Data, Uncertainty and Error Analysis

Louis Bouchard

UCLA Department of Chemistry and Biochemistry
E-mail address: lsbouchard@ucla.edu

Contents

Preface

Los Angeles, CA

December 9, 2022

This is a collection of lecture notes, problem and solutions on the topic of experimental measurements, data uncertainty and error analysis. These notes were assembled over a period of approximately four years while teaching chemistry 114 at UCLA, a physical chemistry laboratory course that includes separate lectures in addition to the laboratory sessions. The course was initially based on the excellent textbooks by J.R. Taylor, "An Introduction to Error Analysis" and Hughes & Hase, "Measurements and their Uncertainties", but was later revised to focus on the probabilistic foundations of classical measurements. I do not consider these notes to be a substitute to the book of Taylor, which I recommend to any newcomer for its strong pedagogical value. Instead, I view these notes as providing a more in-depth coverage of the probabilistic foundation. While the reader is assumed to know calculus, no knowledge of probability theory is assumed; the required concepts are introduced as needed in these notes.

I make frequent use of MATLAB while teaching the course because it is important for young students to learn scientific computing. For those who can't afford MATLAB, a free software alternative can be download, called GNU Octave. Many of the MATLAB examples herein should work on GNU Octave either directly or with a small amount of conversion effort. An entire chapter is dedicated to MATLAB sessions where the students are walked through several examples. At UCLA these MATLAB sessions are done during class time at the Science Learning Center. Problems and solutions

are included at the end of many chapters. The students should be aware that more advanced analysis techniques exist (such as statistical learning) and are not covered here due to the short (1-quarter) nature of this course. I would encourage the reader wanting to learn more, to read books on modern multivariate statistical techniques. Special thanks go to Alison Ly, a UCLA undergraduate student, for redrawing most of the figures in this document.

Louis Bouchard

Chapter 1

Experimental Uncertainties

Every measurement contains some amount of uncertainty due to a variety of experimental factors (e.g. temperature fluctuations, Johnson noise, shot noise, sample motion, vibrations, environmental fields, etc.). The uncertainty is due to fluctuations in the physical quantities being measured and in the measurement apparatus. Uncertainties can be thought as representing the noise magnitude which affects our signal. The goal of the experimentalist is to reduce uncertainties to an acceptable minimum either by repeating the measurement several times or by designing a better experiment. Our goal here is to understand where uncertainties come from and how to characterize them.

1.1. Types of uncertainties

There are two main classes of uncertainties encountered by the experimentalist:

- **Systematic errors:** these errors originate because there is a bias in the system. For example, performing the same experiment on a different day could mean the ambient temperature is different, which could then lead to a drift in some currents in the system. A good piece of equipment should be designed to take into account temperature variations, for example. However, not all instrumentation is designed to compensate for environmental factors. Another example could be the measurement of a magnetic field using a magnetometer. However, the presence of large metal objects nearby could affect the magnetic field.

- **Random errors:** these errors arise from random fluctuations in the electronics or the physical measurement under study. For example, if you are measuring the voltage across the terminals of a load, there will be random fluctuations in the voltage as function of time. These could be due to Johnson noise, which is due to thermal fluctuations of electrons in the resistance of the load and leads to random voltage fluctuations. Johnson noise is also generally present in the measuring apparatus. Vibrations could also give rise to random errors.

Note: Sometimes the distinction between systematic and random error may depend on time scales. The systematic errors can be randomly fluctuating quantities that change so slowly that they appear static on the timescale of (rapid) measurements. Random errors are those due to fluctuations that are rapid compared to the timescale of measurement. In the latter case, a series of consecutive measurements appears to fluctuate randomly.

1.2. Histograms

The quantities measured in the laboratory are random variables. A random variable is not a variable in the usual sense. It is instead characterized by a probability distribution function that encodes all relevant statistical information about the random variable.

Suppose that X is a random variable. For example, let X be the diameter of CDs produced in a factory. If we example n CDs and measure their diameters, we collect n measurements of X:

$$\{x_1, x_2, \ldots, x_n\}.$$

These values are all different because no two CDs are perfectly identical. Because the values fluctuate from measurement to measurement, it is convenient to view X as a random variable. We shall denote a random variable by its capital letter, X and its value by a lowercase letter x. The two are related by $x = X(\omega)$. Here, ω notes a particular outcome of a random experiment. For example, if the experiment consists of rolling a die, there are 6 possible outcomes: $\omega \in \{1, 2, 3, 4, 5, 6\}$; multiple outcomes of die roll are denoted $\omega_1, \omega_2, \ldots, \omega_n$. For the CDs ω denotes a particular instance of diameter measurement; ω_j denotes the j-th measurement of the diameter. Suppose that we measure 10,000 CDs; we would get a list of diameter such as:

```
119.73
117.10
122.76
119.20
119.12
```

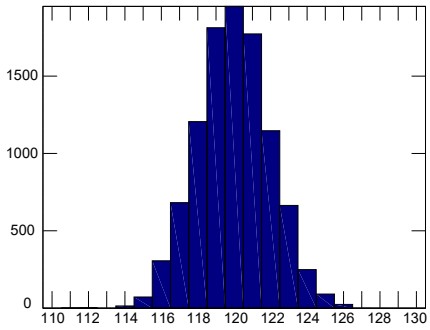


Figure 1.1. Histogram is a discrete approximation of the probability distribution of a random variable. It is obtained from experimental data.

```
117.94
122.35
121.26
120.97
122.66
.
.
.
```

The list $\{x_i = X(\omega_i)\}_{i=1}^{n}$ contains a total of 10,000 values ($i = 1, \ldots, n$, where $n = 10,000$). The nominal diameter of a CD is 120 mm.

Now suppose that we we plot a histogram of these values as follows: bin the horizontal axis into 20 bins (i.e. from 110 to 130), with each bin of width 1. These bins could be centered on those values, for example: [109.5, 110.5], ..., [119.5, 120.5], [120.5, 121.5], [121.5, 122.5], ..., etc. For each bin, count the number of times X falls within the interval of the bin, i.e. # of times you find a value in the above list that falls within that interval. Obviously, the longer the list (the larger n is), the larger this count will be.

Plot this value (frequency of occurrence) vs diameter, as shown in Fig. 1.1.[1] This histogram is a (discrete) approximation of the probability density function (PDF), which is a function that describes the distribution function of

[1]This plot was generated in GNU Octave using the following sequence of commands: `x=120+2*randn([10000 1]);` and `figure; hist(x,linspace(110,130,21));` The first command creates a "fake" data set that simulates the acquisition of experimental data with random error (more specifically, a random variable x with mean 120 plus normally-distributed random noise with standard deviation of 2).

the random variable X. The finer the bins are, and the larger n is, the more closely this histogram approximates a continuous function.

The histogram is also called empirical distribution. Mathematically, we partition the horizontal axis in N "bins", defined by the intervals $(r_k, r_{k+1}]$, where $\cdots < r_0 < r_1 < r_2 < \ldots$, $r_i \in \mathbb{R} \cup \{-\infty, +\infty\}$. Recall that $\{x_i = X(\omega_i)\}_{i=1}^n$ is our set of measurements of X (data points) and n the number of points. In terms of this dataset the histogram is the function:

$$\hat{h}(x) = \frac{1}{n} \sum_{k=1}^{N} \frac{\mathbb{I}_{(r_k, r_{k+1}]}(x)}{(r_{k+1} - r_k)} \sum_{i=1}^{n} \mathbf{1}_{\{x_i \in (r_k, r_{k+1}]\}}$$

where $\mathbf{1}_{\{x_i \in (r_k, r_{k+1}]\}}$ is an indicator function that equals 1 when $x_i \in (r_k, r_{k+1}]$ and 0 otherwise. Similarly, $\mathbb{I}_{(r_k, r_{k+1}]}(x)$ is an indicator function that equals 1 when $x \in (r_k, r_{k+1}]$ and 0 otherwise. $\sum_{i=1}^{n} \mathbf{1}_{\{x_i \in (r_k, r_{k+1}]\}}$ counts the number of times a result $x_i = X(\omega_i)$ falls into the bin $(r_k, r_{k+1}]$. The coefficient $\sum_{k=1}^{N} \mathbb{I}_{(r_k, r_{k+1}]}(x)$ ensures that $x \in (r_k, r_{k+1}]$ (k is a counter that loops over all bins, one at a time). In Problem 34, we prove that the histogram converges to the PDF in the limit of large n.

The term "empirical distribution" often refers to the expression:

$$p(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)$$

where $p(x)$ is a PDF and $\{x_i\}_{i=1}^n$ is the set of experimentally measured data points. Here, $\delta(x)$ is the Dirac delta function. This will be discussed in Section 2.4.

1.3. Methods for reporting uncertainties

1.3.1. Method 1. When reporting a measured quantity x and its uncertainty δx, we write

$$(\text{measured value of } x) = x \pm \delta x$$

For example, the length of a rod is measured to be (1.0 ± 0.1) m. In this course we will present common ways to estimate the value of δx and which value of x to report. In a certain sense, this notation means that most values measured experimentally will fall within the interval $[x - \delta x, x + \delta x]$.

In reality, however, the measured value of x is the sum of the true value of x, x_{true}, systematic errors (*bias*) plus a random error (ξ):

$$(\text{measured value of x}) = \underbrace{x_{true} + bias}_{\text{"constant"}} + \underbrace{\xi}_{\substack{\text{rapidly} \\ \text{fluctuating}}}$$

where ξ is a random variable whose outcomes fall within the interval $[-\delta x, \delta x]$ to a large extent. If the value of the bias is known, it can always be subtracted from the measurement to obtain $x_{true} \pm \delta x$. In some cases it is possible to design experiments to specifically measure *bias*.

1.3.2. Method 2 (notation). Another method for reporting uncertainties uses brackets to list the digits that are uncertain:

$1.234(55)$ m is a more-compact way of writing 1.234 ± 0.055 m.

1.3.3. Report your uncertainties with 1 or 2 significant figures. Experimental uncertainties should be rounded to 1 or 2 significant figures. For example,

$$(10.000 \pm 0.123) \text{ m}$$

should really be rounded to

$$(10.0 \pm 0.1) \text{ m}$$

or

$$(10.00 \pm 0.12) \text{ m}.$$

Two significant figures are normally used in precision measurements. In most other cases, we keep only one significant figure. In this course, we will stick to 1 figure because we are not doing precision measurements.

We round the uncertainty to the 1 significant figure and then report the value of x to the same digits. For example, writing

$$(30.129 \pm 0.1) \text{ m}$$

does not make sense because the number of significant figures to the right of the period differ for each number. We should instead report

$$(30.1 \pm 0.1) \text{ m},$$

where the value of x has been rounded to the same decimal place as the error.

In most calculations, it is advisable to keep many significant figures during the calculation and only round off at the end, when the final uncertainty has been determined. This is to avoid round-off errors.

When comparing two different reported values, we can determine if these values differ significantly or not by looking at the interval defined by the uncertainties. For example, if someone reports two different lengths,

$$(10 \pm 5) \text{ m and } (18 \pm 5) \text{ m},$$

we see that the error bars overlap: the first interval is

$$[5, 15]$$

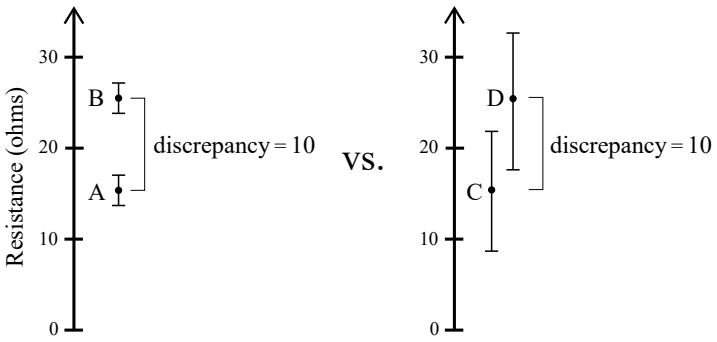


Figure 1.2. Discrepancy between two measurements. We must look at the errors bars, not just the nominal value. (left) No overlap (right) Some overlap. In the second case, we cannot say for sure that the two measured values are significantly different because of the overlap.

while the second is

$$[13, 23].$$

Since the error bars overlap, we cannot say for sure that the two numbers differ significantly. On the other hand if the numbers reported are

$$(10 \pm 1) \text{ m and } (18 \pm 1) \text{ m},$$

which correspond to intervals

$$[9, 11] \text{ and } [17, 19].$$

In this case, we can say that the two values are significantly different from each other.

Figure 1.2 illustrates the error bars for the case where resistance A is measured to be $(15 \pm 1)\ \Omega$ and resistance B is measured to be $(25 \pm 2)\ \Omega$ versus the case where resistance C is measured to be $(16 \pm 8)\ \Omega$ and resistance D is $(26 \pm 9)\ \Omega$.

Since the error bars do not overlap, values A and B are said to be significantly different from each other. Whereas measurements C and D do not differ significantly because of the substantial overlap of their error bars. Even though the x_{best} values differ by 10, in one case, the values are significantly different whereas in the other case they are not.

1.4. Random Errors, Systematic Errors and Mistakes

There are three main types of errors encountered in experiments:

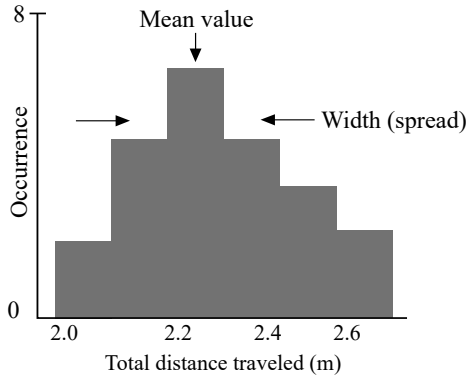


Figure 1.3. Histograms tell us about accuracy (if the true value is known) and precision. The latter is related to the spread of the distribution.

- **Random errors:** These errors are random fluctuations that influence the precision (or "spread") of the measurement, as shown in Fig. 1.3. The width of the distribution provides a measure of the random error. The common method for reducing random errors is by repeating the measurement many times and taking the average. We will see later that this method reduces the random error by a factor $\sqrt{n}$, where n is the number of measurements. The random scatter in the data can be of technical origin (due to the apparatus) or fundamental noise (e.g. Johnson noise, shot noise).
- **Systematic errors:** These errors are caused by a bias in the system or a mis-calibration of the instrument and typically cause the result to "tilt" or "shift" or "drift" in some direction away from the accepted or predicted value. Such errors can sometimes be difficult to detect or correct. To diagnose systematic errors, we need to know the "true value" of a measurement. Identification of the systematic "shift" can be accomplished by performing an experiment with known conditions and parameters. Correction of the systematic error may involve a simple subtraction of the shift or drift, or changes in the apparatus or experiment.
- **Mistakes:** These are user errors such as writing down the wrong value, misreading the scales on the instrument, confusion over the units, or malfunctions of the apparatus.

1.5. Uncertainty of a single measurement

Suppose that we want to determine the uncertainty of a single measurement, so we can write $x_{best} \pm \delta x$, where x_{best} here is the results of a single measurement. The random error can be obtained as follows:

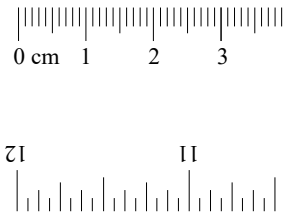


Figure 1.4. It is customary to take the error bars associated with analog measurements (from a ruler, here) to be half of the smallest division (millimeters, here).

- **From the measurement device:** If you are using a digital device, use the manufacturer's specs or a full last digit, e.g.

$$(1.56 \pm 0.01)\ \Omega$$

For analog devices, report half of a division, e.g.

$$\text{measure } 8.5 \text{ m} \qquad \rightarrow \qquad \text{report } (8.50 \pm 0.05) \text{ m}$$

Here, there is no problem adding a zero to 8.5 because with the analog device, we are able to "eyeball" the extra digit by estimating the position between the two divisions. An example of analog device is shown in Fig. 1.4.
It is also possible to estimate the uncertainty yourself if the divisions are large enough. In that case, you can estimate the last digit based on the position between two divisions and estimate the uncertainty in your procedure. For example, the analog device may only have divisions every 1 cm. But you may be able to estimate up to the millimeter by eyeballing the measurement. This procedure may not necessarily give you precision to the 1 mm scale, but you may be able to estimate it within 3 mm. Suppose the measurement is 5 cm according to the divisions available on the analog device, and you estimate the last digit to be 5.3 cm. You would then report 5.4 ± 0.3 cm.

- **Using the standard deviation:** The standard deviation can be used to represent the error in a single measurement. The problem is that we need many measurements to obtain the standard deviation, i.e. we need to repeat the measurement n times and compute the "sample standard deviation"

$$\hat{\sigma}_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_X)^2} \qquad \text{(sample standard deviation)} \tag{1.1}$$

where $\hat{\mu}_X$ is the mean X estimated from the arithmetic sum

$$\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} x_i. \qquad \text{(sample mean)}$$

We use a hat notation ($\hat{\sigma}$, $\hat{\mu}$) to indicate that the quantities (σ, μ) are estimated from data $\{x_i = X(\omega_i)\}_{i=1}^{n}$. The quantity $\hat{\mu}_X$ is called the "sample mean" because it is the calculation of the mean from the samples $x_1, \ldots, x_n$. Other accepted names for these terms include "population standard deviation" and "population mean".[2]

Note: One problem with the second method is that the measurement precision may be limited by the finite resolution of the measurement device. For example, it is possible that repeated measurements all give the same answer:

$$10.0 \text{ m}, 10.0 \text{ m}, 10.0 \text{ m}, 10.0 \text{ m}, \ldots, 10.0 \text{ m}.$$

If that is the case, use the first method (estimate the error bar from the smallest division of the measurement device.)

1.6. Standard error

If you report for x_{best} a mean value calculated from repeated measurements, the error in the measurement of x_{best} is the standard error, not the standard deviation. The standard error is defined as the ratio of the sample standard deviation to $\sqrt{n}$

$$\alpha = \frac{\hat{\sigma}_{n-1}}{\sqrt{n}}$$

If we collect n data points, each with uncertainty $\hat{\sigma}_{n-1}$ ("sample standard deviation" of a single measurement), and calculate the sample mean $\hat{\mu}_X$, the uncertainty in the sample mean is α.

We report

$$\hat{\mu}_X \pm \alpha = \hat{\mu}_X \pm \frac{\hat{\sigma}_{n-1}}{\sqrt{n}}$$

The standard error is also called the "standard deviation of the mean".

1.7. Reporting experimental results for a mean value

When doing experiments, we typically repeat a measurement many times to reduce its random error. The number reported is the sample mean. (Even if, as an experimentalist, you don't do these repeats yourself, it is likely

[2] It may seem ridiculous to collect n measurements just to compute the error bar of a single measurement; true, but it is nonetheless the correct way to obtain the error in a single measurement, when repeated measurements are possible. If you can afford to perform n measurements, report the mean value instead; in which case, the error in the mean is called the standard error, which is defined below.

that your measurement apparatus does this averaging for you.) Here is how this procedure is handled (either manually by you or automatically by your experimental apparatus):

- Given a random variable X to measure, collect experimental data as n samples
$$x_1, x_2, \ldots, x_n$$
where $x_i = X(\omega_i)$.
- Calculate the sample mean $\hat{\mu}_X$ as the arithmetic sum $\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} x_i$ while keeping all the significant figures.
- Calculate the standard error (error in the mean $\hat{\mu}_X$), $\alpha = \frac{\sigma_{n-1}}{\sqrt{n}}$ where
$$\sigma_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_X)^2},$$
while keeping all the significant figures.

- Decide how many significant figures to retain for the error. If many data points are used, keep 2 significant figures. Otherwise keep only 1. In this course, we will stick to retaining only 1 significant figure for the error bar.
- Round the mean to the appropriate decimal place.

Reporting errors:
If the experimental measurement of a random variable X can be repeated many times (e.g., $x_1 = X(\omega_1), x_2 = X(\omega_2), \ldots, x_n = X(\omega_n)$), the most common way to report the experimental value of X is:
$$(\text{measured } \ X) = \hat{\mu}_X \pm \frac{\hat{\sigma}_{n-1}}{\sqrt{n}}$$
where $\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\hat{\sigma}_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_X)^2}$. This assumes that $\hat{\sigma}_{n-1}$ is a meaningful (non-zero) value, i.e. it is constructed from non-identical measurements (such as 10.0, 10,0, 10.0, 10.0, etc).
If it is not possible to repeat the experimental measurement (due to time constraint or other reason), then report the single measured data point, x_1, plus uncertainty as stated by the manufacturer or using a pre-calculated value for the standard deviation.[a]

[a] Although calculating the standard deviation does require repeated measurements, this task can be done at a more convenient time.

1.8. Precision vs. Accuracy

Precision and accuracy are different concepts.

1.8.1. Accuracy. Suppose your know the "true value" of a physical quantity. For example, the speed of light in vacuum is a universal physical constant that is defined to be equal to

$$c = 299,792,548 \text{ m/s}.$$

This figure is exact since the length of the meter is defined from this constant and the international standard for time. The "true value" can be obtained from other means. The accuracy is a measure of how close is the mean value of your measurement from the true value.

1.8.2. Precision. Loosely speaking, *precision* is defined as the "spread" or scatter of values around the mean. A precise measurement corresponds to a small spread, whereas an imprecise measurement corresponds to a large spread.

In Fig. 1.5, four distinct cases are illustrated: accurate and precise measurement, accurate and imprecise measurement, inaccurate but precise measurement and inaccurate and imprecise measurement. The graphs shown are obtained by plotting histograms of the frequency of a given experimental result versus the value measured.

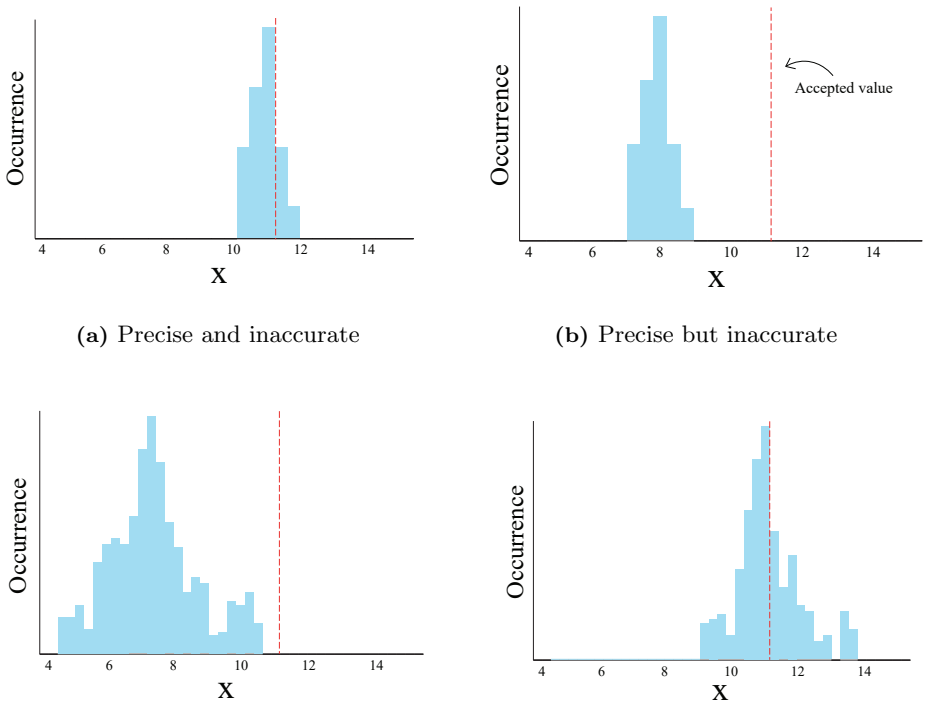
If the "true value" corresponds to the dotted line, an accurate measurement of x means that the average value of the measurement, which can be thought of as the "center of mass" of the histogram, lies close to the dotted line. This is true, regardless of the sharpness or spread around this mean value. An inaccurate measurement is one where the measured values cluster far away from the dotted line.

1.8.3. Fitting data to a model. As scientists, we often need to validate theory with experiments. Physical quantities are related to one another via some physical law (a mathematical formula). In which case, there is an equation (model) available to fit your data to. By fitting data to your model, you can see if the model suitably describes the physical situation. (And there are situations, of course, where the model may not be known and analysis of the data requires you to identify a suitable model.)

Experimental data comes with error bars. Plotting error bars on a graph enables us to decide if experimental data is consistent with a given model. Let us look at the example of Hooke's law which describes the extension of a spring from its equilibrium position when attaching a mass to its end.

Example: Hooke's law. Hooke's law says that the force F on the spring is linearly proportional to the displacement x from equilibrium ($x = 0$):

$$F = kx$$



(a) Precise and inaccurate

(b) Precise but inaccurate

(c) Imprecise and inaccurate

(d) Imprecise but accurate

Figure 1.5. Precision vs accuracy. Repeated measurements of an object's length x (arbitrary units). The nominal length (true value) is 11.3, as indicated by the red dashed line.

If we attach a mass m to the spring, the force acting on the spring is the gravitational force on the mass m:

$$F = mg,$$

where g is the gravitational acceleration constant, approximately $g \sim 9.8\ \mathrm{m/s^2}$. The situation is illustrated in the figure below. Writing the Hooke's law in the form

$$x = \frac{F}{k} = \frac{mg}{k} = (\frac{g}{k})m$$

immediately suggests a possible experiment for measuring the value of k: we attach different masses m and measure the corresponding displacement x. The slope of this graph yields g/k, from which we can obtain k (Fig. 1.6). Figure 1.6 also shows an example where linear fit does not describe the data well and a quadratic component must be added to the model.

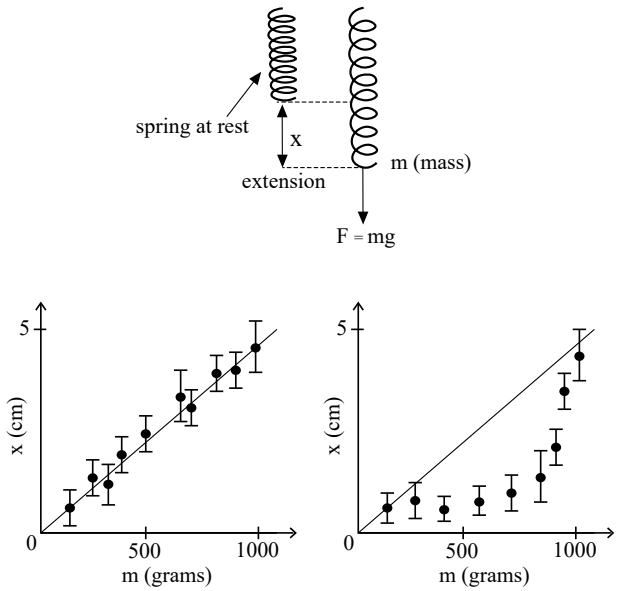


Figure 1.6. The plot on the left (right) shows a scenario where the experimental data are well-described by a linear (quadratic) trend.

1.8.4. Fractional Uncertainties. It is sometimes convenient to express values

$$(\text{measured } x) = x_{best} \pm \delta x$$

in terms of fractional uncertainty

$$(\text{fractional uncertainty}) = \frac{\delta x}{|x_{best}|}$$

If we multiply by 100 we get the percentage uncertainty:

$$(\text{percentage uncertainty}) = \frac{\delta x}{|x_{best}|} \times 100$$

If we multiply by $1,000,000$ we get "parts per million":

$$(\text{parts per million (ppm)}) = \frac{\delta x}{|x_{best}|} \times 10^6.$$

and similarly for "parts per billion" (ppb). The main advantage of fractional uncertainty is that it is a dimensionless quantity (it has no units).

Examples:

$$10 \pm 1 \text{ corresponds to } 10 \pm 10\%$$
$$99 \pm 1 \text{ corresponds to } 99 \pm 1\%$$

1.9. Rules for Rounding and Significant digits

1.9.1. Significant Digits. Here are accepted rules for reporting significant digits

- All non-zero digits are significant:

$$2.998 \times 10^8 \text{ m/s has 4 significant figures}$$

- All zeroes between non-zero digits are significant

$$6.022\ 141\ 79 \times 10^{23} \text{ mol}^{-1} \text{ has 9 significant figures}$$

- Zeroes to the left of the first non-zero digits are not significant

$$0.51 \text{ MeV has 2 significant figures}$$

- Zeroes at the end of a number to the right of the decimal point are significant

$$1.60 \times 10^{-19} \text{ C has 3 significant figures}$$

- If a number ends in zeroes without a decimal point, the zeroes might be significant:

$$270\ \Omega \text{ could have 2 or 3 significant figures}$$

  To avoid confusion, report instead:

$$2.70 \times 10^2\ \Omega \text{ or } 2.7 \times 10^2\ \Omega$$

1.9.2. Rounding rules. To round a number at the N-th digit, where N here will be taken as the tenths digit position for illustrative purposes, the generally accepted practice is described below:

(1) If the next digit $(N+1)$ is 4 or lower, leave N unchanged:

$$6.6\textcircled{2} \times 10^{-34} \text{ becomes } 6.6 \times 10^{-34} \text{ (2 sig figs)}$$

since '2' is lower than 4.

(2) If the next digit $(N+1)$ is 6 or higher, increase N by 1:

$$5.6\textcircled{7} \times 10^{-8} \text{ becomes } 5.7 \times 10^{-8}.$$

since '7' is greater than 6.

(3) If the digit after last one to be retained $(N+1)$ is 5

(a) Leave last digit (N) unchanged if digit N is even:

$$9.\textcircled{4}5 \text{ becomes } 9.4 \text{ (2 sig figs)}$$

since '4' is even.

(b) Increase last digit (N) by 1 if digit N is odd:

$$9.\textcircled{7}5 \text{ becomes } 9.8 \text{ (2 sig figs)}$$

since '7' is odd.

(4) If the digit after last one to be retained $(N+1)$ is 5 but there are non-zero digits to the right of 5, round N up:

$$10.75\textcircled{01} \text{ becomes } 10.8 \text{ (2 sig figs)}$$

since there are non-zero digits to the right of the '5'. Here, the parity of the N-th digit '7' does not affect the decision to round.

Rounding the result of addition and subtraction: We round-off the result to the same number of decimal places as the number with the least number of decimal places:

$$1.23 + 45.\textcircled{6} = 46.\textcircled{8}$$

(We have dropped the '3' in 1.23 because 45.6 does not contain such a level of precision; hence the '3' becomes meaningless when adding.)

This rounding rule is not arbitrary. It is grounded in the theory and methods of error propagation (see Chapter 3). We will see that if two numbers are added, $Z = X + Y$, the rules of error propagation give $\alpha_Z = \sqrt{\alpha_X^2 + \alpha_Y^2}$, where α_i, $i = X, Y, Z$ is the error in X, Y, Z, respectively. In the above example, suppose that α_X and α_Y differ by at least an order of magnitude, i.e. $\alpha_X = 10^n \alpha_Y$, $n \geq 1$. Then, $\alpha_Z = \alpha_X \sqrt{1 + 10^{-2n}} \approx \alpha_X$ is a good approximation since $\sqrt{1 + 0.01} \approx 1$. If the error bars are assumed to be 1 unit of the last significant digit, $\alpha_X = 10^m$, $\alpha_Y = 10^{m-n}$, then the error in Z is at the same digit, 10^m. The uncertainty in Z is therefore determined by the number with the least number of decimal places.

1.9.2.1. *Rounding the result of multiplication and division.* We keep the same number of significant figures as the component with the least number of significant figures:

$$\underbrace{\textcircled{1.2}}_{\text{2 sig figs}} \times 345.6 = 414.72 = \underbrace{\textcircled{4.1}}_{\text{2 sig figs}} \times 10^2$$

We will be learning in Chapter 3 about methods of error propagation.

1.10. Problems

Problem 1. Partial differentiation: (a) Find the total differential of the function $f(x, y) = y \exp(x + y)$. (The total differential of $f(x, y)$ is defined as $\mathrm{d}f = \frac{\partial f}{\partial x}\mathrm{d}x + \frac{\partial f}{\partial y}\mathrm{d}y$.)

(b) Find the first and second partial derivatives of the function $f(x, y) = 2x^3y^2 + y^3$, i.e. calculate $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, $\frac{\partial^2 f}{\partial x^2}$, $\frac{\partial^2 f}{\partial y^2}$, $\frac{\partial^2 f}{\partial x \partial y}$, $\frac{\partial^2 f}{\partial y \partial x}$.

(c) Suppose we have a function $f(x, y)$ where $y = y(x)$. The total derivative of f with respect to x is

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\partial f}{\partial x} + \left(\frac{\partial f}{\partial y}\right)\frac{\mathrm{d}y}{\mathrm{d}x}.$$

Find the total derivative of $f(x, y) = x^2 + 3xy$ with respect to x, given that $y = \sin^{-1} x$.

Solution. (a)

$$\frac{\partial f}{\partial x} = y\exp(x+y), \quad \frac{\partial f}{\partial y} = \exp(x+y) + y\exp(x+y).$$
$$\mathrm{d}f = [y\exp(x+y)]\mathrm{d}x + [(1+y)\exp(x+y)]\mathrm{d}y.$$

(b)

$$\frac{\partial f}{\partial x} = 6x^2y^2, \quad \frac{\partial f}{\partial y} = 4x^3y + 3y^2, \quad \frac{\partial^2 f}{\partial x^2} = 12xy^2, \quad \frac{\partial^2 f}{\partial y^2}4x^3 + 6y,$$
$$\frac{\partial^2 f}{\partial x \partial y} = 12x^2y = \frac{\partial^2 f}{\partial y \partial x}$$

(c)

$$\frac{\partial f}{\partial x} = 2x + 3y, \quad \frac{\partial f}{\partial y} = 3x, \quad \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{(1-x^2)^{1/2}}$$

and so

$$\frac{\mathrm{d}f}{\mathrm{d}x} = 2x + 3y + 3x\frac{1}{(1-x^2)^{1/2}} = 2x + 3\sin^{-1}x + \frac{3x}{(1-x^2)^{1/2}}.$$

■

Problem 2. Find the gradient of the following functions:
(a) $r = \sqrt{x^2+y^2+z^2}$
(b) $f(x, y, z) = x^2 + y^3 + z^4$
(c) $f(x, y, z) = x^2y^3z^4$
(d) $f(x, y, z) = e^x \sin(y)\log(z)$

Solution. (a) $\nabla r = \hat{\mathbf{r}}$. (Since $\partial_x r = (1/2)(1/r)2x = x/r$, etc.)
(b) $\partial_x f = 2x$, $\partial_y f = 3y$, $\partial_z f = 4z$. So that $\nabla f = (2x, 3y, 4z)$.
(c) $\partial_x f = 2xy^3z^4$, $\partial_y f = 3x^2y^2z^4$, $\partial_z f = 4x^2y^3z^3$. So that

$$\nabla f = (2xy^3z^4, 3x^2y^2z^4, 4x^2y^3z^3).$$

(d) $\partial_x f = e^x \sin(y)\log(z)$, $\partial_y f = e^x \cos(y)\log(z)$, $\partial_z f = e^x \sin(y)z^{-1}$.
So that $\nabla f = e^x(\sin(y)\log(z), \cos(y)\log(z), \sin(y)z^{-1})$. ■

Problem 3. For the three questions below let $\mathbf{r}-\mathbf{r}'$ be the separation vector from a fixed point (x',y',z') to the point (x,y,z) and $|\mathbf{r}-\mathbf{r}'|$ be its length. If ∇ indicates derivative with respect to the unprimed variables $\mathbf{r}$, show that:

(e) $\nabla|\mathbf{r}-\mathbf{r}'|^2 = 2(\mathbf{r}-\mathbf{r}')$

(f) $\nabla(1/|\mathbf{r}-\mathbf{r}'|) = -(\mathbf{r}-\mathbf{r}')/|\mathbf{r}-\mathbf{r}'|^3$

(g) What is the general formula for $\nabla(|\mathbf{r}-\mathbf{r}'|^n)$?

(h) In problem (f) above you have computed $\nabla(1/|\mathbf{r}-\mathbf{r}'|) = -(\mathbf{r}-\mathbf{r}')/|\mathbf{r}-\mathbf{r}'|^3$. Now calculate the quantity $\nabla'(1/|\mathbf{r}-\mathbf{r}'|)$, where ∇' denotes the derivative with respect to the primed variables $\mathbf{r}'$ (instead of the unprimed variables $\mathbf{r}$).

