omega L}\right)^2 \tag{108}$$

First plot $|Y|^2$ vs. ω^2 for your high frequency points. The slope will be C^2 and the intercept $1/R^2$. Second, plot $|Y|^2$ vs. $1/\omega^2$ for your low frequency points. The slope will be $1/L^2$ and the intercept $1/R^2$. Finally note that the minimum value of $|Y|^2$ occurs when $\omega = \omega_0 = 1/\sqrt{LC}$, and check that your values obey this requirement. This gives an independent check which is available for any resonant circuit; **or**, if you are unhappy with either your high- or low-frequency data, you can use it to replace one of those measurements. Another check is given by the width of the resonance (see appendix), defined as the frequency difference between the points where the phase difference between current and voltage is 45°.

With all this in mind, see if you can determine (without peeking inside the box) what arrangement of components is contained in some of the lettered boxes (A-F), and what the magnitudes of the components are. In some cases you may find that your measurements will be affected by the input impedance of the oscilloscope, or by some stray capacitance that you may be unaware of. This is particularly true at frequencies above, say, about 50-100 kHz.

Examine a total of 3 boxes, as follows:

- **Either** box A or B.
- **Either** box D or F.
- **Either** box C or E.

Each box might be a single component (capacitor, resistor, inductor), two components in series or in parallel, or one of the resonant circuits in the Appendix (if it's a parallel resonant circuit, it's the

"realistic" kind with the resistor in series with the inductor, not with all three components in parallel). Recall that an inductor has resistance, so it effectively has a resistor in series with it already. Your initial examination of each box should cover a wide range of frequencies (from tens of Hz to 100 kHz) as well as a DC resistance measurement, to start deducing what you're looking at.

6.9 Secondary Experiment: Nonlinear Elements

Boxes J and K contain two-terminal nonlinear networks, that is, networks for which the magnitude of the response will *not* be proportional to the magnitude of the stimulus. Hence the current through such a device will not be proportional to the voltage across it, so that the graph of current vs voltage (the I -V characteristic) will not be a straight line.

For this section, we would like your report to include the description of the procedure and circuit you used (of course) in the Apparatus and Procedure section, then, in Data Analysis and Results, drawings (with reasonably accurate markings of the voltage scale) of what you observed on the oscilloscope, and a discussion of the comparison to your expectations.

Try using a circuit like that of Figure 27 to display the I -V characteristic of the device on the oscilloscope screen, with I on the vertical axis and V on the horizontal axis. You should use a sensing resistor of 1,000 Ω or greater to limit the current in the nonlinear element.

With the oscillator amplitude at zero, there should be a spot at the center of the screen. As you turn the amplitude up, the shape of the I--V characteristic should appear. Make a sketch of it in your notebook, noting the values of the current and voltage where it starts to become nonlinear. Discuss the behavior in light of the above discussion. What happens as you raise the frequency from very low frequencies to somewhat higher ones? Can you explain what is going on?

6.9.1 Box J: A diode.

Box J contains an ordinary silicon diode, a junction between *p-type* and *n-type* regions of a semiconducting crystalline material. It is commonly called a *p-n junction*. It possesses a basic asymmetry: current will flow easily from the *p-type* to the *n-type* (the forward direction), but with difficulty in the reverse direction. An ordinary diode is usually represented like Figure 28.

Figure 28. Symbol and construction of a diode.

It's hard to imagine any product of our society more ubiquitous than the *p-n* junction. This simple device has become the world's basic electronic building block. An ordinary transistor consists of two such junctions back-to-back, and of course nearly every box of electronics now contains zillions of transistors.

The simplest model of such a junction results from an analysis of the statistical physics of electrons in semiconductors.[46] This model predicts the current I that results when a voltage V is applied to the junction:[47]

$$I = I_S(e^{eV/kT} - 1) \tag{109}$$

The current will increase rapidly with V for positive V, but will rapidly approach the constant reverse saturation current I_S when V is negative. The value of I_S depends on properties of the semiconducting material from which the junction is made, and ideally is very small. The value of kT/e is about 26 millivolts at room temperature.[48] Hence for V greater than about $+100$ millivolts the "1" may be neglected, and the logarithm of the current should vary linearly with the voltage.

Try arranging your setup so that the voltage across the diode appears on the horizontal axis of the oscilloscope screen, while the current through it, which will be proportional to the voltage across the sensing resistor, appears on the vertical axis. (Be sure that the oscilloscope input is set on DC and not AC). **Be sure to not let the resistance fall below 1,000 Ω or you may burn out the diode.** Note that if the oscillator produces a sinusoidal voltage, the voltage (and current) will extend in both the positive and negative directions. Does the general shape of what you observe correspond to that predicted by Eq. (109)? If you have the time and inclination, try doing some dc measurements to see whether the logarithm of I varies linearly with V over some range of positive V.

6.9.2 Box K: A Zener diode.

Box K contains a Zener diode, which behaves like an ordinary diode if V is greater than some negative voltage called the Zener voltage. If V is made more negative than the Zener voltage, however, the diode begins rapidly to conduct again, owing to an "avalanche breakdown" effect (for some Zeners), or a "Zener breakdown" effect (for others). The details of the mechanism, and the value of the Zener voltage, are determined in the manufacturing process. Zeners are available for a selection of Zener voltages ranging from 2 to 200 volts. The symbol for a Zener diode is like Figure 29.

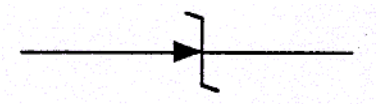

Figure 29. The symbol for a Zener diode.

Try looking at the Zener diode using the same setup as that for the ordinary silicon diode. You should see a "Z"-shaped characteristic on the screen. In your notebook, sketch I vs. V for both positive and negative voltages. **Be sure to not let the resistance fall below 1,000 Ω or you may burn out the diode.**

[46] See, for example, Kittel's classic text, *Introduction to Solid State Physics*.
[47] This equation has come to be known as the *Ebers-Moll* Equation.
[48] k is Boltzmann's constant, e is the electron charge and T is the absolute temperature.

6.9.3 Epilogue

In this lab, even more than the others, it is well to *think* before doing anything. Spend about 80 per cent of your time thinking, and only 20 per cent doing.

There's a good story Richard Feynman tells near the beginning of *Surely You're Joking, Mr. Feynman!*[49] When he was a boy during the depression he taught himself to fix radios. A poor man---"*The guy is obviously poor*," says Feynman---asks Feynman to look at his radio. Here's Feynman talking:

I say, "*What's the trouble with the radio?*"

He says, "*When I turn it on it makes a noise, and after a while the noise stops and everything's all right, but I don't like the noise at the beginning*."

I think to myself: "*What the hell! If he hasn't got any money, you'd think he could stand a little noise for a while.*"

And all the time, on the way to his house, he's saying things like, "*Do you know anything about radios? How do you know about radios---you're just a little boy!*"

He's putting me down the whole way, and I'm thinking, "*So what's the matter with him? So it makes a little noise.*"

But when we got there I went over to the radio and turned it on. Little noise? My God! No wonder the poor guy couldn't stand it. The thing began to roar and wobble---WUH BUH BUH BUH BUH---a tremendous amount of noise. Then it quieted down and played correctly. So I started to think: "*How can that happen?*"

I started walking back and forth, thinking, and I realize that one way it can happen is that the tubes are heating up in the wrong order---that is, the amplifier's all hot, the tubes are ready to go, and there's nothing feeding in, or there's some back circuit feeding in, or something's wrong in the beginning part---the RF part---and therefore it's making a lot of noise, picking up something. And when the RF circuit's finally going, and the grid voltages are adjusted, everything's all right.

So the guy says, "*What are you doing? You come to fix the radio, but you're only walking back and forth!*"

I say, "*I'm thinking!*" Then I said to myself, "*All right, take the tubes out, and reverse the order completely in the set.*" (Many radio sets in those days used the same tubes in different places---212's, I think they were, or 212-A's). So I changed the tubes around, stepped to the front of the radio, turned the thing on, and it's as quiet as a lamb: it waits until it heats up, and then plays perfectly---no noise.

[49] I recommend this book. It was published in 1985 by W. W. Norton. The passage quoted here appears on pages 19-20 of the book.

When a person has been negative to you, and then you do something like that, they're usually a hundred percent the other way, kind of to compensate. He got me other jobs, and kept telling everybody what a tremendous genius I was, saying, "*He fixes radios by thinking!*" The whole idea of thinking, to fix a radio---a little boy stops and thinks, and figures out how to do it---he never thought that was possible.

6.10 Appendix: A Review of Resonant Circuits

6.10.1 The series resonant circuit.

The complex impedance of this circuit is

$$Z = R + j\omega L + \frac{1}{j\omega C} = R + j(\omega L - \frac{1}{\omega C}) = | Z | e^{j\theta} \tag{110}$$

Hence if the current through the circuit is $I_0 \cos(\omega t)$, the voltage across it will be

$$V_{AB} = I_0 | Z | \cos(\omega t + \theta) \tag{111}$$

where

$$| Z | = \left(R^2 + (\omega L - \frac{1}{\omega C})^2 \right)^{\frac{1}{2}} \tag{112}$$

and

$$\tan\theta = \frac{1}{R}(\omega L - \frac{1}{\omega C}) \tag{113}$$

Note how the phase θ varies with frequency.