Solution. For (e), the x component of the gradient is $\partial_x((x-x')^2+(y-y')^2+(z-z')^2) = 2x$, etc. so that $\nabla|\mathbf{r}-\mathbf{r}'|^2 = 2(\mathbf{r}-\mathbf{r}')$. For (f) the x component is $\partial_x 1/\sqrt{(x-x')^2+(y-y')^2+(z-z')^2} = (-1/2)((x-x')^2+(y-y')^2+(z-z')^2)^{-3/2}(2x)$ so that $\nabla(1/|\mathbf{r}-\mathbf{r}'|) = -(\mathbf{r}-\mathbf{r}')/|\mathbf{r}-\mathbf{r}'|^3$. For (g) we have $\nabla|\mathbf{r}-\mathbf{r}'|^n = n\frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}|\mathbf{r}-\mathbf{r}'|^{n-1}$. For (h) the only difference is a negative sign. ■

Problem 4. Show that the Taylor expansion, up to quadratic terms in $x-2$ and $y-3$, of $f(x,y) = y\exp(xy)$ about the point $x=2$, $y=3$, is:

$$f(x,y) = e^6\Big\{3+9(x-2)+7(y-3)+ \\ (2!)^{-1}\Big[27(x-2)^2+48(x-2)(y-3)+16(y-3)^2\Big]\Big\}$$

(Show all your calculations)

Solution. The Taylor expansion of a function of two variables is:

$$f(x+h,y+k) = f(x,y)+(h,k)\cdot\nabla f(x,y)+\frac{1}{2}[h\,k]\begin{bmatrix}\partial^2_{xx}f & \partial^2_{xy}f\\ \partial^2_{yx}f & \partial^2_{yy}f\end{bmatrix}\begin{bmatrix}h\\k\end{bmatrix}+O(|(h,k)|^3)$$

To obtain the stated result, plug in $x=2$ and $y=3$ in the Taylor expansion formula and take $h=x-2$ and $k=y-3$ as the step size, making use of the following derivatives (evaluated at the point $x=2$ and $y=3$):

$$\frac{\partial f}{\partial x} = y^2\exp(xy),\quad \frac{\partial f}{\partial y} = \exp(xy)+xy\exp(xy),\quad \frac{\partial^2 f}{\partial x^2} = y^3\exp(xy),$$
$$\frac{\partial^2 f}{\partial y^2} = 2x\exp(xy)+x^2y\exp(xy),\quad \frac{\partial^2 f}{\partial x\partial y} = 2y\exp(xy)+xy^2\exp(xy).$$

■

Problem 5. Shorthand notation: $f_x \equiv \frac{\partial f}{\partial x}$, $f_{xy} = \frac{\partial^2 f}{\partial x\partial y}$, $f_{xx} = \frac{\partial^2 f}{\partial x^2}$, etc. The multivariable Taylor expansion can be used to study the behavior of

functions near extrema. All stationary points have $f_x = f_y = 0$ and these points may be classified as:

(1) minima if both f_{xx} and f_{yy} are positive and $f_{xy}^2 < f_{xx}f_{yy}$,

(2) maxima if both f_{xx} and f_{yy} are negative and $f_{xy}^2 < f_{xx}f_{yy}$,

(3) saddle point if f_{xx} and f_{yy} have opposite signs or $f_{xy}^2 > f_{xx}f_{yy}$.

Prove that the function $f(x, y) = x^4 + y^4$ has a minimum at the origin, but that $g(x, y) = x^4 + y^3$ has a saddle point there.

Solution. We have $f_x \equiv \partial_x f = 4x^3$ and $f_y \equiv \partial_y f = 4y^3$, both of which are zero at the origin. Thus the origin is a stationary point. To check that this is a maximum, minimum or not, we need the second derivatives: $f_{xx} = 12x^2$, $f_{yy} = 12y^2$, $f_{xy} = 0$. Both f_{xx} and f_{yy} are zero at the origin, so the test is inconclusive. However, both are positive in a small neighborhood of the origin and $f_{xy}^2 < f_{xx}f_{yy}$. Thus, we have a minimum. ■

Problem 6. Show that the function $f(x, y) = x^3 \exp(-x^2 - y^2)$ has a maximum at the point $(\sqrt{3/2}, 0)$, a minimum at $(-\sqrt{3/2}, 0)$. What about the nature of the point at the origin?

Solution. Setting the first two partial derivatives to zero to locate the stationary points, we find

$$\frac{\partial f}{\partial x} = (3x^2 - 2x^4)\exp(-x^2 - y^2) = 0, \quad \frac{\partial f}{\partial y} = -2yx^3 \exp(-x^2 - y^2) = 0.$$

For the second equation to be satisfied we require $x = 0$ or $y = 0$ and for the first one to be satisfied we require $x = 0$ or $x = \pm\sqrt{3/2}$. Hence the stationary points are at $(0, 0,$ $(\sqrt{3/2}, 0)$ and $(-\sqrt{3/2}, 0)$. We now find the second partial derivatives:

$$f_{xx} = (4x^5 - 14x^3 + 6x)\exp(-x^2 - y^2)$$
$$f_{yy} = x^3(4y^2 - 2)\exp(-x^2 - x^2)$$
$$f_{xy} = 2x^2y(2x^2 - 3)\exp(-x^2 - y^2)$$

We then substitute the pairs of values of x and y for each stationary point and find that at $(0, 0)$

$$f_{xx} = 0, \quad f_{yy} = 0, \quad f_{xy} = 0$$

and at $(\pm\sqrt{3/2}, 0)$

$$f_{xx} = \mp\sqrt{3/2}\exp(-3/2), \quad f_{yy} = \mp\sqrt{3/2}\exp(-3/2), \quad f_{xy} = 0.$$

Hence applying the above three criteria, we find that $(0, 0)$ is an undetermined stationary point, $(\sqrt{3/2}, 0)$ is a maximum and $(-\sqrt{3/2}, 0)$ is a minimum. The function is shown in Figure 1.7. ■

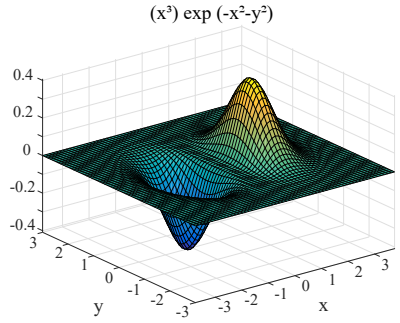


Figure 1.7. Plot of the function f.

Problem 7. Matrix inverse. (a) Derive the formula for matrix inverse of a 2x2 matrix:

$$\mathbf{A}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

(b) Check that it satisfies the *definition* of the matrix inverse, namely check that $\mathbf{A^{-1}A} = \mathbf{AA^{-1}} = \mathbf{I}$ where $\mathbf{I}$ is the 2×2 unit matrix.

(c) Prove that taking the inverse of n×n matrices $\mathbf{A}, \mathbf{B}$ reverses the order of matrices:

$$(\mathbf{AB})^{-1} = \mathbf{B^{-1}A^{-1}}$$

(d) Prove that (k is a non-zero scalar)

$$(k\mathbf{A})^{-1} = k^{-1}\mathbf{A}^{-1}$$

where $\mathbf{A}$ is an invertible matrix.

(e) Prove that taking the inverse of an invertible matrix twice recovers the original matrix:

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A}$$

Problem 8. Manipulation of matrices and matrix operations. (a) For two arbitrary matrices $\mathbf{A}$ and $\mathbf{B}$, write down the *matrix products* $\mathbf{AB}$ and $\mathbf{BA}$. What are the conditions on $\mathbf{A}$ and $\mathbf{B}$ for the matrix product to *exist* (and be well-defined)?

(b) In general, does $\mathbf{AB}$ equal $\mathbf{BA}$?

(c) Check that for matrices $\mathbf{A}, \mathbf{B}$ and scalar c the following property holds:

$$(\mathbf{A}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{A}$$

where $\mathbf{A}^T$ denotes *matrix transpose*.

(d) Prove that the transpose operation preserves the matrix addition:

$$(\mathbf{A} + \mathbf{B})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}} + \mathbf{B}^{\mathrm{T}}$$

(e) Prove that taking the tranpose of a product of matrices reverses the orders of the matrices

$$(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$$

(f) Prove that a scalar is left invariant by the transpose operation:

$$(c\mathbf{A})^{\mathrm{T}} = c\mathbf{A}^{\mathrm{T}}$$

(g) Prove that the dot product of two column vectors $\mathbf{a}$ and $\mathbf{b}$ can be computed as

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^{\mathrm{T}}\mathbf{b}$$

where $\mathbf{a}^{\mathrm{T}}$ denotes the transpose of $\mathbf{a}$ (i.e. a row vector). Verify this by writing out explicitly the matrix product.

(h) Prove that the transpose of an invertible matrix is also invertible and its inverse is the transpose of the inverse of the original matrix:

$$(\mathbf{A}^{\mathrm{T}})^{-1} = (\mathbf{A}^{-1})^{\mathrm{T}}$$

Problem 9. Compute the following dimensionless quantity and find the correct error bars (pay attention to the order of operations indicated by the brackets):

$$[(0.0045 \pm 0.0005) \text{ plus } (0.3 \pm 0.9)] \text{ divided by } (1.5 \pm 0.1) = ?$$

Solution. From the first operation (addition) we have

$$\alpha_Z = \sqrt{(\alpha_A)^2 + (\alpha_B)^2} = \sqrt{(0.0005)^2 + (0.9)^2} = 0.9$$

and

$$Z_{best} = 0.0045 + 0.3 = 0.3045$$

which we round according to the error $= 0.3045 \pm 0.9 = 0.3 \pm 0.9$. Then we perform the second operation (division). There, we have

$$Z'_{best} = \frac{0.3}{1.5} = 0.2$$

and the error bar

$$\frac{\alpha_{Z'}}{Z'} = \sqrt{\left(\frac{\alpha_A}{A}\right)^2 + \left(\frac{\alpha_B}{B}\right)^2}$$

evaluates to

$$\alpha_{Z'} = 0.2\sqrt{\left(\frac{0.9}{0.3}\right)^2 + \left(\frac{0.1}{1.5}\right)^2} = 0.6.$$

And the final result is 0.2 ± 0.6. ■

Problem 10. Suppose that a random experiment consists of measuring the length L of an object. (a) Explain the possible sources of randomness in this measurement.

(b) The "state space" of this random experiment consists of all possible values L can take (without regards to the way L is measured, i.e. assume infinite precision). Is this state space continuous or discrete?

(c) Random variables depend on how they are defined relative to an experiment. Define X as the *random variable* which consists of the value L measured with infinite precision. Is X a continuous or discrete random variable?

(d) Let Y be the random variable which corresponds to the length L rounded to the precision of the measuring instrument. Is Y a discrete or continuous random variable?

(e) Let Z by the random variable which equals 1 if L is greater than 10 mm and 0 if L is less than or equal to 10 mm. (Such random variables are called "indicator functions", because they act as logical bits which are "on" when certain conditions are met and "off" otherwise.) Is Z a discrete or continuous random variable?

Solution. (a) *Any* experiment performed in the lab suffers from random errors (fluctuations). Their origin could be any of: rapid thermal fluctuations, mechanical vibrations, operator (the act of measurement can introduce errors), fluctuations in the electrical power supply, etc.

(b) Continuous. (Since the values of L are not denumerable.)

(c) Continuous. (X can take an indenumerable number of values.)

(d) Discrete. (The number of values Y can take can be infinite, but still denumerable.

(e) Discrete, since Z can only take 2 possible values: 0 and 1. ■

Problem 11. Let f be a differentiable real-valued function on $\mathbb{R}^3$ and v_p a tangent vector of $\mathbb{R}^3$ at a point $p \in \mathbb{R}^3$. Then

$$v_p[f] = \frac{\mathrm{d}}{\mathrm{d}t}(f(p+tv))|_{t=0} \qquad \text{directional derivative}$$

is the (directional) derivative of f with respect to v_p.

For example, suppose that $f = x_1x_2x_3$, $p = (1, -4, 2)$ and $v = (1, 1, 0)$, where x_1, x_2, x_3 are the coordinate functions of $\mathbb{R}^3$.

(a) Find $p + tv$ and show that $f(p + tv) = 2(1 + t)(-4 + t)$. Show that $v_p[f] = -6$ by direct computation of the limit.

(b) If $v_p = (v_1, v_2, v_3)$ is a tangent vector of $\mathbb{R}^3$ at a point p, then

$$v_p[f] = \sum_{i=1}^{3} v_i \frac{\partial f}{\partial x_i}(p).$$

Prove this statement from the above definition (limit).

(c) Use the definition of directional derivative in (b) to show that $v_p[f] = -6$.
(d) Let f and g be differentiable functions on $\mathbb{R}^3$, v_p and w_p tangent vectors at a point p, and a and b real numbers. Prove the following 3 properties:

$$(i) \quad (av_p + bw_p)[f] = av_p[f] + bw_p[f],$$
$$(ii) \quad v_p[af + bg] = av_p[f] + bv_p[g],$$
$$(iii) \quad v_p[fg] = v_p[f] \cdot g(p) + f(p) \cdot v_p[g].$$

Solution. (a)

$$p + tv = (1 + t, -4 + t, 2), \qquad f(p + tv) = 2(1 + t)(-4 + t)$$
$$v_p[f] = \frac{\mathrm{d}}{\mathrm{d}t} 2(1 + t)(-4 + t)|_{t=0} = 2(2t - 3)|_{t=0} = -6.$$

(b) Let $p = (p_1, p_2, p_3)$. Then,

$$f(p + tv) = f(p_1 + tv_1, p_2 + tv_2, p_3 + tv_3).$$

Since $(\mathrm{d}/\mathrm{d}t)(p_i + tv_i) = v_i$, by putting $x_i = p_i + tv_i$ and using the chain rule we obtain the desired result.
(c)

$$\frac{\partial f}{\partial x_1}(p) = x_2 x_3(p) = -8,$$
$$\frac{\partial f}{\partial x_2}(p) = x_1 x_3(p) = 2,$$
$$\frac{\partial f}{\partial x_3}(p) = x_1 x_2(p) = -4.$$

we obtain

$$v_p[f] = 1(-8) + 1(2) + 0(-4) = -6.$$

■

Problem 12. Report the following numerical distance correctly, with error bars: $X+Y$, where $X = 110.125 \pm 0.003$ m and $Y = 85.6 \pm 0.2$ m.

Solution. Whether we use the $\sigma_Z = \sigma_X + \sigma_Y$ method or the quadrature method, the error in $X + Y$ is dominated by the error in Y (i.e. $\sigma_Y \gg \sigma_X$). Thus the error bar is 0.2 m. Next, we round 110.125+85.6=195.725 to the tenths digit, giving (195.7 ± 0.2) m. ■

Problem 13. How many significant figures are there in this expression: 3000000000 liters.

Solution. Anywhere between 1 and 10. ■

Problem 14. Express the following result in proper rounded form, with suitable error bars: mass=19.1234 g with uncertainty 0.6789 g.

Solution. First we need to round the error bar to 1 or 2 sig figs. Let's do 1 sig fig: 0.7 g. Then we round the mass to this figure: 19.1 g. The result is: 19.1(7) g. ■

Problem 15. Your experiment yielded the following measurement:

$$(4.1234 \ \pm \ 0.4321) \text{ Joules.}$$

Report this number with proper error bars and appropriate significant figures.

Solution. Either (4.1 ± 0.4) J or (4.12 ± 0.43) J. ■

Problem 16. How many significant figures are there in each of the following expressions?

(i) 0.00082 s
(ii) 0.14800 psi
(iii) 6.24×10^{6} l
(iv) -754.090×10^{-27} J
(v) 50 cm
(vi) 50 m

Solution. (i) 2
(ii) 5
(iii) 3
(iv) 6
(v) 1 or 2
(vi) 1 or 2 ■

Problem 17. Express the following result in proper rounded form, with suitably truncated error bars: mass=8.4857 g with uncertainty 0.2554 g.

Solution. First we need to truncate the error bar to 1 or 2 sig figs. Let's do 1 sig fig: 0.3 g. Then we round the mass to this figure: 8.5 g. The result is: 8.5(3) g. ■

Problem 18. You measure the length of an object with a ruler (or measuring tape) whose smallest division is 1 mm. You measure the length 5 times with results in mm of 123.4, 123.5, 124.6, 123.7, 123.8 mm (the last digit you have estimated by eyeballing). What is the average length and the uncertainty in length?

Solution. Because this is an "analog" device (ruler's smallest division is 1 mm), we should take 1/2 or the smallest division as our error. Namely, 0.5 mm. Thus, the result is 123.8(5). ■

Problem 19. Compute the following dimensionless quantity and find the correct error bars (pay attention to the order of operations indicated by the brackets):

$$[(0.0045 \pm 0.0005) \text{ plus } (0.3 \pm 0.9)] \text{ divided by } (1.5 \pm 0.1) = ?$$

Solution. From the first operation (addition) we have

$$\alpha_Z = \sqrt{(\alpha_A)^2 + (\alpha_B)^2} = \sqrt{(0.0005)^2 + (0.9)^2} = 0.9$$

and

$$Z_{best} = 0.0045 + 0.3 = 0.3045$$

which we round according to the error $= 0.3045 \pm 0.9 = 0.3 \pm 0.9$. Then we perform the second operation (division). There, we have

$$Z'_{best} = \frac{0.3}{1.5} = 0.2$$

and the error bar

$$\frac{\alpha_{Z'}}{Z'} = \sqrt{\left(\frac{\alpha_A}{A}\right)^2 + \left(\frac{\alpha_B}{B}\right)^2}$$

evaluates to

$$\alpha_{Z'} = 0.2\sqrt{\left(\frac{0.9}{0.3}\right)^2 + \left(\frac{0.1}{1.5}\right)^2} = 0.6.$$

And the final result is 0.2 ± 0.6. ■

Chapter 2

Probability

Experimental measurements in the laboratory are random variables (rv). Every time you measure a physical quantity you get a different number because of random fluctuations (random errors). The random fluctuations are called random errors. Thus, a statistical description of experimental measurements is needed.

Here, we introduce tools to study random variables. Random variables can be continuous or discrete, depending on whether they take continuous or discrete values, respectively. An example of a continuous random variable is the length of an object. Length is a random variable which can take positive real values in a continuous interval. An example of a discrete random variable is the number of counts within a time interval. Counts can only take discrete values $(1, 2, 3, \dots)$, in this case, the positive integers.

2.1. Continuous Random Variables

As mentioned in the previous chapter a random variable X is not a simple variable; it is better described by associating it with a function that encodes all of its statistical properties.[1] We associate with X a probability density function (PDF), $p(x)$. As a matter of convention, we shall use capital letters (X) to denote random variables and lowercase letters (x) for the value of X in some particular experiment ω, i.e., $x = X(\omega)$.

[1]Think of an experiment performed on Monday. The value measured on Tuesday may be slightly different than the one obtained on Monday because of random errors. Same story for measurements performed on subsequent days — these values will also be different due to fluctuations.

2.2. Probability Density Function

The probability density function (PDF) of a random variable X, denoted $p(x)$, is everywhere non-negative: $p(x) \geq 0$ and is normalized to 1:

$$\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1.$$

We note that the PDF refers to a particular random variable (say, X). This is sometimes emphasized by adding a subscript, e.g., $p_X(x)$ instead of $p(x)$. When working with a single random variable we do not need the subscript because it should be clear that there is only one possible random variable that $p(x)$ refers to. However, when the problem involves more than one random variable, we should use a subscript to avoid confusion between the different PDFs.

2.3. Cumulative Distribution Function

We define the *cumulative distribution function* (CDF) as the integral of the PDF:

$$\boxed{\mathbb{P}(X \leq x) \equiv \int_{-\infty}^{x} p(x')\mathrm{d}x'}$$

The CDF is the probability that X takes a value less than or equal to x. The quantity $\{X \leq x\}$ is an example of a *random event*; the function $\mathbb{P}(\cdot)$ associates a number between 0 and 1 to this random event. The notation $\{X \leq x\}$ is shorthand for the set $\{\omega : X(\omega) \leq x\}$, i.e. the set of all random outcomes ω such that $X(\omega) < x$. We note that if $p(x)$ is continuous, then there is no distinction between $\mathbb{P}(X \leq x)$ and $\mathbb{P}(X < x)$. When discontinuities are present, we should be careful about the equality.

From this definition, we can solve for p in terms of P:[2]

$$p(x) = \left.\frac{\mathrm{d}\mathbb{P}(X \leq x)}{\mathrm{d}x}\right|_x.$$

We note that

$$\mathbb{P}(a \leq X \leq b) = \int_a^b p(x)\mathrm{d}x = \left(\int_{-\infty}^b - \int_{-\infty}^a\right) p(x)\mathrm{d}x = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a).$$

2.3.1. Interpretation of PDF. The PDF, $p(x)$, has the following interpretation. The quantity $p(x)\mathrm{d}x$ is the probability that the random variable X lies in the interval $[x, x + \mathrm{d}x]$:

$$p(x)\mathrm{d}x = \mathbb{P}(x \leq X \leq x + \mathrm{d}x),$$

[2] To differentiate the integral with respect to x, apply the Leibniz formula (see Section 12.3) for differentiation of integrals. In the expression $\int_{-\infty}^{x} p(x')\mathrm{d}x'$, the only dependence on x comes from the upper limit of the integral. Thus, $\frac{\mathrm{d}}{\mathrm{d}x}\int_{-\infty}^{x} p(x')\mathrm{d}x' = p(x)$.

where $\mathrm{d}x$ is an infinitesimally small quantity and $\mathbb{P}(\cdot)$ denotes the probability of the event $\cdot$ occurring. The quantity $p(x)\mathrm{d}x$ by itself is rarely used, except under the integral sign. Instead, one integrates this expression to find the probability that X will take some value in a finite interval $[a, b]$:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b p(x)\mathrm{d}x.$$

The last step follows from[3]

$$\mathbb{P}(x \leq X \leq x + \mathrm{d}x) = \mathbb{P}(X \leq x + \mathrm{d}x) - \mathbb{P}(X \leq x) = \mathrm{d}\mathbb{P}(X \leq x),$$

integrating $\mathbb{P}(x \leq X \leq x + \mathrm{d}x) = \mathrm{d}\mathbb{P}(X \leq x)$ from a to b yields $\int_a^b \mathrm{d}\mathbb{P}(X \leq x) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = \mathbb{P}(a \leq X \leq b)$ whereas integrating $p(x)\mathrm{d}x$ yields $\int_a^b p(x)\mathrm{d}x$. Since the two are equal, we have that $\mathbb{P}(a \leq X \leq b) = \int_a^b p(x)\mathrm{d}x$.

2.3.2. Limit value of CDF. We note that the CDF approaches 1 in the limit of large x. This follows from the normalization condition on the PDF.

2.4. Experimental Data: The Empirical Distribution

Suppose that our knowledge of the rv X is not its PDF, $p(x)$ but instead a series of data points obtained experimentally:

$$x_1 = X(\omega_1), x_2 = X(\omega_2), \ldots, x_n = X(\omega_n).$$

(An equivalent description that will be used in subsequent chapters is to take n independent rv's $X_1, \ldots, X_n$ of the same distribution as X and fix ω. The order in which rv's are measured is immaterial since they are assumed independent. Fixing ω implies that all random variables are measured simultaneously. The data is $\{x_i = X_i(\omega)\}_{i=1}^n$.)

We define the empirical PDF as follows:

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i),$$

where $\delta(x)$ is the Dirac delta function. It is trivial to verify that $\int \hat{p}(x)\mathrm{d}x = 1$ and $\hat{p}(x) \geq 0$. The CDF corresponding to $\hat{p}(x)$ is:

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} \hat{p}(x)\mathrm{d}x = \frac{1}{n}\#\{i : x_i \leq x\}.$$

Here, $\#\{i : x_i \leq x\}$ denotes the number of data points x_i satisfying the condition $x_i \leq x$. The empirical distribution $\hat{p}(x)$ is an approximation to the true PDF $p(x)$. This fact follows from the Law of Large Numbers (see Problem 35).

[3] In calculus, recall that $\mathrm{d}f(x) = f(x + \mathrm{d}x) - f(x)$.

2.5. Mean Value of Continuous Random Variable

Let X be a continuous rv. The mathematical expectation of X, denoted $\mathbb{E}(X)$, is defined as:

$$\boxed{\mathbb{E}(X) \equiv \int_{-\infty}^{\infty} x\, p(x) \mathrm{d}x.}$$

where the integral is over all values taken by the rv X (here, over the entire real line). If the random variable takes values in the interval $[0, 1]$ then the limits of the integral range from 0 to 1.

That is, to obtain the mean value of X, we replace the rv X by a regular variable x that represents its value, then multiply by $p(x)$ and integrate over all x.

Other names for $\mathbb{E}[X]$ include "mean value" (of X), or "expectation value" or "average value". Other symbols you may encounter in the literature include $\overline{X}$, μ_X, $\langle X \rangle$ or $m(X)$.

We note that this expression differs from the *sample mean* $\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} x_i$. The sample mean is an *estimate* of the mean.[4] Substitution of the empirical distribution (Eq. 2.1)

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i) \tag{2.1}$$

into the above definition for $\mathbb{E}[X]$ gives the sample mean:

$$\int_{-\infty}^{\infty} x \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)\mathrm{d}x = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Here $\{x_i\}_{i=1}^{n}$ denotes experimental measurements of X.

2.6. Indicator Functions

An *indicator function*, $\mathbf{1}_{\{x<y\}}(x)$ is a function that takes the value 1 when $x < y$ and 0 otherwise. Indicator functions can also be applied to random events. Let X be a rv and A is a random event. The indicator function for the random event $X \in A$ is defined as:

$$\mathbf{1}_A(X) = \begin{cases} 1 & \text{if } X \in A \\ 0 & \text{otherwise} \end{cases}$$

where $X \in A$ refers to the value of the rv X taken after a random experiment. Another notation for $\mathbf{1}_A(X)$ you may encounter is $\mathbf{1}_{X\in A}$. You may also encounter $\mathbb{I}_{X\in A}$ or $\chi_A(X)$ instead of $\mathbf{1}_{X\in A}$. Taking the mathematical

[4] More specifically, the sample mean is the best estimate of the mean in the sense of least squares.

expectation of $\mathbf{1}_{X\in A}$ and applying the definition of probability,

$$\mathbb{E}[\mathbf{1}_{X\in A}] = \int_{-\infty}^{\infty} \mathbf{1}_{x\in A}(x)p(x)\mathrm{d}x = \int_A p(x)\mathrm{d}x = \mathbb{P}(A).$$

where A is a random event and the integral $\int_A$ means integral over all points x that meet the condition $x \in A$ (for example, X could be a coordinate, and $A = (-\infty, y]$ indicates an event where the coordinate is less than y). Indicator functions are useful when dealing with experimental measurements. See Problems 34, 29 and 36 for example uses of the indicator function.

2.7. Variance

The *variance* of X, denoted $var(X)$ or σ_X^2, is defined as the sum of square differences between X and its mean, $\mu_X \equiv \mathbb{E}[X]$, weighted by the PDF:

$$\boxed{\sigma_X^2 \equiv var(X) = \int_{-\infty}^{\infty} p(x)(x-\mu_X)^2\mathrm{d}x}$$

The square can be expanded to give $\int_{-\infty}^{\infty} p(x)(x^2 + \mu_X^2 - 2x\mu_X)\mathrm{d}x$ and thus

$$\sigma_X^2 = \mathbb{E}[X^2] - (\mu_X)^2.$$

The square root of the variance is called the standard deviation and is commonly denoted σ.

2.8. Example PDFs

2.8.1. Point Distribution. Let X be a rv and $p(x)$ its distribution (PDF). The simplest known PDF is one concentrated at a single point x_0:

$$p(x) = \delta(x - x_0).$$

It is trivial to verify that $p(x) \geq 0$ and $\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$. The CDF is easily found:

$$\boxed{\mathbb{P}(X < a) = \int_{-\infty}^{a} \delta(x - x_0)\mathrm{d}x = \theta(a - x_0),}$$

where

$$\theta(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

is the Heaviside step function. We note that the Heaviside function can be expressed in terms of the indicator function as $\theta(x) = \mathbf{1}_{(0,\infty)}(x)$. We also note that $\delta(x) = \frac{\mathrm{d}}{\mathrm{d}x}\theta(x)$.

2.8.2. Discrete Distribution. Let X be a rv that can take values $\{x_i\}_{i=1}^N$ in a set $\mathcal{X}$. N is the number of possible values that X can take. The PDF

$$p(x) = \sum_{i=1}^{N} p_i \delta(x - x_i), \qquad p_i \geq 0, \qquad \sum_{i=1}^{N} p_i = 1$$

is called discrete distribution because it can be used to describe discrete rv. The set of number $\{p_i\}_{i=1}^N$ is called the probability mass function (PMF). The x_i represent the discrete values taken by the rv X. p_i is the probability of observing the discrete outcome $x_i \in \mathcal{X}$. The mean of X is:

$$\mu_X \equiv \mathbb{E}[X] = \int_{-\infty}^{\infty} x \sum_{i=1}^{N} p_i \delta(x - x_i) \mathrm{d}x = \sum_{i=1}^{N} p_i x_i.$$

The variance is

$$var(X) = \sum_{i=1}^{N} p_i (x_i - \mu_X)^2.$$

2.8.3. Distribution After Rescaling Of Random Variable. Let X be a rv. What is the distribution of $2X$? Since we are multiplying all values of X by 2, we at least expect the mean to be twice as large. What about the remaining details of its distribution? First of all we note that:

$$\mathbb{P}(2X < a) = \mathbb{P}(X < a/2) = \int_{-\infty}^{a/2} p_X(x) \mathrm{d}x.$$

Next, we differentiate this integral with respect to a to get the PDF:

$$\frac{\mathrm{d}}{\mathrm{d}a} \mathbb{P}(2X < a) = p_X(a/2) \cdot \frac{1}{2}.$$

We conclude that the PDF of $2X$ is half as high and twice as spread out compared to the distribution of X. If the mean of X is μ_X then the mean of $Y = 2X$ is

$$\mathbb{E}[Y] = \frac{1}{2} \int_{-\infty}^{\infty} y p_X(y/2) \mathrm{d}y = \frac{1}{2} \int_{-\infty}^{\infty} (2x) p_X(x) (2\mathrm{d}x) = 2\mu_X.$$

The variance is:

$$var(Y) = \frac{1}{2} \int_{-\infty}^{\infty} (y - \mu_Y)^2 p_X(y/2) \mathrm{d}y = \frac{1}{2} \int_{-\infty}^{\infty} (2x - 2\mu_X)^2 p_X(x) (2\mathrm{d}x) = 2^2 var(X).$$

2.8.4. Cauchy Distribution. Let X be a rv with the Cauchy (or Lorentzian) distribution. Its PDF is defined as:

$$\boxed{p_X(x) = \frac{1}{\pi} \frac{1}{(1 + x^2)}.}$$

The CDF is:

$$\mathbb{P}(X < x) = \int_{-\infty}^{x} \frac{1}{\pi} \frac{1}{(1+x^2)} \mathrm{d}x.$$

We know from calculus that the derivative of $\tan^{-1}(x)$ is $1/(1+x^2)$. Therefore, the last expression can be integrated:

$$\boxed{\mathbb{P}(X < x) = \frac{1}{\pi} \tan^{-1}(x) + \frac{1}{2}.}$$

2.8.5. Rayleigh Distribution. Let X be a rv with Rayleigh distribution ($X \sim$ Rayleigh). The Rayleigh distribution has a PDF:

$$\boxed{p(x) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)},}$$

where $x \geq 0$ and σ is a parameter of the distribution. The CDF can be shown to be:

$$\boxed{\mathbb{P}(X < x) = 1 - e^{-x^2/(2\sigma^2)},}$$

where $x \geq 0$. The reader can check that the mean of X is $\sigma\sqrt{\pi/2}$ and its variance is $\sigma^2 \frac{4-\pi}{2}$.

2.8.6. Gaussian (Normal) Distribution. The normal distribution $\mathcal{N}(\mu, \sigma^2)$ with parameters μ and σ^2 is defined by the density:

$$\boxed{p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},}$$

where $x \in (-\infty, \infty)$. The prefactor $\frac{1}{\sqrt{2\pi\sigma^2}}$ is such that $p(x)$ adds up to 1:

$$\int_{-\infty}^{\infty} p(x) \mathrm{d}x = 1.$$

This can be verified using the well-known result for a Gaussian integral $\int_{-\infty}^{\infty} e^{-ax^2} \mathrm{d}x = \sqrt{\pi/a}$, where $a > 0$. As an exercise, you should check that this PDF is normalized to 1, the mathematical expectation of $X \sim \mathcal{N}(\mu, \sigma^2)$ is μ and its variance is σ^2, i.e. $\mathbb{E}X = \mu$ and $\mathbb{E}(X-\mu)^2 = \sigma^2$.

This probability density is plotted below. It is centered on μ and the width is proportional[5] to σ.

If a rv X follows a Gaussian distribution (Fig. 2.1) with mean μ and variance σ^2 we write $X \sim \mathcal{N}(\mu, \sigma^2)$. For a Gaussian distribution, the CDF is called the *error function*. See Figure 2.2.

[5] In fact, the full width at half maximum of the Gaussian is $2\sqrt{2\log 2}\sigma \approx 2.355\sigma$. You can check this by finding the values of x for which $\frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$ or $\frac{1}{2} = e^{-\frac{x^2}{2\sigma^2}}$, since the maximum of the function is $\frac{1}{\sqrt{2\pi\sigma^2}}$ (set $x = 0$). Taking logs of both sides gives $x = \pm\sqrt{2\sigma^2 \log 2}$.

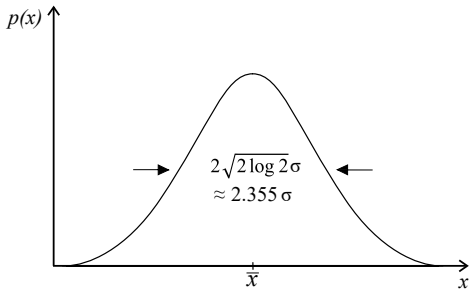


Figure 2.1. Gaussian (bell shaped) distribution. The PDF has full width at half-maximum of approximately 2.355σ.

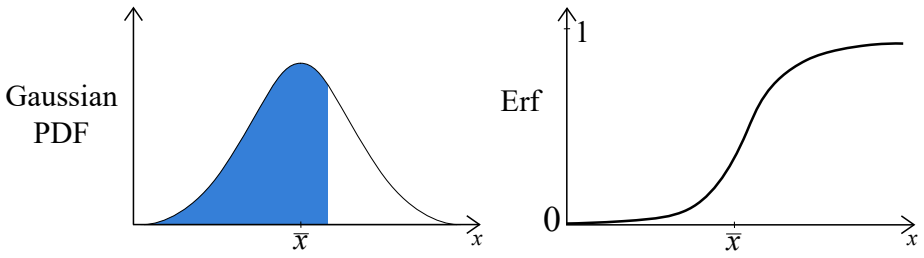


Figure 2.2. Error function is defined as the (cumulative) area under the Gaussian PDF.

$$\boxed{\operatorname{erf}(x)_{\mu,\sigma} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x.} \tag{2.2}$$

The error function is an integral which cannot be evaluated analytically. Instead it must be solved numerically. Values of the error function can be obtained from tables, calculators or computer programs. The error function for standard normal rv (mean 0, variance 1) is often tabulated in books. In MATLAB the command `normcdf(x,mu,sigma)` will return values for $\operatorname{erf}(x)_{\mu,\sigma}$. See Section 2.9 for a discussion of the error function.

2.8.7. Log-Normal Distribution. In Section 2.8.6, we have introduced the error function as the CDF of the Gaussian PDF (Eq. 2.2):

$$\operatorname{erf}(x)_{\mu,\sigma} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-(x-\mu)^2/2\sigma^2} \mathrm{d}x.$$

If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = e^X$, then

$$\mathbb{P}(Y < y) = \mathbb{P}(e^X < y) = \mathbb{P}(X < \log y),$$

which leads to the CDF:

$$\boxed{\mathbb{P}(Y<y)=\int_{-\infty}^{\log y}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}\mathrm{d}x=\mathrm{erf}(\log y)_{\mu,\sigma}.} \tag{2.3}$$

This is called the log normal distribution ($Y\sim$ log-normal). You can check, using the Leibniz formula (see Section 12.3) for differentiation, that the PDF of the log-normal distribution is:

$$\boxed{p_Y(y)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(\log(y)-\mu)^2/2\sigma^2}\cdot\frac{1}{y}.}$$

2.9. Tabulated Values of Error Function

It is important to be able to use tabulated values of the error function. Let $X\sim\mathcal{N}(\mu,\sigma^2)$. Then,

$$\begin{aligned}\phi_{\mu,\sigma}(x)\equiv\mathrm{erf}(x)_{\mu,\sigma}=\mathbb{P}(X<x)&=\underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{x}e^{-(\xi-\mu)^2/(2\sigma^2)}\mathrm{d}\xi}_{\text{let }\zeta=(\xi-\mu)/\sigma,\,\mathrm{d}\zeta=\mathrm{d}\xi/\sigma}\\&=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{(x-\mu)/\sigma}e^{-\zeta^2/2}\mathrm{d}\zeta=\Phi\left(\frac{x-\mu}{\sigma}\right),\end{aligned}$$

where $\Phi(\cdot)$ denotes the *normalized error function*:

$$\boxed{\Phi(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\zeta^2/2}\mathrm{d}\zeta.}$$

The latter is the CDF of the standard normal distribution, $\mathcal{N}(0,1)$. z is known as the z-score:

$$z=\frac{x-\mu}{\sigma}. \tag{2.4}$$

As an example, Eq. (2.3) can be expressed in this notation as:

$$\mathrm{erf}(\log y)_{\mu,\sigma}=\Phi\left(\frac{\log y-\mu}{\sigma}\right).$$

You should beware that there exist other conventions for the error function. For example, MATLAB and EXCEL softwares define the error function as:

$$\mathrm{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\mathrm{d}t. \tag{2.5}$$

This is related to the normal CDF according to:

$$\phi_{\mu,\sigma}(x)=\mathrm{erf}(x)_{\mu,\sigma}=\Phi\left(\frac{x-\mu}{\sigma}\right)=\frac{1}{2}\left(1+\mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right).$$

This expression can be used to calculate $\text{erf}(x)_{\mu,\sigma}$ using data from tables, where μ is the mean of the population and σ is the standard deviation of the population. First, we form the "z-score" (Eq. 2.4). Then we use tabulated values of the error function for this particular value of z.

For example, suppose that a manufacturer produces electrical resistors whose nominal value is $(100\pm2)\ \Omega$, where $2\ \Omega$ is the standard deviation (both could be estimated, for example, using sample mean and sample variance). Assuming that the distribution of the resistance X is Gaussian (i.e. $X \sim \mathcal{N}(100, 2^2)$), what is the probability that choosing a resistor at random will yield a resistance of $95\ \Omega$ or less? We want to show that

$$\mathbb{P}(X \leq 95\ \Omega) = \text{Erf}(95)_{100,2} \approx 0.0062.$$

Method 1 uses MATLAB:

```
>> normcdf(95,100,2)

ans =

    0.0062
```

Method 2 uses tabulated values of $\Phi(z)$: The z-score is:

$$z = \frac{x-\mu}{\sigma} = \frac{95-100}{2} = -2.5,$$

which is negative. Unfortunately, tables of error function do not list negative z values. However, notice that negative z values can be obtained from positive ones:

$$\boxed{\Phi(-z) = 1 - \Phi(z).}$$

Here, for positive $z = 2.5$ the value $\Phi(2.5)$ is 0.993790. Taking $1 - \Phi(2.5)$ gives 0.00621, the result we sought. Most books on statistics will have such a table. Tables can also be generated in MATLAB by typing:

```
normcdf(linspace(0,3,50)',0,1)
```

The results $\{(x, \Phi(x))\}$, $x \in [0, 3]$ are shown in Table 2.1.

2.10. The z-score

Let's view the z-score as a random variable:

$$Z(\omega) = \frac{X(\omega) - \mu}{\sigma}$$

where $\mu \equiv \mathbb{E}[X]$, $\sigma \equiv \sqrt{var(X)}$ and $Z \sim \mathcal{N}(0, 1)$. The statement that $Z \sim \mathcal{N}(0, 1)$ follows automatically when X is normal with mean μ and

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0	0.5000	0.5056	0.6934	1.0112	0.8440	1.5169	0.9353	2.0225	0.9784	2.5281	0.9943
0.0337	0.5134	0.5393	0.7052	1.0449	0.8520	1.5506	0.9395	2.0562	0.9801	2.5618	0.9948
0.0674	0.5269	0.5730	0.7167	1.0787	0.8596	1.5843	0.9434	2.0899	0.9817	2.5955	0.9953
0.1011	0.5403	0.6067	0.7280	1.1124	0.8670	1.6180	0.9472	2.1236	0.9831	2.6292	0.9957
0.1348	0.5536	0.6404	0.7391	1.1461	0.8741	1.6517	0.9507	2.1573	0.9845	2.6629	0.9961
0.1685	0.5669	0.6742	0.7499	1.1798	0.8810	1.6854	0.9540	2.1910	0.9858	2.6966	0.9965
0.2022	0.5801	0.7079	0.7605	1.2135	0.8875	1.7191	0.9572	2.2247	0.9869	2.7303	0.9968
0.2360	0.5933	0.7416	0.7708	1.2472	0.8938	1.7528	0.9602	2.2584	0.9880	2.7640	0.9971
0.2697	0.6063	0.7753	0.7809	1.2809	0.8999	1.7865	0.9630	2.2921	0.9891	2.7978	0.9974
0.3034	0.6192	0.8090	0.7907	1.3146	0.9057	1.8202	0.9656	2.3258	0.9900	2.8315	0.9977
0.3371	0.6320	0.8427	0.8003	1.3483	0.9112	1.8539	0.9681	2.3596	0.9909	2.8652	0.9979
0.3708	0.6446	0.8764	0.8096	1.3820	0.9165	1.8876	0.9705	2.3933	0.9917	2.8989	0.9981
0.4045	0.6571	0.9101	0.8186	1.4157	0.9216	1.9213	0.9727	2.4270	0.9924	2.9326	0.9983
0.4382	0.6694	0.9438	0.8274	1.4494	0.9264	1.9551	0.9747	2.4607	0.9931	2.9663	0.9985
0.4719	0.6815	0.9775	0.8358	1.4831	0.9310	1.9888	0.9766	2.4944	0.9937	3.0000	0.9987

Table 2.1. Numerical values of the error function $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2} \mathrm{d}x$.

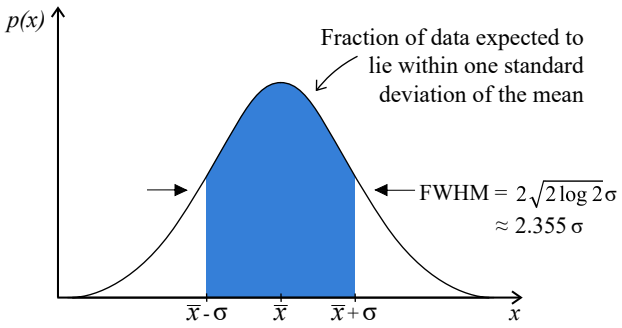


Figure 2.3. Confidence limits. The Gaussian distribution has a full width at half-maximum of approximately 2.355σ.

variance σ^2. In that case,

$$\mathbb{P}(X \le x) = \mathbb{P}(\sigma Z + \mu \le x) = \mathbb{P}(Z \le \frac{x-\mu}{\sigma}) \tag{2.6}$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{-\zeta^2/2} \mathrm{d}\zeta = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

This is result identical to the one in the previous section, but its derivation did not require us to change variables of integration. Using the probability function $\mathbb{P}(\cdot)$ can sometimes save you a step.

2.11. Confidence Limits and Error Bars

Recall the Gaussian probability density which has a bell shape centered on $\mathbb{E}[X] = \mu_X$ and full width at half-maximum $\approx 2.355\sigma$ (Fig. 2.3). The area

under the curve bounded by the interval $x \in [\mu_X - \sigma, \mu_X + \sigma]$ is given by:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mu_X - \sigma}^{\mu_X + \sigma} e^{-\frac{(x-\mu_X)^2}{2\sigma^2}} \mathrm{d}x = \mathrm{erf}(\mu_X + \sigma)_{\mu_X,\sigma} - \mathrm{erf}(\mu_X - \sigma)_{\mu_X,\sigma} \approx 0.683$$

where

$$\mathrm{erf}(x)_{\mu,\sigma} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x.$$

About 2/3 of the total area under the curve is within $\pm\sigma$ of the mean. Recall that:

$$(\text{value of } x) = x_{best} \pm \delta x$$

where we often take $\delta x = \sigma$. This choice for δx corresponds to the "68% confidence level". This means that we are confident, at the 68% level, that were we to take another measurement, the value would lie within one standard deviation of the mean. There are other possible conventions for choosing δx. Common choices for δx are:

$$\begin{aligned} \pm\sigma &\rightarrow 68\% \text{ level} \\ \pm 2\sigma &\rightarrow 95\% \text{ level} \\ \pm 3\sigma &\rightarrow 99.7\% \text{ level} \end{aligned}$$

2.11.1. Example: From CDF to PDF. It is important to be able to convert from PDF to CDF and vice versa. Suppose that we have a CDF:

$$\mathbb{P}(Y < a) = \int_{-\infty}^{a} \frac{1}{\pi} \frac{\mathrm{d}y}{(1+y^2)} \tag{2.7}$$

To get the PDF from this CDF we use the formula

$$\frac{\mathrm{d}\mathbb{P}(Y \leq a)}{\mathrm{d}a} = p_Y(a).$$

The result is:

$$p_Y(y) = \frac{1}{\pi} \frac{1}{(1+y^2)},$$

(We renamed a as y.) The differentiation is always with respect to the upper bound of the integral. Another way to look at it is to write $F(x) = \mathbb{P}(X \leq x)$ and

$$\frac{\mathrm{d}F(x)}{\mathrm{d}x} = p_X(x) \quad \text{or} \quad \frac{\mathrm{d}F(a)}{\mathrm{d}a} = p_X(a).$$

Inspection of the Leibniz formula (see Section 12.3) for differentiation shows that the differentiation step is completely trivial and amounts to simply identifying the integrand. This is consistent with the definition of CDF:

$$\mathbb{P}(Y < a) = \int_{-\infty}^{a} p_Y(y)\mathrm{d}y. \tag{2.8}$$

2.12. Calculating Probabilities: Single Variable

Probabilities of random events of the type $\{a \leq X \leq b\}$ are calculated by integrating the PDF from a to b:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b p(x)\mathrm{d}x.$$

More generally, we deal with *random events*, which are statements of the type $\{X \in A\}$ where A is a set of points. The quantity $\mathbb{P}(X \in A)$ is a number between 0 and 1, which gives the probability that the rv X will take values in the set A:

$$\mathbb{P}(X \in A) = \int_{\{x|x\in A\}} p(x)\mathrm{d}x,$$

where $\{x|x \in A\}$ is the set of points x that belong to the set A. The integral is a Riemann summation over the set of points $\{x|x \in A\}$ on the real line. This notation is useful because we can transform the statement $\{x \in A\}$ into any equivalent statement, including one that involves a change of variables. For example, the two following statements are equivalent:

$$\{X < a\} = \{\log(X) < \log(a)\}.$$

This is useful if another rv Y is defined as $Y = \log(X)$. In that case, evaluating the probability of $\{Y < b\}$, $b = \log(a)$, gives the same numerical result as evaluating the probability of $\{X < a\}$.

2.12.1. Average of $f(X)$. The average (or mean, or expectation value) of a function f of a rv X is defined as:

$$\boxed{\mathbb{E}[f(X)] \equiv \int_{-\infty}^{\infty} f(x)p(x)\mathrm{d}x.}$$

2.13. Statistical Moments, Deviation and Dispersion

2.13.1. Moments: Mean, Variance, Skewness, Kurtosis. Let X be a rv. Take $f(x) = x^n$ in the above formula. This gives the *n-th moment* of X:

$$\mathbb{E}[X^n] \equiv \int_{-\infty}^{\infty} p(x)x^n\mathrm{d}x.$$

The case $n = 1$ (*first moment*) is called the mathematical expectation or mean value of X:

$$\mathbb{E}[X] \equiv \int_{-\infty}^{\infty} xp(x)\mathrm{d}x. \qquad \text{“mathematical expectation”}$$

We assumed that X takes values in the range $(-\infty, \infty)$. When X takes values in a set $\mathcal{X}$ the limits of integration in all our integrals must be changed

accordingly:

$$\boxed{\mathbb{E}[X] \equiv \int_{\mathcal{X}} xp(x)\mathrm{d}x.}$$

We shall often write as shorthand:

$$\mu_X \equiv \mathbb{E}[X].$$

As we have seen in the previous lecture, the *variance* of X is defined as:

$$\sigma^2 \equiv \int_{-\infty}^{\infty} p(x)(x-\mu_X)^2\mathrm{d}x,$$

which also equals $\sigma^2 = \mathbb{E}[X^2] - \mu_X^2$. Thus, the variance is the second moment of X minus the square of the average of X. Variance is also known as the second *central* moment of X.

The n-th *central moment* of X is defined as:

$$\mathbb{E}[X-\mu_X]^n \equiv \int_{-\infty}^{\infty} p(x)(x-\mu_X)^n\mathrm{d}x.$$

Why are moments important? Moments describe the probability distribution. There is a theorem of mathematics that says if we know the moments of all orders, we can reconstruct the entire distribution function. You already know how to obtain the sample mean and variance. The mean is just the center of mass of the distribution whereas the variance is related to its width (about the mean).

Also of interest are the skewness (3rd central moment)

$$Skew[X] = \frac{\mathbb{E}[X-\mu_X]^3}{[\mathbb{E}(X-\mu_X)^2]^{3/2}} = \frac{\mathbb{E}[X-\mu_X]^3}{\sigma^3},$$

and the kurtosis (4th central moment):

$$Kurt[X] = \frac{\mathbb{E}[X-\mu_X]^4}{[\mathbb{E}(X-\mu_X)^2]^2} = \frac{\mathbb{E}[X-\mu_X]^4}{\sigma^4}.$$

The skewness measures the asymmetry of the distribution about its mean whereas the kurtosis is often used to assess by how much a distribution deviates from the bell-shape. For example, if a distribution looks like a bell shape but has much longer tails, the kurtosis will reflect this.

2.13.2. Median, Percentile. The median of a rv X is the value of x_{50} such that

$$\mathbb{P}(X \geq x_{50}) = \mathbb{P}(X \leq x_{50}) \equiv \int_{-\infty}^{x_{50}} p(x)\mathrm{d}x = \frac{1}{2}.$$

The median is a special case of *percentile*. The 10-th percentile of X is the value of x_{10} such that:

$$\mathbb{P}(X \leq x_{10}) \equiv \int_{-\infty}^{x_{10}} p(x)\mathrm{d}x = 0.10.$$

The n-th percentile of X is the value x_n such that:

$$\boxed{\mathbb{P}(X \leq x_n) \equiv \int_{-\infty}^{x_n} p(x)\mathrm{d}x = \frac{n}{100}.}$$

2.13.3. Mode. The mode is the value that appears most often in a set of data values. If X is a discrete rv, the mode is the value that is most likely to be sampled. For example, in a sample $\{1, 1, 6, 7, 5, 9, 10, 1\}$ the mode is 1. In a sample $\{1, 1, 6, 5, 7, 7\}$ there are two modes: 1 and 7. A distribution with more than one mode is called multimodal. The most extreme case of a multimodal distribution occurs for uniform distributions, where all values occur equally often. This definition can be adapted for continuous rv by discretizing the PDF into a histogram and finding the value(s) of x for which the histogram is highest.

Another definition of mode for continuous distribution is the set of local maxima. When the PDF of a continuous distribution has multiple local maxima those are called the modes of the distribution (any peak is a mode). It may be tempting to define the mode of a PDF $p(x)$ as the set of points x for which $\mathrm{d}p(x)/\mathrm{d}x = 0$; however, this method does not always work. There are shapes of PDFs that have a mode, but at which the derivative of the PDF is not zero. The Laplace distribution being an obvious example:

$$p(x) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right).$$

The mode is μ but the derivative at μ does not exist since the derivative of $|x|$ does not exist at $x = 0$. We note that a continuous rv's mode is not the value of X most likely to occur, as was the case for discrete rv. Furthermore, for some densities, even when the derivative is 0, it doesn't imply there's a mode there. Consider the beta density as an example, where setting $p'(x) = 0$ will find a local minimum rather than a maximum.

2.13.4. Average Absolute Deviation (AAD). We have seen that the center of a distribution can be quantified by the mathematical expectation (mean), the mode and the median. There are likewise many possible descriptors of the dispersion of a rv. The variance is one example. Another example is the average absolute deviation (AAD). AAD of a data set is the average of the absolute deviations from a central point. The central point can be a mean, median, mode or any other point of reference. The two most

common AADs are the mean absolute deviation and the median absolute deviation (MAD).

Let X be a rv. The mean absolute deviation of a random sample $\{x_i = X(\omega_i)\}_{i=1}^n$ of X is

$$MAD(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} |x_i - m(X)|,$$

where $m(X)$ is the reference value (typically, the mean or median). This arithmetic average provides an estimate of the expectation value $\mathbb{E}|X - m(X)|$ given a random sample. The median absolute deviation is defined similarly, except that we compute the median of $|X - m(X)|$ instead of its mean.

More generally, a dispersion can be defined by

$$\mathcal{D}_p(x_1, \ldots, x_n) = \sqrt[p]{\mathbb{E}|X - m(X)|^p} \approx \sqrt[p]{\frac{1}{n}\sum_{i=1}^{n} |x_i - m(X)|^p},$$

where $p = 0, 1, 2, \ldots, \infty$ and $\{x_1, \ldots, x_n\}$ is a random sample of X. For $p = \infty$ we get the maximum absolute deviation. For $p = 1$ we get the average absolute deviation. For $p = 2$ we get the mean squared error.

2.13.5. Remark: $\langle f(X)\rangle$ is NOT the Same as $f(\langle X\rangle)$.

2.13.5.1. *Example 1:* Suppose that the kinetic energy, $K(v) = \frac{1}{2}mv^2$, of an object of mass m is to be calculated using experimentally measured values of the velocity v. Thus, v is a rv. Since v is a random variable, $K(v)$ is also a random variable. We may denote it as V. You determine from experiments that the velocities, V, are Gaussian-distributed around 100 m/s, with a standard deviation of 1 m/s, i.e. $p(v) = \frac{1}{\sqrt{2\pi}}e^{-(v-100)^2/2}$. What is the average kinetic energy, $\mathbb{E}[K(V)]$? You expect that $\mathbb{E}[K(V)]$ should be close to $K(100) = \frac{1}{2}m(100)^2$. However, the exact value of $\mathbb{E}[K(V)]$ will depend on the distribution $p(v)$. We need to calculate:

$$\begin{aligned}\mathbb{E}[K(V)] =& \int_{-\infty}^{\infty} \frac{1}{2}mv^2 \frac{1}{\sqrt{2\pi}} e^{-(v-100)^2/2} \mathrm{d}v \\ =& \frac{m}{2\sqrt{2\pi}} \int_{-\infty}^{\infty} v^2 e^{-(v-100)^2/2} \mathrm{d}v \approx \frac{m}{2} 1.0001 \cdot 10^4.\end{aligned}$$

So fairly close to $K(100)$ but slightly higher.

2.13.5.2. *Example 2:* Suppose that $f(\theta) = \cos(\theta)$ and $p(\theta) = 1$ for $\theta \in [-\frac{1}{2}, \frac{1}{2}]$ and $p = 0$ elsewhere (uniform distribution). Denote the random variable as Θ. Using $p(\theta)$ you can easily check that $\overline{\Theta} = \int_{-1/2}^{1/2} \theta \, \mathrm{d}\theta = 0$. The

average of $\cos(\Theta)$ is:

$$\overline{\cos(\Theta)} = \int_{-1/2}^{1/2} \cos(\theta)\mathrm{d}\theta \approx 0.9589.$$

Note: it is not equal to 1 even though the average of Θ is 0.

2.13.5.3. *Example 3:* In physics and chemistry, the notations $\langle r\rangle$ and $\overline{r}$ are used interchangeably to denote the mathematical expectation. Consider the dipole-dipole interaction between two electric dipoles. The energy of interaction depends on $1/r^3$, where r is the distance separating the two point dipoles. It is easy to show that in general, $\left\langle \frac{1}{r^3}\right\rangle \neq \frac{1}{\langle r\rangle^3}$. r is a rv due to molecular diffusion. It has an average value $\overline{r}$ and deviation δr:

$$r = \overline{r} + \delta r,$$

where $\overline{r}$ is deterministic and δr is random with zero mean. The average of $1/r^3$ is Taylor-expanded about the mean $\overline{r}$:

$$\left\langle \frac{1}{r^3}\right\rangle = \left\langle \frac{1}{(\overline{r}+\delta r)^3}\right\rangle = \left\langle \frac{1}{\overline{r}^3}\right\rangle - \left\langle \frac{3}{\overline{r}^4}\delta r\right\rangle + \left\langle \frac{12}{\overline{r}^5}(\delta r)^2\right\rangle + O(|\delta r|^3)$$

The first term is $1/\overline{r}^3$ since $\overline{r}$ is deterministic. In the second term, $\frac{3}{\overline{r}^4}$ can come out of the angle bracket because it is a deterministic quantity. Similarly for $\frac{12}{\overline{r}^5}$ in the third term. Thus,

$$\left\langle \frac{1}{r^3}\right\rangle = \frac{1}{\langle r\rangle^3} - \frac{3}{\langle r\rangle^4}\langle\delta r\rangle + \frac{12}{\langle r\rangle^5}\langle(\delta r)^2\rangle + O(|\delta r|^3)$$

and you can see that $\left\langle \frac{1}{r^3}\right\rangle$ is in general different from $\frac{1}{\langle r\rangle^3}$. Since $\langle\delta r\rangle = 0$ we have:[6]

$$\left\langle \frac{1}{r^3}\right\rangle = \frac{1}{\langle r\rangle^3} + \underbrace{\frac{12}{\langle r\rangle^5}\langle(\delta r)^2\rangle + O(|\delta r|^3)}_{\text{extra terms (nonzero)}}.$$

We sometimes see in the literature $\frac{1}{\langle r\rangle^3}$ in lieu of $\left\langle \frac{1}{r^3}\right\rangle$. This is technically incorrect. However, for small values of $\langle|\delta r|\rangle/\langle r\rangle$, it is a good approximation.

2.13.6. Jensen's Inequality. A topic related to the previous section is Jensen's inequality. Let $\varphi(x)$ be a convex function, i.e.

$$\varphi(tx_1 + (1-t)x_2) \leq t\varphi(x_1) + (1-t)\varphi(x_2), \qquad x_1 < x_2, \quad t \in [0,1]$$

This can be generalized for $\lambda_1 + \cdots + \lambda_n = 1$, $\lambda_i \geq 0$ as:

$$\varphi(\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n) \leq \lambda_1\varphi(x_1) + \lambda_2\varphi(x_2) + \cdots + \lambda_n\varphi(x_n),$$

for any $x_1, \ldots, x_n$. Let X be a rv. Then,

$$\boxed{\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]}$$

[6] For our definition $r = \overline{r} + \delta r$ to hold, we need $\langle\delta r\rangle = 0$.

Proof:

$$\varphi(\mathbb{E}[X]) = \varphi\left(\int_0^1 xp(x)\mathrm{d}x\right) = \lim_{n\to\infty} \varphi\left(\sum_{j=0}^{2^n} 2^{-n}(j\cdot 2^{-n}p(j\cdot 2^{-n}))\right)$$
$$\leq \lim_{n\to\infty} \sum_{j=0}^{2^n} 2^{-n}\varphi\left(j\cdot 2^{-n}p(j\cdot 2^{-n})\right) = \int_0^1 \varphi(xp(x))\mathrm{d}x = \mathbb{E}[\varphi(X)].$$

As an example application of this inequality we have:

$$(\mathbb{E}[|X-\mu_X|])^2 \leq \mathbb{E}[|X-\mu_X|^2] = var(X).$$

Taking the square root of both sides:

$$\mathbb{E}[|X-\mu_X|] \leq \sqrt{var(X)}.$$

We conclude that the mean absolute deviation from the mean is less than or equal to the standard deviation.

2.13.7. Remark: Discrete Random Variables as Special Case. Suppose that we roll a die and a rv X (i.e. value of the top face of die) takes values in a discrete set, such as $\mathcal{X} = \{1,2,3,4,5,6\}$. It is said to be a discrete rv because the set $\mathcal{X}$ is countable. In the general case X may take values in a discrete set $\{x_1,\ldots,x_N\}$. Let $p_i \geq 0$ be the probability of observing the value x_i. Define the PDF in terms of Dirac delta functions and PMF $\{p_i\}_{i=1}^N$:

$$p(x) = \sum_{i=1}^N p_i\delta(x-x_i)$$

Since the PDF is normalized, we must have:

$$\int_{-\infty}^{\infty} p(x)\mathrm{d}x = \int_{-\infty}^{\infty}\sum_{i=1}^N p_i\delta(x-x_i)\mathrm{d}x = \sum_{i=1}^N p_i = 1.$$

All of our previous definitions hold if we replace integrals by summations. For example:

$$\mu_X \equiv \mathbb{E}[X] = \int_{-\infty}^{\infty} x\cdot p(x)\mathrm{d}x = \int_{-\infty}^{\infty} x\cdot\sum_{i=1}^N p_i\delta(x-x_i)\mathrm{d}x = \sum_{i=1}^N p_i x_i.$$

The variance:

$$var(X) = \int_{-\infty}^{\infty}(x-\mu_X)^2 p(x)\mathrm{d}x = \sum_{i=1}^N p_i(x_i-\mu_X)^2.$$

Similarly,

$$\mathbb{E}f(X) = \int_{-\infty}^{\infty} f(x)\cdot p(x)\mathrm{d}x = \sum_{i=1}^N p_i f(x_i).$$

Here, $x_i \in \mathcal{X}$ are the possible values X can take, whereas $p_i \equiv \mathbb{P}(X = x_i)$ are the corresponding probabilities.

2.14. Two (Continuous) Random Variables

If we are to compute the average of an expression that is a function of more than one rv, we need to use the *joint probability density* $p_{XY}(x, y)$, which is everywhere non-negative ($p_{XY}(x, y) \geq 0$) and integrates to 1:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{XY}(x, y)\mathrm{d}x\mathrm{d}y = 1.$$

The joint PDF is obtained from the joint CDF analogously to the single-variable case:

$$p_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y}\mathbb{P}(X < x, Y < y). \tag{2.9}$$

The average of a function $g(X, Y)$ would be:

$$\mathbb{E}(g(X, Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{XY}(x, y)g(x, y)\mathrm{d}x\mathrm{d}y.$$

Given a joint probability density function, $p_{XY}(x, y)$, the *marginal density* function for X is obtained by integrating over y:

$$\boxed{p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x, y)\mathrm{d}y.}$$

Similarly, the marginal density for Y is obtained by integrating over all x:

$$\boxed{p_Y(y) = \int_{-\infty}^{\infty} p_{XY}(x, y)\mathrm{d}x.}$$

Note: you can easily check that both marginals $p_X(x)$ and $p_Y(y)$ are bona fide densities, i.e. nonnegative and normalized to 1.

2.15. Statistical Independence

The marginal density is a useful concept if you are asked to check whether or not two rv are statistically independent. Two rv X and Y are *statistically independent* if the joint probability density is written as the product of densities of each variable:

$$\boxed{p_{XY}(x, y) = p_X(x) \cdot p_Y(y),}$$

where $p_X(x)$ and $p_Y(y)$ are the marginal densities of X and Y, respectively. They can be computed from $p_{XY}(x, y)$ by integrating.

There are at least two consequences of statistical independence that we can immediately point out. First, one concerns expectation values. Consider

the average of a function $g(X,Y)$ of two rv X and Y:

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{XY}(x,y)g(x,y)\mathrm{d}x\mathrm{d}y$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_X(x)p_Y(y)g(x,y)\mathrm{d}x\mathrm{d}y.$$

If g factors as a product of a function of X times a function of Y, for example $g(X,Y) = XY$ then the expectation value of XY is equal to the product of expectation values of X and that of Y:

$$\mathbb{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_X(x)p_Y(y)xy\,\mathrm{d}x\mathrm{d}y = \int_{-\infty}^{\infty} p_X(x)x\mathrm{d}x \cdot \int_{-\infty}^{\infty} p_Y(y)y\mathrm{d}y$$
$$= \mathbb{E}(X)\cdot\mathbb{E}(Y).$$

Thus, the expectation value of a product of rv's factorizes as a product of expectation values for each rv:

$$\boxed{\mathbb{E}(XY) = \mathbb{E}(X)\cdot\mathbb{E}(Y).}$$

The second consequence pertains to the calculation of probabilities in general, which also factors as a product. For example, the joint probability:

$$\begin{aligned}\mathbb{P}(X\in A, Y\in B) &= \int_{\{(x,y)|x\in A, y\in B\}} p_{XY}(x,y)\mathrm{d}x\mathrm{d}y\\ &= \int_{\{(x,y)|x\in A, y\in B\}} p_X(x)p_Y(y)\mathrm{d}x\mathrm{d}y\\ &= \int_{\{x|x\in A\}} p_X(x)\mathrm{d}x \cdot \int_{\{y|y\in B\}} p_Y(y)\mathrm{d}y\\ &= \mathbb{P}(X\in A)\cdot\mathbb{P}(Y\in B).\end{aligned}$$

As a special case, take the intervals $A = (-\infty, x]$ and $B = (-\infty, y]$ and we get the result that the CDFs also factorize:

$$\boxed{\mathbb{P}(X\le x, Y\le y) = \mathbb{P}(X\le x)\cdot\mathbb{P}(Y\le y).}$$

2.16. Calculating Probabilities: Two Variables

Probabilities of an event A are calculated by integrating the PDF over the relevant set of points which make the event A true. That is, for a single rv X:

$$\boxed{\mathbb{P}(X\in A) = \int_{\{x|x\in A\}} p_X(x)\mathrm{d}x,} \tag{2.10}$$

where $\{x|x \in A\}$ denotes the set of all points x such that $x \in A$. For example, if $A = [a, b]$ (interval), we have:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b p_X(x)\mathrm{d}x.$$

If A is the union of two disjoint (non-overlapping) intervals $[a, b]$ and $[c, d]$, i.e. $A = [a, b] \cup [c, d]$, then the probability of X taking a value in A is the sum of two integrals:

$$\mathbb{P}(X \in A) = \int_a^b p_X(x)\mathrm{d}x + \int_c^d p_X(x)\mathrm{d}x.$$

For two or more rv's we integrate the joint probability density over all such points (x, y) that meet the sought criterion:

$$\mathbb{P}(X \in A, Y \in B) = \int\int_{\{(x,y)|x\in A, y\in B\}} p_{XY}(x, y)\mathrm{d}x\mathrm{d}y,$$

where $\{(x, y)|x \in A, y \in B\}$ denotes the set of all points (x, y) such that $x \in A$ and $y \in B$.

In general, for a given set of points over which the probability needs to be calculated, we must translate what this means in terms of the upper and lower limits of integration. Let us look at some specific examples. Let (X, Y) be a 2-dimensional (bivariate) rv with joint density $p_{XY}(x, y)$. The probability that the vector (X, Y) will lie in the first quadrant of the 2D plane is:

$$\begin{aligned}\mathbb{P}(X > 0, Y > 0) = \int\int_{\{(x,y)|x>0, y>0\}} p_{XY}(x, y)\mathrm{d}x\mathrm{d}y \\ = \int_0^\infty \int_0^\infty p_{XY}(x, y)\mathrm{d}x\mathrm{d}y.\end{aligned}$$

Suppose again that we have a random experiment involving two rv X and Y. The probability that the outcome will lie in one of the first two quadrants:

$$\begin{aligned}\mathbb{P}(X > 0) = \mathbb{P}(X > 0, Y \in (-\infty, \infty)) = \int\int_{\{(x,y)|x>0\}} p_{XY}(x, y)\mathrm{d}x\mathrm{d}y \\ = \int_{-\infty}^\infty \left(\int_0^\infty p_{XY}(x, y)\mathrm{d}x\right)\mathrm{d}y.\end{aligned}$$

2.16.1. Product of X and Y. Let X and Y be independent rv's and let $Z = XY$. The PDF of Z is:

$$\boxed{p_Z(z) = \int_{-\infty}^\infty p_X(x)p_Y(z/x)\frac{1}{|x|}\mathrm{d}x.}$$

Proof:

$$\begin{aligned}\mathbb{P}(Z\le z) =&\mathbb{P}(XY\le z) = \mathbb{P}(XY\le z, X>0) + \mathbb{P}(XY\le z, X\le 0)\\ =&\mathbb{P}(Y\le z/X, X>0) + \mathbb{P}(Y\ge z/X, X\le 0)\\ =&\int_0^\infty p_X(x)\int_{-\infty}^{z/x} p_Y(y)\mathrm{d}y\mathrm{d}x + \int_{-\infty}^0 p_X(x)\int_{z/x}^\infty p_Y(y)\mathrm{d}y\mathrm{d}x\end{aligned}$$

Differentiating with respect to z, we get the PDF:

$$\int_0^\infty p_X(x)p_Y(z/x)\frac{1}{x}\mathrm{d}x - \int_{-\infty}^0 p_X(x)p_Y(z/x)\frac{1}{x}\mathrm{d}x = \int_{-\infty}^\infty p_X(x)p_Y(z/x)\frac{1}{|x|}\mathrm{d}x.$$

See also Problem 26.

2.16.2. Sum of X and Y. Here is another application of the Leibniz formula (see Section 12.3). Suppose we have two rv's X, Y with joint density $p_{XY}(x,y)$. What is the density of their sum, $X+Y$? Since the density is the derivative of the CDF:

$$\begin{aligned}p_{X+Y}(z) = \frac{\mathrm{d}}{\mathrm{d}z}\mathbb{P}(X+Y<z) =&\frac{\mathrm{d}}{\mathrm{d}z}\int_{\{(x,y)|x+y<z\}} p_{XY}(x,y)\mathrm{d}x\mathrm{d}y\\ =&\frac{\mathrm{d}}{\mathrm{d}z}\int_{\{(x,y)|x<z-y\}} p_{XY}(x,y)\mathrm{d}x\mathrm{d}y\\ =&\int_{-\infty}^\infty \frac{\partial}{\partial z}\left(\int_{-\infty}^{z-y} p_{XY}(x,y)\mathrm{d}x\right)\mathrm{d}y\\ =&\int_{-\infty}^\infty p_{XY}(z-y,y)\mathrm{d}y.\end{aligned}$$

This is as far as we can go without further information about X, Y. If X and Y are independent, the joint PDF factorizes into a product, $p_{XY}(z-y,y) = p_X(z-y)\cdot p_Y(y)$, and the last operation becomes a *convolution*:

$$\boxed{p_{X+Y}(z) = \int_{-\infty}^\infty p_X(z-y)\cdot p_Y(y)\mathrm{d}y.}$$

Thus, the PDF of $Z = X+Y$ is the convolution of the PDFs of X and Y, whenever X and Y are statistically independent.

2.17. Several Variables

Suppose we have several rv's $X_1, X_2, \ldots, X_n$. Probabilistic expressions involving these rv's can be evaluated if we know the joint distribution:

$$\mathbb{P}(X_1<b_1,\ldots,X_n<b_n) = \int_{-\infty}^{b_1}\mathrm{d}x_1\cdots\int_{-\infty}^{b_n}\mathrm{d}x_n\, p_{X_1,\ldots,X_n}(x_1,\ldots,x_n),$$

where $p_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$ is the joint PDF of $X_1,\ldots,X_n$ and $\mathbb{P}(X_1<b_1,\ldots,X_n<b_n)$ is the joint CDF.

We can also ask about the probability of the following event:

$$\{X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n\}.$$

Then, using the joint PDF:

$$\mathbb{P}\{X_1 \in B_1, \ldots, X_n \in B_n\} = \int_{B_1} \mathrm{d}x_1 \cdots \int_{B_n} \mathrm{d}x_n \, p_{X_1,\ldots,X_n}(x_1, \ldots, x_n).$$

As in the single-variable case, we can write probabilities over intervals in terms of the CDFs. For example, let X, Y be two rv's and let $B_1 = (a_1, a_2)$ and $B_2 = (b_1, b_2)$. Then,

$$\begin{aligned}
\mathbb{P}(X \in B_1, Y \in B_2) =& \mathbb{P}(a_1 < X < a_2, b_1 < Y < b_2) \\
=& \int_{a_1}^{a_2} \mathrm{d}x \int_{b_1}^{b_2} \mathrm{d}y \, p_{XY}(x, y) \\
=& \int_{a_1}^{a_2} \mathrm{d}x \left(\frac{\partial}{\partial x} \mathbb{P}(X < x, Y < b_2) - \frac{\partial}{\partial x} \mathbb{P}(X < x, Y < b_1) \right) \\
=& \mathbb{P}(X < a_2, Y < b_2) - \mathbb{P}(X < a_1, Y < b_2) \\
& - \mathbb{P}(X < a_2, Y < b_1) + \mathbb{P}(X < a_1, Y < b_1).
\end{aligned}$$

We have made use of Eq. (2.9), i.e.

$$p_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y} \mathbb{P}(X < x, Y < y),$$

and invoked the fundamental theorem of calculus (twice).

2.18. Additional Properties of rv's

2.18.1. Linearity of the Expectation Value. Let X and Y be rv's and a, b constants. From the linearity of the expectation value operator:

$$\begin{aligned}
\mathbb{E}[aX + bY] =& \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (ax + by) p_{XY}(x, y) \mathrm{d}x \mathrm{d}y \\
=& a \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x p_{XY}(x, y) \mathrm{d}x \mathrm{d}y + b \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y p_{XY}(x, y) \mathrm{d}x \mathrm{d}y \\
=& a \int_{-\infty}^{\infty} x p_X(x) \mathrm{d}x + b \int_{-\infty}^{\infty} y p_Y(y) \mathrm{d}y \\
=& a\mathbb{E}[X] + b\mathbb{E}[Y],
\end{aligned}$$

where $p_{XY}(x, y)$ is the joint probability density of x and y.[7] $p_X(x)$ is the marginal density of X. Similarly for $p_Y(y)$. This can be generalized to any

[7] Note: while the exact form of $p_{XY}(x, y)$ may be unknown, its knowledge was not required to demonstrate linearity.

number of rv's, e.g. for $X = X_1 + X_2 + \cdots + X_n$ it also follows that

$$\boxed{\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n].}$$

It is also trivial to show that the same result holds in the case of discrete rv's.

2.18.2. Scaling Property of the Variance. From the definition of the variance of a rv X (let $\mu_X \equiv \mathbb{E}[X]$),

$$var(X) \equiv \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}[X^2] - 2\mathbb{E}[X\mu_X] + (\mu_X)^2 = \mathbb{E}[X^2] - \mu_X^2$$

we deduce that

$$\boxed{var(aX) = a^2 var(X).}$$

2.18.3. Variance of a Product of Independent Random Variables. Let v be the velocity of a particle and t the time variable. If both are statistically independent rv's, the mean displacement factors as a product of means, $\mathbb{E}[vt] = \mathbb{E}[v] \cdot \mathbb{E}[t]$, whereas its variance is

$$\begin{aligned} var(vt) =& \mathbb{E}[(vt)^2] - (\mathbb{E}[vt])^2 \\ =& \mathbb{E}[v^2 t^2] - (\mathbb{E}[v])^2 (\mathbb{E}[t])^2 \\ =& \mathbb{E}[v^2] \cdot \mathbb{E}[t^2] - (\mathbb{E}[v])^2 (\mathbb{E}[t])^2. \end{aligned}$$

Thus, by statistically independence, we can express the mean and variance of the displacement $x = vt$ in terms of the mean and variance of v and t.