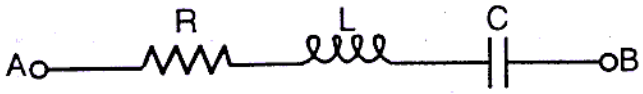

Figure 30. A resonant series RLC circuit.

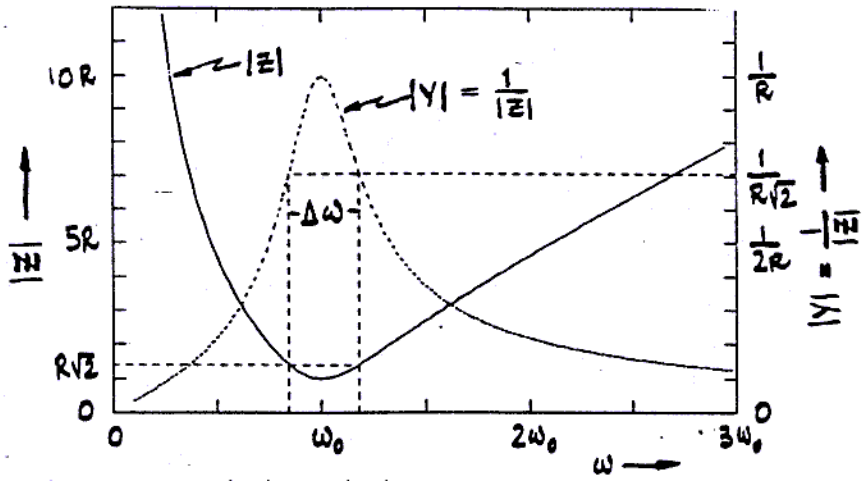

Figure 31. |Z| and |Y| vs. ω for a resonant series RLC circuit.

At resonance, the voltage is in phase with the current ($\theta = 0$), and the impedance of the circuit takes on its *minimum* value of *R*. This will happen when the angular frequency is

$$\omega_0 \equiv \text{resonant frequency} = \frac{1}{\sqrt{LC}} \tag{114}$$

A graph of the impedance amplitude |*Z*| as a function of the angular frequency ω will look something like Figure 31.

The two values of the frequency for which the voltage is $\pm 45^{\circ}$ out of phase with the current define the *width* of the resonance.[50] At these frequencies, which we denote by ω_1 and ω_2, the magnitude of the impedance is $\sqrt{2}$ times its minimum value of *R*, which occurs when $\theta = \pm 45^{\circ}$. The difference between ω_2 and ω_1 is related to the values of *L* and *R*:

$$\omega_2 - \omega_1 \equiv \Delta\omega = \frac{R}{L} \tag{115}$$

In addition, the *quality factor* "*Q*" of the circuit is given by

$$\text{quality factor} = Q = \frac{\omega_0}{\Delta\omega} = \frac{\nu_0}{\Delta\nu} = \frac{\omega_0 L}{R} = 2\pi\frac{\nu_0 L}{R} \tag{116}$$

In general, the separation of *any* two frequencies, ω_a and ω_b, for which $|Z| = \alpha R$, where α is some constant greater than 1, is related to the component values:

$$\omega_b - \omega_a = \frac{R}{L}(\alpha^2 - 1)^{\frac{1}{2}} \tag{117}$$

[50] It is the plot of |*Y*| vs ω, rather than |*Z*| vs ω, that has the appearance of a "resonance" curve. It is the *current* amplitude, rather than the *voltage* amplitude, that will go through a maximum at the resonant frequency for this circuit.

If we can measure the frequency dependence of $|Z|$, we may use Eqs. (114) and (115), together with the knowledge that the impedance at resonance is equal to R, to deduce values for each of the components of this resonant circuit.

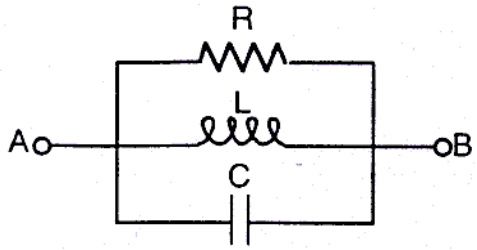

Figure 32. The parallel resonant RLC circuit.

6.10.2 The parallel resonant circuit.

Figure 32 shows an idealized version of the parallel resonant circuit; all three components are in parallel. All physical inductors have resistance of their own, however, so there will always be a resistance in series with the inductor. So you would have to add an extra, second resistor to this circuit, on the branch with the inductor. We have not given you any circuits this complicated. Instead, if we have given you a parallel resonant circuit at all, it will look like Figure 34 instead; the only resistance is that associated with the inductance. Now, we will derive the admittance for both the idealized and physical circuit, and show you an interesting relation between them.

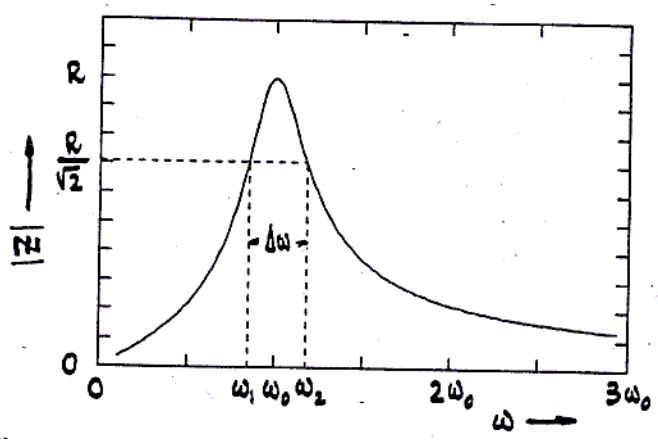

Figure 33. $|Z|$ vs. ω for a parallel resonant RLC circuit.

The complex admittance of the idealized parallel circuit is

$$Y = \frac{1}{Z} = \frac{1}{R} + \frac{1}{j\omega L} + j\omega C = \frac{1}{R} + j(\omega C - \frac{1}{\omega L}) = |Y|e^{j\phi} \tag{118}$$

Now, if the total current through the circuit is $I_0\cos(\omega t)$, the voltage across it will be

$$V_{AB} = \frac{I_0}{|Y|}\cos(\omega t - \phi) \tag{119}$$

where

$$|Y| = \left(\frac{1}{R^2} + (\omega C - \frac{1}{\omega L})^2 \right)^{\frac{1}{2}} \tag{120}$$

and

$$\tan\phi = R(\omega C - \frac{1}{\omega L}) \tag{121}$$

In this case, a graph of $|Z$ vs ω will look something like Figure 33.

Again the resonant frequency is given by Eq. (114), at which point the impedance of the circuit is equal to its *maximum* value of *R*. The width of the resonance is defined, just as for the series resonant circuit, by $\Delta\omega$, the difference between those values of the frequency for which the voltage is $\pm 45^o$ out of phase with the current. At these points the impedance $|Z|$ will be $R/\sqrt{2}$.

Reasoning similar to that for the series configuration leads to

$$\omega_2 - \omega_1 \equiv \Delta\omega = \frac{1}{RC} \tag{122}$$

In this case the quality factor *Q* is given by

$$Q = \frac{\omega_0}{\Delta\omega} = \frac{R}{\omega_0 L} \tag{123}$$

As with the series circuit, the values of the components can be deduced from a knowledge of the frequency dependence of $|Z|$.

There is one complication that arises with the parallel circuit. The usual parallel configuration of an inductor and a capacitor does *not* look like Figure 32, since any real inductor will have a resistance that is effectively in *series* with it, rather than in parallel. Thus, an actual parallel resonant circuit should be represented as in Figure 34.

Here the resistance *r* is the resistance of the inductor, as might be measured, for example, with an ohmmeter. If the "*Q*" of this circuit is not too small, that is, if $r \ll \omega L$ near the resonant frequency, then this circuit is *equivalent* to that shown in Figure 32. The physical circuit behaves near resonance according to the same formula as the idealized circuit, but in place of the idealized resistance *R* there is an *effective* resistance equal to

$$R_{eff} = \frac{\omega_0^2 L^2}{r} \tag{124}$$

Remember that you will be trying to find *r*, not R_{eff} !

Here's the proof: The admittance of the circuit of Figure 34 is

$$Y = j\omega C + \frac{1}{r + j\omega L} \tag{125}$$

and the square of the magnitude of the admittance is therefore

$$|Y|^2 = Y^* Y = \frac{(\omega r C)^2 + (1 - \omega^2 LC)^2}{r^2 + (\omega L)^2} \tag{126}$$

If we assume that $r^2 \ll (\omega L)^2$, we may neglect the term r^2 in the denominator, and the square of the admittance becomes

$$|Y|^2 = \frac{r^2 C^2}{L^2} + (\omega C - \frac{1}{\omega L})^2 \tag{127}$$

This expression has the same form as that of Eq. (126), and in fact will be *identical* to it if we let

$$\frac{1}{R} \equiv \frac{rC}{L} \tag{128}$$

Hence the two circuits have the same resonant frequency, and indeed have the same frequency response. Setting $\omega_0^2 = 1/LC$, we see that the quality factor *Q* is given by

$$Q = \frac{R}{\omega_0 L} = \frac{\omega_0 L}{r} \tag{129}$$

Note that the correspondence will be exact only in the limit that *Q* is large, that is, if $\omega_0 L \gg 1$.