2.18.4. Variance Between Pairs of Random Variables: The Covariance. The *covariance* of X and Y is defined as (let $\mu_X \equiv \mathbb{E}[X]$ and $\mu_Y \equiv \mathbb{E}[Y]$):

$$\boxed{cov(X, Y) \equiv \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X \cdot \mu_Y.}$$

We note that the covariance of two independent rv's is zero since $\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y] = \mu_X \cdot \mu_Y$.

2.18.5. Variance of the Sum of Two Random Variables. Using the covariance, we may write the variance of the sum $X + Y$ as

$$\begin{aligned} var(X + Y) =& \mathbb{E}[(X + Y - \mathbb{E}[X + Y])^2] \\ =& \mathbb{E}[(X - \mu_X)^2] + \mathbb{E}[(Y - \mu_Y)^2] + 2\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\ =& var(X) + var(Y) + 2cov(X, Y). \end{aligned} \tag{2.11}$$

If X and Y are statistically independent, $cov(X, Y) = 0$, and $var(X + Y) = var(X) + var(Y)$, i.e. the error in $X + Y$ is related to the errors in X and Y by adding the variances.

2.18.6. Corollary: Adding Experimental Errors. Suppose X and Y are independent rv's with standard deviations σ_X and σ_Y, respectively. Let $Z = X + Y$. Then the variances add quadratically:

$$\sigma_Z = \sqrt{\sigma_X^2 + \sigma_Y^2}.$$

This formula is often used in the analysis of laboratory experimental results. It is only valid in the case where X and Y are independent. How can you verify of X and Y are statistically independent?

2.18.7. Sample Covariance. In Formula (2.11) the covariance must be added to get the error in the sum of two rv's, $X + Y$. The covariance is zero if the two rv's are statistically independent. It is difficult to check for independence. However, it is easy to check for statistical correlation by computing the sample covariance. Suppose that the following pairs are measured experimentally $\{(x_i, y_i)\}_{i=1}^n$. This random sample is described by the empirical joint PDF:

$$\hat{p}_{XY}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)\delta(y - y_i).$$

Substitution into the definition of covariance:

$$\begin{aligned} cov(X, Y) \equiv & \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\ = & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)\delta(y - y_i) \mathrm{d}x\mathrm{d}y \\ = & \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_X) \cdot (y_i - \mu_Y) \end{aligned}$$

where μ_X and μ_Y are the means of X and Y, respectively. Since we have experimental data at our disposal, we take them to be sample means. This is normally adjusted by replacing $1/n$ by $1/(n-1)$ on the basis that a degree of freedom has been lost due to our use of experimental data to obtain statistical estimators for the means ($\hat{\mu}_X$ and $\hat{\mu}_Y$):

$$\boxed{cov_{n-1}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu}_X) \cdot (y_i - \hat{\mu}_Y).}$$

This formula provides us with an explicit prescription for computing the covariance of X and Y from experimental data. One may as well directly use Formula (2.11), since it enables us to determine the amount of covariance between them, and add its contribution to the error estimate, if needed.

2.18.8. Correlation Coefficient. A concept that is related to the covariance is the correlation coefficient:

$$\boxed{\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y},}$$

where $\sigma_X = \sqrt{var(X)}$ and $\sigma_Y = \sqrt{var(Y)}$. ρ takes values in the range $-1 \leq \rho(X,Y) \leq 1$. The correlation coefficient is a measure of linear dependence between X and Y. It is more useful than the covariance in the sense that ρ is a dimensionless quantity which is normalized to the magnitude of X and Y. A value of $\rho = 1$ means that X and Y are correlated. A value of $\rho = -1$ means they are anti-correlated. A value of $\rho = 0$ means they are uncorrelated. Please note that if X and Y are independent then $\rho = 0$. However, the converse is not true. That is, $\rho = 0$ does not always imply that X and Y are independent.

The reader can easily check[8] that if $Y = aX + b$ (a, b constants) we have $cov(X,Y) = a \cdot var(X)$ and

$$\rho(X,Y) = \frac{a}{|a|} = \begin{cases} 1 & a > 0 \\ -1 & a < 0 \end{cases}.$$

Thus, the correlation coefficient is a measure of linear dependence. In this example, $\rho = 1$ if $Y = aX + b$ and $a > 0$ (X and Y are correlated), whereas $\rho = -1$ if $a < 0$ (X and Y are anti-correlated). This result is independent of the magnitude of a; it only depends on its sign. For example, $Y = 0.00001 \cdot X$ and $Y = 10000 \cdot X$ both give $\rho = 1$.

Q: Can you find examples of rv's X and Y where ρ is *not* equal to -1, 0 or 1 but some intermediate value (say 0.5)? What is the meaning of a correlation coefficient that is not equal to 0 or 1?

2.18.9. Linear Correlation. Suppose that two random variables X and Y depend on each other linearly:

$$Y = a + bX.$$

The correlation coefficient becomes:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{cov(X, a+bX)}{\sigma_X \cdot \sigma_Y} = \frac{b \cdot cov(X,X)}{\sigma_X \cdot \sigma_Y} = \frac{b \cdot \sigma_X}{\sigma_Y}$$

Therefore, the slope b is related to the value of the correlation coefficient (as well as the variances of X and Y):

$$b = \rho(X,Y)\frac{\sigma_Y}{\sigma_X}.$$

[8] Start by the numerator: $cov(X,Y) = cov(X, aX+b) = \mathbb{E}[(X - \mu_X)(aX + b - \mu_{aX+b})]$, but since $\mu_{aX+b} = a\mu_X + b$, this reduces to $cov(X, aX+b) = \mathbb{E}[(X-\mu_X)(aX - a\mu_X)] = a\mathbb{E}[(X-\mu_X)^2] = a \cdot var(X)$.

2.18.10. Sample Correlation Coefficient. Let X, Y be rv's with mean μ_X and μ_Y, respectively. Let $x_1, x_2, \ldots, x_n$ be measurements of X. Similarly for Y. The correlation coefficient can be estimated from experimental data, $\{(x_i, y_i)\}_{i=1}^n$, using the *sample correlation coefficient*:

$$\boxed{r_{X,Y} = \frac{\frac{1}{n-1}\sum_{i=1}^n (x_i - \hat{\mu}_X)\cdot(y_i - \hat{\mu}_Y)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i - \hat{\mu}_X)^2}\cdot\sqrt{\frac{1}{n-1}\sum_{i=1}^n (y_i - \hat{\mu}_Y)^2}},}$$

where n is the number of data points and $\hat{\mu}_X$ is the sample mean

$$\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^n x_i,$$

and similarly for $\hat{\mu}_Y$. They are uncorrelated if $r_{X,Y} = 0$. You can also check for possible correlation between X and Y using a scatter plot. This is done by plotting the set of ordered pairs $\{(x_i, y_i)\}$ as points on the same graph.

2.18.11. Uncorrelated but not Independent. Let X and Y be rv's related by $Y = X^2$. Let μ_X be the mean of X, μ_{X^2} its second moment, etc. Clearly, these rv's are not independent of each other. However,

$$cov(X, Y) = cov(X, X^2) = \mathbb{E}[(X - \mu_X)(X^2 - \mu_{X^2})] = \mu_{X^3} - \mu_X\mu_{X^2}.$$

If the distribution of X is such that $\mu_{X^3} = \mu_X\mu_{X^2}$ (for example, if the mean and skewness are zero, which is the case for a zero-mean normal distribution), then $cov(X, Y) = 0$ and the rv's X and Y are uncorrelated even though they are clearly dependent on each other.

This can easily be illustrated in MATLAB. Let's create two plots. One for the equation $Y = X + \eta$ (linear case), where X and η are independent standard normal rv's:

```
1 X=randn([1 10000]);
2 Y=X+randn([1 10000]);
```

and one for $Y = X^2 + \eta$ (quadratic case), where X and η are independent standard normal rv's.

```
1 X=randn([1 10000]);
2 Y=X.^2+randn([1 10000]);
```

You can think of the linear case as in the familiar form $Y = a + bX$, but for the special case of $a = 0$, $b = 1$, and noise added (η). Same for the quadratic equation, it has noise added to it, as a way to simulate the outcome of a random experiment.

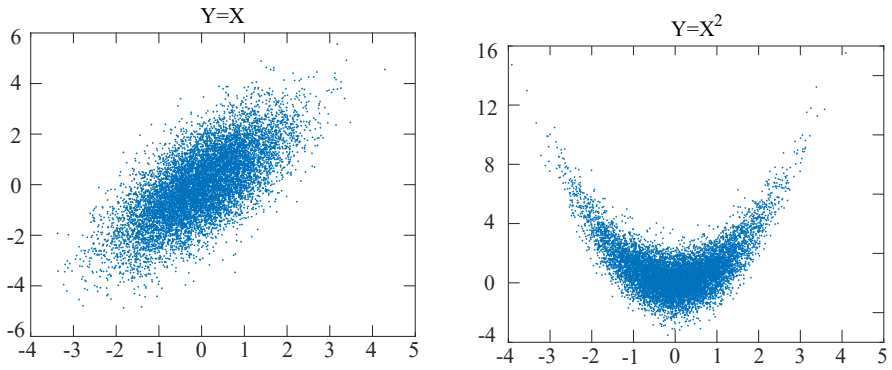


Figure 2.4. Correlations between rv's (X, Y).

In each case, a plot is generated by typing `figure;plot(X,Y,'.')` (see Fig. 2.4). The sample correlation coefficient is obtained using the `corrcoef` command, e.g. for the linear case

```
>> corrcoef(X,Y)

ans =

    1.0000    0.7025
    0.7025    1.0000
```

whereas for the quadratic case we get:

```
>> corrcoef(X,Y)

ans =

    1.0000   -0.0089
   -0.0089    1.0000
```

The diagonal elements are 1 (since X is perfectly correlated to X, as is Y correlated to Y). The off-diagonal elements are the correlation coefficients of X and Y. In the linear case, we have strong (≈ 0.7) correlation between X and Y. We would expect 1 without the additive noise, η (you can check this by reducing the amplitude of the additive noise). In the quadratic case, the correlation coefficient is nearly 0, as it should because the Gaussian rv X has zero skewness and zero mean.

2.19. Calculating Probabilities

If you are asked to compute the probability of a random event involving X, your first reflex should be to write down an integral of the PDF, $p_X(x)$,

over the set of points that represent this event. Recall the formula (2.12) introduced in the previous lecture:

$$\boxed{\mathbb{P}(X \in A) = \int_{\{x|x\in A\}} p_X(x)\mathrm{d}x,} \tag{2.12}$$

In two dimensions you do the same thing except that the joint PDF is to be used, e.g.

$$\begin{aligned}\mathbb{P}(a \le X \le b, c \le Y \le d) &= \int_{\{(x,y)|a\le x\le b, c\le y\le d\}} p_{XY}(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_a^b \int_c^d p_{XY}(x,y)\mathrm{d}x\mathrm{d}y.\end{aligned}$$

This is for the specific case where the random even is $\{a \le X \le b, c \le Y \le d\}$. For a general random event, we integrate over the set of points (x, y) such that the event is true. It is not possible to enumerate all possible events we may encounter. Some examples are:

$$\{X \in A, Y \in B\}, \quad \{X/Y < a\}, \quad \{X + Y > b\},$$
$$\{a < \cos(X) \cdot \log(Y) < b\}, \text{ etc.}$$

In each case, we integrate the joint PDF of X and Y, $p_{XY}(x, y)$, over the set of points (x, y) that obey the conditions specified in the event.

To summarize the procedure involved when calculating probabilities, there are two main steps. The first step is to write down the right hand side, but keeping in mind that the random event will need to be expressed in a form suitable for integration. The second step involves writing the integral in a form that can be solved. This sometimes involves a change of variables, if the integration region needs to be simplified.

2.19.1. Single-Variable Case. In the first step, we often invoke some algebraic manipulations in order to transform the logical statement $X \in A$ into a form that allows us to apply the information known to us. Let use revisit the example already covered in Sections 2.9 and 2.8.7. Let $Y = e^X$ and $X \sim \mathcal{N}(\mu, \sigma^2)$. You are asked to find the distribution of Y given the distribution X (a normal law in the present case). At first sight, you may think that $\mathbb{P}(Y < y)$ cannot be easily calculated because you are not given the distribution of Y. However, the distribution of X is provided. So your goal is to transform the statement $Y < y$ into one that involves X instead. In Section 2.8.7 we worked out the case of the log-normal distribution, $Y = e^X$ where $X \sim \mathcal{N}(\mu, \sigma^2)$.

2.19.2. Two Variables Case. Another example is $Y = U/V$ where U and V are independent standard normal variables, i.e $U \sim \mathcal{N}(0, 1)$ and

$V \sim \mathcal{N}(0,1)$. The CDF of Y is:

$$\mathbb{P}(Y<a) = \mathbb{P}(U/V<a) = \iint\limits_{\{(u,v)|u/v<a\}} p_U(u)p_V(v)\mathrm{d}u\mathrm{d}v$$

Effecting a change of variables $Y = U/V$, $Z = V$ (inverse: $V = Z$, $U = ZY$) under the integral sign and using the Jacobian of the transformation:

$$\mathrm{d}u\mathrm{d}v = \left|\frac{\partial(u,v)}{\partial(y,z)}\right| \mathrm{d}y\mathrm{d}z, \qquad \text{where } \frac{\partial(u,v)}{\partial(y,z)} \equiv \begin{Vmatrix} \partial_y u & \partial_z u \\ \partial_y v & \partial_z v \end{Vmatrix} = \begin{Vmatrix} z & y \\ 0 & 1 \end{Vmatrix} = z$$

where $\|\cdot\|$ denotes "matrix determinant", whereas $|\cdot|$ denotes absolute value. Then,

$$\mathbb{P}(Y<a) = \iint\limits_{\{(y,z)|y<a\}} p_U(yz)p_V(z)|z|\mathrm{d}z\mathrm{d}y$$
$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left\{\int_{-\infty}^{a} e^{-\frac{1}{2}y^2z^2}e^{-\frac{1}{2}z^2}|z|\mathrm{d}y\right\}\mathrm{d}z.$$

Using the change of variables $w = \frac{1}{2}z^2$, $\mathrm{d}w = z\mathrm{d}z$ and replacing $\int_{-\infty}^{\infty}\mathrm{d}z$ by $2\int_0^{\infty}\mathrm{d}z$ (this replacement is allowed since its integrand is positive):

$$\mathbb{P}(Y<a) = \frac{1}{\pi}\int_{-\infty}^{a}\left\{\int_0^{\infty} e^{-\frac{1}{2}z^2(1+y^2)}z\mathrm{d}z\right\}\mathrm{d}y$$
$$= \frac{1}{\pi}\int_{-\infty}^{a}\left\{\int_0^{\infty} e^{-w(1+y^2)}\mathrm{d}w\right\}\mathrm{d}y = \int_{-\infty}^{a}\frac{1}{\pi}\frac{\mathrm{d}y}{(1+y^2)}.$$

This is known as the Lorentzian (or Cauchy) distribution. The PDF of the Lorentzian distribution is obtained by differentiating with respect to a:

$$p_Y(y) = \frac{1}{\pi}\frac{1}{(1+y^2)}.$$

In another example we can ask what is the probability that a rv X takes on exactly the value x:

$$\mathbb{P}(X=x) = \lim_{\mathrm{d}x\to 0}\mathbb{P}(x<X\le x+\mathrm{d}x) = \lim_{\mathrm{d}x\to 0}\int_x^{x+\mathrm{d}x} p(x)\mathrm{d}x = \lim_{\mathrm{d}x\to 0} p(x)\mathrm{d}x = 0$$

provided that $p(x)$ is continuous. If $p(x)$ is discontinuous at x, this result is not necessarily zero. In this course, we will not be dealing with discontinuous probability functions.

2.20. Probability of Mutually Exclusive Random Events

If random events A_1, A_2 and A_n are disjoint sets, i.e. $A_i \cap A_j = \emptyset$, then the probability of any of the A_i events is a sum of probabilities:

$$\boxed{\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_n).} \tag{2.13}$$

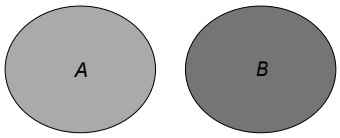


Figure 2.5. Mutually exclusive random events A and B.

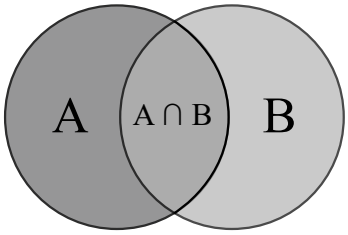


Figure 2.6. Events that are not mutually exclusive share common outcomes (as represented here by the overlap region).

Such a set of random events is called mutually exclusive events. The "union" $A_1 \cup A_2 \cup \cdots \cup A_n$ of random events is equivalent to a "logical OR" operation, i.e.

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \mathbb{P}(A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_n).$$

Suppose that we measure the number of radioactive counts within a 1-second time interval. When we make the statement "12 or fewer counts were observed" (during a 1-second time interval),

$$\text{12 or fewer counts } = \text{1 count or 2 counts or ... or 12 counts.}$$

In other words, let X be a rv that represents the # of counts (in the 1-second time interval).

$$\{X \leq 12\} = \{X = 1\} \cup \{X = 2\} \cup \{X = 3\} \cup \cdots \cup \{X = 12\}.$$

Decomposing the event $\{X \leq 12\}$ as a union of mutually exclusive random events, i.e. $\{X = 1\} \cap \{X = 2\} = \emptyset$, etc., offers some advantages when calculating the probabilities of events. It enables us to invoke formula (2.13).

Two mutually exclusive events A and B have no overlap can be represented as shown in Fig. 2.5. What should we do if the random events are not mutually exclusive? For simplicity, consider only 2 events, A and B. Mutual exclusivity means that $A \cap B = \emptyset$. If the intersection is nonzero, then we have the situation illustrated in the Venn diagram (Fig. 2.6).

In this case, we should avoid overcounting by subtracting the intersection:

$$\boxed{\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).}$$

Let us look at the example of rolling a die. Let X be the numerical result of the die roll (# appearing on the top face). Define two random events, A and B, as follows: $A = \{X \leq 3\}$ and $B = \{X \text{ is odd}\}$. It is helpful to decompose these random events into a union of mutually exclusive "elementary events":

$$A = \{X \leq 3\} = \{X = 1\} \cup \{X = 2\} \cup \{X = 3\},$$

and

$$B = \{X \text{ is odd}\} = \{X = 1\} \cup \{X = 3\} \cup \{X = 5\}.$$

The union of A and B is:

$$A \cup B = \{X = 1\} \cup \{X = 2\} \cup \{X = 3\} \cup \{X = 5\}.$$

The intersection of A and B is:

$$A \cap B = \{X = 1\} \cup \{X = 3\}.$$

If the die is unbiased, i.e. $\mathbb{P}(X = x_i) = 1/6$ for $x_i \in \{1, 2, \ldots, 6\}$, then $\mathbb{P}(A \cup B) = 2/3$, $\mathbb{P}(A) = 1/2$, $\mathbb{P}(B) = 1/2$ and $\mathbb{P}(A \cap B) = 1/3$. This verifies the formula $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ for this particular choice of random events.

2.21. Discrete Random Variables

A discrete rv X takes values in a discrete set $\mathcal{X} = \{x_i\}_{i=1}^N$, where N is the number of possible values X can take (cardinality of the set $\mathcal{X}$) and $x_i \in \mathcal{X}$ are the possible values. The word discrete refers to the "state space". $\mathcal{X}$, which is countable (discrete) and in the present case, contains N elements (N can also be infinite). Discrete rv's can be described using the PDF formed with Dirac delta functions:

$$p_X(x) = \sum_{i=1}^{N} p_i \delta(x - x_i), \quad p_i \geq 0, \quad \sum_i p_i = 1.$$

In this section we explicitly state the main formulas pertaining to the properties of discrete rv's by way of discrete sums and the "probability mass function" or PMF. Either description is valid.

2.21.1. Properties of Discrete Random Variables. The rv is defined by the probability distribution $\{p_i\}$ (also known as the "probability mass function" or PMF), where $p_i \geq 0$ for all i. The normalization condition is

$$\sum_{i=1}^{N} p_i = 1$$

p_i: probability that rv X takes the value x_i.

N: can be finite or infinite; in any case, the $\{p_i\}$ must sum to 1.

We note that from the definition of p_i as the probability that X takes the discrete value x_i, and the fact that the events $\{X = x_i\}$ are mutually exclusive random events, it follows that

$$\begin{aligned}\mathbb{P}(X \leq x_j) =& \mathbb{P}(\{X = x_1\} \cup \{X = x_2\} \cup \cdots \cup \{X = x_j\}) \\ =& \sum_{i=1}^{j} \mathbb{P}(\{X = x_i\}) = \sum_{i=1}^{j} p_i.\end{aligned}$$

The *mean* value of X is

$$\mathbb{E}[X] = \sum_{i=1}^{N} p_i x_i.$$

The *n-th moment* of X is

$$\mathbb{E}[X^n] = \sum_{i=1}^{N} p_i x_i^n.$$

The *mean* or mathematical expectation of $f(X)$ is

$$\mathbb{E}[f(X)] \equiv \sum_{i=1}^{N} p_i f(x_i).$$

The *variance* of X is (let $\mu_X = \mathbb{E}X$)

$$\sigma^2 = \mathrm{Var}(X) = \mathbb{E}[X - \mu_X]^2 = \sum_{i=1}^{N} p_i (x_i - \mu_X)^2.$$

The variance is also equal to $\mathbb{E}[X^2] - (\mu_X)^2$. When calculating averages of functions of rv's, we proceed by replacing the rv X by its value x_i, multiplying the expression by p_i and summing over all i. For example,

$$\mathbb{E}[\exp(-X)] = \sum_i p_i \exp(-x_i), \qquad \mathbb{E}[g(X)] \equiv \sum_i p_i g(x_i).$$

This is analogous to the continuous case covered in the previous lecture where $\int p(x)$ replaces $\sum_i p_i$:

$$\mathbb{E}[\exp(-X)] = \int_{-\infty}^{\infty} p(x) \exp(-x) \mathrm{d}x, \qquad \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} p(x) g(x) \mathrm{d}x.$$

When we deal with two discrete rv's X and Y, the joint probability can be written

$$p_{ij} \equiv \mathbb{P}(X = x_i \cap Y = y_j)$$

where $\mathbb{P}(X = x_i \cap Y = y_j)$ denotes the probability of X taking the value x_i and Y taking the value y_j. The p_{ij} are normalized to 1:

$$\sum_{i,j} p_{ij} = 1.$$

The average of a function $g(X, Y)$ is:

$$\mathbb{E}[g(X, Y)] = \sum_{ij} p_{ij} g(x_i, y_j).$$

As before, if X and Y are statistically independent, the mean of XY equals the product of the means of X and Y:

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{ij} x_i y_j \mathbb{P}(X = x_i \cap Y = y_j) \\ &= \sum_i x_i \mathbb{P}(X = x_i) \sum_j y_j \mathbb{P}(Y = y_j) = \mathbb{E}[X] \cdot \mathbb{E}[Y],\end{aligned}$$

and similarly, we have:

$$\mathbb{E}[X^n Y^m] = \mathbb{E}[X^n] \cdot \mathbb{E}[Y^m].$$

A consequence of this result is

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)],$$

because sufficiently nice functions f, g can be expanded as a power series, enabling us to apply the result $\mathbb{E}[X^n Y^m] = \mathbb{E}[X^n] \cdot \mathbb{E}[Y^m]$ to each term of the expansion.

2.21.2. Poisson Distribution. The Poisson distribution is a discrete probability distribution which is frequently used to describe counts of rare events. The main assumptions are:

- The events counted are *rare* events.
- All events are statistically independent.
- Average count rate does not change over time.

The typical application of this distribution is *radioactive counting* (for example, with a Geiger counter), where the average count $\overline{n}$ in a given time is given by the formula:

$$\overline{n} = \lambda\tau$$

where λ is average count rate and τ is time interval. For example: $\lambda = 1.5$ s^{-1}, $\tau = 10$ s gives $\overline{n} = 15$. $\overline{n}$ does not have to be an integer number.

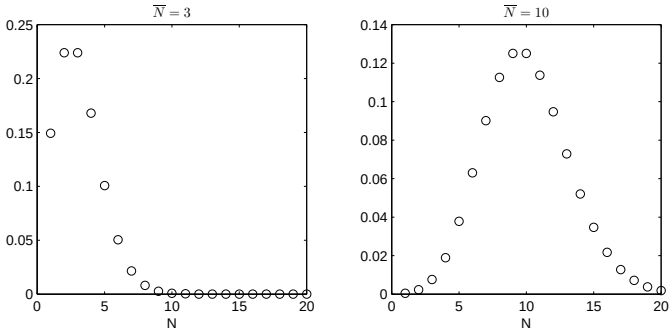


Figure 2.7. Poisson distribution (PMF) for two different parameter values.

2.21.2.1. *Probability Mass Function.* The probability distribution is completely described by a single parameter, $\overline{n}$,[9]

$$\boxed{\mathbb{P}(n;\overline{n}) = \frac{e^{-\overline{n}}\overline{n}^n}{n!}, \qquad n = 0, 1, 2, \ldots}$$

In the usual interpretation, $P(0;\overline{n})$ gives the probability of observing 0 counts in a time interval τ, $P(1;\overline{n})$ gives the probability of observing 1 count, etc. It is easy to check that this PMF is normalized to 1:

$$\sum_{n=0}^{\infty} \mathbb{P}(n;\overline{n}) = \sum_{n=0}^{\infty} \frac{e^{-\overline{n}}\overline{n}^n}{n!} = e^{-\overline{n}} \sum_{n=0}^{\infty} \frac{\overline{n}^n}{n!} = e^{-\overline{n}} e^{\overline{n}} = 1,$$

since the Taylor expansion of $\exp(x)$ is $\sum_{k=0}^{\infty} x^k/k!$. Figure 2.7 shows plots of the Poisson distribution for $\overline{n} = 3$ and $\overline{n} = 10$. Notice that the distribution looks more like a Gaussian at large $\overline{n}$.

These plots were generated in MATLAB using the commands:

```
1  Nbar=2;N=0:20;figure;plot(N,exp(-Nbar)*(Nbar.^N)./factorial(N),'o');
2  Nbar=5;N=0:20;figure;plot(N,exp(-Nbar)*(Nbar.^N)./factorial(N),'o');
3  Nbar=10;N=0:20;figure;plot(N,exp(-Nbar)*(Nbar.^N)./factorial(N),'o');
4  figure;ezplot('exp(-((x-10)^2)/(2*10))',[0,20]);
```

Two more properties of the Poisson distribution which you can easily verify are:

$$\text{average of } n = \sum \mathbb{P}(n;\overline{n})n = \overline{n}, \qquad var(n) = \sigma^2 = \overline{n} = \lambda\tau.$$

Thus, the mean and variance are both equal to $\overline{n}$. The standard deviation, $\sigma = \sqrt{\overline{n}}$, gives the error in the measurement. For the mean, the proof is

[9]This is in contrast to the Gaussian distribution, whose description requires two parameters: $\overline{X}$ and σ^2.

trivial and left as an exercise. For the variance, the proof is easy but requires more steps.[10]

2.21.2.2. *Error Bars of a Measurement.* If the experiment yields a mean count of n, the best estimate of the error[11] in this quantity is $\sqrt{n}$. We report this measurement as

$$\boxed{n \pm \sqrt{n}.}$$

While it may seem that the error grows with n, the fractional uncertainty actually decreases with n:

$$\frac{\delta n}{n} = \frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}}$$

i.e., larger n result in smaller fractional uncertainty.

2.21.2.3. *Poisson Counts.* This example was taken from the book by Hughes & Hase and modified. A bridge cannot hold too many cars at once due to the potential for structural damage. A particular bridge is designed to hold less than 13 cars (safe level) per time interval (1 min.). In a random sample, the total number of cars recorded crossing the bridge in 10 hours was 1980.

Q: What is the average number of cars crossing per minute and its error? A: $\lambda = \frac{1980}{10\times 60} = 3.30$ cars/min.

$$\delta\lambda = \frac{\sqrt{1980}}{10 \times 60} = 0.07 \text{ cars/min.}$$

Q: What is the probability that during a random one-minute interval 13 cars will be observed crossing? A: $\overline{n} = \lambda\tau = 3.30$, with $n = 13$

$$\mathbb{P}(n = 13; 3.3) = \frac{e^{-3.3}3.3^{13}}{13!} = 3.3 \times 10^{-5}$$

Q: What is the probability that the bridge will fail (due to too many cars crossing)? A:

$$\begin{aligned}\mathbb{P}(\text{13 or more cars}) =& 1 - \mathbb{P}(\text{12 or fewer cars}) \\ =& 1 - \{\mathbb{P}(0; 3.3) + \mathbb{P}(1; 3.3) + \cdots + \mathbb{P}(12; 3.3)\} \\ =& 4.2 \times 10^{-5}.\end{aligned}$$

During 1 minute of observation, this is the probability that the bridge will fail.

[10] A proof can be found at: `https://proofwiki.org/wiki/Variance_of_Poisson_Distribution`

[11] Sometimes, all we have is 1 count. While this may not be the mean count, it is all that we have. The best we can do in this case is report $n \pm \sqrt{n}$.

2.21.2.4. *Poisson Distribution in the Limit of Large Means.* For large means, the Poisson distribution converges to a Gaussian distribution:

$$\boxed{\frac{e^{-\overline{n}}\overline{n}^n}{n!} \approx \frac{1}{\sqrt{2\pi\mu_X}} \exp\left(-\frac{(x-\mu_X)^2}{2\mu_X}\right)}$$

where:

$$\begin{aligned} n \text{ (discrete)} &\to x \text{ (continuous)} \\ \overline{n} &\to \mu_X \text{ (variance, mean)} \\ \sqrt{\overline{n}} &\to \sqrt{\mu_X} \text{ (standard deviation)} \end{aligned}$$

The proof makes use of Stirling's approximation

$$\log n! \approx n \log n - n + O(\log n),$$

and

$$\frac{|n-\overline{n}|}{\overline{n}} \ll 1.$$

These two conditions (Stirling approx. and $\overline{n}$ close to n) imply that our proof is valid in the limit of large means. If the mean is not large, the Stirling approximation cannot be used.

$$\begin{aligned} \frac{e^{-\overline{n}}\overline{n}^n}{n!} &= \exp\{-\overline{n} - \log n! + n\log\overline{n}\} \\ &= \exp\{-\overline{n} - n\log n + n + n\log\overline{n}\} \\ &= \exp\{(n-\overline{n}) + n\log(\overline{n}/n)\} \\ &= \exp\left\{(n-\overline{n}) + n\log\left[1+\left(\frac{\overline{n}-n}{n}\right)\right]\right\} \\ &\approx \exp\left\{-\frac{(\overline{n}-n)^2}{2n}\right\} \approx \exp\left\{-\frac{(\overline{n}-n)^2}{2\overline{n}}\right\} \end{aligned}$$

The first step was to invoke the Stirling's approximation, $\log n! \approx n\log n - n$. The second step was to expand about mean ($\overline{n}$) for large $\overline{n}$). Then we Taylor expanded $\log(1+x) \approx x - x^2/2 + O(x^3)$. In the last step we have used the approximation $n \approx \overline{n}$ near the mean for the denominator in the argument of the exp.

The prefactor $\frac{1}{\sqrt{2\pi\overline{n}}}$ could have been recovered had we used the slightly more accurate form of the Stirling's formula:

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n,$$

and of course, followed by the application of $\frac{|n-\overline{n}|}{\overline{n}} \ll 1$ to justify replacing $\sqrt{2\pi n}$ by $\sqrt{2\pi\overline{n}}$. See Problem 40.

2.21.3. Statistical Independence (Discrete Case). The notion of statistical independence is the same as in the continuous case. Let X and Y be two rv's. Independence of X and Y means that the joint probability factors as a product:

$$p_{ij}^{XY} = p_i^X \cdot p_j^Y, \text{ for all } i, j \tag{2.14}$$

where p^{XY} is the joint PMF for X and Y. p_i^X is a marginal PMF:

$$p_i^X \equiv \sum_j p_{ij}^{XY}.$$

Likewise, p_j^Y is also a marginal PMF:

$$p_j^Y \equiv \sum_i p_{ij}^{XY}.$$

2.21.4. Example 1. Consider a random experiment that involves rolling a die

$$X \in \{1, 2, 3, 4, 5, 6\}$$

and tossing a coin

$$Y \in \{H, T\}.$$

You are asked to determine whether or not X is statistically independent from Y. Intuitively, this should be the case (i.e., why would a coin toss affect the outcome of rolling a die?).

For this random experiment, there are 12 possible "elementary" outcomes:

$$\begin{aligned}
(X,Y) &= (1,H), & (X,Y) &= (1,T),\\
(X,Y) &= (2,H), & (X,Y) &= (2,T),\\
(X,Y) &= (3,H), & (X,Y) &= (3,T),\\
(X,Y) &= (4,H), & (X,Y) &= (4,T),\\
(X,Y) &= (5,H), & (X,Y) &= (5,T),\\
(X,Y) &= (6,H), & (X,Y) &= (6,T).
\end{aligned}$$

To get the joint PMF we must repeat this experiment many times and record the results. Suppose that we repeat the experiment 10,000 times and count the number of times each elementary outcome occurs. Let's do this in MATLAB:

```
>> X=randi([1 2],[1 10000]);
>> Y=randi([1 6],[1 10000]);
```

We then plot a 2D histogram (see Fig. 2.8):

```
>> figure; h=histogram2(X,Y)
```

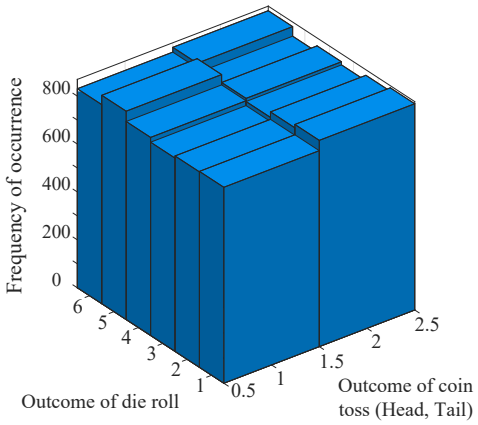


Figure 2.8. Bivariate histogram of the die & coin experiment.

```
h =

  Histogram2 with properties:

            Data: [10000x2 double]
          Values: [2x6 double]
         NumBins: [2 6]
       XBinEdges: [0.5000 1.5000 2.5000]
       YBinEdges: [1x7 double]
        BinWidth: [1 1]
   Normalization: 'count'
       FaceColor: 'auto'
       EdgeColor: [0.1500 0.1500 0.1500]
```

The histogram is an approximation to the PMF. As you can see, this distribution is uniform. The values used to plot the histogram are:

```
>> h.Values

ans =

   818   816   807   827   870   831
   861   847   824   837   811   851
```

Dividing by the number of experiments performed (10,000):

```
>> h.Values/10000

ans =
```

X\Y	heads	tails	sum
1	1/12	1/12	1/6
2	1/12	1/12	1/6
3	1/12	1/12	1/6
4	1/12	1/12	1/6
5	1/12	1/12	1/6
6	1/12	1/12	1/6
sum	1/2	1/2	

Table 2.2. Joint PDF for an experiment that involves rolling a die ($X \in \{1, 2, 3, 4, 5, 6\}$) and tossing a coin ($Y \in \{H, T\}$).

```
    0.0818    0.0816    0.0807    0.0827    0.0870    0.0831
    0.0861    0.0847    0.0824    0.0837    0.0811    0.0851
```

gives an approximation to the joint PMF. Each entry is approximately equal to 1/12. A PMF with entries equal to 1/12 can be represented as a matrix (see Table 2.2). The marginal PMFs for X and Y are obtained by summing along rows and columns, respectively. You can check that for the data shown in Table 2.2), Eq. (2.14) holds for all i, j. Therefore, X and Y are statistically independent.

We also know that statistical independence implies the variables are uncorrelated. Let's check this by computing the correlation coefficient:

```
>> corrcoef(X,Y)

ans =

    1.0000   -0.0115
   -0.0115    1.0000
```

The MATLAB command `corrcoef` computes the matrix of correlation coefficients, $\begin{bmatrix} \rho(X,X) & \rho(X,Y) \\ \rho(Y,X) & \rho(Y,Y) \end{bmatrix}$. The diagonal elements should be 1 (since X is fully correlated with X; same for Y) whereas the off-diagonal elements should be zero. Indeed, the off-diagonal elements are two orders of magnitude smaller than 1, indicating the lack of correlation between X and Y.

A counter-example illustrating statistical dependence would be the joint PMF shown in Table 2.3, which differs from Table 2.2 only in the second row. Namely, when the result from rolling the die is 2, the coin toss always yields "tails". (Don't try too hard to imagine how this can possibly happen in the lab; it is perhaps easier to imagine that a magician is doing the

X\Y	heads	tails	sum
1	1/12	1/12	1/6
2	0/12	1/6	1/6
3	1/12	1/12	1/6
4	1/12	1/12	1/6
5	1/12	1/12	1/6
6	1/12	1/12	1/6
sum	1/2	1/2	

Table 2.3. Joint PDF for an experiment that involves rolling a die ($X \in \{1,2,3,4,5,6\}$) and tossing a coin ($Y \in \{H,T\}$). This joint PMF is the same as that of Table 2.2 except for the second row.

experiment for you.) Because $p_{2,1}^{XY} = 0 \neq p_2^X \cdot p_1^Y = \frac{1}{6} \cdot \frac{1}{2}$, we are unable to prove statistical independence of X and Y.

2.21.5. Example 2. The joint distribution of the bivariate rv (X,Y) is given by

$$p_{XY}(x_i,y_j) = \begin{cases} k\dfrac{|x_i|}{2^{y_j}} & x_i = -1,1; y_j = 1,2,3,\ldots \text{ (to infinity)} \\ 0 & \text{otherwise} \end{cases}$$

(a) k is a constant. Find the value of k.

$$\sum_i \sum_j p_{ij} = k2 \sum_j \frac{1}{2^{y_j}} = k2 \cdot 1 = 2k. \qquad k = 1/2.$$

(b) Find the marginal probability distributions of X and Y.

$$p_X(x_i) = \sum_j p_{ij} = \sum_j \frac{1}{2}|x_i|\frac{1}{2^{y_j}} = \frac{1}{2}|x_i|. \qquad x_i = -1,1.$$

$$p_Y(y_j) = \sum_i p_{ij} = \frac{1}{2^{y_j}}. \qquad y_j = 1,2,3,\ldots$$

(c) Are X and Y statistically independent?

Forming the product of marginal distributions,

$$p_X(x_i)p_Y(y_j) = \frac{1}{2}|x_i| \cdot \frac{1}{2^{y_j}} = p_{XY}(x_i,y_j)$$

Hence X and Y are independent.

2.21.6. Cross-Correlation in Image Analysis. The concept of covariance leads to the cross-correlation analysis. Cross-correlation is a type of covariance that involves comparing two signals (or images) together through pixel-by-pixel multiplication of a window (or ROI) that is translated across different regions of a target signal (or image). If the two signals (or images)

are of the same size, a single value is obtained. If the sliding window is smaller than the target signal (or image), the output is a function of the translation coordinate(s).

The 2D cross-correlation of a $M \times N$ matrix, X, and a $P \times Q$ matrix, H, is a matrix C, of size $M + P - 1$ by $N + Q - 1$. Its elements are given by:

$$C(x,y) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X(m,n)\overline{H}(m-x, n-y), \qquad \begin{cases} -(P-1) \le x \le M-1, \\ -(Q-1) \le y \le N-1, \end{cases}$$

where the bar denotes complex conjugation. Likewise, a 1D signal can also be compared to another signal, for purposes of comparison or pattern recognition. The true cross-correlation sequence of two random samples $\{x_n\}$ and $\{y_n\}$ is

$$R_{xy}(m) = \mathbb{E}[x_{n+m} y_n^*] = \mathbb{E}[x_n y_{n-m}^*],$$

where $-\infty < n < \infty$ and asterisk denotes complex conjugation. The raw cross-correlation is computed as:

$$\hat{R}_{xy}(m) = \begin{cases} \sum_{n=0}^{N-m-1} x_{n+m} y_n^*, & m \ge 0, \\ \hat{R}_{xy}^*(-m), & m < 0. \end{cases}$$

In MATLAB these two commands are implemented as `xcorr2` and `xcorr`, respectively. For more information including examples, see the MATLAB documentation:
`https://www.mathworks.com/help/signal/ref/xcorr2.html`
`https://www.mathworks.com/help/matlab/ref/xcorr.html`

2.22. Conditional Probability and Conditional Expectation

2.22.1. Conditional densities. The conditional density of X given Y is defined as

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)},$$

where $p_{X,Y}(x,y)$ is the joint PDF of X and Y, $p_Y(y)$ is the marginal PDF of Y. This is a consequence of the formula for conditional probability,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

with $A = \{X = x\}$ and $B = \{Y = y\}$, i.e.

$$\mathbb{P}(x < X \le x + \mathrm{d}x | y < Y \le y + \mathrm{d}y)\mathrm{d}x = \frac{\mathbb{P}(x < X \le x + \mathrm{d}x, y < Y \le y + \mathrm{d}y)\mathrm{d}x\mathrm{d}y}{\mathbb{P}(y < Y \le y + \mathrm{d}y)\mathrm{d}y}.$$

An interpretation of $p_{X|Y}(x|y)$ is obtained by integrating it:

$$\mathbb{P}(a < X \leq b|Y = y) = \int_a^b p_{X|Y}(x|y)\mathrm{d}x$$

(i.e. the probability that $X \in [a, b]$ given that $Y = y$). However, the left hand side $\{Y = y\}$ is an event with probability zero, which is ill-defined. We instead use a limit to circumvent this difficulty:

$$\mathbb{P}(a < X \leq b|Y = y) = \lim_{\epsilon \to 0} \mathbb{P}(a < X \leq b||Y - y| < \epsilon).$$

The conditional expectation of X given $Y = y$ is defined as

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} x p_{X|Y}(x|y)\mathrm{d}x.$$

A consequence of this definition is:

$$\begin{aligned}
\mathbb{E}[Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y p_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y \left[p_{Y|X}(y|x)p_X(x)\right]\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} y p_{Y|X}(y|x)\mathrm{d}y\right] p_X(x)\mathrm{d}x \\
&= \int_{-\infty}^{\infty} \mathbb{E}[Y|X = x]p_X(x)\mathrm{d}x.
\end{aligned}$$

An analogous concept of "conditional probability mass" exists for the case of discrete rv's. See Problems 57 and 58 for more on conditional densities.

2.22.2. Conditional Expectation. We may also calculate expectation values under some condition. This is the same idea as calculating the normal expectation value, except that we use the conditional density instead of the regular density. For (a) the condition involves a random event H. If X is a rv, H is a random event and $p_{X|H}(x)$ is the conditional density of X under the condition H, the expectation value of X under the condition H is:

$$\mathbb{E}[X|H] \equiv \int x p_{X|H}(x)\mathrm{d}x,$$

where the integral is over all possible values of X (i.e., the "range" of X). As an example, the event H could be $H = \{Y = 10\}$, or it could be $H = \{Y = y\}$ (where the value y remains unspecified). You can check that $\mathbb{E}[X|H]$ is still linear in X. For (b) the condition is a rv, e.g., $\mathbb{E}[X|Y]$. The conditional expectation $\mathbb{E}[X|Y]$ is obtained by calculating $\mathbb{E}[X|Y = y]$. The result will be a function of y. Then replace y by the rv Y. Notice that the end result for $\mathbb{E}[X|Y]$ is itself a rv. In other words, to get $\mathbb{E}[X|Y]$ we use $\mathbb{E}[X|Y = y]$ together with $y = Y(\omega)$. The meaning of $\mathbb{E}[X|Y]$ is "the

function of Y that best approximates X".

From these definitions many properties follow. For example, it is easy to check that

$$\mathbb{E}[\mathbb{E}[Y|X]] = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} y\, p_{Y|X}(y|x)\, \mathrm{d}y \right] p_X(x)\, \mathrm{d}x = \int_{-\infty}^{\infty} y\, p_Y(y)\, \mathrm{d}y = \mathbb{E}[Y].$$

We used the definitions of conditional probability $p_{Y|X}(y|x)p_X(x) = p_{XY}(x,y)$ and marginal density $p_Y(y) = \int_{-\infty}^{\infty} p_{XY}(x,y)\mathrm{d}x$. More properties of conditional expectations can be found at this site (click on *Proof* to obtain the proofs of various results):

`https://www.randomservices.org/random/expect/Conditional.html`

In Problem 59 you are asked to calculate expectation values under some condition.

2.23. Signal Averaging Reduces Relative Error

An important concept in experimental science is that of "signal averaging". This is done to reduce the noise error, or equivalently, to reduce the size of the error bars relative to the measurement. Suppose X is a rv and we perform n measurements of X. We obtain the data set $\{x_i = X(\omega_i)\}_{i=1}^n$. Another way to view this experiment is to consider the simultaneous measurement of n independent rv's $X_1, \ldots, X_n$, each of which has the same distribution as X: each has mean μ_X and variance ($var(X_i) < \infty$). We form the average:

$$\overline{X_{av}}(\omega) = \frac{X_1(\omega) + X_2(\omega) + \cdots + X_n(\omega)}{n},$$

where a single value of ω is used, as we assumed simultaneous measurement of $X_1, \ldots, X_n$. The variance of this sum is

$$var(\overline{X_{av}}) = \frac{\sum_{i=1}^n var(X_i)}{n^2} \propto \frac{1}{n},$$

where we used the property $var(aX) = a^2 var(X)$. The noise is the square root of the variance $\sigma = \sqrt{var(\overline{X_{av}})}$. Thus, signal averaging reduces the noise from random errors.

We have just derived the formula for standard error: the mean of $\overline{X_{av}}(\omega) = \frac{X_1(\omega)+X_2(\omega)+\cdots+X_n(\omega)}{n}$ is μ_X (where $\mu_X = \mathbb{E}X$) and its standard deviation is $\sqrt{var(X)/n} = \sigma_X/\sqrt{n}$. We recognize this as the standard error.

If the X_i are iidrv the signal-to-noise ratio (SNR) is defined as:

$$\text{SNR} = \frac{\text{signal}}{\text{noise}} = \frac{\overline{X_{av}}}{\sqrt{var(\overline{X_{av}})}} \propto \sqrt{n}$$

Thus, SNR improves with signal averaging.

2.24. Some Theorems on Random Variables

2.24.1. Normal Linear Transform Theorem. The normal linear transform theorem is:

$$\boxed{\alpha + \beta\mathcal{N}_1(\mu, \sigma^2) = \mathcal{N}_2(\alpha + \beta\mu, \beta^2\sigma^2).}$$

(We denoted $\mathcal{N}$ with subscripts 1 and 2 to emphasize that they are different rv's, i.e., the rv on the right hand side is created from the rv on the left hand side.)

Proof: Let $Y = \alpha + \beta X$, where $X \sim \mathcal{N}_1(\mu, \sigma^2)$. Write

$$\mathbb{P}(Y < a) = \mathbb{P}(\alpha + \beta X < a) = \mathbb{P}(X < (a-\alpha)/\beta) = \int_{-\infty}^{(a-\alpha)/\beta} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \mathrm{d}x.$$

and make a substitution of variables $y = \alpha + \beta x$, $\mathrm{d}y = \beta \mathrm{d}x$ to get an integral of the form $\int_{-\infty}^{a} p_Y(y)\mathrm{d}y$:

$$= \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-((y-\alpha)/\beta-\mu)^2/2\sigma^2} (\mathrm{d}y/\beta) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi\beta^2\sigma^2}} e^{-(y-\alpha-\beta\mu)^2/2\beta^2\sigma^2} \mathrm{d}y,$$

which is the CDF of a normal rv with mean $\alpha + \beta\mu$ and variance $\beta^2\sigma^2$. In the special case $\mu = 0$, $\sigma^2 = 1$ we have:

$$\alpha + \beta\mathcal{N}_1(0, 1) = \mathcal{N}_2(\alpha, \beta^2).$$

2.24.2. Normal Sum Theorem. The normal sum theorem is:

$$\boxed{\mathcal{N}_3(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) = \mathcal{N}_1(\mu_1, \sigma_1^2) + \mathcal{N}_2(\mu_2, \sigma_2^2),}$$

where on the right-hand-side is the sum of two statistically independent rv's. In other words, let $X \sim \mathcal{N}_1(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}_2(\mu_2, \sigma_2^2)$ be independent rv's. (We denoted $\mathcal{N}$ with subscripts 1, 2 and 3 to emphasize that they are different rv's.) We are asking what is the distribution of the new rv $U = X + Y$. The proof of this involves handling a 2D integral of the joint PDF of X and Y:

$$\mathbb{P}(X + Y < a) = \frac{1}{2\pi\sigma_1\sigma_2} \iint_{\{(x,y)|x+y<a\}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}} \mathrm{d}x\mathrm{d}y.$$

With some effort, you should be able to simplify this double-integral and show that it is the CDF of the rv $X+Y$ (use the change of variables $V=X+Y$, $Z=Y$).

2.24.3. Sum of Independent Gaussians. Suppose that $X_1,\ldots,X_n$ are iidrv and normal, say, $\mathcal{N}(0,1)$. The distribution of their sum is also normal. This can be proven by induction, using the result from Section 2.24.2.

2.24.4. Sum of Two Independent Cauchy's. Let X,Y be independent Cauchy rv's. What is the distribution of their sum $X+Y$? We start with the CDF:

$$\mathbb{P}(X+Y<a)=\iint\limits_{\{(x,y)|x+y<a\}}\frac{1}{\pi^2}\frac{1}{(1+x^2)}\frac{1}{(1+y^2)}\mathrm{d}x\mathrm{d}y.$$

Let us effect a change of variables: $u=x+y$ and $v=y$. The inverse is $y=v$ and $x=u-v$. The area element is:

$$\mathrm{d}x\mathrm{d}y=\left|\frac{\partial(x,y)}{\partial(u,v)}\right|\mathrm{d}u\mathrm{d}v=\left|\det\begin{pmatrix}1&1\\0&1\end{pmatrix}\right|\mathrm{d}u\mathrm{d}v=\mathrm{d}u\mathrm{d}v.$$

In the new coordinates, the integral is:

$$\mathbb{P}(X+Y<a)=\frac{1}{\pi^2}\int_{-\infty}^{a}\mathrm{d}u\int_{-\infty}^{\infty}\mathrm{d}v\frac{1}{(1+(u-v)^2)}\frac{1}{(1+v^2)}.$$

To solve the integral over v go to `www.wolframalpha.com` and type:

```
integrate (1/(1+(u-v)^2))*(1/(1+v^2)) from v=-infinity to infinity
```

The result is:

$$\int_{-\infty}^{\infty}\mathrm{d}v\frac{1}{(1+(u-v)^2)}\frac{1}{(1+v^2)}=\frac{2\pi}{4+u^2}.$$

This gives the CDF in integral form. Differentiating with respect to a gives the PDF:

$$p_{X+Y}(a)=\frac{1}{\pi}\frac{2}{(4+a^2)}=\frac{1}{\pi}\frac{1}{2(1+(a/2)^2)}.$$

The general Cauchy distribution has PDF $p(x)=\left[\pi\gamma\left(1+\left(\frac{x-x_0}{\gamma}\right)^2\right)\right]^{-1}$, where x_0 is the center and γ is the scale parameter (related to the width). Thus, adding two Cauchy rv's centered at 0 with "scale parameter" of $\gamma=1$ results also in a Cauchy distribution centered at 0, but with $\gamma=2$ (i.e. it is twice as broad as the $\gamma=1$ case).

2.24.5. Central Limit Theorem. One of the most important theorems in probability theory is the central limit theorem (CLT). The CLT describes many important physical phenomena observed in nature that arise from the sum of many independent random effects (e.g. microscopic forces). Loosely

speaking, the central limit theorem states that regardless of the distribution of these random effects, the limiting distribution is Gaussian.

Let $(X_1, \ldots, X_n)$ be a sequence of iidrv (independent identically distributed rv), each having mean μ_X and variance σ^2. Then,

$$\lim_{n\to\infty} \text{rv} \left\{ \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_X) \right\} \xrightarrow{d} \frac{1}{\sqrt{n}} \mathcal{N}(0, \sigma^2).$$

This is equivalent to saying that the arithmetic average $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges[12] (in distribution) to a *normal law* with mean μ_X and variance $\dfrac{\sigma^2}{n}$:

$$\boxed{\lim_{n\to\infty} \text{rv} \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i \right\} \xrightarrow{d} \mathcal{N}\left(\mu_X, \frac{\sigma^2}{n}\right).}$$

The quantities $\{X_i\}$ are rv's. Their sum is also a rv; the CLT states that the sum will be Gaussian-distributed.

Note: that the arithmetic average should also have the mean μ_X comes as no surprise. But also that the variance $\frac{\sigma^2}{n}$ scales as $1/n$ should come as no surprise if you recall the definition of standard error, which states that the error in the mean scales as $1/\sqrt{n}$.

The central limit theorem is very important in the physical sciences because many physical measurements yield Gaussian distributions as a result of the effects of many small additive forces. For example, the Brownian motion of a particle is the result of many small collisions with solvent molecules, yielding a Gaussian distribution for the net displacement.

The CLT is illustrated in Fig. 2.9. The histogram on the left represents the probability distribution of a single rv, X_1. The histogram in the center is the distribution of the average of two such iidrv $\frac{1}{2}(X_1 + X_2)$. The histogram on the right is the distribution for the average of 10 iidrv, $\frac{1}{10}(X_1 + X_2 + \cdots + X_{10})$. As can be seen, while each rv has a uniform (flat) distribution, as the histogram on the left shows, the more rv's we average, the closer the distribution of the average approaches a normal (bell-shaped) distribution.

2.24.5.1. *Random Walk in One Dimension.* The random walk, which is an application of the CLT, is important in the physical sciences. Brownian motion is a limit of the random walk.

---

[12] Convergence in distribution means that the distribution functions converge pointwise:

$$\lim_{n\to\infty} \mathbb{P}(\overline{X}_n \le x) = F_n(x),$$

where $F_n(x)$ denotes the CDF of the normal random variable with mean μ and variance σ^2/n and $\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$.

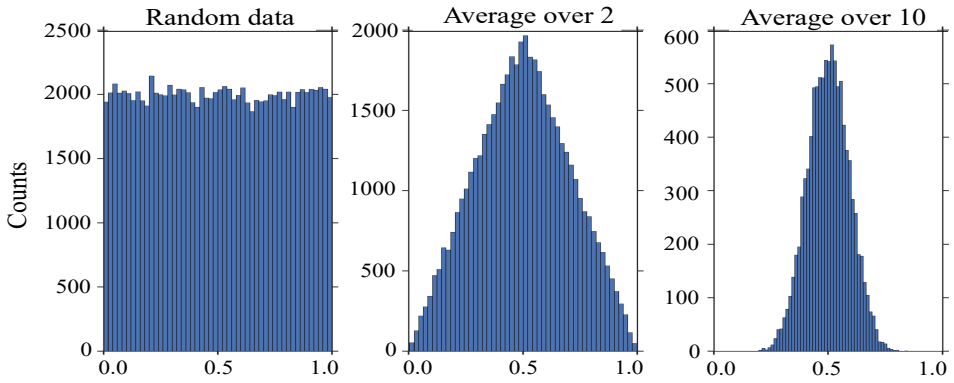


Figure 2.9. Illustration of the central limit theorem.

Let X_i be the rv which denotes the displacement at the i-th step

$$X_i \in \{-\sigma, +\sigma\}$$

and each outcome occurs with equal probabilities, i.e.

$$p_\sigma = \frac{1}{2} \qquad \text{and} \qquad p_{-\sigma} = \frac{1}{2}.$$

These displacements at different points in time are assumed to be statistically independent. After n such steps the net displacement is

$$S_n = X_1 + \cdots + X_n,$$

where $\overline{X_i} = \frac{1}{2}\sigma + \frac{1}{2}(-\sigma) = 0$. Therefore, $\overline{S_n} = 0$. Also, $\mathrm{var}X_i = \frac{1}{2}(\sigma^2) + \frac{1}{2}(-\sigma)^2 = \sigma^2$. Then,

$$\mathrm{var}X_i = \overline{X_i^2} - \overline{X_i}^2 = \sigma^2.$$

By statistical independence of the X_i's:

$$\overline{S_n^2} = \sum_{i=1}^{n} \mathrm{var}X_i = n\sigma^2,$$

The total duration of the random walk is $t = n\Delta t$. We have

$$\overline{S_n^2} = \left(\frac{\sigma^2}{\Delta t}\right) t = 2D\, t.$$

The quantity $D = \frac{\sigma^2}{2\Delta t}$ is called the *diffusion constant* (or diffusion coefficient). D has units of length square divided by time (e.g. m^2/s).

By the central limit theorem, we have that

$$\boxed{\lim_{n\to\infty} S_n = X_1 + \cdots + X_n \sim \mathcal{N}(0, n\sigma^2)}$$

$\lim S_n \sim \mathcal{N}(0, n\sigma^2)$ means that its PDF is

$$p_{S_n}(x,t) = \frac{1}{\sqrt{4\pi Dt}} \exp\left[-\frac{x^2}{4Dt}\right].$$

This PDF is also called transition probability density.[13] It is the probability of finding the particle at (x,t) under the condition that it was at position $x = 0$ at time $t = 0$.

The average position of the random walker after the n-th step is zero: $\overline{S_n} = 0$. This means that if we repeat the random walk experiments, say 10,000 times, the average position after n steps will be zero. It *does not* mean that the random walker automatically returns to the origin. It is merely a statement about the random walk when the walk is repeated many times.

On the other hand, the result $\overline{S_n^2} = 2D\,t$ about the mean square displacement being proportional to t (or root mean square displacement, x, being proportional to $\sqrt{t}$) should be contrasted to the case of ballistic motion for which $x = vt$ (displacement proportional to t). The different powers of t reflect the fact that in diffusional motion, there is a lot of back-and-forth, leading to a shorter displacement over time.

2.24.5.2. *Random Walk in Three Dimensions.* In three dimensions, the displacement is a 3-components vector $R = (X, Y, Z)$. If the components X, Y, Z are statistically independent of each other, the joint probability density is a product:

$$p_{XYZ}(x,y,z,t) = p_X(x,t)p_Y(y,t)p_Z(z,t),$$

where $p_X(x,t) = \frac{1}{\sqrt{4\pi Dt}} \exp\left[-\frac{x^2}{4Dt}\right]$, etc. (for each component). Applying the result of the previous section for each direction (component), we obtain the joint probability density for the 3D random walk:

$$\boxed{p_{XYZ}(x,y,z,t) = \frac{1}{(4\pi Dt)^{3/2}} \exp\left(-\frac{r^2}{4Dt}\right)}$$

where $r^2 = x^2 + y^2 + z^2$, $\vec{r} = (x,y,z)$.

In 3D the mean square displacement (MSD) is[14]

$$\mathbb{E}[r^2] = \mathbb{E}[x^2 + y^2 + z^2] = \mathbb{E}[x^2] + \mathbb{E}[y^2] + \mathbb{E}[z^2] = 2Dt + 2Dt + 2Dt = 6Dt.$$

In the general case of d dimensions, $\vec{r} = (x_1, \ldots, x_d)$, the MSD is:

$$\mathbb{E}[r(t)^2] = \mathbb{E}[x_1^2 + \cdots + x_d^2] = 2dDt.$$

---

[13] A transition probability density is written $p(x,t|y,s)$ to the denote the probability of finding a particle at position x at time t given that it was initially at position y at some earlier time s.

[14]
$$\mathbb{E}[r(t)^2] \equiv \frac{1}{(4\pi Dt)^{3/2}} \iiint_{\mathbb{R}^3} r^2 \exp\left(-\frac{r^2}{4Dt}\right) \mathrm{d}^3 r$$

.

2.25. Importance Sampling

2.25.1. Law of Large Numbers (LLN). Let $X_1, \ldots, X_n$ be a sequence of iiddv each with mean μ. We form the sample mean:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + \cdots + X_n).$$

Then, for any $\epsilon > 0$,

$$\boxed{\lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| > \epsilon) = 0.}$$

This is the *weak law of large numbers* (WLLN). For a proof of the WLLN, see Problem 36. $\overline{X}_n$ is the *sample mean*. The *strong law of large numbers* is[15]

$$\boxed{\mathbb{P}\left(\lim_{n\to\infty} |\overline{X}_n - \mu| > \epsilon\right) = 0.}$$

Both weak and strong LLN are statements about how the sample mean converges to the real mean. However, there is an important difference: the weak LLN tells us how sequences of probabilities ($\mathbb{P}$) converge whereas the strong LLN tells us how the sequence of rv $\overline{X}_n$ behaves in the limit. The CLT, on the other hand, is a much stronger[16] statement: the sample mean (arithmetic average) rv converges (in distribution) to a normal law. The central limit theorem (CLT) should not be confused with the LLN.

2.25.1.1. *WLLN In Words.* The statement $\lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| > \epsilon) = 0$ for any $\epsilon > 0$ simply means that if we take any sequence of iidrv's $X_1, \ldots, X_n$, their arithmetic average tends to their mathematical expectation as $n \to \infty$. Thus, we can approximate mathematical expectations (which may be difficult to compute, especially if the distribution is unknown), using arithmetic averages formed using experimental data. The larger n is, the better the approximation of the mathematical expectation.

Since X_i are rv's, the WLLN formulation also applies to functions of rv's, $Y_i = f(X_i)$, since the latter are also rv's. If desired, f could be almost any formula. For example, suppose that X_i is the acceleration of a particle, a_i (i.e., $X_i \equiv a_i$). The force is mass times acceleration: $F(X_i) = ma = mX_i$. The WLLN allows us to compute not only the average acceleration, $\mathbb{E}[X_i]$, but the average force, $\mathbb{E}[F(X_i)]$, by simply renaming $Y_i = f(X_i)$ and applying it to the sequence $\{Y_i\}$.

[15] The proof of the strong law requires the Borel-Cantelli lemma, which we have not covered.

[16] This is a stronger statement because it is a statement about the entire distribution law of a rv, not just its mean and variance.

2.25.2. Expectation With Respect To Probability Measure. The mathematical expectation of X was defined as:

$$\mathbb{E}[X] = \int_{\mathbb{R}} xp(x)\mathrm{d}x,$$

where $p(x)$ is the PDF of X and the integral is taken over the range of X (here, $\mathbb{R}$). $p(x)\mathrm{d}x$ is the probability measure. It will be useful to use the notation $\mathbb{E}_p[X]$ to emphasize that the PDF used to calculate the mathematical expectation is p. This way, there is no ambiguity as to which probability measure is used. For example, if the PDF of Y is $q(y)$, then we write:

$$\mathbb{E}_q[Y] = \int yq(y)\mathrm{d}y.$$

(Note: y is a dummy integration variable; the choice of symbol is unimportant.) The WLLN states that if Y is distributed according to $q(y)$, then its mathematical expectation can be approximated by the arithmetic average:

$$\mathbb{E}_q[Y] \approx \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

where the Y_i's are sampled according to the distribution q. If instead the PDF of Y had been some other function f, we would have written $\mathbb{E}_f[Y]$ for $\int yf(y)\mathrm{d}y$. The two numerical values $\mathbb{E}_f[Y]$ and $\mathbb{E}_q[Y]$ can, of course, be different, since f and q may be different functions.

As far as the WLLN is concerned, it is meant to enable us to approximate mathematical expectations of rv's (or functions of rv's) by arithmetic averages constructed using experimental data. We present several examples below to illustrate applications of the WLLN. The WLLN is best explained by working out specific examples.

2.25.3. Numerical Integration by Monte-Carlo Method. Monte-Carlo methods can be used to estimate the numerical value of integrals. For example, suppose we want to compute the integral:

$$I = \int_a^b h(x)\mathrm{d}x$$

which is the same as

$$I = \int_a^b \frac{h(x)(b-a)}{(b-a)}\mathrm{d}x = \int_a^b u(x)p(x)\mathrm{d}x = \mathbb{E}_p[u(X)]$$

where $\mathbb{E}_p$ denotes the mathematical expectation with respect to the PDF $p(x)$ and

$$u(x) = h(x)(b-a), \qquad p(x) = \frac{1}{(b-a)}.$$

Thus, $p(x)$ is the PDF of a uniformly distributed rv. This integral can be evaluated by generating random numbers $X_1, \ldots, X_n$ that are iidrv and uniformly distributed on the interval $[a, b]$. By the WLLN, the following estimator converges to I:

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} u(X_i) \to \mathbb{E}_p[u(X)] = \int_a^b u(x)p(x)\mathrm{d}x = I.$$

This gives us a way to evaluate integrals by generating random numbers. For multi-dimensional integrals, this method has important advantages. Namely, the generation of random numbers followed by estimation of I is less computationally intensive than the direct numerical integration (by quadratures) of the multi-dimensional integral.

There is no special reason to pick the uniform distribution. In fact, any distribution $p(x)$ can be used. In some cases, special choices of $p(x)$ may be advantageous. For example, sampling the domain $[a, b]$ may be a waste of time if most of the points within that interval correspond to regions where $h(x)$ is zero or nearly zero. In that case, we instead want to sample regions of the domain where $|h(x)| > 0$ is concentrated.

In other words, let

$$I = \int_a^b h(x)\mathrm{d}x = \int_a^b \frac{h(x)}{p(x)} p(x)\mathrm{d}x = \mathbb{E}_p[u(X)], \qquad u(x) = \frac{h(x)}{p(x)}$$

and choose $p(x)$ such that the "peaks" of p correspond to the peaks of h. The numerical value of I can be estimated by sampling iidrv $X_1, \ldots, X_n$ according to $p(x)$ and computing the sum:

$$I \approx \frac{1}{n} \sum_{i=1}^{n} u(X_i) = \frac{1}{n} \sum_{i=1}^{n} \frac{h(X_i)}{p(X_i)}, \qquad X_i \sim p(x).$$

2.25.4. Change of Distribution. Suppose that X is a rv with PDF $p(x)$ and we want to calculate the average of a function, $f(X)$ of X. Let $q(x) > 0$ be another PDF on the same probability space as p. Then,

$$\mathbb{E}_p[f(X)] = \int f(x)p(x)\mathrm{d}x = \int \frac{f(x)p(x)}{q(x)} q(x)\mathrm{d}x = \mathbb{E}_q\left(\frac{f(X)p(X)}{q(X)}\right).$$

Here, $\mathbb{E}_p[f(X)]$ denotes the mathematical expectation of f calculated using the PDF $p(x)$ for X, whereas $\mathbb{E}_q\left(\frac{f(X)p(X)}{q(X)}\right)$ is the expectation of fp/q calculated by associating the PDF $q(x)$ to X instead.

In the first case, $X_1, \ldots, X_n$ random numbers are sampled from the distribution $p(x)$ and the integral is estimated as:

$$\mathbb{E}_p[f(X)] \approx \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Convergence is assured by the LLN. In the second case, $X_1, \ldots, X_n$ are sampled from the distribution $q(x)$ and the integral is estimated as

$$\mathbb{E}_q\left(\frac{f(X)p(X)}{q(X)}\right) \approx \frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)p(X_i)}{q(X_i)}.$$

The correction factor p/q is called the likelihood ratio. This method, of course, requires us to be able to compute the ratio p/q for any value of X that we may sample. The idea is to choose an importance distribution q that leads to faster convergence than the nominal distribution p. We generally want to choose q such that its spikes correspond to those of fp; in fact, it can be shown that q should be proportional to fp. Choosing q can be done using an "educated guess" or by random sampling of the function.

2.25.5. Calculation of Probabilities. Since probabilities are expectation values of indicator functions, the WLLN can also be used to speed up the calculation of probabilities. This is especially useful for rare events. For example, suppose we want to calculate the probability of an event $\{X \in A\}$, where the PDF for X is $p(x)$. From experimental measurements $X_1, \ldots, X_n$, this would normally be approximated by

$$\mathbb{P}(X \in A) = \mathbb{E}_p[\mathbf{1}_{X\in A}] = \int_{\mathbb{R}} \mathbf{1}_A(x)p(x)\mathrm{d}x \approx \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_A(X_i) = \frac{\#\text{ draws in A}}{n}$$

where $\mathbf{1}_A(x)$ is the indicator function of A, i.e. it is a function that equals 1 when $x \in A$ and 0 otherwise. However, if $p(x)$ is such that this event rarely happens, we are going to need n very large or else the result will be zero.

On the other hand, the WLLN enables us to reweigh the integral, if we can find a better distribution $q(x)$ that samples values that are closer to the set A:

$$\mathbb{E}_p[\mathbf{1}_{X\in A}] = \mathbb{E}_q\left[\frac{\mathbf{1}_A(X)p(X)}{q(X)}\right] \approx \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{1}_A(X_i)p(X_i)}{q(X_i)}. \qquad X_i \sim q(x)$$

If the event is rare, the ratio p/q will be small (yielding the correct probability for the rare event), whereas the summation will count several non-zero terms, giving a more accurate answer (for the same n).

A special case of $\mathbb{P}(X \in A)$ is the estimation of the CDF, $\mathbb{P}(X < x)$, which can be expressed as $\mathbb{E}_p[\mathbf{1}_{\{X<x\}}]$, and which equals to $\mathbb{E}_q\left[\frac{\mathbf{1}_{\{X<x\}}p(X)}{q(X)}\right]$. And if $q(x)$ is a better distribution than $p(x)$, we can use the latter formula, together with the WLLN, to approximate the CDF by a summation.

2.25.6. Generalization to d-Dimensions. All of the above formulas are valid in d-dimensions. $x \in \mathbb{R}$ and $X \in \mathbb{R}$ are replaced by $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{X} \in \mathbb{R}^d$,

respectively. Likewise, $\mathrm{d}x$ stands for $\mathrm{d}^d\mathbf{x}$, the d-dimensional volume element in the d-dimensional integral.

2.26. Comparing Distributions

Given some random samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ of two rv's X and Y, respectively. It is natural to compare them to see if there is a difference between them. We have already mentioned that one may compare sample means, sample variances and sample covariance. However, these statistical quantifiers are the lowest order moments (first, second) of the distributions. They do not provide a complete comparison. Two rv's are identical if and only if their distributions match. To compare distributions, we must use distance metrics. In this section we discuss a number of popular methods: the Kolmogorov-Smirnov test, the cross entropy, the Bhattacharyya distance, Wasserstein metric and the Kullback-Leibler divergence.

In mathematics a metric on a set X is a function $d : X \times X \to [0, \infty)$ that obeys the following conditions for all $x, y, z \in X$: 1) $d(x, y) = 0$ if and only if $x = y$. 2) $d(x, y) = d(y, x)$ (symmetry). 3) $d(x, y) \leq d(x, z) + d(z, y)$ (triangle inequality).

2.26.1. Kolmogorov-Smirnov test.

2.26.2. Cross entropy.

2.26.3. Bhattacharyya distance.

2.26.4. Wasserstein metric.

2.26.5. Kullback-Leibler divergence and the Relative Entropy.

2.26.5.1. *Entropy.* Suppose we have a rv X taking values x in the set $\mathcal{X}$ each with probability $p(x)$. The Shannon entropy

$$H[X] = -\sum_{x \in \mathcal{X}} \mathbb{P}(X = x) \log \mathbb{P}(X = x) \qquad \text{(discrete rv } X\text{)}$$

or

$$H[X] = -\int p(x) \log p(x) \mathrm{d}x \qquad \text{(continuous rv } X\text{)}$$

quantifies the "lack" of information about the system described by $p(x)$. For example, if we have a system that can be found in 6 possible states with probabilities $(1, 0, 0, 0, 0, 0)$, the entropy is lowest ($H = 0$). On the other hand, if the probability distribution is $(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$ (uniform distribution) the entropy is maximal. The uniform distribution shows the system can be found in any of its 6 states with equal probability; therefore we do not know anything about its state. In the case of the first distribution we know exactly which state the system is in (the first state). If entropy

quantifies the lack of information, the negative of the entropy quantifies information.

There are other measures of entropy. The Renyi entropy measure is

$$H_\alpha^R[X] = \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} \mathbb{P}(X = x)^\alpha.$$

The Tsallis entropy measure is

$$H_\alpha^T[X] = c \frac{\sum_{x \in \mathcal{X}} \mathbb{P}(X = x)^\alpha - 1}{1-\alpha}.$$

Here, $\alpha > 0$ is a positive parameter and c is a positive constant depending on the particular units used. Both of these families include the Shannon measure as a special case in the limit $\alpha \to 1$, where $H_1^R(p) = H_1^T(p) = H(p)$. We may interchangeably write $H(p)$ for $H[X]$ and vice versa, since X is defined by its distribution p. From this, we see that

$$H(p) = -\frac{1}{n} \log p(x^n)$$

where $p(x^n) = \prod_{i=1}^n p(x_i)$. This expression for $H(p)$ is called the empirical entropy of the empirical probability distribution.

2.26.5.2. *Empirical entropy.* The above definitions presume that we know the distributions. Suppose that instead we have data $x_1, x_2, \ldots, x_n$ all taking values in the discrete set $\mathcal{X}$. The empirical PMF is:

$$p(x) = \frac{1}{n} \sum_{i=1}^n \delta_x(x_i),$$

where $\delta_x(x_i)$ is the Kronecker delta function and $x \in \mathcal{X}$. Using the definition of entropy:

$$H(p) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -\sum_{x \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^n \delta_x(x_i) \log p(x) = -\frac{1}{n} \sum_{i=1}^n \log p(x_i).$$

In the last step we have interchanged the order of the two sums and used

$$\sum_{x \in \mathcal{X}} \delta_x(x_i) \log p(x) = \log p(x_i).$$

2.26.5.3. *KL Divergence.* Suppose that we have two PDFs $q(x)$ and $p(x)$ defined on the same probability space (i.e. the range of values is the same $x \in \mathcal{X}$, and the set of all possible random events is identical) with PDFs $q(x)$ and $p(x)$. Here we assume that the range is $\mathcal{X} = (-\infty, \infty)$. The relative entropy between q and p is defined by:

$$D_{KL}[p(x) : q(x)] = \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) \mathrm{d}x.$$

This can easily be generalized to discrete rv's by taking

$$p(x) = \sum_{i=1}^{N} p_i \delta(x - x_i), \qquad q(x) = \sum_{i=1}^{N} q_i \delta(x - x_i),$$

where N is the number of possible values $x \in \mathcal{X}$ the rv can take. This gives:

$$D_{KL}[p : q] = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right) = \sum_{i=1}^{N} p_i \log \left(\frac{p_i}{q_i} \right).$$

We note that D_{KL} is not symmetric, i.e. $D_{KL}[p : q] \neq D_{KL}[q : p]$, and nor does it satisfy the triangle inequality. Therefore, it is not technically a metric. It is possible to make it symmetric by taking the sum $D_{KL}[p : q] + D_{KL}[q : p]$ in order to obtain a metric.