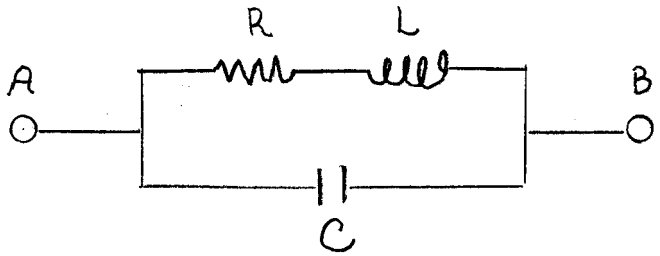

Figure 34: The real parallel resonant circuit.

7. ABSORPTION OF GAMMA RAYS BY MATTER

In this experiment, your main project will be to measure the rate at which gamma-rays of a certain energy are attenuated by lead, as a function of its thickness, and to determine whether the form of this dependence is exponential as predicted by the theory of how high-energy photons interact with matter. This will be the main subject of your report; your instructor may tell you to include other elements, so be sure you know what you're expected to include in your report, in your lab notebook, and in any associated homework set your instructor may assign.

The experiment was designed with the following learning goals in mind:

- To teach you some physics related to high-energy particle interactions that isn't covered in the rest of the required curriculum;
- To familiarize you with the kind of instrumentation and data analysis techniques used in experimental high-energy physics;
- To give you practice in the correct use of probability and statistics for the case of a Poisson process; and
- To give you further practice in the ways you can use a digital oscilloscope to analyze electronic signals.

The radio-isotope used, cesium-137 (^{137}Cs) has a half-life of approximately 30 years. Because of this long half-life, the intensity of the radioactivity of these sources does not change measurably during the course of the experiment, making them convenient to study.

For all kinds of detectors, a gamma-ray must be converted to a high-energy electron or (if its energy is high enough) an electron/positron pair in order to be detected. At low energies, this conversion takes place mostly by photoelectric absorption; at intermediate energies, by Compton scattering; and at very high energies, by e+/e- pair production in the field of the nucleus. In the Compton-dominated regime (very roughly 50 to 5000 keV, with each end of the range having a dependence on the atomic number of the medium), matter can be very transparent, and only a fraction of the photons incident on a detector might register.

Charged particles are usually detected by the ionization they produce while traveling through matter. One of the simplest detectors, the Geiger-Mueller (G-M) tube, will be used in this experiment.

7.1 About the Geiger-Mueller Tube

A Geiger-Mueller tube is a thin-windowed, conducting cylinder filled with a rare gas such as argon, and having a fine wire down the center. A positive high voltage is applied between the central wire and the cylinder wall. Alpha and beta particles emerging from external radioactive sources may penetrate the window and ionize the interior gas through collisions. Gamma rays, by contrast,

will typically undergo Compton scattering in the window, causing the emission of a high energy electron into the gas, which in turn will ionize the gas.

Once a gas molecule is ionized, the resulting electron will be attracted to the central wire (the anode), which is at a positive potential with respect to the wall (the cathode). As these electrons approach the central anode they gain sufficient energy to produce additional ionization by collisions; the ionization potential of argon is 15.7 electron-volts. The secondary electrons produced in this way are themselves accelerated and ionize more atoms, resulting in an avalanche of charges and a large current pulse. The avalanche stops when enough of slow-moving ions form a sheath around the anode, partially cancelling the formerly high electric field.

The number of electrons ultimately reaching the anode bears little relation to the initial amount of ionization deposited in the gas. However, the number of electrons, and hence the pulse size, increases roughly linearly with increases in the applied voltage.[51]

In these experiments we will be using a commercial counter that records negative pulses that have exceeded a fixed amplitude, approximately -0.80 volts (the counter box also contains the high voltage power supply for the tube). So once the high voltage is set to produce pulses that exceed this threshold, all events that leave energy in the tube will be recorded, and the count rate will be roughly independent of further increases in the voltage.

A final note of caution. As you increase the count rate (by bringing the source closer to the detector) or as you increase the high voltage, the average current drawn by the tube increases. Since the high voltage to the tube is delivered through a large resistor, the voltage drop across this resistor may become significant enough to reduce the actual high voltage at the anode. As a consequence, the pulse amplitude may actually *decrease* as you increase the rate or the high voltage.

7.2 About the health hazards of radiation

Radiation, in the broad sense, refers to any flux of subatomic particles, including photons (covering the entire electromagnetic spectrum from low frequencies up through X rays to high-energy gamma rays), all types of charged particles such as electrons and positrons (beta particles), protons, alpha particles (helium nuclei) and ions, and uncharged particles such as neutrons.

Such particles, particularly those of higher energy, can damage biological tissue through ionization, that is, through the breaking of chemical bonds. To first order this damage is cumulative and linear with the received dose. It takes extremely high doses to to produce damage that you can feel (as radiation burns, *etc*.), and at lower doses most kinds of damage can be repaired by the the body's natural mechanisms; but damage to genetic material has the potential to reproduce itself in the form of cancers. So even at low doses that you would not notice, the risk of cancer later in life can increase. On the other hand, most cancer is not cause by radiation exposure, and most cancer that is caused by radiation is caused by cosmic rays and natural radioactive isotopes in the ground. For example, a CT (compute tomography) scan involves enough ionizing radiation to give about

[51] See the appendix in this chapter for details.

1 1 in 2000 chance of a fatal cancer later in the patient's life; but that is on top of a roughly 1 in 5 chance from all other (mostly natural) causes.

Radioactive sources, such as the beta and gamma emitters used in this experiment, are described by two kinds of numbers. The first relates simply to the number of particles emitted by the source, also called its *activity*, and is usually measured in *Curies*. One Curie is defined to be 3.7×10^{10} disintegrations per second; this is approximately the activity of 1 gram of radium in equilibrium with its decay products.

The second type of descriptive number is more complex, and relates to the effects produced by the radiation, such as ionization or biological damage. Included in this second kind of number is the *roentgen*, which is the amount of gamma radiation producing a certain amount of ionization per cm^3 of air through which it travels, and the *rem* (for roentgen equivalent mammal), which is the amount of any radiation which when absorbed by mammalian tissue, will produce the same biological effects as the absorption of 1 roentgen of X ray or gamma ray radiation. The need for the rem is related to the effect of heavy particles (protons, neutrons, alphas), which cause nuclei to recoil and thus do more damage to tissue per amount of energy deposited than electrons (and gamma-rays and x-rays, which of course do their damage by producing energetic electrons).For the sources we use, the rem is the amount of radiation that deposits (loses) 100 ergs in each gram of tissue it traverses.

The ^{137}Cs sources we shall use are typically on the order of 10 micro-Curies. This is relatively harmless, as we can show. Background radiation from cosmic rays amounts to about 0.015 millirem per hour (mr/hr) so that in 140 hours one will accumulate a "whole body dose" of about 2 mr. It is also the case that a transcontinental flight will give each passenger about 2 mr. Suppose you do the worst possible thing: swallow the 10 micro-Curie ^{137}Cs source. Assuming that it will not dissolve in your system (it is sealed in plastic), it turns out that this will give you a whole body dose of around 0.1 mr/hr, assuming that about half the emitted gamma rays are absorbed, and that those that are absorbed are distributed uniformly throughout your body. Thus, if you keep it in your stomach for 20 hours you will accumulate a dose of around 2 mr --- about the same as that provided by natural cosmic rays in about 140 hours at sea level, or in making one transcontinental flight.

Of course, we do *not* recommend that you swallow any of our sources. One precaution which you should exercise, however, is not to bring any of the sources close to your eye. It is possible that the cornea can be damaged in this way.

7.3 Powering and reading out the G-M tube.

The SPECTECH counter both provides the high voltage to the G-M tube, and reads its output pulses; the output pulses, in fact, consist in temporary drops in the charge (and therefore the voltage) on the inner wire due to the current associated with each pulse. The signal can also be split and passed on to an oscilloscope that can show the shapes of the pulses. Figure 35 illustrates this setup. Everything within the dashed line is inside the SPECTECH counter box, including the high voltage power supply, the resistor associated with the counter electronics, and two blocking capacitors that keep the DC high voltage away from the counter and oscilloscope while passing

through the transient signals associated with each pulse. The resistor shown prevents the high voltage supply from charging the inner wire back up so quickly that the pulse signal isn't observed as a dip in voltage. The diagram also shows the coaxial cables running from the G-M tube to the counter box and from the counter box to the oscilloscope. These cables include a grounded shield as shown, and contribute to the overall capacitance of the system (see below).

Other connections to the G-M tube can bed used, but this is a common arrangement, and has the advantage that the exposed shell (cathode) of the tube is at ground potential.

7.4 Preliminary experiments

As with all experiments, check with your instructor as to what experiments should be written up in the lab report, and whether any others should be handed in as homework or written up in your lab notebook.

The ^{137}Cs source emits both monoenergetic gamma rays and fast electrons (beta particles) with a continuous spectrum of energies. The source capsule is designed so the the beta particles can only emerge from the unlabeled side of the source, where there is little plastic to stop them, while gammas emerge in all directions; if you want to observe only the gamma rays, you should have the labeled side toward the G-M tube. For all of the following experiments, except the dead time experiments, you should b e working in the gamma-only configuration.