2.26.5.4. *Relationship to cross-entropy.* Cross-entropy is closely related to relative entropy or KL-divergence that computes distance between two probability distributions. For example, in between two discrete PMFs, the relation between them is:

$$H(p, q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x) \qquad \text{cross entropy}$$

$$H(p) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \qquad \text{entropy}$$

$$D_{KL}[p : q] = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right) \qquad \text{relative entropy}$$

$$H(p, q) = H(p) + D_{KL}[p : q]$$

Expressing the KL divergence in the form

$$D_{KL}[p : q] = \sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} p(x) \log q(x)$$

yields the interpretation of the KL divergence to be something like the following: if P is the "true" distribution, then the KL divergence is the amount of information "lost" when expressing it via Q.

2.26.6. Density Estimation. So far we have worked dealt with *parametric statistics* meaning that we assumed knowledge of the PDF in order to compute statistics involving rv's. For example, concepts such as mean and variance were defined in terms of PDFs. The PDF is either given to us, or it is estimated from the data by fitting its parameters (e.g. mean, variance) to the histogram. This procedure has limited capabilities, as it requires choosing a model for the PDF. *Non-parametric statistics* makes no assumptions about the form of the PDF. The density function (PDF) is instead derived

from the data. Recall that (rescaled) histograms are a discrete approximation to the PDF. In this section, we will show that non-parametric estimates of the density can be constructed using a sum of kernel functions.

2.26.6.1. *Kernels.* Kernels are best described informally as "bump functions". An example is the Gaussian function, also known as *radial basis function*

$$K_x(y) = K(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}},$$

which is a fundamental solution to the heat equation describing the response to a point source of heat in thermodynamics. Another example is the polynomial kernel:

$$K_x(y) = K(x, y) = (x \cdot y + 1)^d.$$

Kernels in statistics must be non-negative, real-valued integrable functions $K_x : \mathbb{R} \to \mathcal{X}$ satisfying symmetry, $K_x(-y) = K_x(y)$ and normalization, $\int_{-\infty}^{\infty} K_x(y)\mathrm{d}y = 1$.

A *reproducing kernel* K_x operates on a Hilbert space H of functions that are defined on a set X. A function $K : X \times X \to \mathbb{R}$ defined by the inner product on H:

$$K(x, y) = \langle K_x, K_y \rangle_H$$

that has the property of taking a function f and evaluating it at x:

$$\langle f, K_x \rangle_H = f(x),$$

is called *reproducing* because it maps a function f to its value $f(x)$. An example is the Dirac measure $\delta_x(y)$ and the Hilbert space $L^2(\mathbb{R})$:

$$\langle f, \delta_x \rangle_{L^2} = \int_{-\infty}^{\infty} f(y) \cdot \overline{\delta_x(y)}\mathrm{d}y = \int_{-\infty}^{\infty} f(y)\delta(x - y)\mathrm{d}y = f(x).$$

2.26.6.2. *Kernel Density Estimation.* Kernel density estimation (KDE) is method for estimating the probability density function of a rv. It can also be viewed as a data smoothing technique where inferences about the population are made (PDF), based on a finite data sample (histogram).

Let $(x_1, x_2, \ldots, x_n)$ be independent and identically distributed samples drawn from some univariate distribution with an unknown density f at any given point x. We are interested in estimating the shape of this function f. Its kernel density estimator is

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big)$$

where K is the kernel — a non-negative function — and $h > 0$ is a smoothing parameter called the bandwidth. We note that $K_x(y)$ in the previous section is now denoted $K(y - x)$.

A kernel with subscript h is called the scaled kernel and defined as $K_h(x) = (1/h) \cdot K(x/h)$. The KDE can be though of as a weighted average, where the weight is:

$$w_i = \frac{1}{h} K\Big(\frac{x - x_i}{h}\Big).$$

The choice of bandwidth h matters in practice. Wider bandwidths smooth out the data more (low variance). Narrower bandwidths result in noisier data (high variance). Obviously, if we pick too low a bandwidth, the density estimation has a generally greater bias because the moving average (trendline) is less responsive to changes in the data points.

Suppose that we measure a signal Y_i that is the sum of the underlying signal $f(x_i)$ and some additive noise ξ_i:

$$Y_i = f(x_i) + \xi_i$$

where one usually assumes that

$$\xi_i \sim \mathcal{N}(0, \sigma^2).$$

Here, x_i represents some internal variables that are not directly measured. We denote them as x_i rather than X_i, to emphasize that those variables have already been "fixed" at the time of the measurement, i.e.

$$f(x_i) = \mathbb{E}[f(X_i)|X_i = x_i].$$

Taking the conditional expectation given $X_i = x_i$ we find:

$$\mathbb{E}[Y_i|X_i = x_i] = \mathbb{E}[f(X_i)|X_i = x_i].$$

We will obtain in the next section an expression for $\mathbb{E}[Y_i|X_i = x_i]$.

2.26.6.3. *Kernel Regression.* The problem of kernel regression can be summarized as follows. We want to estimate the conditional expectation $\mathbb{E}[Y|X = x]$, which is a function of x. First note that:

$$\mathbb{E}[Y|X = x] = \int y p_Y(y|x) \mathrm{d}y = \int y \frac{p_{XY}(x, y)}{p_X(x)} \mathrm{d}y.$$

However, this requires knowledge of the densities. We use the following kernel density estimates:

$$\hat{p}_{XY}(x, y) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) K_h(y - y_i), \quad \hat{p}_X(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i),$$

where the hat denotes estimate. We get the following estimate:

$$\begin{aligned}\hat{\mathbb{E}}[Y|X=x] &= \int \frac{y\sum_{i=1}^{n} K_h(x-x_i)K_h(y-y_i)}{\sum_{j=1}^{n} K_h(x-x_j)}\mathrm{d}y,\\ &= \frac{\sum_{i=1}^{n} K_h(x-x_i)\int yK_h(y-y_i)\mathrm{d}y}{\sum_{j=1}^{n} K_h(x-x_j)},\\ &= \frac{\sum_{i=1}^{n} K_h(x-x_i)y_i}{\sum_{j=1}^{n} K_h(x-x_j)},\end{aligned}$$

We have used the reproducing property of kernels:

$$\int yK_h(y-y_i)\mathrm{d}y = y_i.$$

2.27. Problems

Problem 20. The height of a person is measured over time, every month from birth to The data set consists of the person's age in months and her height in centimeters. The summary statistics for the data are provided below:

x = age, in months

y = height, in centimeters

$\overline{x} = 44 \qquad s_x = 8.5 \qquad \overline{y} = 82 \qquad s_y = 4.1$

Also, the correlation coefficient between x and y is $r = 0.860$

(a) What is the slope of the LSRL? (Round to the nearest hundredth.)
(b) What is the y-intercept of the LSRL? (Round to the nearest hundredth.)
(c) Find the equation of the least-squares regression line (with y as the response variable)
(d) What percentage of the variation in predicted height can be explained for by the LSRL

Problem 21. Suppose that we have an amplifier that takes a voltage and amplifies it by a factor of 10×, i.e. $f(x) = 10x$. Suppose that we feed this amplifier the following input voltages:

$$X = \{2.53, 2.55, 2.45, 2.49, 2.50, 2.52, 2.47, 2.48, 2.56, 2.49\}$$

(a) What is the sample variance at the *output* of the amplifier?

(b) Suppose that we have two rv's, X and Y and they are statistically independent. Furthermore, suppose that $var(X)$=2.7 and $var(Y)$=2.5. Compute the value of $var(X+Y)$ and $var(X-Y)$.

(c) Given that $var(X)$=2.7, $var(Y)$=2.5 and $\rho(X,Y)$=0.9 (correlation coefficient), what is $var(X+Y)$ and $var(X-Y)$?

(d) If X, Y and Z are statistically independent and $var(X)$=1.7, $var(Y)$=2.3 and $var(Z)$=1.4. What is $var(0.3X + 0.7Y + 0.5Z)$?

Solution. (a) First, calculate the sample variance of X and then multiply by 100. Then multiply each X_i by 10 and then calculate the sample variance of the multiplied values. The sample variance of the X's is 0.001249. Multiply this by 100 to get 0.1249. Multiplying each X by 10 and taking the sample variance we get 0.1249, which is the same as the first method. From this we confirmed the validity of the formula $var(aX) = a^2 var(X)$.

(b) By statistical independence we have

$$var(X + Y) = var(X) + var(Y) = 2.7 + 2.5 = 5.2.$$

Then from

$$var(aX + bY) = a^2 var(X) + b^2 var(Y),$$

with $a = 1$ and $b = -1$ we have

$$var(X-Y) = (1)^2 var(X)+(-1)^2 var(Y) = var(X)+var(Y) = 2.7+2.5 = 5.2.$$

From this, we conclude that when X and Y are statistically independent, $var(X + Y) = var(X - Y) = var(X) + var(Y)$.

(c) From the definition of the correlation coefficient

$$\rho(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}}$$

we have $cov(X, Y) = \rho(X, Y)\sqrt{var(X)var(Y)}$. Then, inserting this into:

$$var(X + Y) = var(X) + var(Y) + 2cov(X, Y)$$

we get:

$$var(X + Y) = var(X) + var(Y) + 2\rho(X, Y)\sqrt{var(X)var(Y)}$$

from which we can obtain a numerical value:

$$var(X + Y) = 2.7 + 2.5 + 2(0.9)[(2.7)(2.5)]^{1/2} = 9.877.$$

For $var(X - Y)$ we have:

$$var(X - Y) = var(X) + var(-Y) + 2\rho(X, -Y)\sqrt{var(X)var(-Y)}$$

and since $var(-Y) = (-1)^2 var(Y) = var(Y)$, we have:

$$var(X - Y) = var(X) + var(Y) + 2\rho(X, -Y)\sqrt{var(X)var(Y)}$$

Now, using the property $cov(aX, bY) = ab \cdot cov(X, Y)$, we see that

$$\rho(aX, bY) = \frac{cov(aX, bY)}{\sqrt{var(aX)var(bY)}} = \frac{ab \cdot cov(X, Y)}{|a||b|\sqrt{var(X)var(Y)}} = \frac{ab}{|a||b|}\rho(X, Y).$$

For $a = 1$ and $b = -1$,

$$\rho(1 \cdot X, -1 \cdot Y) = \frac{(1) \cdot (-1)}{|1||-1|}\rho(X, Y) = -\rho(X, Y).$$

Hence,

$$var(X - Y) = var(X) + var(Y) - 2\rho(X, Y)\sqrt{var(X)var(Y)}$$

and then we have

$$var(X - Y) = 2.7 + 2.5 - 2(0.9)[(2.7)(2.5)]^{1/2} = 0.5235.$$

We conclude that $var(X + Y)$ is not equal to $var(X - Y)$ when X and Y are correlated to some extent.

(d) If X, Y and Z are statistically independent, then we can write:

$$var(aX + bY + cZ) = a^2var(X) + b^2var(Y) + c^2var(Z),$$

where a, b, c are constants. Hence,

$$\begin{aligned} var(0.3X + 0.7Y + 0.5Z) &= (0.3)^2var(X) + (0.7)^2var(Y) + (0.5)^2var(Z) \\ &= (0.3)^2 1.7 + (0.7)^2 2.3 + (0.5)^2 1.4 = 1.630. \end{aligned}$$

■

Problem 22. Suppose that X is a rv with distribution $p_X(x)$ and $Y = g(X)$ is another rv related to X via a continuous differentiable function g. Prove that the density of Y can be written as:

$$p_Y(y) = \int_{-\infty}^{\infty} p_X(x)\delta(y - g(x))\mathrm{d}x.$$

Solution. Starting with the CDF:

$$\mathbb{P}(Y < a) = \mathbb{P}(g(X) < a) = \int\limits_{\{x:g(x)<a\}} p_X(x)\mathrm{d}x = \int_{-\infty}^{\infty} \mathbf{1}_{g(x)<a}(x)p_X(x)\mathrm{d}x.$$

Using the fact that the Dirac delta function is the derivative of the Heaviside function:

$$\delta(x) = \frac{\mathrm{d}}{\mathrm{d}x}\theta(x), \qquad \theta(x) := \mathbf{1}_{x>0}(x)$$

And if the origin is shifted to x_0, we may change variables to $x = \tilde{x} - x_0$:

$$\delta(\tilde{x} - x_0) = \frac{\mathrm{d}}{\mathrm{d}\tilde{x}}\theta(\tilde{x} - x_0), \qquad \theta(\tilde{x} - x_0) := \mathbf{1}_{\tilde{x}>x_0}(\tilde{x})$$

Taking the derivative with respect to a we get the PDF, $p_Y(a)$:

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}a}\mathbf{1}_{a>g(x)}(x)p_X(x)\mathrm{d}x = \int_{-\infty}^{\infty} \delta(a - g(x))p_X(x)\mathrm{d}x.$$

■

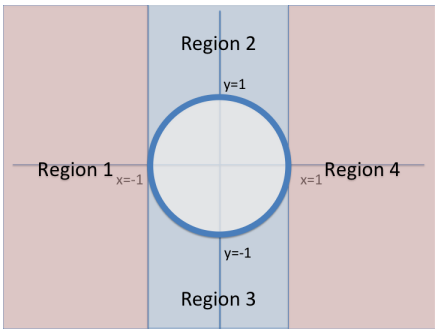


Figure 2.10. Partition of the region $x^2 + y^2 > 1$ into 4 regions.

Problem 23. Find the probability distribution function of the rv $Z = X^2 + Y^2$ in terms of the distribution of X and Y.

Solution. The probability that the vector (X, Y) lies outside the unit circle $\{(x, y)|x^2 + y^2 = 1\}$ is:

$$\mathbb{P}(X^2 + Y^2 > 1) = \iint\limits_{\{(x,y)|x^2+y^2>1\}} p_{XY}(x, y)\mathrm{d}x\mathrm{d}y.$$

This can be calculated explicitly by splitting the integration domain ($\mathbb{R}^2 -$ {unit disc}) into 4 regions (Fig. 2.10).

$$\mathbb{P}(X^2+Y^2 > 1) = \underbrace{\int_{-\infty}^{\infty}\left(\int_{-\infty}^{-1} p_{XY}(x, y)\mathrm{d}x\right)\mathrm{d}y}_{\text{Region 1}} + \underbrace{\int_{-\infty}^{\infty}\left(\int_{1}^{\infty} p_{XY}(x, y)\mathrm{d}x\right)\mathrm{d}y}_{\text{Region 4}}$$
$$+ \underbrace{\int_{-1}^{1}\left(\left\{\int_{-\infty}^{-\sqrt{1-x^2}} + \int_{\sqrt{1-x^2}}^{\infty}\right\} p_{XY}(x, y)\mathrm{d}y\right)\mathrm{d}x}_{\text{Regions 2 and 3}}.$$

Another way to calculate this would be to convert $p_{XY}(x, y)$ to polar coordinates $p_{R,\Theta}(r, \theta)$ and integrate from $r = 1$ to ∞ while letting θ range from 0 to 2π. ■

Problem 24. Find the probability distribution function of the rv $Z = \sqrt{X^2 + Y^2}$ in terms of the distribution of X and Y.

Solution. Suppose that $X, Y \sim \mathcal{N}(0, \sigma^2)$ (zero-mean Gaussians) are independent rv's. Consider the transformation to polar coordinates:

$$R = \sqrt{X^2 + Y^2}, \quad \Theta = \tan^{-1}(Y/X).$$

The inverse transformation is:

$$x = r\cos\theta, \quad y = r\sin\theta.$$

What is the distribution of Θ and R? Let us do R. The CDF of R is found by writing:

$$\mathbb{P}(R < r) = \iint\limits_{\{(x,y)|\sqrt{x^2+y^2}<r\}} \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/(2\sigma^2)} \mathrm{d}x\mathrm{d}y.$$

It will be convenient to transform to polar coordinates. The Jacobian of the transformation yields the new area element:

$$\mathrm{d}x\mathrm{d}y = \left|\frac{\partial(x,y)}{\partial(r,\theta)}\right| \mathrm{d}r\mathrm{d}\theta,$$

where

$$\frac{\partial(x,y)}{\partial(r,\theta)} = \begin{Vmatrix} \partial_r x & \partial_\theta x \\ \partial_r y & \partial_\theta y \end{Vmatrix} = \begin{Vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{Vmatrix} = r.$$

Then,

$$\mathbb{P}(R < r) = \frac{1}{2\pi\sigma^2} \int_0^{2\pi} \mathrm{d}\theta \int_0^r \mathrm{d}r e^{-r^2/(2\sigma^2)} \cdot r = \frac{1}{\sigma^2} \int_0^r \mathrm{d}r e^{-r^2/(2\sigma^2)} \cdot r.$$

The corresponding PDF is obtained by differentiating with respect to r:

$$\boxed{p_R(r) = \frac{r}{\sigma^2} e^{-r^2/(2\sigma^2)}.}$$

We have recovered the Rayleigh distribution, by constructing the rv $R = \sqrt{X^2 + Y^2}$, where $X, Y \sim \mathcal{N}(0, \sigma^2)$.

The derivation of the distribution for Θ is trivial. Recall that $\Theta = \tan^{-1}(Y/X)$. Then,

$$\mathbb{P}(\Theta < \theta) = \iint\limits_{\{(x,y)|\tan^{-1}(y/x)<\theta\}} \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/(2\sigma^2)} \mathrm{d}x\mathrm{d}y.$$

Transformation to polar coordinates gives:

$$\mathbb{P}(\Theta < \theta) = \int_0^\theta \mathrm{d}\theta \int_0^\infty \mathrm{d}r \frac{1}{2\pi\sigma^2} e^{-r^2/(2\sigma^2)} r.$$

The integral over r can be solved with the substitution $w = r^2/(2\sigma^2)$, $\mathrm{d}w = r\mathrm{d}r/\sigma^2$. Thus, our CDF is:

$$\mathbb{P}(\Theta < \theta) = \frac{1}{2\pi} \int_0^\theta \mathrm{d}\theta = \frac{\theta}{2\pi},$$

where $\theta \in [0, 2\pi]$. The PDF is that of a uniform distribution:

$$\boxed{p_\Theta(\theta) = \frac{1}{2\pi},}$$

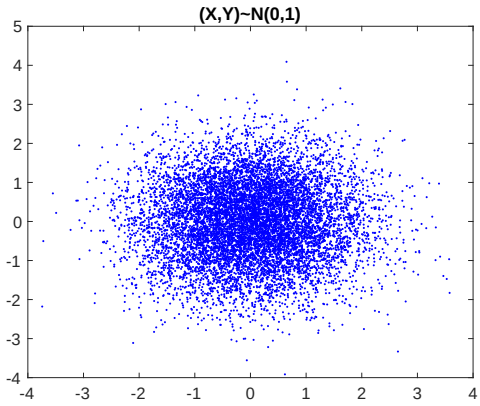


Figure 2.11. Scatter plot of the ordered pairs $\{(X_i, Y_i)\}_{i=1}^{10,000}$, where $X_i, Y_i \sim \mathcal{N}(0,1)$ are all independent.

with $\theta \in [0, 2\pi]$. Thus, R is Rayleigh whereas Θ is uniform. This can be seen in Fig. 2.11, which is a scatter plot of the pairs (X, Y), where $X, Y \sim \mathcal{N}(0,1)$. This plot was generated in MATLAB as follows:

```
>> X=randn([1 10000]); Y=randn([1 10000]);
>> figure;plot(X,Y,'.b');title('(X,Y)~N(0,1)');
>> set(gca,'fontsize',16);
```

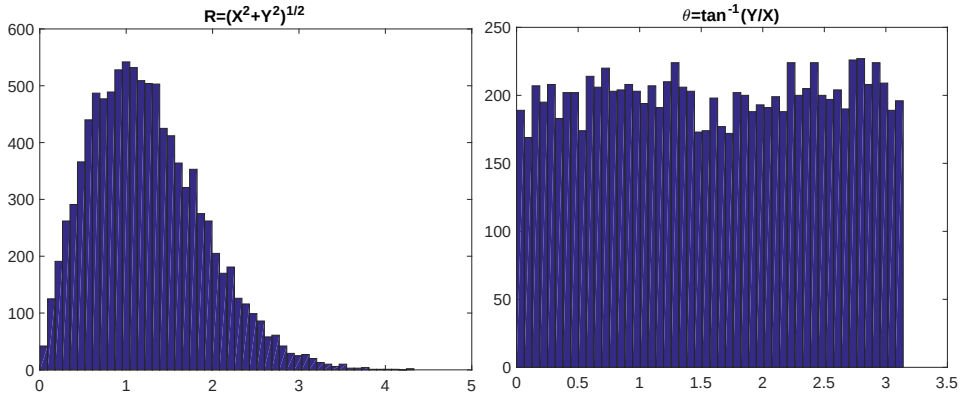
The distributions of R and Θ can be plotted by taking the pairs (X, Y) and generating R, Θ. Histograms of R and Θ are shown in Fig. 2.12. It can be seen that R is Rayleigh and Θ is uniform. These plots were generated in MATLAB using the following commands:

```
>> R=sqrt(X.^2+Y.^2); theta=atan(Y./X);
>> figure;hist(R,50);
>> set(gca,'fontsize',16);
>> title('R=(X^2+Y^2)^{1/2}');
>> figure;hist(theta+pi/2,50);set(gca,'fontsize',16);
>> title('\theta=tan^{-1}(Y/X)');
```

■

Problem 25. The median of a finite list of numbers is the "middle" number, when those numbers are listed in order from smallest to greatest. (A set of an even number of observations has no distinct middle value and the median is usually defined to be the arithmetic mean of the two middle values.)

(a) Prove that given a random sample $x_1, \ldots, x_n$ (take n as odd, so there is a middle value) of a rv X, the median is the value x_{50} that is the middle data point in the ordered list of the random sample.



(a) R (b) Θ

Figure 2.12. Histograms of R and Θ, as generated from a sequence of rv's $\{(X_i, Y_i)\}_{i=1}^{10,000}$, where $X_i, Y_i \sim \mathcal{N}(0, 1)$ are all independent.

(b) Explain the relationship between median and mean. When would you use one vs the other?

Solution. (a) The median is defined as the value x_{50} satisfying:

$$\int_{-\infty}^{x_{50}} p(x)\mathrm{d}x = \frac{1}{2}.$$

Substituting the empirical distribution

$$p(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i),$$

into the definition of median:

$$\frac{1}{2} = \int_{-\infty}^{x_{50}} \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)\mathrm{d}x = \frac{1}{n}(n/2),$$

i.e. for this integral to equal $1/2$ it must evaluate to $(n/2)/n = 1/2$. In other words, half the terms in the summation contribute. Which terms? The integral is over the range $(-\infty, x_{50}]$, i.e. begins at $-\infty$ and ends at x_{50}. Integration will therefore pick out all the terms labeled x_i that are found in the interval $(-\infty, x_{50}]$. Each term is a Dirac delta function that integrates to 1. Thus, it is a counter of sorts. Once we have found the midway point of the ordered list, the corresponding value x_{50} is called the median.

(b) The median, like the mean, attempts to produce some kind of average of a random sample. The media ignores the extreme and outlier values since it only picks the central value. The mean is affected by outliers. ■

Problem 26. We have learned that given two independent rv's X and Y, we may form a new rv Z that is the sum of X and Y, i.e. $Z = X + Y$, and that the probability density of Z is the convolution of the densities of X and Y, i.e.

$$p_{X+Y}(a) = \int_{-\infty}^{\infty} p_X(a-y)p_Y(y)\mathrm{d}y$$

or in terms of CDFs:

$$\begin{aligned}\mathbb{P}(X+Y \le a) &= \iint\limits_{\{(x,y):x+y\le a\}} p_X(x)p_Y(y)\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{a-y} p_X(x)\mathrm{d}x\right)p_Y(y)\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\mathbb{P}(X \le a-y)p_Y(y)\mathrm{d}y\end{aligned}$$

Please note: limits of integration $(-\infty, \infty)$ should be replaced by the domain of definition of the rv if different from $(-\infty, \infty)$.

(a) Suppose that X and Y are independent and let $X \sim Uni(0,1)$, $Y \sim Uni(0,1)$ (uniformly distributed over the interval [0,1]), i.e. PDF is $p_X(x) = 1$ for $0 \le x \le 1$ and same for $p_Y(y)$. What is the PDF of $X + Y$?

(b) Show that if $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ then $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

(c) Suppose that you play 2 lotteries. In the first lottery you either win \$1000 with probability 1/2 or lose (with probability 1/2) and get nothing. In the second lottery you are guaranteed of winning *something*; however the payout is less: the payout follows a Rayleigh distribution with mode equal to \$100. (The Rayleigh PDF is $p_Y(y) = (y/\sigma^2)e^{-x^2/(2\sigma^2)}$, where σ is the mode, the mean is $\sigma\sqrt{\pi/2}$.). You can assume that X and Y are independent. What is the PDF describing the total payout from both lotteries? Plot the PDF. What is the average amount you'd expect to win?

(d) Consider Newton's law, $F = ma$, where m is mass and a is acceleration. Given the distributions of m and a as $\mathcal{N}(10,1)$ and $\mathcal{N}(10, 0.1)$, respectively. What is the distribution of F?

(e) Find the mode of the following PDF, which approximates the thumb length X in inches in a particular country:

$$p(x) = \begin{cases} \frac{\pi}{4}\sin\left(\frac{\pi(x-2)}{2}\right), & 2 \le x \le 4 \\ 0, & \text{elsewhere} \end{cases}$$

Solution. (a) The convolution is

$$p_{X+Y}(a) = \int_0^1 p_X(a-y)p_Y(y)\mathrm{d}y = \int_0^1 p_X(a-y)\mathrm{d}y = \int_0^1 \mathbf{1}_{[0,1]}(a-y)\mathrm{d}y,$$

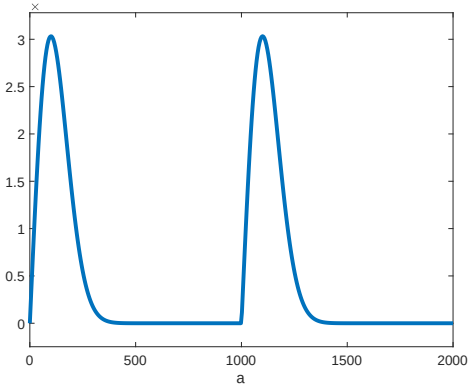
where $\mathbf{1}_A(y)$ is the indicator function over the set A. The latter results in an integral equal to zero unless $a-y \in [0,1]$, or $-y \in [0,1]-a$ or $y \in a+[-1,0]$. The overlap between $a+[-1,0]$ and the limits of integration $[0,1]$ can be split into 2 regions: $a \in [0,1]$ and $a \in [1,2]$. In the first region the overlap progressively increases; in the second region it decreases. Performing the integral we obtain the tent function:

$$p_{X+Y}(a) = \int_0^1 \mathbf{1}_{[0,1]}(a-y)\mathrm{d}y = \begin{cases} \int_0^a \mathrm{d}y = a & 0 \le a \le 1 \\ \int_{a-1}^1 \mathrm{d}y = 2-a & 1 < a \le 2 \\ 0 & \text{otherwise} \end{cases}$$

(c) Let $Z = X + Y$ and $p_X(x) = p_0\delta(x-x_l) + p_1\delta(x-x_w)$, with $x_l = \$0$ and $x_w = \$10^3$. $p_Y(y)$ is given to us with $\sigma = \$100$. The PDF of Z is the convolution ($a > 0$):

$$\begin{aligned} p_{X+Y}(a) =& \int_0^\infty p_X(a-y)p_Y(y)\mathrm{d}y \\ =& \int_0^\infty [p_0\delta(a-y) + p_1\delta(a-y-x_w)]\frac{y}{\sigma^2}e^{-y^2/(2\sigma^2)}\mathrm{d}y \\ =& p_0\frac{a}{\sigma^2}e^{-a^2/(2\sigma^2)} + p_1\frac{(a-x_w)}{\sigma^2}e^{-(a-x_w)^2/(2\sigma^2)}\mathbf{1}_{[x_w,\infty)}(a), \end{aligned}$$

where $\sigma = \$100$, $x_w = \$10^3$, $p_1 = 1/2$ and $p_0 = 1/2$.



The mean value:

$$\begin{aligned}\mathbb{E}Z =&\frac{p_0}{\sigma^2}\int_0^\infty a^2 e^{-a^2/(2\sigma^2)}\mathrm{d}a + \frac{p_1}{\sigma^2}\int_0^\infty (a-x_w)ae^{-(a-x_w)^2/(2\sigma^2)}\mathbf{1}_{[x_w,\infty)}(a)\mathrm{d}a\\ =&p_0\sigma\sqrt{\frac{\pi}{2}} + \frac{p_1}{\sigma^2}\int_0^\infty a'(a'+x_w)e^{-(a')^2/(2\sigma^2)}\mathrm{d}a'\\ =&p_0\sigma\sqrt{\frac{\pi}{2}} + \frac{p_1}{\sigma^2}(5\times 10^5)(20+\sqrt{2\pi}) \approx \$562.50\end{aligned}$$

which is right about somewhere between the two peaks, as we would expect the average to be, based on the center-of-mass of this PDF. (We have used `wolframalpha.com` to obtain a numerical value for this integral in the last line.)

(d) Let $Z = XY$. The PDF of Z is:

$$p_Z(z) = \int_{-\infty}^{\infty} p_X(x)p_Y(z/x)\frac{1}{|x|}\mathrm{d}x \qquad (*)$$

Plugging in the distributions for X and Y: $\mathcal{N}(10,1)$ and $\mathcal{N}(10,0.1)$, we have:

$$p_Z(z) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi(0.1^2)}}\int_{-\infty}^{\infty} e^{-(x-10)^2/2}e^{-(z/x-10)^2/2(0.1^2)}\frac{1}{|x|}\mathrm{d}x$$

A proof is:

$$\begin{aligned}\mathbb{P}(Z\le z) =&\mathbb{P}(XY\le z) = \mathbb{P}(XY\le z, X>0) + \mathbb{P}(XY\le z, X\le 0)\\ =&\mathbb{P}(Y\le z/X, X>0) + \mathbb{P}(Y\ge z/X, X\le 0)\\ =&\int_0^\infty p_X(x)\int_{-\infty}^{z/x} p_Y(y)\mathrm{d}y\mathrm{d}x + \int_{-\infty}^0 p_X(x)\int_{z/x}^\infty p_Y(y)\mathrm{d}y\mathrm{d}x\end{aligned}$$

Differentiating with respect to z, we get the PDF:

$$\begin{aligned}p_Z(z) =&\int_0^\infty p_X(x)p_Y(z/x)\frac{1}{x}\mathrm{d}x - \int_{-\infty}^0 p_X(x)p_Y(z/x)\frac{1}{x}\mathrm{d}x\\ =&\int_{-\infty}^\infty p_X(x)p_Y(z/x)\frac{1}{|x|}\mathrm{d}x\end{aligned}$$

(e) The mode here can be found by setting the derivative to zero: $p'(x) = 0$. In the nonzero region $2 \le x \le 4$ the derivative of $\frac{\pi}{4}\sin\left(\frac{\pi(x-2)}{2}\right)$ is $p'(x) = \frac{\pi^2}{8}\cos\left(\frac{\pi(x-2)}{2}\right)$. Setting the derivative equal to zero we must solve $\cos\left(\frac{\pi(x-2)}{2}\right) = 0$. Taking the inverse cosine, $\frac{\pi(x-2)}{2} = \frac{\pi}{2} + k\pi$, $k\in\mathbb{Z}$, or $x-2 = 1+2k$ and $x = 3+2k$. The solution in the interval $2\le x\le 4$ is $x = 3$. ■

Problem 27. The Poisson's law with parameter a ($a > 0$) is defined by:

$$\mathbb{P}[k \ events] = e^{-a}\frac{a^k}{k!},$$

where $k = 0, 1, 2, \ldots$. With $a = \lambda\tau$, where λ is the average number of events per units time and τ is the length of the interval $(t, t+\tau)$, the probability of k events in τ is

$$\mathbb{P}(k; t, t+\tau) = e^{-\lambda\tau}\frac{(\lambda\tau)^k}{k!}.$$

This equation assumes that λ is independent of t. If λ depends on t, the product $\lambda\tau$ gets replaced by the integral $\int_t^{t+\tau}\lambda(\xi)\mathrm{d}\xi$, and the probability of k events in the interval $(t, t+\tau)$ is

$$\mathbb{P}(k; t, t+\tau) = \exp\left[-\int_t^{t+\tau}\lambda(\xi)\mathrm{d}\xi\right]\frac{1}{k!}\left[\int_t^{t+\tau}\lambda(\xi)\mathrm{d}\xi\right]^k.$$

The parameter λ is called the rate parameter. $\lambda(t)$ is the rate function. Suppose that a company manufactures superconducting wire. Upon close examination of the product on the assembly line, you find that the defect density along the length of the wire is not uniform. For wire strips of length D, the defect density $\lambda(x)$ along the wire length x varies as

$$\lambda(x) = \lambda_0 + \frac{1}{2}(\lambda_1 - \lambda_0)\left(1 + \cos(\frac{2\pi x}{D})\right), \quad \lambda_1 > \lambda_0$$

for $0 \le x \le D$ due to greater wire contamination at the edges $x = 0$ and $x = D$.

(i) What is the meaning of $\lambda(x)$ in this case?

(ii) What is the average number of defects for a wire strip of length D?

(iii) Find an expression for the probability of k defects on a wire strip of length D?

Solution. (i) Bearing in mind that $\lambda(x)$ is a defect density, i.e., the average number of defects per unit length at x, we conclude that $\lambda(x)\Delta x$ is the average number of defects in the tape from x to $x + \Delta x$.

(ii) Given the definition of $\lambda(x)$ we conclude that the average number of defects along the whole wire is merely the integral of $\lambda(x)$, i.e.,

$$\int_0^D \lambda(x)\mathrm{d}x = \int_0^D \left[\lambda_0 + \frac{1}{2}(\lambda_1 - \lambda_0)\left(1 + \cos\frac{2\pi x}{D}\right)\right]\mathrm{d}x = \frac{\lambda_0 + \lambda_1}{2}D = \Omega.$$

(iii) Assuming the Poisson law holds, use the equation with x and Δx (distances) replacing t and τ (times). Thus,

$$\mathbb{P}(k; x, x+\Delta x) = \exp(-\int_x^{x+\Delta x}\lambda(\zeta)\mathrm{d}\zeta)\frac{1}{k!}(\int_x^{x+\Delta x}\lambda(\zeta)\mathrm{d}\zeta)^k.$$

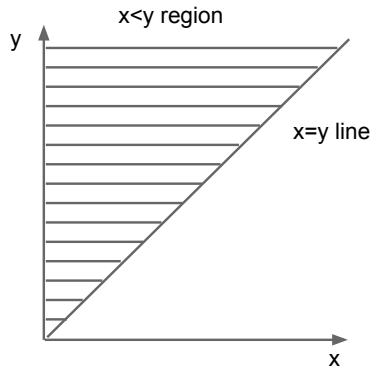


Figure 2.13. Integration in the $x < y$ region.

In particular, with $x = 0$ and $x + \Delta x = D$, we obtain

$$\mathbb{P}(k; 0, D) = \Omega^k \frac{e^{-\Omega}}{k!}$$

■

Problem 28. Let X and Y be independent rv's having the exponential distribution with parameters λ and μ respectively. (Recall that if rv X has exponential distribution with parameter $\lambda > 0$, its CDF is $\mathbb{P}(X < x) \equiv 1 - e^{-\lambda x}$, $x \geq 0$, whose density is $\mathrm{d}\mathbb{P}(X < x)/\mathrm{d}x = \lambda e^{-\lambda x}$.) Let $U = \min\{X, Y\}$, $V = \max\{X, Y\}$ and $W = V - U$. Find the probability $\mathbb{P}(U = X) = \mathbb{P}(X \leq Y)$. Show that U and W are statistically independent.

Solution. First you should realize that the logical statements $U = X$ and $X \leq Y$ mean the same thing. Indeed, $U = \min\{X, Y\}$ less than or equal to X implies that both $X \leq X$ (if $\min\{X, Y\} = X$) and $Y \leq X$ (if $\min\{X, Y\} = Y$). The former ($X \leq X$) is a trivial statement which is true at all times. Thus, it can be ignored. The only non-trivial statement left is $Y \leq X$, hence the equivalence of the two statements $U = X$ and $X \leq Y$. If the two statements are equivalent, then their probabilities are also equal: $\mathbb{P}(U = X) = \mathbb{P}(X \leq Y)$.

$\mathbb{P}(U = X) = \mathbb{P}(X \leq Y)$ can be computed since it is in terms of X and Y whose distributions are known. Since $\mathbb{P}(X \leq Y)$ involves both X and Y we must integrate the joint PDF of X and Y over the set of all points (x, y) such that $x < y$ is satisfied:

$$\mathbb{P}(X \leq Y) = \int_{\{(x,y)|x<y\}} p_{XY}(x, y)\, \mathrm{d}x\mathrm{d}y.$$

Let's integrate along horizontal strips, as shown in Fig. 2.13. Thus,

$$\mathbb{P}(X \leq Y) = \int_0^y \mathrm{d}x \int_0^\infty \mathrm{d}y\, p_{XY}(x, y) = \int_0^y \mathrm{d}x \int_0^\infty \mathrm{d}y\, p_X(x) p_Y(y),$$

where in the second equality we invoked the statistical independence of X and Y and wrote the integrand as a product of densities in X and Y. Now, we invoke the shorthand notation $\mathbb{P}(X \leq y) = \int_0^y p_X(x)\mathrm{d}x$ and use the fact that $p_Y(y) = \mu e^{-\mu y}$ and rewrite this as:

$$\mathbb{P}(X \leq Y) = \int_0^\infty \mathbb{P}(X \leq y)\mu e^{-\mu y}\mathrm{d}y = \int_0^\infty (1 - e^{-\lambda y})\mu e^{-\mu y}\mathrm{d}y = \frac{\lambda}{\mu + \lambda}.$$

For $w > 0$, $\mathbb{P}(U \leq u, W > w) = \mathbb{P}(U \leq u, W > w, X \leq Y) + \mathbb{P}(U \leq u, W > w, X > Y)$.[17] Thus, there are two terms to calculate. For the first one:

$$\begin{aligned}
\mathbb{P}(U \leq u, W > w, X \leq Y) =& \mathbb{P}(X \leq u, Y > X + w) \\
=& \iint_{\{(x,y)|x \leq u, y > x+w\}} p_{XY}(x,y)\,\mathrm{d}x\mathrm{d}y \\
=& \int_0^u \mathrm{d}x\, \lambda e^{-\lambda x} \underbrace{\int_{x+w}^\infty \mathrm{d}y\, \mu e^{-\mu y}}_{-e^{-\mu y}]_{x+w}^\infty} \\
=& \int_0^u \lambda e^{-\lambda x} e^{-\mu(x+w)}\mathrm{d}x \\
=& \frac{\lambda}{\lambda + \mu} e^{-\mu w}(1 - e^{-(\lambda+\mu)u})
\end{aligned}$$

and similarly, $\mathbb{P}(U \leq u, W > w, X > Y) = \frac{\mu}{\lambda+\mu} e^{-\lambda w}(1 - e^{-(\lambda+\mu)u})$. Hence, for $0 \leq u \leq u + w < \infty$, we have an expression which factorizes into the product of a function of u with a function of w. Hence U and W are independent:

$$\mathbb{P}(U \leq u, W > w) = (1 - e^{-(\lambda+\mu)u}) \left(\frac{\lambda}{\lambda + \mu} e^{-\mu w} + \frac{\mu}{\lambda + \mu} e^{-\lambda w} \right).$$

■

Problem 29. A coin is flipped n times. The outcome is a rv X, which can take the value *heads* or *tails* ($X = heads$ or $X = tails$). For n measurements, there are n such rv's (and corresponding results): $X_1, X_2, \ldots, X_n$. The coin is possibly biased. Therefore, the probabilities of each outcome are not necessarily $1/2$. Instead they are given in term of a parameter $-1/2 \leq \theta \leq 1/2$ quantifying the bias:

$$\mathbb{P}(X = heads) = 1/2 + \theta, \qquad \mathbb{P}(X = tails) = 1/2 - \theta.$$

(a) Explain how the numerical value of the bias parameter, θ, can be determined experimentally (empirically) by flipping the coin several times, i.e.

[17] Since the two events $\{U \leq u, W > w, X \leq Y\}$ and $\{U \leq u, W > w, X > Y\}$ are mutually exclusive whereas the event $\{X \leq Y\} \cup \{X > Y\}$ is always true. Recall that two events A and B are mutually exclusive if there is no overlap: $A \cap B = \emptyset$.

find an explicit formula for $\hat{\theta}_n$, the estimator of θ, in terms of $X_1, \ldots, X_n$. Show that, under specific circumstances, $\hat{\theta}_n$ converges to θ in probability, i.e. show that $\mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon) \to 0$, as $n \to \infty$ for any $\epsilon > 0$.

(b) Denote the number of times where you get heads as the result by H (and H is a rv, because its value may differ each time this experiment is done). Write down an explicit expression for H in terms of the experimental data. Find the mathematical expectation of H.

(c) Find the variance of H. For which value(s) of θ is the variance a minimum? A maximum?

(d) Calculate the "signal-to-noise ratio" of H. Explicitly give the dependence of SNR on n.

(e) For a fixed value of n, find the conditions for which the SNR is 1) infinite and 2) undetermined/undefined. Give a physical explanation of those two different situations.

(f) Find the limiting (n large) distribution of H.

Solution. (a) The probability $\mathbb{P}(X = heads)$ can be determined by counting the number of heads, i.e. let f_H be the empirical probability

$$f_H = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i = heads\}},$$

where $\mathbf{1}_{\{X_i = heads\}}$ equals 1 if $X_i = heads$ and 0 otherwise. Taking the mathematical expectation we get

$$\mathbb{E} f_H = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\mathbf{1}_{\{X_i = heads\}},$$

where

$$\mathbb{E}\mathbf{1}_{\{X_i = heads\}} = \sum_{\{x_i : x_i = heads\}} \mathbb{P}(X_i = heads) = \mathbb{P}(X_i = heads)$$

Therefore (the X_i are iidrv, with the same distribution as X),

$$\mathbb{E} f_H = \mathbb{P}(X = heads).$$

By the law of large numbers (LLN), f_H converges to $\mathbb{P}(X_i = heads)$ as n increases. Now, since $\mathbb{P}(X = heads) = 1/2 + \theta$, which is also equal to $\mathbb{E} f_H$, we take our estimator $\hat{\theta}$ to be:

$$\hat{\theta}_n = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i = heads\}} \right) - \frac{1}{2},$$

which implies that $\hat{\theta}_n$ converges to $\mathbb{P}(X = heads) - \frac{1}{2}$, as n increases. However, $\mathbb{P}(X = heads) - \frac{1}{2}$ is also equal to θ, by the LLN. Thus, $\hat{\theta}_n \to \theta$.

(b)

$$H = \sum_{i=1}^{n} \mathbf{1}_{\{X_i = heads\}}$$

where $\mathbf{1}_{X_i=heads}$ equals 1 when $X_i = heads$ and 0 when $X_i = tails$. Taking expectation value:

$$\mathbb{E}H = \sum_{i=1}^{n} \mathbb{E}\mathbf{1}_{\{X_i=heads\}} = \sum_{i=1}^{n} \mathbb{P}(\{X_i = heads\}) = \sum_{i=1}^{n} (1/2+\theta) = n(1/2+\theta).$$

(c) Variance:

$$var(H) = \sum_{i=1}^{n} var(\mathbf{1}_{\{X_i=heads\}}) = \sum_{i=1}^{n} \mathbb{E}[(\mathbf{1}_{\{X_i=heads\}})^2] - [\mathbb{E}(\mathbf{1}_{\{X_i=heads\}})]^2$$

$$= n(1/2+\theta) - n(1/2+\theta)^2.$$

since $(\mathbf{1}_{\{X_i=heads\}})^2 = \mathbf{1}_{\{X_i=heads\}}$. The variance reaches a maximum when $\theta = 0$ and a minimum when $\theta = \pm 1/2$.

(d) Find the dependence of SNR on n:

$$\text{SNR} = \frac{n(1/2+\theta)}{\sqrt{n}\sqrt{1/4-\theta^2)}} \propto \sqrt{n}.$$

(e) SNR is undetermined when $\theta = -1/2$ (probability of heads=0). 1) SNR is infinite when $\theta = 1/2$ (probability of heads=1).

(f) By the CLT, the limiting distribution is Gaussian. The mean is $n(1/2+\theta)$ and variance is $n(1/2+\theta) - n(1/2+\theta)^2$. ■

Problem 30. Consider a die which is unbiased. (a) You roll the die once. What is the probability of getting a "1" vs a "4"?

(b) You roll the die twice. What is the probability of getting a total of "2" (i.e. "1" on both trials) versus the probability of getting a total of "7" ("Total" means you add the two results together.)

(c) You roll the die 10,000 times and record the results. What is the probability distribution of the mean (i.e. the mean of all the results), its first moment and variance?

Solution. (a) 1/6 and 1/6

(b) 2: 1/6 times 1/6 = 1/36

7: 6 times 1/6 times 1/6 = 1/6

(c) by the CLT the distribution converges to the normal law, $\mathcal{N}(3.5, \sigma^2/10000)$, where the value of σ^2 is:

$$\sigma^2 = \sum_{i=1}^{6} (x_i - \mu)^2 p_i = 2.917$$

■

Problem 31. What power of t (time) does the root-mean-square displacement in a 1D random walk depend on? How does this differ from the case of ballistic motion. Explain.

Solution. For random walk the root mean square displacement is proportional to $\sqrt{t}$ whereas for ballistic motion it depends on t. The $\sqrt{t}$ dependence can be explained by the large number of "back-and-forth" steps in the random walk. ■

Problem 32. Consider the normal (Gaussian) distribution with parameters μ and σ^2, i.e. let $X \sim \mathcal{N}(\mu, \sigma^2)$. Show all calculations.

(a) Calculate moments of all orders ($n = 0, 1, 2, 3, \dots$) for X.

(b) Calculate the central moments of all orders for X.

(c) Define a new function $K(t) = \log \mathbb{E}(e^{tX})$, and Taylor expand $K(t)$ in powers of t:

$$K(t) = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!}.$$

Find a general expression for the coefficients κ_n.

(d) Define a new function $M(t) = \exp(K(t))$. Show how the moments can be obtained from $M(t)$ in terms of the κ_n's.

(e) Show how the density of X, $p(x)$, can be constructed from a knowledge of the statistical moment, or from the central moments, or from the κ_n's.

(f) Explain why the method in (e) of reconstructing $p(x)$ is important from an experimental science standpoint.

Solution. (a) The moments of odd orders are all zero because the integral of an odd function (n-th moment of X, where n is odd) times an even function (Gaussian PDF) vanishes because the integrand is odd. On the other hand, the moment $\mathbb{E}(X^n)$, where n is even are non-zero. They are calculated as follows:

$$\mathbb{E}(X^n) = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \cdot x^n \, \mathrm{d}x,$$

where we use

$$\int_0^{\infty} e^{-ax^2} x^n \mathrm{d}x = \frac{(n-1)!!}{2^{n/2+1} a^{n/2}} \sqrt{\frac{\pi}{a}}$$

for n even. The result is:

$$\mathbb{E}(X^n) = \sigma^n (n-1)!!.$$

The first few moments are:

order	moment	central moment
$n=1$	μ	0
$n=2$	$\mu^2+\sigma^2$	σ^2
$n=3$	$\mu^3+3\mu\sigma^2$	0
$n=4$	$\mu^4+6\mu^2\sigma^2+3\sigma^4$	$3\sigma^4$
$n=5$	$\mu^5+10\mu^3\sigma^2+15\mu\sigma^4$	0

see `https://en.wikipedia.org/wiki/Normal_distribution#Moments`

(b) See solution to (a).

(c) Expand $e^{tX}=\sum_{i=0}^{\infty}\frac{t^iX^i}{i!}$ and take the average, $\mathbb{E}e^{tX}=\sum_{i=0}^{\infty}\frac{t^i\mathbb{E}(X^i)}{i!}$. On the other hand, take the exponential of $K(t)$,

$$e^{K(t)}=e^{\sum_{n=1}^{\infty}\kappa_n\frac{t^n}{n!}}=1+\sum_{n=1}^{\infty}\kappa_n\frac{t^n}{n!}+\frac{1}{2}\left(\sum_{n=1}^{\infty}\kappa_n\frac{t^n}{n!}\right)^2+\frac{1}{3!}\left(\sum_{n=1}^{\infty}\kappa_n\frac{t^n}{n!}\right)^3+\ldots.$$

We can now identify the like powers of t (let $\mu_r=\mathbb{E}X^r$):

$$\begin{aligned}
t^1:&\mu_1=\kappa_1 && \mu_1=\kappa_1\\
t^2:&\mu_2/2=\kappa_2/2+\kappa_1^2/2 && \mu_2=\kappa_2+\kappa_1^2\\
t^3:&\mu_3/6=\kappa_3/6+\kappa_1\kappa_2/2+\kappa_1^3/6 && \mu_3=\kappa_3+3\kappa_1\kappa_2+\kappa_1^3\\
&\vdots && \vdots
\end{aligned}$$

This can be "inverted" to give:

$\kappa_1=\mu_1$

$\kappa_2=\mu_2-\mu_1^2$

$\kappa_3=\mu_3-3\mu_2\mu_1+2\mu_1^3$

$\vdots$

(d) $M(t)=\mathbb{E}(e^{tX})=\mathbb{E}\sum_{i=0}^{\infty}t^iX^i/i!=\sum_{i=0}^{\infty}t^i\mathbb{E}(X^i)/i!$. The moments are obtained by differentiation with respect to t and setting $t=0$:

$$\mu_r\equiv\mathbb{E}(X^r)=\left.\frac{\mathrm{d}^r}{\mathrm{d}t^r}M(t)\right|_{t=0}.$$

(e) Consider the quantity $\mathbb{E}(e^{tX})$,

$$M(t)=\sum_{r=0}^{\infty}\frac{t^r\mu_r}{r!}=\mathbb{E}(e^{tX})\equiv\int_{-\infty}^{\infty}e^{tx}p(x)\mathrm{d}x.$$

We can solve for $p(x)$ by invoking the inversion formula:

$$\frac{1}{2\pi i}\lim_{T\to\infty}\int_{\gamma-iT}^{\gamma+iT} e^{-st}M(s)\mathrm{d}s = \frac{1}{2\pi i}\lim_{T\to\infty}\int_{\gamma-iT}^{\gamma+iT} e^{-st}\left[\int_{-\infty}^{\infty} e^{sx}p(x)\mathrm{d}x\right]\mathrm{d}s.$$

$$= \frac{1}{2\pi}\lim_{T\to\infty}\int_{-\infty}^{\infty}\int_{-T}^{T} e^{\gamma(x-t)}e^{i(x-t)\tau}p(x)\mathrm{d}x\mathrm{d}\tau.$$

Then integrating over τ and invoking $\delta(x-a) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i(x-a)t}\mathrm{d}t$:

$$= \int_{-\infty}^{\infty} e^{\gamma(x-t)}\delta(x-t)p(x)\mathrm{d}x = p(t).$$

(f) This is important because if we know all the moments, we can reconstruct $p(x)$. The moments can be estimated from experimental data. ■

Problem 33. The scores in a chemistry class from 2016 were as follows (out of 100):

40.311 33.386 44.142 65.631 41.066 47.051 42.322 50.752 30.730 28.777 50.885 45.143 29.997 34.107 31.045 39.684 25.157 38.825 41.838 35.716 26.620 44.827 50.506 63.251 32.622 59.843 56.967 50.783 51.961 39.746 50.895 36.447 26.660 49.376 29.302 37.166 33.532 33.627 34.030 34.816 52.107 58.384 50.539 37.568 39.806 54.394 42.399 40.042 47.231 21.915

(a) If the course policy is to assign 'A' grades to the top 10-percentile of the class and 'F' to the rest, how many students obtained an F?

(b) Draw a histogram of the exam results.

(c) Calculate the mean, standard deviation and median of the exam and indicate those quantities on the histogram. (Explain how those quantities are calculated from the data.)

(d) In units of standard deviation (σ), how far is the 10-percentile from the mean?

(e) Reconstruct the PDF of this rv X (score), using the numerical data.

(f) Suppose that the 2017 scores were:

59.378 102.006 54.660 39.713 61.877 46.731 17.570 45.646 71.654 19.959 58.948 57.506 78.838 31.859 20.175 31.766 39.408 41.096 31.092 52.754 53.712 67.778 66.991 37.362 57.768 72.032 48.005 78.559 46.742 84.157 66.175 90.976 72.627 40.335 19.464 60.673 51.911 34.235 35.143 39.269 48.814 83.537 50.505 40.340 47.480 58.682 72.354 56.195 74.103 50.013

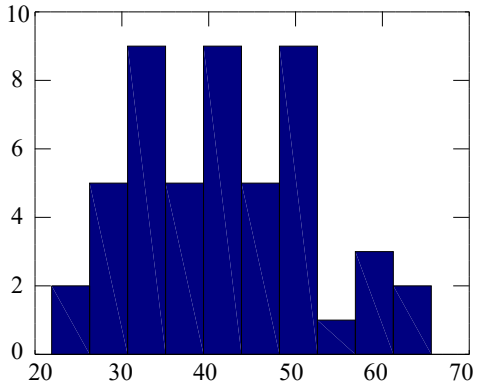


Figure 2.14. Histogram.

Plot histogram and calculate distribution parameters. Are these scores significantly different from those of 2016? (Why/why not?)

Solution. (a) The top 10-percentile is called the 90th percentile, and is the value of x_{90} such that

$$\int_{-\infty}^{x_{90}} p(x)\mathrm{d}x = 0.90.$$

Numerically, there is a function in MATLAB called `prctile` that will compute this for us. I get 55.68 for the above data. This means that all scores below this get an 'F'; we count 45 of those.

(b) See Figure 2.14.

(c) Mean = 41.679 (use formula for sample mean), std = 10.372 (use formula for sample standard deviation), median = 40.177 (order the numbers and pick the middle one).

(d) In MATLAB, we simply type `(prctile(d,90)-mean(d))/std(d)` and obtain 1.3500.

(e) From the raw data we can calculate the moments of the distribution:

r	raw moment	central moment
1	41.679	9.4502×10^{-15}
2	1842.5	105.43
3	8.5908×10^4	325.47
4	4.1974×10^6	2.6805×10^4
5	2.1355×10^8	2.0898×10^5

and using the inversion formula, we can obtain $p(x)$.

(f) The mean/std are 53.371 ± 19.136. Compare this to 41.679 ± 10.372. These two numbers are not significantly different because their error bars overlap considerably. ■

Problem 34. Prove, using the law of large numbers, that the histogram of a rv X converges to its PDF, $p(x)$.

Solution. Let X have CDF $F(x)$. Let $X_1, X_2, \ldots, X_n$ be a random sample of F. Define the indicator function $\mathbf{1}_{(-\infty,x]}(y)$ to be equal to 1 if $y \leq x$ and zero otherwise. Then,

$$\mathbb{E}\mathbf{1}_{(-\infty,x]}(X_i) = \int_{-\infty}^{x} p(x_i)\mathrm{d}x_i = \mathbb{P}(X_i \leq x) = F(x).$$

For each n, the histogram of the random sample is:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(X_i)$$

Its expectation value is:

$$\mathbb{E}F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\mathbf{1}_{(-\infty,x]}(X_i) = F(x).$$

It then follows from the law of large numbers that $F_n(x)$ converges to $F(x)$. If the CDFs converge, the PDFs also converge. ■

Problem 35. Prove, using the law of large numbers, that the empirical distribution of random variable X, $\hat{p}(x)$ converges to its PDF, $p(x)$.

Solution. The solution is identical to that of Problem 34. The empirical distribution $\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)$ has the empirical CDF:

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} \hat{p}(x)\mathrm{d}x = \frac{1}{n}\#\{i : x_i \leq x\},$$

where $\#\{i : x_i \leq x\}$ denotes the number of data points x_i satisfying the condition $x_i \leq x$. Let's denote the random variables as X_i and x_i, their corresponding values. Since n data points are used to construct this CDF let us denote it as $F_n(x)$. Its expectation value is

$$\mathbb{E}F_n(x) = \frac{1}{n}\mathbb{E}\#\{i : X_i \leq x\} = \frac{1}{n}\mathbb{E}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(X_i) = F(x),$$

where $F(x)$ is the CDF of $p(x)$ and $\#\{i : X_i \leq x\} = \sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(X_i)$. It then follows from the law of large numbers that $F_n(x)$ converges to $F(x)$. Since the CDFs converge, the PDFs also converge, as the PDF is obtained from the CDF by differentiation. ■

Problem 36. Consider the weak law of large numbers (WLLN): Let X_1, $X_2, \ldots$ be iidrv with mean μ and variance $\sigma^2 < \infty$. Then, $(1/n)\sum_{i=1}^{n} X_i$ converges to μ in probability.

(a) Prove the WLLN.

(b) Illustrate it using a numerical example, i.e. choose $\epsilon > 0$, generate random sequences $X_1, \ldots, X_n$, compute the sample mean $\overline{X}_n$, record this value as m_1. Generate a second random sequence, and obtain the sample mean as m_2. Repeat this process many times (r times) and plot a histogram of the sample means $(m_1, m_2, \ldots, m_r)$. Then increase n and repeat this process. You should now have several histograms as function of n. Then plot the probability $\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon)$ as a function of n and show that it converges to 0 as n increases. Since we are dealing with experimental data, the probability should be calculated empirically:

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) = \frac{1}{r}\sum_{j=1}^{r} \mathbf{1}_{|m_j - \mu| \geq \epsilon}$$

where $\mathbf{1}_{|m_j-\mu|\geq\epsilon}$ is an "indicator function", i.e. equals 1 when $|m_j - \mu| \geq \epsilon$ and equals 0 otherwise.

Solution. (a) Weak law: let $\overline{X}_n = (1/n)(X_1 + X_2 + \cdots + X_n)$, $var(\overline{X}_n) = (1/n^2)n \cdot var(X_1) = \sigma^2/n$, and $\overline{x}_n = \frac{x_1+\cdots+x_n}{n}$. Then,

$$\begin{aligned}
\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) &= \int_{\{(x_1,\ldots,x_n)||\overline{x}_n-\mu|\geq\epsilon\}} p_1(x_1)\cdot\cdots\cdot p_n(x_n)\mathrm{d}x_1\ldots\mathrm{d}x_n \\
&\leq \int_{\{|\overline{x}_n-\mu|\geq\epsilon\}} \frac{(\overline{x}_n-\mu)^2}{\epsilon^2} p_1(x_1)\cdot\cdots\cdot p_n(x_n)\mathrm{d}x_1\ldots\mathrm{d}x_n \\
&\leq \int_{\mathbb{R}^n} \frac{(\overline{x}_n-\mu)^2}{\epsilon^2} p_1(x_1)\cdot\cdots\cdot p_n(x_n)\mathrm{d}x_1\ldots\mathrm{d}x_n \\
&= \frac{var(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ as } n \to \infty.
\end{aligned}$$

This completes the proof of the WLLN. The first inequality is justified because $|\overline{x}_n - \mu| \geq \epsilon$, and therefore, $1 \leq \frac{|\overline{x}_n-\mu|}{\epsilon}$, and consequently (squaring both sides), $1 \leq \frac{|\overline{x}_n-\mu|^2}{\epsilon^2}$. The second equality is justified because the integral is everywhere non-negative. Therefore, extending the region of integration from the restricted set $\{(x_1, \ldots, x_n) : |\overline{x}_n - \mu| \geq \epsilon\}$ to the whole space $\mathbb{R}^n$ leads to an upper bound. This proof assumes the existence of the variance σ^2 of X_i.

(b) There are many possible solutions here. Here is mine. I used this code in MATLAB to generate the random numbers and required plots:

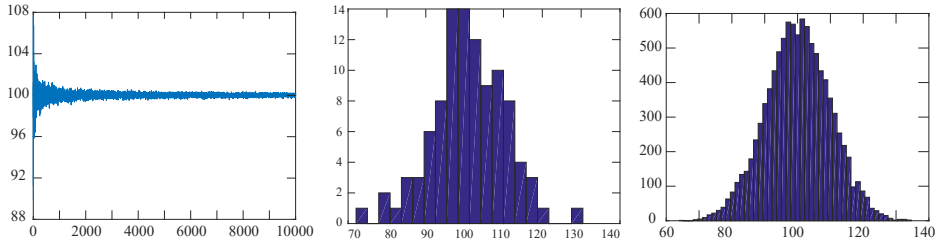


Figure 2.15. Law of large numbers illustrated.

```
m=10000; r=10000;
X=10*randn([m,r])+100;
for j=1:r,
    Xn(j)=(1/j)*squeeze(sum(X(1:j,j),1));
end;
figure;plot(Xn);
figure;hist(X(100,1:100),20);
figure;hist(X(10000,1:10000),50);
```

Here we generated random variables $\sim \mathcal{N}(100, 100)$. The first plot (Fig. 2.15) illustrates the LLN because the arithmetic averages are shown to converge to the true mean (100) as the number of terms in the sum increases. The histograms show that with only a few terms, we do not get a nice Gaussian, whereas using 10,000 terms, we get a nice bell curve. (If you chose a distribution other than normal, these histograms should reflect the chosen distribution.) ■

Problem 37. Derive the probability distribution of a biased random walk (i.e. let $p_\sigma = 1/2 + \delta$ and $p_{-\sigma} = 1/2 - \delta$ for some bias $0 \leq \delta \leq 1/2$).

Solution. By the CLT, the distribution will be Gaussian, of course. The mean step size is $\mu_i = \mathbb{E}X_i = \sigma(p_\sigma - p_{-\sigma}) = \sigma(1/2 + \delta - 1/2 + \delta) = 2\sigma\delta$. So the total displacement

$$X_{tot} = X_1 + X_2 + \cdots + X_n$$

has expectation value

$$\mathbb{E}X_{tot} = 2\sigma\delta n,$$

instead of 0. (i.e. it "drifts" linearly with time at constant speed $2\sigma\delta$.)

The variance is $var(X_i) = \mathbb{E}(X_i - \mu_i)^2 = p_\sigma(\sigma - 2\sigma\delta)^2 + p_{-\sigma}(-\sigma - 2\sigma\delta)^2 = (1/2 + \delta)\sigma^2(1 - 2\delta)^2 + (1/2 - \delta)\sigma^2(1 + 2\delta)^2 = \sigma^2(1/2 + \delta)(1 - 4\delta + 4\delta^2) + \sigma^2(1/2 - \delta)(1 + 4\delta + 4\delta^2) = \sigma^2[1 - 4\delta^2]$. The total variance is:

$$var(X_{tot}) = var(X_1) + \cdots + var(X_n) = \sigma^2[1 - 4\delta^2]n,$$

as opposed to $\sigma^2 n$. Thus, the variance is reduced. When $\delta = \pm 1/2$ (meaning steps are always to the left, or always to the right), then the variance is zero

because the path is no longer random, but instead becomes deterministic. ■

Problem 38. Prove that in 3D the mean square displacement is $6Dt$, and in the general case of d dimensions, it is equal to $2dDt$ (a direct calculation of the d-dimensional integral requires the spherical volume element in d-dim, which includes some Gamma functions).

Solution. In 3D the mean square displacement is

$$\begin{aligned}\mathbb{E}(r(t)^2) \equiv & \frac{1}{(4\pi Dt)^{3/2}} \int_{\mathbb{R}^3} r^2 \exp\left(-\frac{r^2}{4Dt}\right) \mathrm{d}^3 r \\ = & \frac{1}{(4\pi Dt)^{3/2}} \int_{\mathbb{R}^3} r^2 \exp\left(-\frac{r^2}{4Dt}\right) r^2 \mathrm{d}r \mathrm{d}(\cos\theta) \mathrm{d}\phi \\ = & \frac{4\pi}{(4\pi Dt)^{3/2}} \int_0^\infty r^4 \exp\left(-\frac{r^2}{4Dt}\right) \mathrm{d}r\end{aligned}$$

To integrate this we use the famous result $\int_{-\infty}^{\infty} e^{-ax^2} \mathrm{d}x = \sqrt{\pi/a}$, differentiate wrt a twice: $\int_{-\infty}^{\infty} x^4 e^{-ax^2} \mathrm{d}x = \frac{3}{4}\sqrt{\pi} a^{-5/2}$.

$$\frac{4\pi}{(4\pi Dt)^{3/2}} \int_0^\infty r^4 \exp\left(-\frac{r^2}{4Dt}\right) \mathrm{d}r = \frac{4\pi}{(4\pi Dt)^{3/2}} \frac{3}{8} \sqrt{\pi} (4Dt)^{5/2} = 6Dt.$$

In the general case of d dimensions, the mean square displacement is:

$$\mathbb{E}(r(t)^2) = \mathbb{E}(x_1^2 + \cdots + x_d^2) = 2dDt.$$

There is no d-dimensional integral needed here, as each $\overline{x_1^2}$ contributes $2Dt$, and there are d such terms, for a total of $2dDt$. ■

Problem 39. Prove that for the Poisson distribution the mean and variance are both equal to the parameter of the distribution.

Solution. Proofs can be found here:
`http://filestore.aqa.org.uk/subjects/AQA-MS03-W-2-SM.PDF`
`https://proofwiki.org/wiki/Variance_of_Poisson_Distribution` ■

Problem 40. Prove that Poisson distribution converges to a Gaussian in the limit of large n. However, obtain the coefficient of the exponential as well (the prefactor), making use of the slightly more accurate Stirling's formula.

Solution. The calculation we did previously was:

$$\begin{aligned}\frac{e^{-\overline{n}}\overline{n}^n}{n!} &= \exp\{-\overline{n} - \log n! + n\log\overline{n}\} \\ &= \exp\{-\overline{n} - n\log n + n + n\log\overline{n}\} \\ &= \exp\{(n-\overline{n}) + n\log(\overline{n}/n)\} \\ &= \exp\left\{(n-\overline{n}) + n\log\left[1+\left(\frac{\overline{n}-n}{n}\right)\right]\right\} \\ &\approx \exp\left\{-\frac{(\overline{n}-n)^2}{2n}\right\} \approx \exp\left\{-\frac{(\overline{n}-n)^2}{2\overline{n}}\right\}\end{aligned}$$

The prefactor $\frac{1}{\sqrt{2\pi\overline{n}}}$ is recovered by using

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

Now, replace $\frac{e^{-\overline{n}}\overline{n}^n}{n!}$ by $\frac{1}{\sqrt{2\pi n}}\frac{e^{-\overline{n}}\overline{n}^n}{\left(\frac{n}{e}\right)^n} = \frac{1}{\sqrt{2\pi n}}e^{-\overline{n}}\overline{n}^n e^{-n\log n+n}$. This gives the desired result with the correct prefactor. ■

Problem 41. Let $X \sim \mathcal{N}(0,1)$. Generate iid random numbers on a computer (say, 10,000 numbers). Those are the different realizations of X, i.e. $x_1, x_2, \ldots, x_{10,000}$. Next, consider another random variable, Y.

(a) Let $Y = X + 1$, so that we now have 10,000 pairs of points: (x_1, y_1), (x_2, y_2), …, $(x_{10,000}, y_{10,000})$. Plot these 10,000 pairs $\{(x_i, y_i)\}$ as dots on scatter plot. Fit a straight line. What slope do you get? From the data, calculate the sample correlation coefficient. Is Y correlated to X? Why?

(b) Let X be as previously defined. Let Z be distributed identically to X, but independent of X. Generate random numbers on a computer to obtain pairs $\{(x_i, z_i)\}$ of random numbers. Define a new random variable $Y = X + Z$. Is Y correlated to X? Why? (Plot XY pairs on a scatter plot, fit a straight line, calculate $r_{X,Y}$.)

(c) Let X be as defined previously. Let $Y \sim \mathcal{N}(0,1)$. Generate random numbers for X and Y, and plot the resulting pairs $\{(x_i, y_i)\}$ on a scatter plot. Are X and Y correlated? Why?

Solution. (a) Y is correlated to X ($r = 1$). On a scatter plot, we should see a perfect straight line (no deviation from it).

(b) $cov(X,Y) = cov(X, X+Z) = cov(X,X) + cov(X,Z) = var(X) = 1$, hence $r = 1$. Here on a scatter plot there will be random deviations from a straight line due to Z. However, fitting a straight line will still give a slope of 1.

(c) Totally uncorrelated, since X and Y are independent. (Scatter plot looks random.) ■

Problem 42. Choose a distribution we have *not* used in class. Fix (choose) the parameters of the distribution. Let X be a random variable distributed accordingly. Calculate the mean and variance of X analytically (i.e. using the distribution function). Use a computer to generate random numbers according to the distribution of X. (How do you generate such random numbers?) Plot the histogram of X, compare to the PDF or PMF of X (plot both on the same graph). Calculate numerically the mean and variance of X (using the random numbers you generated). Compare to the true values of mean and variance obtained from the PDF or PMF.

Problem 43. Let X be the result of rolling a die. Generate n random numbers on a computer and obtain the random sample $X_1, X_2, \ldots, X_n$. Take the arithmetic average: $\overline{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$. Plot $\overline{X}_n$ versus n. What do you conclude? What theorem does this exercise illustrate?

Solution. This illustrates the law of large numbers. See the article for the plot:
`https://en.wikipedia.org/wiki/Law_of_large_numbers` ■

Problem 44. Let Y be a Poisson rv with parameter λ. Prove that Y can be written as the sum

$$Y = X_1 + X_2 + \cdots + X_n,$$

where X_i are independent identically distributed rv's, also with the Poisson distribution. What should be the lambda parameter of the X_i?

Solution. Let's do the case of two rv's. Let $Z = X + Y$ where X and Y are Poisson, with parameters λ and μ, respectively. Then the PMF of Z is:

$$\mathbb{P}(Z = z) = \sum_{x=0}^{z} \frac{e^{-\lambda}\lambda^x}{x!}\frac{e^{-\mu}\mu^{z-x}}{(z-x)!} = \frac{e^{-(\lambda+\mu)}}{z!}\sum_{x=0}^{z}\binom{z}{x}\lambda^x\mu^{z-x}$$

Thus, Z is also Poisson, but with mean $\lambda + \mu$. This can be extended to n Poisson variables. Their sum will also be Poisson. If $Y = X_1 + \cdots + X_n$ has parameter λ, then each X_i must have parameter λ/n. ■

Problem 45. Let $X_1, \ldots, X_n$ be a sequence of independent random variables with CDF's F_n (X_i has CDF F_i, $i = 1, \ldots, n$). Let X be a random variable with CDF F. The sequence X_n is said to *converge in distribution* if the CDF's converge pointwise, i.e.,

$$\lim_{n\to\infty} F_n(x) = F(x),$$

at all points x for which F is continuous.

(a) Show that convergence of the CDF's also implies the PDF's. i.e. let f_i be the PDF of independent rv's X_i ($i = 1, \ldots, n$) and f be the PDF of X.

Prove that convergence of the CDF's implies:

$$\lim_{n\to\infty} f_n(x) = f(x)$$

for all x.

(b) Prove that the sequence of independent rv's $X_i \sim \mathcal{N}(1/n, 1)$ converges *in distribution* to a standard normal random variable.

Solution. Since

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[x - \frac{1}{n}\right]^2\right) \to \exp\left(-\frac{x^2}{2}\right),$$

it follows that X_n converges in distribution to X $\mathcal{N}(0, 1)$. ■

Problem 46. Let $X_n \sim \mathcal{N}(0, 1/n)$ and let $X = 0$. Prove that for any $\epsilon > 0$,

$$\mathbb{P}(|X_n| > \epsilon) \to 0$$

as $n \to \infty$. This is an example of *convergence in probability*, i.e. $\mathbb{P}(|X_n| > \epsilon) \to 0$ implies that X_n converges in probability to X (=0), since $\mathbb{P}(|X_n| > \epsilon) = \mathbb{P}(|X_n - X| > \epsilon)$.

Solution. First we note that $\mathbb{P}(|X_n| > \epsilon) = \mathbb{P}(|X_n|^2 > \epsilon^2)$. The latter is the integral:

$$\int\limits_{\{x_n^2 > \epsilon^2\}} p(x_n)\mathrm{d}x_n \le \int\limits_{\{x_n^2 > \epsilon^2\}} \frac{x_n^2}{\epsilon^2} p(x_n)\mathrm{d}x_n \le \int_{\mathbb{R}} \frac{x_n^2}{\epsilon^2} p(x_n)\mathrm{d}x_n$$

$$= \frac{\mathbb{E}X_n^2}{\epsilon^2} = \frac{var(X_n)}{\epsilon^2} = \frac{1}{n\epsilon^2} \to 0$$

■

Problem 47. Let X_i be iidrv with uniform distribution over the interval [0,1]. Take the sum $S_n = X_1 + X_2 + \cdots + X_n$. Find the distribution of X_n analytically (i.e. find its CDF and PDF). Show numerically (i.e. by generating random numbers on a computer) the histogram of S_n from $n = 1, 2, \ldots, 10$. What do you conclude?

Solution. This is straightforward and will be left as an exercise (simply generate random numbers in MATLAB to construct S_n, and plot using the `hist` function). S_1 has the uniform distribution. S_2 has the "tent" distribution. etc. whereas S_n for large n looks more and more Gaussian as n increases, thanks to the CLT. Convergence to a Gaussian is very fast and does not require n to be very large. ■

Problem 48. Suppose that X has a PDF, $p(x) = \frac{1}{2}\sin(x)$, where $x \in [0, \pi]$, and equals zero elsewhere. Calculate its mean and variance. Calculate its

skewness and kurtosis. Compare skewness and kurtosis to those of a normal distribution (with same mean and variance).

Solution. We will do the first two moments (others are obtained similarly):

$$\mathbb{E}(X) = \int_0^\pi \frac{1}{2}\sin(x)x\mathrm{d}x = \frac{\pi}{2}.$$

$$\mathbb{E}(X - \frac{\pi}{2})^2 = \int_0^\pi \frac{1}{2}\sin(x)(x - \frac{\pi}{2})^2\mathrm{d}x = \frac{1}{4}(\pi^2 - 8).$$

Those results can be obtained from WolframAlpha by typing:

```
integrate x*(1/2)*sin(x) from 0 to Pi
integrate ((x-Pi/2)^2)*(1/2)*sin(x) from 0 to Pi
```

■

Problem 49. Let $X \sim \mathcal{N}(0,1)$. What is the distribution of $Y = X^3 + 5$?