7.4.1 Observe the pulse shape and determine the number of electrons per pulse.

Obtain a ^{137}Cs source to place near the G-M tube. Connect the Geiger-Mueller tube to the GM output of the SPECTECH counter, and connect channel 1 of the oscilloscope to the SIGNAL output of the counter, as illustrated in Figure 35. Record and plot both the amplitude of the pulses and the count rate as you increase the tube voltage over the region 400-1000 Volts (max). The high voltage where the count rate no longer significantly increases is called the *plateau voltage*; and the corresponding pulse height is called the *threshold amplitude*.

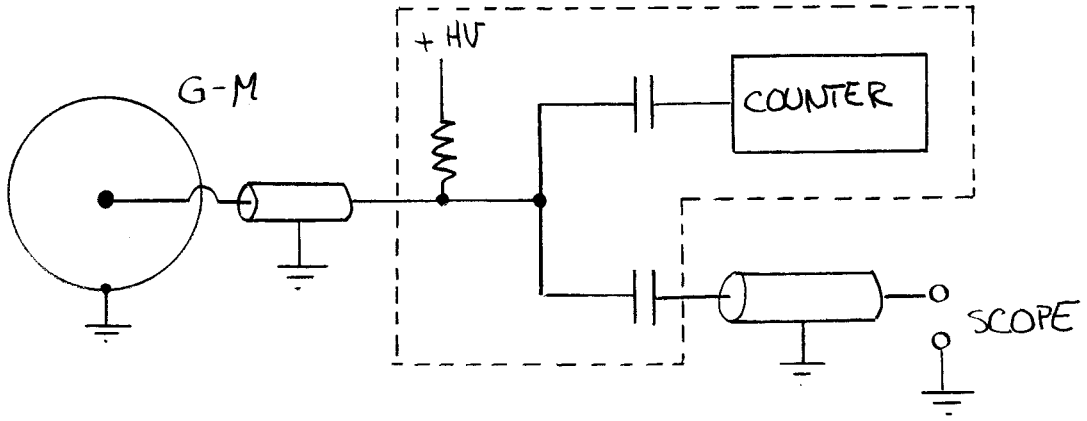

Figure 35. A schematic circuit for counting and observing pulses. The resistor is known as the charging resistor, and the two capacitors block the high voltage form the counting electronics, and from the oscilloscope, but pass the transient pulses.

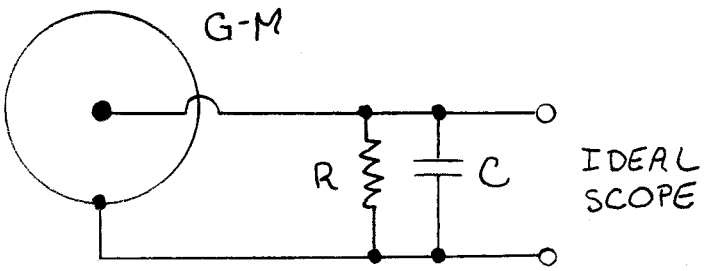

Figure 36. An equivalent circuit for the G-M apparatus. *R* is the effective parallel resistance, given by the parallel combination of the charging resistance and the input resistance of the oscilloscope (1.0 megohms). *C* is the effective parallel capacitance, given by the parallel combinations of the GM tube capacitance, the two coaxial cable capacitances, the input capacitance to ground of the counter circuit, and the input capacitance of the oscilloscope (20 picofarads). Note that the blocking capacitors shown in the schematic diagram do not significantly contribute to the capacitance to ground.

When a charge *Q* is collected by the anode it charges up an effective capacitance *C*, which is the combination of the capacitance of the tube itself between its shell (cathode) and wire (anode), the capacitances of the coaxial cables connecting the signal to the oscilloscope and counter, and the in put capacitances of of the oscilloscope and counter (see Figure 36), all of which are connected in parallel between the anode wire and ground.

The charge subsequently discharges through the parallel resistances (mainly the charging resistor and the input resistances of the oscilloscope and counter), which combined in parallel to form a net resistance *R* (which, as you recall, will be lower than any of the individual resistances that make it up). The voltage across the blocking capacitors remains nearly constant during this process, so the nature of the pulses may be understood by leaving it and the voltage supply out, as shown in Figure 36.

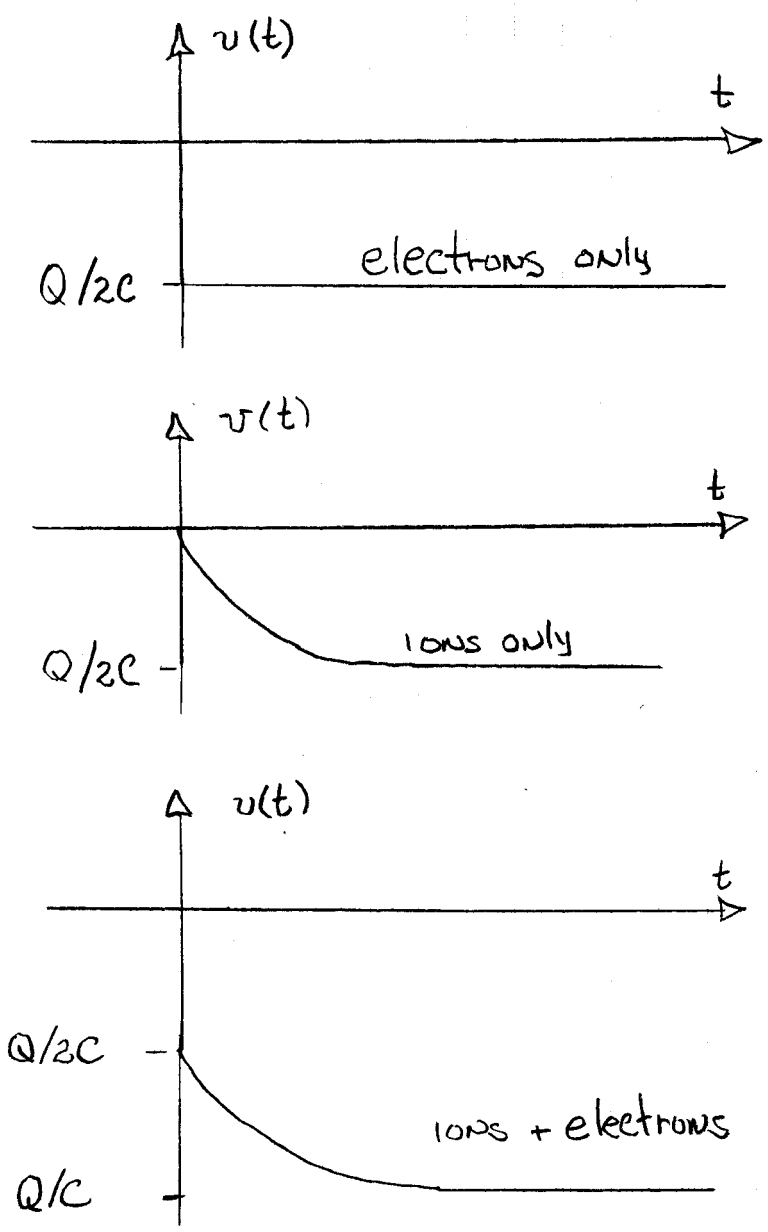

Figure 37. The change in anode voltage if $R = \infty$ (unrealistic).

Figure 37 shows the change in anode voltage in the hypothetical situation where the discharge resistance is infinite. Note that on this time scale, the change in voltage due to the motion of the electrons is essentially instantaneous. However, since the ions move much more slowly, the change in voltage is more gradual.

Figure 38 shows a more realistic situation, where the discharge resistance is finite. In this case, the voltage change from both the electrons and the ions returns to zero with an exponential time constant equal to *RC*. Thus, if we connect a known capacitor C_K in parallel with the output, the resulting time constant and amplitude will be $R(C+C_K)$ and $Q/(C+C_K)$, respectively. If instead,

we connect a known resistor R_K in parallel with the output, the resulting time constant and amplitude will be $[(RR_K)/(R+R_K)]C$ and *Q/C*, respectively. Thus, the external capacitor increases the time constant and decreases the pulse amplitude by the same factor. By contrast, the external resistor reduces the time constant but in principle does not affect the amplitude.

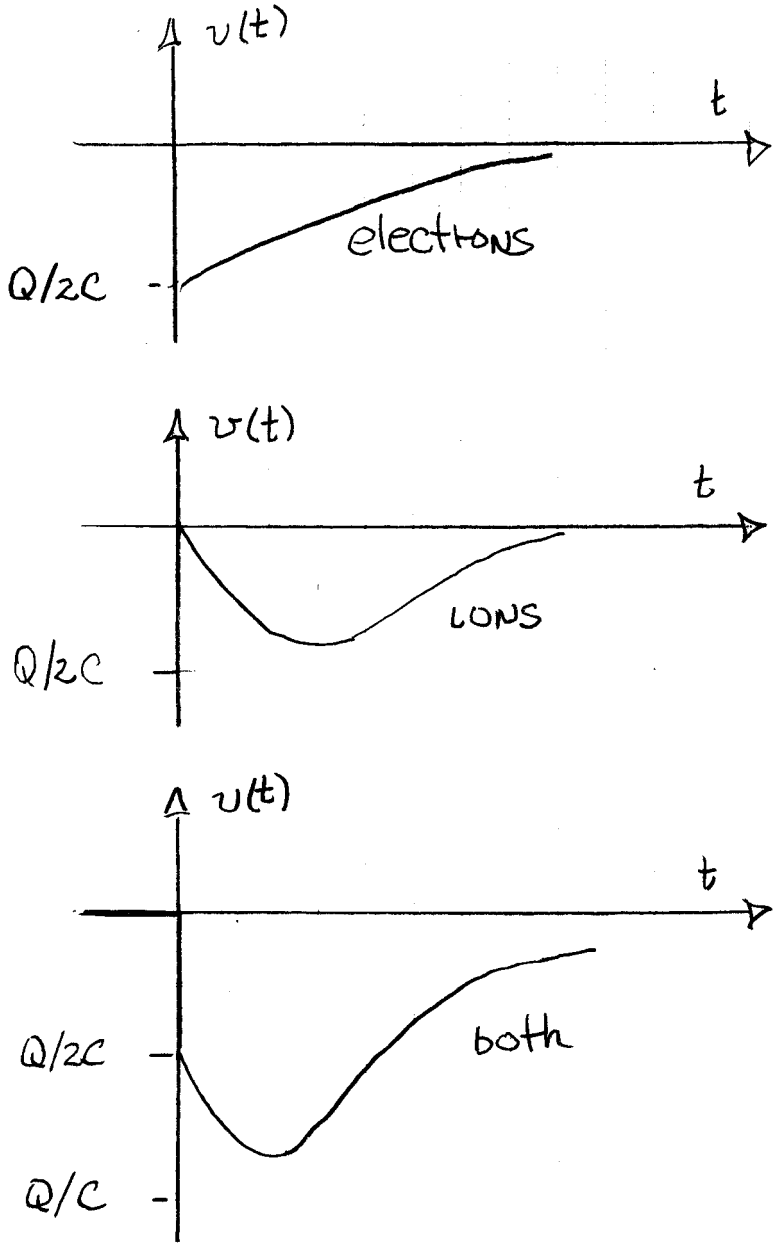

Figure 38. The change in anode voltage for finite *R* (realistic).

Measurements: Connect a known capacitance C_K (say 0.002 μF) in parallel across the input to the counter and measure the amplitude and time constants of the pulses, with and without the

capacitor. From these four values, determine the charge Q, the intrinsic total capacitance *C*, and the intrinsic net resistance *R*.[52]

Now remove the external capacitor and substitute a known resistance (say 10 kΩ). Note how the pulses become much shorter (make sure you understand why). While in principle you could redo your calculation of *Q* with the time constant and amplitude measured in this configuration, in practice you will not get a good answer; this is because the *RC* time constant is now so short that it is comparable to the drift time of the ions to the outer shell (cathode), so the pules shape is not terribly exponential (check this!) and the time constant you would be measuring is incorrect.

For much of the rest of the experiment, particularly measuring dead time (section 7.4.4), having short pulses is desirable, so leave the resistor in place for the rest of your experiments.

7.4.2 Measure the counting rate.

As discussed earlier, for a given radiation source, you will want to set the high voltage to a high enough level that the count rate is independent of the high voltage (but not much higher!). For the following experiments, be sure that this is the case.

In addition to the sealed source, there will be background pulses, mainly from muons (the result of cosmic rqays interacting in the atmosphere) and nearby natural radioactive materials (mostly ^{40}K and ^{232}Th). Measure the background counting rate by moving all sources far from the Geiger tube. Since this can change a little from pace to place and day to day, you should take other, longer background measurements when you do other specific experiments.

7.4.3 Investigate the counting statistics

A radioactive nucleus, such as ^{137}Cs, is unstable. It will not last forever, but will decay. The exact moment at which any particular nucleus will decay is not predictable; it happens by chance. Suppose we pick a nucleus and watch it for a short time interval Δt. There is a small probability *p* that our nucleus will decay in that time interval, a probability that we expect will be proportional to Δt. We can write

$$p = \beta \Delta t \tag{130}$$

where β is the constant of proportionality. This is the most important point about radioactive decay: it has no memory. Its chance of decaying in the next second does not depend on how old it is.

If we have *N* such nuclei present at time *t* = 0, where *N* is a very large number, some of them will decay during the time interval Δt. Let the mean number that decay be ΔN. We expect that ΔN will be equal to *pN*. That is, we expect that

$$\frac{\Delta N}{N} = -\beta \Delta t \tag{131}$$

[52] Note that an exponentially decaying pulse decays to one-half of its peak value at a time $T_{1/2}$ equal to 0.69 times the time constant. This fact is useful in making time constant measurements with an oscilloscope.

The minus sign is there because ΔN is a *decrease* in the value of *N*. If Δt is infinitesimally small, so that we can represent it by *dt*, we can integrate Eq. (131) to find the mean number of nuclei *N*(*t*) remaining at any time *t*:

$$N(t) = N_0 e^{-\beta t} \tag{132}$$

where N_0 is the number of nuclei present at $t = 0$. $1/\beta$ is the *mean lifetime* of a nucleus, and is about 43.3 years for ^{137}Cs.[53]

Now for our case, the rate at which the (unstable) nuclei decrease is also the rate at which decay products are produced. When one ^{137}Cs nucleus decays, a single gamma ray is produced. Hence the gamma rays are produced at a mean rate

$$-\frac{d}{dt}N(t) = \beta N_0 e^{-\beta t} = \beta N(t) \tag{133}$$

If we count gamma rays for a time much shorter than the mean life of the nucleus, *N*(*t*) will remain essentially constant (at N_0), and the gamma rays will be produced at a mean rate that is essentially constant.

Now the geometry of our apparatus is such that we shall be able to count only some fixed fraction of all the gamma rays produced by our radioactive source of ^{137}Cs. Our detector, because it does not totally surround the source and is rather transparent, monitors only a small fraction of *N*, say *n* of them. Thus our detector will count the gamma rays at a mean rate given by

$$\frac{dn}{dt} = \epsilon\beta N \tag{134}$$

where ϵ is the small fraction of the gamma rays caught by our detector.

In any time interval Δt our Geiger tube will detect Δn gamma rays:

$$\Delta n = \epsilon\beta N \Delta t \tag{135}$$

Now Δn will be some integer number of counts; let us call this number *m*. It is characteristic of the probabilistic nature of the decay process that *m* will fluctuate from one counting interval to the next. If the counting interval is one second, for example, we may count 3 particles in the first second, 5 particles in the second, 4 particles in the third, and so on. In some of the time intervals we may even count zero particles.

The probability that *m* particles be counted in Δt is given by the *Poisson Distribution* function:

[53] The *half life* of the nucleus is the time taken for *N* to decrease to half of N_0; the half life is $\ln 2$ times the mean life, or about 30 years for ^{137}Cs.

$$P(m) = e^{-\lambda}\frac{\lambda^m}{m!} \tag{136}$$

where λ is the long-term average count rate (which of course can't be known exactly without counting for an infinite amount of time, but which can be estimated by a suitably long data collection).

The Poisson Distribution function $P(m)$ has the following properties:

1. It is *normalized*:

$$\sum_0^\infty P(m) = 1 \tag{137}$$

This means that the total probability of observing *some* number of particles (including 0) in Δt is unity.

2. The mean (average) value of m is λ:

$$\sum_0^\infty mP(m) = \lambda \tag{138}$$

λ is called the *population* mean. If we could count for an infinite number of time intervals, the mean value of m would be exactly equal to λ. This is trivial from the definition of λ, in that one long collection and and infinite number of short ones should have the same average rate.

3. The population variance [the mean value of $(m-\lambda)^2$ for an infinite number of time intervals] is also λ:

$$\sum_0^\infty (m-\lambda)^2 P(m) = \lambda \tag{139}$$

The population variance is usually denoted by σ^2; hence $\sigma^2 = \lambda$ for the Poisson distribution.

Exercise: Verify the truth of the these assertions starting with Eq. (136).

It turns out that the Poisson distribution function governs a surprising variety of statistical processes. The number of cars passing per hour on an uncrowded road, for example, or the number of raindrops falling per second on a square inch of the earth's surface during a steady light rainfall are both governed by the Poisson distribution. In general, any collection of random events that are uncorrelated (the presence of one count does not encourage or discourage the presence of another) obey the Poisson distribution That's why counting cars when traffic is heavy won't obey this distribution, because you can't have two cars in exactly the same place at the same time, so they are anti-correlated with each other.

The fact that the population variance is equal to the mean for the Poisson distribution means that the standard deviation, which is the square root of the variance, is just the *square root* of the mean.

The standard deviation is a measure of the expected uncertainty in a single measurement. Thus, if the mean number of counts per second is 9.0, we expect that in any one second we are quite likely to observe somewhere between 6 ($9-\sqrt{9}$) and 12 ($9+\sqrt{9}$) counts. What is meant by "quite likely" depends somewhat on the value of λ. For $\lambda = 9.0$, it means about 76 per cent. If λ is large, the Poisson distribution becomes well-approximated by the Gaussian distribution, with $\sigma = \sqrt{\lambda}$:

$$P_G(m) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(m-\lambda)^2/2\sigma^2}, \tag{140}$$

In the case of a Gaussian distribution, "quite likely" means a probability of about 68 per cent, meaning that about 68 percent of the observations are expected to lie between $\lambda - \sqrt{\lambda}$ and $\lambda + \sqrt{\lambda}$. If $\lambda = 100$, for example, about 68 per cent of the counts will fall between 90 and 110.

It is often desired in an experiment to determine λ as accurately as possible. Intuitively we sense that the more measurements we make the more accurately we can determine λ. If we make a single one-second measurement in a situation where the expected number of counts (λ) is 9, we expect a standard deviation of 3 counts, and so can determine the true λ to within a fractional uncertainty of only about 33 percent or so.

10 one-second measurements will total about 90 counts with a standard deviation of about 9.5 counts, or a fractional uncertainty of a little more than 10 percent. 100 one-second measurements will total about 900 counts with a standard deviation of 30 counts, or a fractional uncertainty of about 3.3 percent, and so forth. Of course, there is no need to take 100 measurements for this purpose; it is easiest to take a single measurement over a very long time interval. It is only the total number of counts accumulated that matters.

Hence to reduce the fractional uncertainty by a factor of 10 one must increase the number of detected counts by a factor of 100, which in general means increasing the observation time also by a factor of 100. In short, all other factors being equal, *the signal-to-noise ratio increases only like the square root of the amount of time spent making the measurement.*

Experiment.

Let's confirm that the probability distribution of counts in many 1-second intervals really resembles the theoretical form of the Poisson distribution.

Place a source so that about 5 - 10 counts/second are detected, and record the counts for 100 one-second intervals. Make a plot of the frequency distribution $F(m)$ vs m where $F(m)$ is the number of one-second intervals showing m counts. Calculate $\bar{m}$, the mean for your sample of measurements. This is your best estimate of λ, so use this estimate to calculate values of $P(m)$ for the Poisson distribution. Compare your observed frequency distribution $F(m)$ with the expected frequency distribution $100P(m)$ by using the χ^2 "goodness of fit" test described in the chapter on Chi-square testing" of this manual. In particular, using the graph in the last figure of that chapter, determine the value of α (the confidence level) for the value of χ^2 calculated from your sample. This is the probability of getting your χ^2 value or higher if the Poisson model is correct. is it

greater than about 0.1? If so, the hypothesis that your data are drawn from a Poisson distribution is a reasonable one.[54]

Alternatively, you can look at the same graph from the opposite perspective. You can determine from this graph the value of $\chi^2_{\nu,\alpha}$, where α = 0.15, say. Is your calculated value of χ^2 larger than this? if so, you might consider rejecting, at the 85 per cent confidence level, the hypothesis that your data are drawn from a Poisson distribution.

7.4.4 Measure the counter dead time

The Geiger tube will be disabled for a short period of time after it is discharged by the passage of a particle through it. This is called the "dead time," and is typically several hundred microseconds for tubes such as the one we use. This is the time required for the re-establishment of the high voltage, owing to the time required for the recombination of the ionized gas atoms in the tube. A particle passing through the tube during this time will not be counted. Just like the case of counting cars in heavy traffic, counts become anti-correlated with each other and Poisson statistics become invalid.

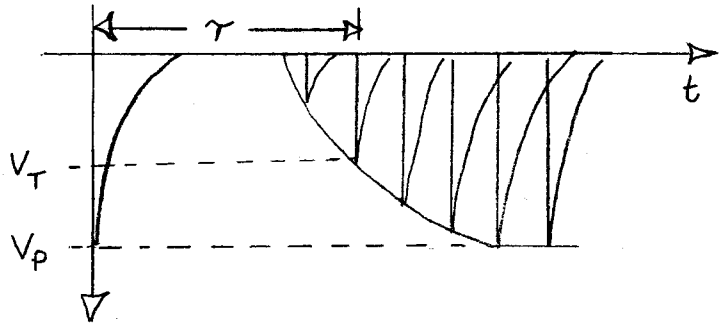

Figure 39. A time-lapse image of pulses. V_P is the pulse amplitude, and V_T is the trigger threshold.

The dead time may be estimated directly by the following method. Observe pulses from the beta particles of the source (which are more numerous) at a high enough voltage so that the pulses are all about twice the counter threshold amplitude (as measured in section 7.4.1). With a sweep time of several hundred microseconds, examine the oscilloscope trace. You should observe something like Figure 39. (The digital oscilloscopes have a persistence mode in which many traces can remain on the screen instead of each replacing the next; this mode is particularly useful for this measurement).

Note from the figure that the tube recovery is not abrupt; pulses detected too soon after the primary pulse will not develop their full amplitude. As a consequence, the effective dead time of the counter electronics will depend upon the counter threshold for triggering the pulses, as well as the nominal pulse height (as determined by the tube high voltage).

[54] This fitting procedure is convenient and gives results that aren't too bad, but in fact it is not correct. The accurate analysis is complicated but interesting, and makes use of the notion of likelihood and random trials with a computer. If you are interested in learning about it, your instructor can provide a resource that discusses it and allows you to run the analysis using Python code.

From your data, estimate the dead time for pulses with amplitude about twice the counter threshold amplitude (which you measured earlier in section 7.4.1).

The dead time may also be estimated indirectly, using the following line of reasoning: at high counting rates the tube will be dead a significant part of the time, thus reducing the apparent counting rate. Imagine that *m* particles per second enter the Geiger tube whose dead time is τ, and that *n* pulses per second are registered. In any one second, the tube will be dead for a time $n\tau$ seconds, and so will be alive for $1-n\tau$ seconds. Hence the number of counts registered during this second will be $n=(1-n\tau)m$. This equation may be solved for either *n* or *m*:

$$n=\frac{m}{1+m\tau} \qquad \text{and} \qquad m=\frac{n}{1-n\tau} \tag{141}$$

Note from the first equation that if $m\tau \gg 1$, *n* will be significantly reduced, while if $m\tau \ll 1$ the dead time effects may be neglected and $n \approx m$.

Now because of the dead time, the counting rate of two samples together will be less than the sum of the rates for the separately counted samples. Suppose that m_a and n_a are the true and apparent counting rates for sample a, and that m_b and n_b are those for sample b, and that m_{ab} and n_{ab} are those for two counted together. Since $m_{ab}=m_a+m_b$, we may write[55]

$$\frac{n_{ab}}{1-n_{ab}\tau}=\frac{n_a}{1-n_a\tau}+\frac{n_b}{1-n_b\tau} \tag{142}$$

This is a quadratic equation for τ, with the exact and approximate (low count rate) solutions

$$\tau n_{ab}=1-\left[1-\frac{n_{ab}(n_a+n_b-n_{ab})}{n_a n_b}\right]^{\frac{1}{2}} \tag{143}$$

$$\tau \approx \frac{n_a+n_b-n_{ab}}{2n_a n_b} \tag{144}$$

Exercise: Solve Eq. (142) to yield Eq. (143).

Now, using your two sources, viewing the beta-ray (unlabeled) side, count the first one, then the two together, then the other. As before, you should set the high voltage so that the pulses have about twice the amplitude of the threshold amplitude. Be very careful not to move a source that should stay in place from one measuring to the next. For best results you should adjust the distance of the source to the G-M tube so that there is a significant, but not overwhelming, dead time effect. You can test this effect by repeating the experiment with different distances between the sources and the detector. Compare this value of τ with that which you measured directly.

[55] Note that the background rate m_0 doesn't appear in this equation, so it is only approximate. If you would like to be more precise, try to work out how the equations $m_{ab}=m_a+m_b$ and equation (142) should be modified to account for background, and then find that for the τ that solves the new equation (142) numerically instead of algebraically.

7.5 The absorption of gamma rays by lead

Now we come to the heart of this experiment. Here are some useful references:

1. Irving Kaplan, *Nuclear Physics*, 2nd ed. (Addison-Wesley, 1963)

2. C. M. Davisson and R. D. Evans, "Gamma-Ray Absorption Coefficients," *Rev. Mod. Phys. 24*, 79 (1952).

3. The National Institute of Standards and Technology XCOM table of x-ray and gamma-ray interaction cross sections, at

https://physics.nist.gov/PhysRefData/Xcom/html/xcom1.html;

This allows you to easily generate tables like Table 5 for any material or compound at all energies.

The gamma rays are produced by a sample of ^{137}Cs, whose decay scheme is illustrated by the energy-level diagram shown in Figure 40.

Charged particles lose energy nearly continuously along their path, usually by ionizing many atoms. The range that a charged particle of a given energy will travel in a given material is therefore pretty much fixed; if you know the energy, you know just how much shielding you need to cut them all out. In contrast, an individual gamma ray does not gradually lose energy along its path, but instead goes right through some material untouched until it interacts strongly, at random, with one electron or atom. We expect that an individual gamma ray is either 1) scattered from the beam, 2) totally absorbed, or 3) continues unaffected. In this experiment, scattering can in principle either increase or decrease the count rate in our detector; photons which would have been headed toward it can be scattered away, but photons which would have missed it can also be scattered back. But to have a definite task, we will assume that scattering and absorption both remove photons from the beam equally.

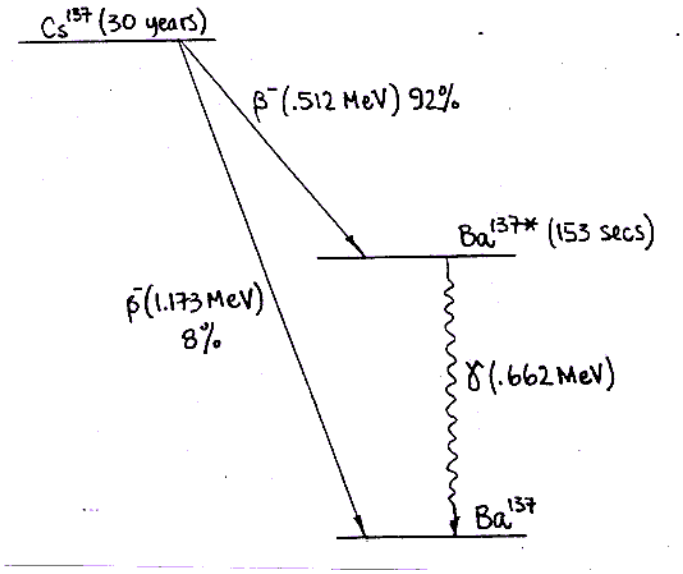

Figure 40. Decay scheme of ^{137}Cs.

Thus, a beam of gamma rays of intensity *I*, incident on a thin slab of absorber of thickness Δx, will change as it passes through the slab by an amount that is proportional both to *I* and to Δx:

$$\Delta I = -\mu I \Delta x \tag{145}$$

The constant of proportionality μ is called the *absorption coefficient*. In general, μ will depend on energy. If the gamma ray photons all have the same energy, then μ is a constant, and Eq. (145) may be integrated to yield

$$I = I_0 e^{-\mu x} \tag{146}$$

Equation (146) gives the beam intensity *I* after it has traversed a total absorber thickness *x*, where I_0 is the incident intensity. If the absorber thickness is measured in cm, μ will have units of cm^{-1}, since $\mu\Delta x$ must be dimensionless. We see that the mathematics of the intensity of a beam of photons decreasing with distance passing through a shielding material and the remaining number of of radioactive nuclei in a sample decreasing with time due decay are the same. This is because there is no memory of how much time has passed or how much shielding material has been passed through—the probability of an interaction in the next bit of material or the next moment in time is a constant.

Other units are occasionally used in conjunction with the term *absorption coefficient*, particularly when making connections between experiment and theory. It is helpful to clarify their use.

We expect that μ will be proportional to the number of nuclei (or number of atoms) per unit volume of the absorber. Hence it is useful to think of the absorption coefficient *per atom*, denoted by the symbol σ. This quantity will be central in theoretical analyses of the interactions between photons and nuclei. The relationship between μ and σ is

$$\mu = N \cdot \frac{1}{A} \cdot \rho \cdot \sigma \tag{147}$$

where N is Avogadro's number (the number of atoms per mole), A is the atomic weight of the absorber (the mass per mole), and ρ is the absorber density (the mass per unit volume). The units of σ will therefore be $\text{cm}^2 / \text{atom}$, or $10^{-24} \text{cm}^2 / \text{atom}$. Because of the units of area, σ is called the *atomic cross-section*.

Furthermore, it is sometimes useful, when comparing absorbers of different materials, to express the absorber thickness in terms of grams/cm^2, the mass of absorber contained in 1 cm^2 of cross-section; this quantity will be ρx. Thus, we may write

$$\mu x = \frac{\mu}{\rho} \cdot \rho x \tag{148}$$

Thus the quantity μ / ρ, called the *mass absorption coefficient*, will be in units of cm^2 per gram, and is also tabulated in Table 5.

Three processes are mainly responsible for gamma ray absorption:

1. Photoelectric absorption, in which the gamma is absorbed completely upon ejecting a bound electron from the atom usually one of the most tightly bound ones, as their binding energy is closest to being in resonance with the gamma ray energy.
2. Compton scattering, in which the gamma ray "bounces" off any of the electrons in the atom ejecting the electron but continuing on with reduced energy and changed direction, and
3. Pair production, in which the gamma-ray passes near the atomic nucleus and converts into an electron and positron (anti-electron). This can occur only when the gamma ray has enough energy (1.022 MeV or higher) to create both particles.

The total absorption coefficient μ is therefore the sum of three terms:

$$\mu = \mu_{\text{photo}} + \mu_{\text{compton}} + \mu_{\text{pair}} \tag{149}$$

Atomic cross-sections (or absorption coefficients per atom) may be theoretically calculated for each of these processes; the results for the absorption and scattering of gamma rays by lead are listed in Table 5, taken from Kaplan's book. The symbols σ_{PE}, σ_C and σ_{PF} denote the atomic cross-sections for photoelectric effect, Compton scattering, and pair production, respectively. Of course, pair production cannot occur unless the photon carries at least as much energy as the rest mass of an electron plus a positron. Photoelectric absorption, as you can see, falls sharply with energy as the gamma-ray energy goes further out of resonance with the binding energy of the K-shell electrons. Thus, lead is particularly good for shielding because of its atomic number, not its mass density.

It is important to note that the *Compton* cross-section pertains to the *scattering* of gamma rays by the target. If the detector is fairly close to the absorber, Compton scattered gamma rays may scatter

into the detector, mimicking the incident rays. As a consequence, including the Compton cross-section in your theoretical predictions may lead to an overestimate of the attenuation of the count rate. On the other hand, once scattered, the gamma-ray has lower energy and is more likely to be photo-electrically absorbed. The most common fate of a 662 keV gamma-ray going through a thick lead layer (multiple centimeters) is to Compton scatter once or twice and then be photo-electrically absorbed.

Photon energy MeV	Photo-electric σ_{PE}	Comp-ton σ_C	Pair form-ation σ_{PF}	Total σ_T	Coefficient per cm, μ, cm^{-1}	Mass coefficient μ/ρ cm^2/gm
0.1022	1782	40.18		1822	59.9	5.30
0.1277	985	38.01		1023	33.6	2.97
0.1703	465	35.04		500	16.4	1.45
0.2554	161	30.70		192	6.31	0.558
0.3405	75.7	27.63		103.3	3.39	0.300
0.4086	47.8	25.74		73.5	2.42	0.214
0.5108	27.7	23.50		51.2	1.68	0.149
0.6811	14.5	20.73		35.2	1.16	0.102
1.022	6.31	17.14		23.45	0.771	0.0682
1.362	3.86	14.81	0.1948	18.87	0.620	0.0549
1.533		13.91	0.3313			
2.043	2.08	11.86	1.247	15.19	0.499	0.0442
2.633						
3.065		9.313	3.507			
4.086	0.369	7.761	5.651	14.28	0.469	0.0415
5.108	0.675	6.698	7.560	14.93	0.491	0.0434
6.130		5.917	9.119			
10.22	0.316	4.115	14.04	18.47	0.607	0.0537
15.32	0.206	3.042	18.00	21.25	0.698	0.0618
25.54	0.122	2.044	23.24	25.41	0.835	0.0739

Table 5. Values of the absorption cross sections and absorption coefficients for lead.[56]

Experiment:

By placing sheets of lead of different thickness between the source and the detector to absorb the gamma rays, measure the counting rate as a function of the absorber thickness ρx. Note that $\rho x = (\rho x a)/a = M/a$, where *M* is the mass of a sheet and *a* is the area of the sheet. (This is a much more accurate than using the measured physical thickness of the material, and you should use this method if possible). Also note that sheets may be combined. For this experiment, place the labeled side (gamma only) side up, because you are interested in the absorption of gamma-rays, not beta rays.

[56] Irving Kaplan, *Nuclear Physics*, 2nd ed. (Addison-Wesley, 1963).

To accurately measure the absorption coefficient, it is necessary to vary the thickness by much more than an absorption length, which in the case of lead is about one centimeter. It is highly recommended in this experiment that your lead thickness vary by at least three absorption lengths, or three centimeters. That is, if your smallest thickness is 3 mm, then your largest should be more than 30 mm. Note that it is crucial not to disturb the relative separation of the source and detector during the experiment. Therefore, the detector must be at least 35 mm from the source so that there will be enough space between them to insert more than 30 mm of lead.

Start by comparing the count rate with no lead at all to the count rate with 1 mm of lead. Usually the rate will go up instead of down, because fluorescence x-rays emitted by the lead are more easily detected by the G-M tube than the gamma-rays that created them (see section 7.6). For this reason, the thinnest lead shield that you should use in your final data set should be around 3 mm, at which point the ratio of gamma rays to fluorescence x-rays seen by the tube has become roughly independent of thickness.

Now, proceed to take measurements with several thicknesses of lead. Do not use thicknesses that are too similar to each other. Take points separated by 5 mm, not 1 mm. Then you will have seven points spanning 30 mm.

Set up your counting system so that you can count a long time without recording more than one number. That is, if you want to count for 100 seconds, do just that. Do not count ten times for 10 seconds each; it's a waste of effort.

Be sure to leave time to count for a much longer period when the lead gets thick; you want a a small fractional error $(\sqrt{N} / N)$ at each thickness. Making sure that you count to N > 1000 for each thickness gives an error of about 3% for each data point. If you look at the count rate at your greatest thickness first, you can estimate right away how long your whole series of measurements could take.

There is another important factor in your observations, and that is background. Take away your source completely, and count for at least five minutes. You will find that your rate is NOT negligible compared with the counting rate with your largest lead thickness. You must determine the background rate with enough accuracy so that it does not dominate your final answer for the absorption length; this means you should accumulate background for at least as long as you accumulate data at your greatest thickness.

First make a graph of $\ln I$ *vs.* x, to see if it looks linear as expected. Be sure to subtract background before plotting. Next, use a least-squares fitting program to fit your data set without subtracting background, to a function which is the sum of an exponential that decays with x and a constant value representing the background.

Your fit should tell you two separate things: 1) Whether the model you used (exponential plus background) is consistent with the data (your *hypothesis test*, which makes use of the value of χ^2), and 2) what the measured absorption coefficient and its uncertainty are (an exercise in *parameter estimation*). Both of these should be discussed in your report.

Plot the data given in Table 5 to illustrate the dependence of μ on the gamma ray energy, and use this plot to compare your measured value of μ with the theoretical value, assuming that the gamma ray energy is 0.662 MeV. Plot only four or five points near that energy, and indicate the range of acceptable values of μ that you have measured.

Does your measured value of μ (complete with its statistical uncertainty) agree with the theoretical value?

Note that the logic of this experiment could be reversed, in that one could use the experimental data to deduce features of the *spectrum* of the observed gamma rays, assuming that the absorption coefficients are already known. A truly exponential decrease of the beam intensity with absorber thickness (a straight line on a ln(*I*) *vs.* (*x*) plot) suggests that the gamma ray photons are monochromatic (mono-energetic), and the energy of the gamma rays can then be deduced from the measured absorption coefficient, using the theory leading to Table 5. What range of gamma-ray energies are considered acceptable by your data from this perspective?

7.6 Optional Experiment: Absorption by Graded Z Shields

You will find a "graded-Z" shield in the cabinet. The purpose of this material is to efficiently shield high energy gamma rays by significantly attenuating their energies as well as scattering them. It consists of thin layers of different Z (atomic number) metals laminated together. For this lab, we have tantalum (Z = 73), tin (Z = 50), and stainless steel (which is mostly iron, Z = 26). These layers are ordered according to Z, and the radiation-blocking properties depend on whether the low-Z or the high-Z side is facing the source.

The ^{137}Cs source emits gamma rays with energy 662 keV, which may then be absorbed by a subsequent material. In this process, a gamma ray interacts with a bound electron (usually a *K* electron) in the material and ionizes the atom. The newly-freed photoelectron flies off with most of the energy of the initial gamma ray. For example, when a 662 keV gamma ray hits an electron, 88 keV of the energy goes into liberating the electron from the atom, and the remaining 574 keV goes into the kinetic energy of the photoelectron. Since the photoelectron has a short scattering length (on the order of μm), it will collide with other atoms and electrons in the material and lose all of this energy as heat. It is in this manner that a significant portion of the incoming gamma ray's energy is dissipated.

After the photoelectron is emitted from the atom, an electron from a shallower orbital (usually an *L* or *M* electron) drops down to fill the vacancy, emitting another photon in the process - this is called fluorescence. Fluorescence photons are emitted isotropically; even if the initial gamma ray were heading straight towards the G-M tube, the resulting fluorescence photon may now go in any direction. If the G-M tube spans a small solid angle, these photons have a low probability of being detected. However, the fluorescence photon is lower in energy than the absorbed photon, and travels fairly unimpeded through the rest of the material. Why doesn't it get quickly absorbed in turn? Because it doesn't have enough energy to liberate another K-shell electron and the cross-section (probability) for it to liberate a more loosely bound electron is much lower.

When you have a single layer of metal covering your source, the story stops here. The emitted photon is free to exit the metal and possibly be detected by your G-M tube. However, with the graded-Z shield, the fluorescence photon encounters another layer of metal with a different absorption cross-section. If this next layer has a lower Z, the fluorescence photon from the prior material is very likely to be absorbed by a K-shell electron in the new material, setting off the same process as before. The fluorescence photon emerging from this interaction is then even lower energy than the previous one. With our graded-Z shield, this process repeats three times, and the resulting photons from the steel layer are so low in energy that they are absorbed by the wall of the G-M tube without sending an electron into the gas.

Take the graded Z-shield and a 1 mm sheet of lead, and put the G-M tube fairly close to the source so that the count rate is high but not approaching the rates at which dead time is significant. Measure the count rate very precisely (using at least 10,000 counts) in the following configurations: lead only, graded Z-shield only (shiny side up and then down), and the graded Z-shield stacked on top of the lead (lead toward the source, shield side toward the tube), trying shiny side both up and down. Rank the rates from highest to lowest, and see if you can explain this entire ranking using the physics we described above. Ask your instructor for help if you are not satisfied with your explanation.

Note that blocking or letting through fluorescence photons isn't the only thing that can change the rate in the tube; Compton electrons can be ejected from the lead, and while fluorescence x-rays have a higher efficiency of detection by the tube that the original gammas, electrons have an even higher efficiency, as they will always trigger the tube if they have enough energy to get through the entrance window.

7.7 Appendix: Why pulse size increases linearly with voltage

A Geiger-Muller tube may be modeled as a capacitor, with capacitance given by $C_{GM} = 2\pi\varepsilon_0 L / \ln(R_> / R_<) \approx 1$ pF. If the tube is charged to a high voltage V, the charge on the wire will be $Q = C_{GM}V$. An electron avalanche of charge q (for electrons, q is negative) will leave behind a positive ion sheath of charge $-q$. If the ion sheath is, say 10% or more of the original wire's charge, the electric field in the neighborhood of the anode will be reduced enough to suppress the avalanche.

If the tube is operated at a higher voltage, the initial static charge on the wire will be proportionally increased, and thus the amount of ion charge required to suppress avalanche will be proportionally increased. But, from the above arguments, the ion charge is just the charge constituting the pulse observed by the electronics. Therefore, the pulse size will be roughly proportional to the applied high voltage. In fact, the pulse size will be of order $v_p \approx 0.1(C_{GM} / C_P)V$, where C_P is the parasitic capacitance arising from cables, oscilloscope input, *etc*.

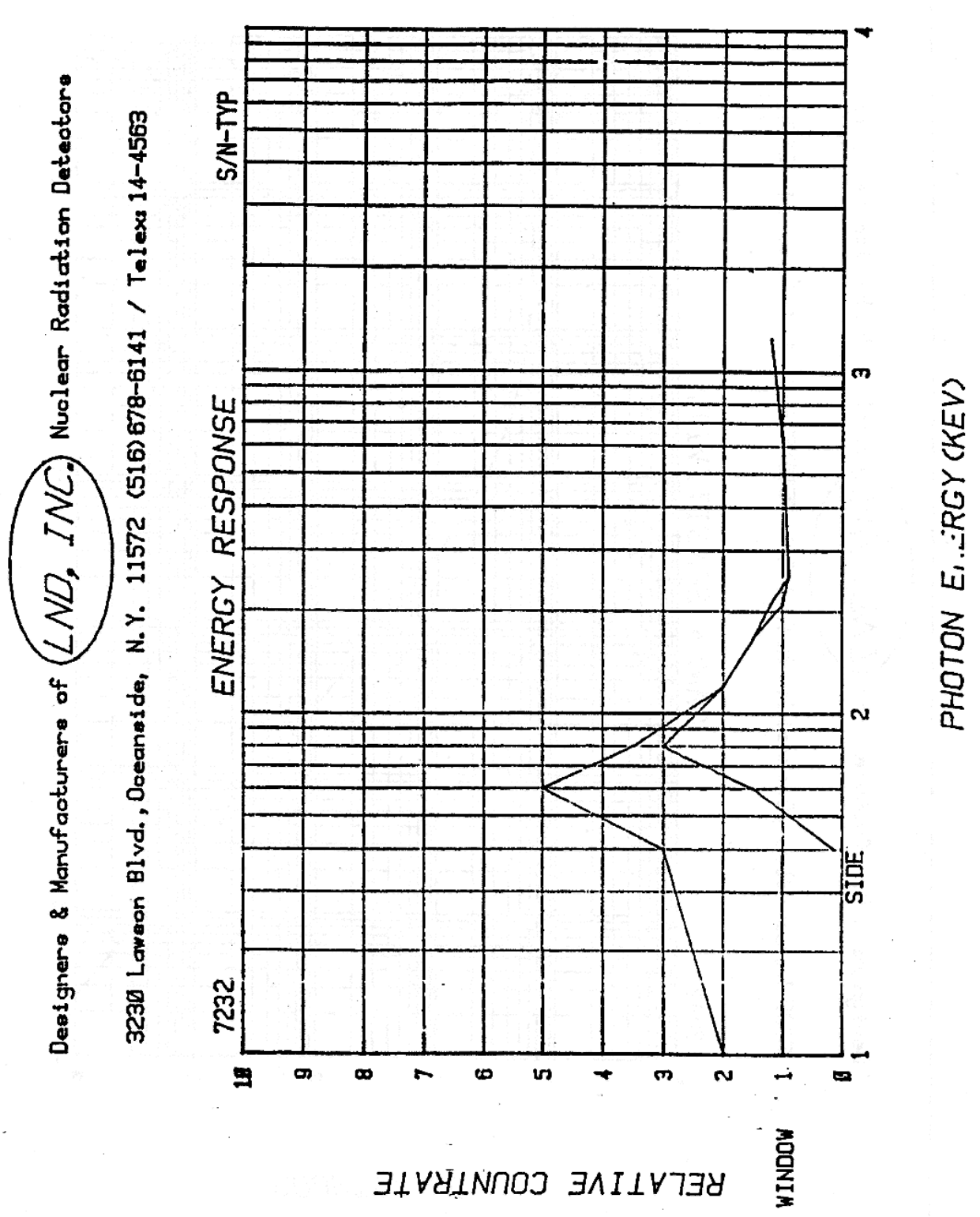

Figure 41. Response of the tube as a function of the photon energy in keV. Note that the peak of the response is close to the lead K-shell x-ray energy (somewhat below 100 keV and the response of the ^{137}Cs energy of 662 keV is much lower. The upper trace is for illumination through the end window and the lower trace is for illumination through the side of the tube.