Solution.

$$\mathbb{P}(Y < y) = \mathbb{P}(X^3 + 5 < y) = \mathbb{P}(X \leq \sqrt[3]{y-5}) = \int_{-\infty}^{\sqrt[3]{y-5}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\mathrm{d}x.$$

It is also ok to calculate its PDF by differentiating the above CDF with respect to y, making use of the Leibniz formula. ■

Problem 50. Calculate the mean and the variance of a random variable X distributed according to the PDF:

$$p(x) = \frac{\gamma}{(x-\mu)^2 + \gamma^2}.$$

Solution. For the mean we have an integral of the type (set $\gamma = 1$, $\mu = 0$ without loss of generality, since I analyze the "tail" of the function here):

$$\int_{-\infty}^{\infty} \frac{1}{x^2+1} x\mathrm{d}x$$

When x is large, this integral behaves like $\int 1/x \sim \log(x)$, which diverges with x. Thus, the mean does not exist. For the variance, we have an integral of the type

$$\int_{-\infty}^{\infty} \frac{1}{x^2+1} x^2\mathrm{d}x \sim \int \mathrm{d}x \sim x \to \infty$$

which also diverges. Thus, it has no variance. ■

Problem 51. The probability of k successes in n trials is ($k = 0, 1, \ldots, n$, $0 \leq p \leq 1$):

$$\mathbb{P}(k \text{ successes}) = \binom{n}{k} p^k (1-p)^{n-k}.$$

B is a random variable distributed as such. Prove that B has mean np and variance $np(1-p)$.

Solution. See https://en.wikipedia.org/wiki/Binomial_distribution ■

Problem 52. Suppose that you have a string instrument (e.g., electric guitar) whose strings, when plucked, behave like oscillators. The potential energy of the string is modeled by an anharmonic oscillator which consists of the sum of quadratic and quartic terms:

$$V(x) = ax^2 + bx^4. \qquad a, b \text{ non-negative constants}$$

The potential V is transferred to kinetic energy, which is then measured by the guitar's pick-up coils and sent to the amplifier. The noise statistics of V are important for the design of the guitar amplifier circuits.

If the position x (x: extension of the center of the string from its equilibrium position) is measured experimentally using an interferometer whose instrument noise is known to be normally distributed with mean μ and variance σ^2, what would you expect the noise statistics of V to look like? (i.e. find the probability distribution of V) Note: you can assume there are no temporal correlations in the noise.

(a) When $b = 0$ and a is nonzero (no anharmonicity).

(b) When $a = 0$ and b is nonzero (anharmonic part only).

Solution. (a) When $V = ax^2$, the probability of $V < v$, $\mathbb{P}(V < v)$, is:

$$\mathbb{P}(ax^2 < v) = \mathbb{P}(-\sqrt{\tfrac{v}{a}} < x < \sqrt{\tfrac{v}{a}}) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\sqrt{\frac{v}{a}}}^{\sqrt{\frac{v}{a}}} e^{-(x-\mu)^2/2\sigma^2} \mathrm{d}x.$$

Differentiating with respect to v gives the PDF, $p_V(v) = \frac{\mathrm{d}\mathbb{P}(V<v)}{\mathrm{d}v}$:

$$\frac{e^{-(\sqrt{v/a}-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \cdot \frac{\mathrm{d}}{\mathrm{d}v}\sqrt{v/a} - (\text{lower limit}) = \frac{e^{-(\sqrt{v/a}-\mu)^2/2\sigma^2} + e^{-(\sqrt{v/a}+\mu)^2/2\sigma^2}}{\sqrt{8av\pi\sigma^2}}$$

(b) When $V = bx^4$ the probability of $V < v$, $\mathbb{P}(V < v)$, is (imaginary roots are discarded, because probabilities are non-negative quantities):

$$\mathbb{P}(bx^4 < v) = \mathbb{P}(-(v/b)^{\frac{1}{4}} < x < (v/b)^{\frac{1}{4}}) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-(v/b)^{1/4}}^{(v/b)^{1/4}} e^{-(x-\mu)^2/2\sigma^2} \mathrm{d}x.$$

The PDF is obtained by differentiation, $p_V(v) = \frac{\mathrm{d}\mathbb{P}(V<v)}{\mathrm{d}v}$:

$$\frac{e^{-((v/b)^{1/4}-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \cdot \frac{\mathrm{d}}{\mathrm{d}v}(v/b)^{1/4} - (\text{lower limit})$$
$$= \frac{1}{v^{3/4}} \frac{e^{-((v/b)^{1/4}-\mu)^2/2\sigma^2} + e^{-((v/b)^{1/4}+\mu)^2/2\sigma^2}}{\sqrt{32\pi\sigma^2}\, b^{1/4}}$$

■

Problem 53. The joint density of X and Y is

$$p_{XY}(x,y) = \begin{cases} 2 & \text{if } 0 \le x \le y \le 1; \\ 0 & \text{otherwise.} \end{cases}$$

Another rv Z is independent from X and Y and has the same distribution as X. (a) Find the covariance matrix of the vector (X, Y, Z). (b) Calculate the first two moments of a new rv that is the sum of all 3, i.e. $X+Y+Z$. (c) Compute $cov(X, Y + Z)$. (d) Compute the covariance matrix of the vector $(X, X + Z, Y + Z)$.

Solution. (a) Let $\mathbf{v} = (X, Y, Z)$. Then, since Z is independent of X and Y, we can immediately put 0's in a few places:

$$cov(\mathbf{v},\mathbf{v}) = \begin{bmatrix} cov(X,X) & cov(X,Y) & cov(X,Z) \\ cov(Y,X) & cov(Y,Y) & cov(Y,Z) \\ cov(Z,X) & cov(Z,Y) & cov(Z,Z) \end{bmatrix}$$
$$= \begin{bmatrix} var(X) & cov(X,Y) & 0 \\ cov(Y,X) & var(Y) & 0 \\ 0 & 0 & var(Z) \end{bmatrix}$$

For $var(X)$ and $var(Y)$, we need the marginal PDFs:

$$p_X(x) = \int_{\{y\in[0,1]|y>x\}} p_{XY}(x,y)\mathrm{d}y = \int_x^1 2\mathrm{d}y = 2\, y|_{y=x}^{y=1} = 2(1-x).$$
$$p_Y(y) = \int_{\{x\in[0,1]|x<y\}} p_{XY}(x,y)\mathrm{d}x = \int_0^y 2\mathrm{d}x = 2\, x|_{x=0}^{x=y} = 2y.$$

where $0 \le x, y \le 1$. Using the marginal PDFs,

$$\mathbb{E}X = \int_0^1 x p_X(x)\mathrm{d}x = \int_0^1 x2(1-x)\mathrm{d}x = \frac{1}{3},$$
$$var(X) = \int_0^1 (x - \frac{1}{3})^2 p_X(x)\mathrm{d}x = \frac{1}{18}.$$
$$\mathbb{E}Y = \int_0^1 y p_Y(y)\mathrm{d}y = \int_0^1 y2y\mathrm{d}x = \frac{2}{3},$$

$$var(Y) = \int_0^1 (y - \frac{2}{3})^2 p_Y(y) \mathrm{d}y = \frac{1}{18}.$$

Finally, for $cov(X, Y)$ we use the joint PDF:

$$cov(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) = \int_0^1 \mathrm{d}x \int_x^1 (x - \frac{1}{3})(y - \frac{2}{3}) 2 \, \mathrm{d}y$$

$$= \int_0^1 \frac{1}{9}(1 - 3x)^2 (1 - x) \mathrm{d}x = \frac{1}{36}.$$

The covariance matrix is:

$$cov(\mathbf{v}, \mathbf{v}) = \begin{bmatrix} \frac{1}{18} & \frac{1}{36} & 0 \\ \frac{1}{36} & \frac{1}{18} & 0 \\ 0 & 0 & \frac{1}{18} \end{bmatrix}.$$

(b) First moment:

$$m_1 = \mathbb{E}(X + Y + Z) = \mathbb{E}X + \mathbb{E}Y + \mathbb{E}Z = 2\mathbb{E}X + \mathbb{E}Y = \frac{2}{3} + \frac{2}{3} = \frac{4}{3}.$$

Second moment:

$$m_2 = \mathbb{E}(X + Y + Z)^2 = \mathbb{E}X^2 + \mathbb{E}Y^2 + \mathbb{E}Z^2 + 2\mathbb{E}XY + 2\mathbb{E}YZ + 2\mathbb{E}XZ.$$

Since Z is independent of X and Y and has the same distribution as X:

$$m_2 = 2\mathbb{E}X^2 + \mathbb{E}Y^2 + 2\mathbb{E}XY + 2\mathbb{E}Y\mathbb{E}Z + 2\mathbb{E}X\mathbb{E}Z.$$

where

$$\mathbb{E}X^2 = \int_0^1 x^2 2(1 - x) \mathrm{d}x = \frac{1}{6}$$

$$\mathbb{E}Y^2 = \int_0^1 y^2 2y \mathrm{d}x = \frac{1}{2}$$

$$\mathbb{E}XY = \int_0^1 \mathrm{d}x \int_x^1 xy 2 \, \mathrm{d}y = \int_0^1 (x - x^3) \mathrm{d}x = \frac{1}{4}.$$

Therefore,

$$m_2 = 2 \cdot \frac{1}{6} + \frac{1}{2} + 2 \cdot \frac{1}{4} + 2 \cdot \frac{2}{3} \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} \cdot \frac{1}{3} = 2.$$

(c) By linearity, and independence of Z from X:

$$cov(X, Y + Z) = cov(X, Y) + cov(X, Z) = cov(X, Y) = \frac{1}{36}.$$

(d) Let $\mathbf{v} = (X, X + Z, Y + Z)$. The covariance matrix is:

$$cov(\mathbf{v}, \mathbf{v}) = \begin{bmatrix} var(X) & cov(X, X + Z) & cov(X, Y + Z) \\ cov(X + Z, X) & var(X + Z) & cov(X + Z, Y + Z) \\ cov(Y + Z, X) & cov(Y + Z, X + Z) & var(Y + Z) \end{bmatrix}$$

$$= \begin{bmatrix} var(X) & cov(X,X)+cov(X,Z) & cov(X,Y)+cov(X,Z) \\ cov(X,X)+cov(Z,X) & var(X+Z) & cov(X,Y)+cov(X,Z)+cov(Y,Z)+cov(Z,Z) \\ cov(Y,X)+cov(Z,X) & cov(X,Y)+cov(X,Z)+cov(Y,Z)+cov(Z,Z) & var(Y+Z) \end{bmatrix}$$

Since Z is independent of X and Y, this simplifies to:

$$= \begin{bmatrix} var(X) & var(X) & cov(X,Y) \\ var(X) & var(X) & cov(X,Y)+var(Z) \\ cov(Y,X) & cov(X,Y)+var(Z) & var(Y) \end{bmatrix} = \begin{bmatrix} \frac{1}{18} & \frac{1}{18} & \frac{1}{36} \\ \frac{1}{18} & \frac{1}{18} & \frac{1}{36}+\frac{1}{18} \\ \frac{1}{36} & \frac{1}{36}+\frac{1}{18} & \frac{1}{18} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{18} & \frac{1}{18} & \frac{1}{36} \\ \frac{1}{18} & \frac{1}{18} & \frac{3}{36} \\ \frac{1}{36} & \frac{3}{36} & \frac{1}{18} \end{bmatrix}$$

■

Problem 54. The covariance matrix of the vector (X, Y, Z) is

$$\begin{bmatrix} cov(X,X) & cov(X,Y) & cov(X,Z) \\ cov(Y,X) & cov(Y,Y) & cov(Y,Z) \\ cov(Z,X) & cov(Z,Y) & cov(Z,Z) \end{bmatrix} = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 4 & -1 \\ 1 & -1 & 4 \end{bmatrix}.$$

(a) Calculate the variance of the rv $X+Y+Z$. (b) Compute $cov(X, Y+Z)$. (c) Compute the covariance matrix of the random vector $(X, X+Z, Y+Z)$.

Solution. (a)

$var(X+Y+Z) = cov(X+Y+Z, X+Y+Z) = 2+4+4+1+1-1-1 = 10.$

(b)

$$cov(X, Y+Z) = cov(X,Y) + cov(X,Z) = 0+1 = 1.$$

(c) Let $\mathbf{v} = (X, X+Z, Y+Z)$. Then,

$$cov(\mathbf{v},\mathbf{v}) = \begin{bmatrix} cov(X,X) & cov(X,X+Z) & cov(X,Y+Z) \\ cov(X+Z,X) & cov(X+Z,X+Z) & cov(X+Z,Y+Z) \\ cov(Y+Z,X) & cov(Y+Z,X+Z) & cov(Y+Z,Y+Z) \end{bmatrix}$$

$$= \begin{bmatrix} cov(X,X) & cov(X,X)+cov(X,Z) & cov(X,Y)+cov(X,Z) \\ cov(Z,X)+cov(X,X) & cov(X,X)+2cov(X,Z)+cov(Z,Z) & cov(X,Y)+cov(X,Z)+cov(Z,Y)+cov(Z,Z) \\ cov(Y,X)+cov(Z,X) & cov(Y,X)+cov(Y,Z)+cov(Z,X)+cov(Z,Z) & cov(Y,Y)+2cov(Y,Z)+cov(Z,Z) \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 2+1 & 0+1 \\ 1+2 & 2+2(1)+4 & 0+1+(-1)+4 \\ 1+0 & 0+(-1)+1+4 & 4+2(-1)+4 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 1 \\ 3 & 8 & 4 \\ 1 & 4 & 6 \end{bmatrix}$$

■

Problem 55. Distribution of the sum of two random variables. (a) Prove that the sum of two discrete and independent rv's (e.g. $X+Y$, where X and Y are independent) has distribution function (PMF) given by the convolution of two PMFs (one for X, one for Y), i..e,

$$\mathbb{P}(X+Y=k) = \sum_{l=0}^{k} \mathbb{P}(X=l)\mathbb{P}(Y=k-l).$$

Let X and Y be two independent rv's. X is Poisson with parameter 2. Y is Poisson with parameter 3. (b) Find the expectation and the variance of the sum $X+Y$. (c) Find the probability mass function (PMF) of the rv $X+Y$.

Solution. (a) Let $Z = X + Y$ (X and Y are independent), then

$$\begin{aligned}\mathbb{P}(X+Y=k) &= \sum_{x+y=k} \mathbb{P}(X=x, Y=y) \\ &= \sum_{x=0}^{k} \mathbb{P}(X=x, Y=k-x) \\ &= \sum_{x=0}^{k} \mathbb{P}(X=x)\mathbb{P}(Y=k-x)\end{aligned}$$

(b) For Z Poisson with parameter λ, $\mathbb{E}Z = var(Z) = \lambda$. Thus, $\mathbb{E}X = 2$, $\mathbb{E}Y = 3$ and $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y = 2+3 = 5$. Next, we have $var(X) = 2$, $var(Y) = 3$ and since X and Y are independent, we have $var(X+Y) = var(X) + var(Y) = 2+3 = 5$. Then,

$$\begin{aligned}\mathbb{P}(X+Y=k) &= \sum_{x=0}^{k} \frac{e^{-\lambda}\lambda^x}{x!}\frac{e^{-\mu}\mu^{k-x}}{(k-x)!} \\ &= \frac{e^{-(\lambda+\mu)}}{k!}\sum_{x=0}^{k}\binom{k}{x}\lambda^x\mu^{k-x} \\ &= \frac{e^{-(\lambda+\mu)}(\lambda+\mu)^k}{k!}\end{aligned}$$

and the sum of two Poisson rv's is also Poisson with additive parameters $\lambda+\mu$. Therefore, $X+Y$ is Poisson with parameter 2+3=5. The distribution is $\mathbb{P}(X+Y=k) = e^{-5}e^k/k!$ for $k = 0, 1, 2, \ldots$. ■

Problem 56. $X_1, X_2, \ldots, X_n$ are independent rv's, such that X_j is Poisson with parameter 2, $j = 1, 2, \ldots, n$. Find the expectation, the variance and standard deviation of the variable:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Solution. If Z is Poisson with parameter λ, $\mathbb{E}Z = var(Z) = \lambda$. Here, X_j is Poisson with parameter 2. We have $\mathbb{E}\overline{X} = \mathbb{E}\frac{1}{n}\sum_{j=1}^{n} X_j = \sum_{j=1}^{n}\frac{1}{n}\mathbb{E}X_j = \frac{1}{n}n\cdot 2 = 2$. Since $X_1, \ldots, X_n$ are independent we have $var(\overline{X}) = \frac{1}{n^2}var(X_1 + X_2 + \cdots + X_n) = \frac{1}{n^2}(var(X_1) + \cdots + var(X_n)) = \frac{1}{n^2}n\cdot 2 = \frac{2}{n}$ and $\sigma_{\overline{X}} = \sqrt{\frac{2}{n}}$. ■

Problem 57. Let $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,4)$ be independent rv's. What is the conditional density of $Z = X+Y$ given $X = 3$ (i.e. under the condition $X = 3$).

Solution. Recall that the sum of two independent Gaussians is also Gaussian with additive means and variances. Since $Z = X + Y$ we have that Z given $X = 3$ has the same distribution as $3 + Y$ given $X = 3$. Since Y is independent from X and $3 + Y \sim \mathcal{N}(3, 4)$ this yields that the rv $Z|X = 3 \sim \mathcal{N}(3, 4)$. Thus, $p_{Z|X=3}(z) = \frac{1}{2\sqrt{2\pi}} \exp(-\frac{1}{8}(z-3)^2)$. ■

Problem 58. $X_1, X_2, \ldots, X_{10}$ are iidrv's with $X_j \sim \mathcal{N}(0, 4)$, $j = 1, \ldots, 10$. Find the conditional density of X_1 under the condition $X_1 + X_2 + \cdots + X_{10} = 3$.

Solution. We have $Y = X_2 + X_3 + \cdots + X_{10} \sim \mathcal{N}(0, 9 \cdot 4) = \mathcal{N}(0, 36)$. Using the definition of conditional probability

$$p_{X_1|X_1+Y=3}(x_1) = \frac{p_{X_1,X_1+Y}(x_1, 3)}{p_{X_1+Y}(3)} = \frac{p_{X_1,Y}(x_1, 3 - x_1)}{p_{X_1+Y}(3)}.$$

Since X_1 and Y are independent, we have

$$p_{X_1,Y}(x_1, 3 - x_1) = p_{X_1}(x_1) p_Y(3 - x_1)$$
$$= \frac{1}{2\sqrt{2\pi}} \exp(-\frac{1}{8}(x_1)^2) \frac{1}{6\sqrt{2\pi}} \exp(-\frac{1}{72}(3 - x_1)^2).$$

We also have $X_1 + Y \sim \mathcal{N}(0, 10 \cdot 4)$ and $p_{X_1+y}(3) = \frac{1}{\sqrt{80\pi}} \exp(-\frac{1}{80} 3^2)$. Finally we find that $p_{X_1|X_1+Y=3}(x_1) = \frac{1}{\sqrt{2\pi \cdot 18/5}} \exp(-\frac{1}{2 \cdot 18/5}(x_1 - 0.3)^2)$. ■

Problem 59. Exercise on conditional expectations: (a) By applying the above definitions, check the trivial case $\mathbb{E}(X|X) = X$. Here, X is a random variable, i.e., $\mathbb{E}[X|X](\omega) = X(\omega)$. (b) Check also that $\mathbb{E}[Y|X] = \mathbb{E}[Y]$ when X and Y are independent. Here, $\mathbb{E}[Y]$ is the random variable taking the constant value $\mathbb{E}[Y]$ for any ω, i.e. $\mathbb{E}[Y|X](\omega) = \mathbb{E}[Y](\omega)$.

Let $X \sim \mathcal{N}(0, 1)$ and $Y \sim \mathcal{N}(0, 4)$ be independent rv's. Let $Z = X + Y$. (c) Compute $\mathbb{E}[Z|X = 3]$ (expectation value calculated using the conditional density of Z under the condition $X = 3$). (d) Calculate $\mathbb{E}[Z|X]$.

Solution. (a) First let's check that $\mathbb{E}[X|X] = X$. First we start with $\mathbb{E}(X|X = x')$, whose definition is $\mathbb{E}[X|X = x'] = \int x p_{X,X=x'}(x) \mathrm{d}x$. Notice that $p_{X,X=x'}(x) = \delta(x - x')$ is the only possible PDF (i.e. the probability that $X = x$ given that $X = x'$ can only be non-zero iff $x = x'$). Hence, $\mathbb{E}[X|X = x'] = x'$. Replace x' by X and get $\mathbb{E}[X|X] = X$.

(b) To prove $\mathbb{E}[Y|X] = \mathbb{E}[Y]$, we write $p_{Y|X=x}(y) = \frac{p_{Y,X}(y,x)}{p_X(x)} = \frac{p_Y(y) p_X(x)}{p_X(x)} = p_Y(y)$ since X and Y are independent. Then, $\mathbb{E}[Y|X = x] = \int y p_{Y|X=x}(y) \mathrm{d}y = \int y p_Y(y) \mathrm{d}y = \mathbb{E}[Y]$. Therefore, $\mathbb{E}[Y|X] = \mathbb{E}[Y]$.

(c) From Problem 57 we have already calculated the conditional density of Z. Using that density, we get that $\mathbb{E}[Z|X = 3] = 3$. (d) Conditional expectation is linear: $\mathbb{E}[Z|X] = \mathbb{E}[X + Y|X] = \mathbb{E}[X|X] + \mathbb{E}[Y|X] = X + \mathbb{E}[Y] = X$. ■

Problem 60. Let $X_1, \ldots, X_{10}$ be iidrv with $X_j \sim \mathcal{N}(0,4)$, $j = 1, \ldots, 10$. Let $S = X_1 + \cdots + X_{10}$. (a) Calculate $\mathbb{E}[X_1|S = 3]$. (b) Calculate $\mathbb{E}[X_1|S]$.

Solution. (a) Using the conditional distribution obtained in Problem 6, we get $\mathbb{E}[X_1|S = 3] = 0.3$. Another solution: by symmetry, for $j = 1, 2, \ldots, 10$ we get $\mathbb{E}[X_1|S = 3] = \mathbb{E}(X_j|S = 3)$. Hence $10\mathbb{E}[X_1|S = 3] = \sum_{j=1}^{10} \mathbb{E}[X_1|S = 3] = \sum_{j=1}^{10} \mathbb{E}[X_j|S = 3] = \mathbb{E}[\sum_{j=1}^{10} X_j|S = 0] = \mathbb{E}[S|S = 3] = 3$. Hence $\mathbb{E}[X_1|S = 3] = 0.3$. (b) In a similar way as in (a), using symmetry we get $\mathbb{E}[X_1|S] = S/10$. ■

Problem 61. Choose a space craft pilot in the nearest galaxy at random and call N the number of accidents during a year for this pilot. The number of accidents N depends on another random variable, P, which quantifies the pilot's skills. The number of accidents given some skillset $P = p$ has $Binomial(4, p)$ distribution, i.e., $N|P = p \sim Binomial(4, p)$. The parameter P among the population of pilots has $P \sim U([0, 1])$ (uniform distribution). (a) Find the marginal distribution of N. (b) Find $\mathbb{E}[N|P]$. (c) Find $\mathbb{E}N$.

Solution. (a) The marginal distribution of N reads as

$$\mathbb{P}(N = n) = \int_0^1 \mathbb{P}(N = n|P = p)p_P(p)\mathrm{d}p = \int_0^1 \binom{4}{n} p^n (1-p)^{4-n}\mathrm{d}p$$

for $n = 0, 1, 2, 3, 4$ (this may be calculated explicitly but it is a bit time consuming). (b) $N|P = p \sim Binomial(4, p)$ we have $\mathbb{E}[N|P = p] = 4p$ thus $\mathbb{E}[N|P] = 4P$. (c) We have $\mathbb{E}N = \mathbb{E}[\mathbb{E}[N|P]] = \mathbb{E}[4P] = 4\mathbb{E}P = 4\int_0^1 p \cdot 1\mathrm{d}p = 4\frac{1}{2}\, p^2\big|_{p=0}^{p=1} = 2$. ■

Problem 62. Let X be a random variable with the following distribution function (PMF):

$$\mathbb{P}(X = 1) = 0.2$$
$$\mathbb{P}(X = 2) = 0.3$$
$$\mathbb{P}(X = 3) = 0.3$$
$$\mathbb{P}(X = 4) = 0.2$$

Find $\mathbb{E}X$, $\mathbb{E}X^2$, the variance and skewness.

Solution. The mean is:

$$\mathbb{E}(X) = 1 * 0.2 + 2 * 0.3 + 3 * 0.3 + 4 * 0.2 = 2.5$$

Second moment:

$$\mathbb{E}(X^2) = 1 * 0.2 + 4 * 0.3 + 9 * 0.3 + 16 * 0.2 = 7.3$$

Variance:

$$\sigma^2 = var(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = 7.3 - (2.5)^2 = 1.05$$

Skewness:

$$\frac{\mathbb{E}(X-\mathbb{E}X)^3}{\sigma^3}$$
$$= \frac{0.2*(1-2.5)^3+0.3*(2-2.5)^3+0.3*(3-2.5)^3+0.2*(4-2.5)^3}{(1.05)^{3/2}} = 0$$

■

Problem 63. A random variable X has binomial distribution $B(3, 0.4)$. See `https://en.wikipedia.org/wiki/Binomial_distribution`
Find $\mathbb{P}(X = 0)$, $\mathbb{P}(X = 2)$ and $\mathbb{P}(X = 10)$. Calculate the standard deviation of X.

Solution. Variance is npq, where $p = 0.4$ and $q = 1 - p = 0.6$. Thus, $npq = 0.72$. Standard deviation is the square root: $\sqrt{0.72} \approx 0.8485$. The PMF is

$$\binom{n}{k}p^k q^{n-k}$$
$$\mathbb{P}(X=0) = \binom{3}{0}p^0q^{3-0} = \frac{3!}{0!(3-0)!}(0.4)^0(0.6)^3 = 0.6^3 = 0.216$$
$$\mathbb{P}(X=2) = \binom{3}{2}p^2q^{3-2} = \frac{3!}{2!(3-2)!}(0.4)^2(0.6)^1 = 0.288$$

$\mathbb{P}(X = 10)$ does not exist since 10>3. ■

Problem 64. Let X be Poisson with parameter 4. For which value $k = 0, 1, \ldots$ does X attain the greatest probability? Calculate or estimate $\mathbb{P}(X \leq 3)$ and $\mathbb{P}(X \geq 5)$.

Solution. For $k = 0, 1, \ldots$ we have $\frac{\mathbb{P}(X=k+1)}{\mathbb{P}(X=k)} = \frac{e^{-4}4^{k+1}/(k+1)!}{e^{-4}4^k/k!} = \frac{4}{k+1}$. Thus, $\frac{\mathbb{P}(X=k+1)}{\mathbb{P}(X=k)} > 1$ for $k = 0, 1, 2$, $\frac{\mathbb{P}(X=k+1)}{\mathbb{P}(X=k)} = 1$ for $k = 3$ and $\frac{\mathbb{P}(X=k+1)}{\mathbb{P}(X=k)} < 1$ for $k = 4, 5, \ldots$ and we have $\mathbb{P}(X = 0) < \mathbb{P}(X = 1) < \mathbb{P}(X = 2) < \mathbb{P}(X = 3) = \mathbb{P}(X = 4) > \mathbb{P}(X = 5) > \ldots$. X attains with the greatest probability values 3 and 4.

■

Problem 65. Find the value of the constant c such that $f : \mathbb{R} \to \mathbb{R}$,

$$p_X(x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{c}{x^2} & \text{if } x \geq 1 \end{cases}$$

is a bona fide PDF of a continuous rv X. Calculate $\mathbb{P}(X \leq 2)$, $\mathbb{P}(X = 2)$, $\mathbb{P}(X \in [2, 3])$. Compute $\mathbb{E}X^2$ and $\mathbb{E}\sqrt{X}$.

Solution. We calculate $1 = \int_{-\infty}^{\infty} p_X(x)\mathrm{d}x = \int_1^{\infty} \frac{c}{x^2}\mathrm{d}x = \int_1^{\infty} cx^{-2}\mathrm{d}x = c\frac{x^{-2+1}}{-2+1}\Big|_1^{\infty} = c(-\frac{1}{x})\Big|_1^{\infty} = c(-\frac{1}{\infty} - (-\frac{1}{1})) = c(0+\frac{1}{1}) = c$. Therefore, $c = 1$. We have $\mathbb{P}(X \leq 2) = \int_{-\infty}^{2} p_X(x)\mathrm{d}x = \int_1^2 \frac{1}{x^2}\mathrm{d}x = (-\frac{1}{x})\Big|_1^2 = -\frac{1}{2} - (-\frac{1}{1}) = \frac{1}{2}$, $\mathbb{P}(X = 2) = \int_2^2 p_X(x)\mathrm{d}x = 0$, $\mathbb{P}(X \in [2,3]) = \int_2^3 p_X(x)\mathrm{d}x = \int_2^3 \frac{1}{x^2}\mathrm{d}x = (-\frac{1}{x})\Big|_2^3 = -\frac{1}{3} - (-\frac{1}{2}) = \frac{1}{6}$. Next, we calculate $\mathbb{E}X^2 = \int_{-\infty}^{\infty} x^2 p_X(x)\mathrm{d}x = \int_1^{\infty} x^2 \frac{1}{x^2}\mathrm{d}x = x|_1^{\infty} = \infty - 1 = \infty$, thus $\mathbb{E}X^2$ does not exist. $\mathbb{E}\sqrt{X} = \int_{-\infty}^{\infty} x^{1/2} p_X(x)\mathrm{d}x = \int_1^{\infty} x^{1/2}\frac{1}{x^2}\mathrm{d}x = \int_1^{\infty} x^{-3/2}\mathrm{d}x = \frac{x^{-3/2+1}}{-3/2+1}\Big|_1^{\infty} = 0 - \frac{1}{-1/2} = 2$, thus $\mathbb{E}\sqrt{X}$ is finite. ■

Problem 66. Compute the fourth moment of the normal random variable.

Solution. Solutions can be found at:

`https://arxiv.org/pdf/1209.4340.pdf`

`https://www.le.ac.uk/users/dsgp1/COURSES/MATHSTAT/6normgf.pdf`

Integrals can be computed explicitly. Let $I_k(a)$ denote:

$$I_k(a) = \frac{1}{\sqrt{2\pi}} \int_a^{\infty} u^k e^{-u^2/2}\mathrm{d}u$$

The $k = 0$ case is given in terms of the standard normal CDF:

$$I_0(a) = \frac{1}{\sqrt{2\pi}} \int_a^{\infty} e^{-u^2/2}\mathrm{d}u = 1 - \Phi(a)$$

The $k = 1$ case is obtained by direct integration:

$$I_1(a) = \frac{1}{\sqrt{2\pi}} \int_a^{\infty} ue^{-u^2/2}\mathrm{d}u = -\frac{1}{\sqrt{2\pi}} \left[e^{-u^2/2}\right]_a^{\infty} = \frac{1}{\sqrt{2\pi}} e^{-a^2/2}.$$

The $k = 2$ case is obtained by integration-by-parts:

$$\begin{aligned} I_2(a) =& \frac{1}{\sqrt{2\pi}} \int_a^{\infty} u^2 e^{-u^2/2}\mathrm{d}u = -\frac{1}{\sqrt{2\pi}} \left[ue^{-u^2/2}\right]_a^{\infty} + \frac{1}{\sqrt{2\pi}} \int_a^{\infty} \left[e^{-u^2/2}\right] \mathrm{d}u \\ =& \frac{1}{\sqrt{2\pi}} ae^{-a^2/2} + (1 - \Phi(a)) \end{aligned}$$

These are solved using integration by parts For $k = 3$, we can also integrate by parts:

$$\begin{aligned} I_3(a) =& \frac{1}{\sqrt{2\pi}} \int_a^{\infty} u^3 e^{-u^2/2}\mathrm{d}u = \frac{1}{\sqrt{2\pi}} \int_a^{\infty} u^2 \left[ue^{-u^2/2}\right] \mathrm{d}u \\ =& -\frac{1}{\sqrt{2\pi}} \left[u^2 e^{-u^2/2}\right]_a^{\infty} + 2\frac{1}{\sqrt{2\pi}} \int_a^{\infty} u \left[e^{-u^2/2}\right] \mathrm{d}u \\ =& \frac{1}{\sqrt{2\pi}} a^2 e^{-a^2/2} + \frac{1}{\sqrt{2\pi}} 2e^{-a^2/2} \end{aligned}$$

For $k = 4$, we have

$$\begin{aligned} I_4(a) =& \frac{1}{\sqrt{2\pi}} \int_a^\infty u^4 e^{-u^2/2} \mathrm{d}u = \frac{1}{\sqrt{2\pi}} \int_a^\infty u^3 \left[u e^{-u^2/2} \right] \mathrm{d}u \\ =& -\frac{1}{\sqrt{2\pi}} \left[u^3 e^{-u^2/2} \right]_a^\infty + 3 \frac{1}{\sqrt{2\pi}} \int_a^\infty u^2 \left[e^{-u^2/2} \right] \mathrm{d}u \end{aligned}$$

The last integral was already solved in the $k = 2$ case. Substituting that results gives:

$$I_4(a) = \frac{1}{\sqrt{2\pi}} a^3 e^{-a^2/2} + 3 \left[\frac{1}{\sqrt{2\pi}} a e^{-a^2/2} + (1 - \Phi(a)) \right]$$

We are, of course, interested in the limit $a \to -\infty$. For a normal $\mathcal{N}(\mu, \sigma^2)$ rv we simply make the substitution $u = \frac{x-\mu}{\sigma}$ and use the above formulae. Specifically,

$$\mathbb{E}X^4 = \int_{-\infty}^\infty x^4 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \mathrm{d}x$$

The substitution $u = \frac{x-\mu}{\sigma}$, $\mathrm{d}u = \mathrm{d}x/\sigma$:

$$\mathbb{E}X^4 = \int_{-\infty}^\infty (\sigma u + \mu)^4 \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \mathrm{d}u$$

Expanding

$$(\sigma u + \mu)^4 = \mu^4 + \sigma^4 u^4 + 4\sigma^3 u^3 \mu + 6\sigma^2 u^2 \mu^2 + 4\mu^3 \sigma u$$

gives

$$\begin{aligned} \mathbb{E}X^4 =& \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty \left[\mu^4 + \sigma^4 u^4 + 4\sigma^3 u^3 \mu + 6\sigma^2 u^2 \mu^2 + 4\mu^3 \sigma u \right] e^{-u^2/2} \mathrm{d}u \\ =& \mu^4 I_0(-\infty) + \sigma^4 I_4(-\infty) + 6\sigma^2 \mu^2 I_2(-\infty) + 4\mu^2 \sigma I_1(-\infty) \\ =& \mu^4 + \sigma^4 \cdot 3 + 6\sigma^2 \mu^2 + 4\mu^2 \sigma \cdot 0 = \mu^4 + 3\sigma^4 + 6\sigma^2 \mu^2 \end{aligned}$$

The fourth moment of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ is: $\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$. ■

Problem 67. Find the formula for $\mathbb{P}(X > t)$ of X and the CDF of X when X has the PDF:

$$p_X(x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{2}{x^3} & \text{if } x \geq 1 \end{cases}$$

Solution. The CDF is $\mathbb{P}(X \leq t) = \int_{-\infty}^t p_X(x) \mathrm{d}x$. For $t < 1$ we have 0 since the PDF is zero in that region. For $t \geq 1$, $\mathbb{P}(X \leq t) = \int_1^t \frac{2}{x^3} \mathrm{d}x = 2 \frac{x^{-3+1}}{-3+1} \Big|_1^t = 2 \left(\frac{t^{-2}}{-2} - \frac{1^{-2}}{-2} \right) = 1 - \frac{1}{t^2}$. Finally, $\mathbb{P}(X > t) = 1 - \mathbb{P}(X \leq t) = 1$ for $t < 1$ and $\frac{1}{t^2}$ for $t \geq 1$. ■

Problem 68. Let rv X be Erlang-distributed with parameters 2 and 5, see for details:

`https://en.wikipedia.org/wiki/Erlang_distribution`

Find the formula for $\mathbb{P}(X > t)$ of X and the CDF of X.

Solution. Erlang(2,5) distribution has PDF $5^2 x e^{-5x}$. Then, $\int x e^{-5x} dx = \int x(-\frac{1}{5}e^{-5x})' \mathrm{d}x$ = - $x\frac{1}{5}e^{-5x}$ + $\int x' \frac{1}{5} e^{-5x} \mathrm{d}x$ = -$\frac{1}{5}xe^{-5x}$ + $\frac{1}{5}\int e^{-5x}\mathrm{d}x$ = -$\frac{1}{5}xe^{-5x}$ - $\frac{1}{5^2}e^{-5x}$ = - $\frac{1}{25}e^{-5x}(5x+1)$. For $t > 0$, $\mathbb{P}(X > t) = \int_t^\infty 25xe^{-5x}\mathrm{d}x = 25\int_t^\infty xe^{-5x}\mathrm{d}x - e^{-5x}(5x+1)\big|_{x=t}^{\infty} = 0 + e^{-5t}(5t+1)$ while for $t \le 0$ we have $\mathbb{P}(X > t) = 1$. The CDF of X is $\mathbb{P}(X \le t) = 1 - \mathbb{P}(X > t)$, which equals 1 for $t < 0$ and $1 - e^{-5t}(5t+1)$ for $t \ge 0$. ■

Problem 69. Let the rv X have the following PMF ($k = 1, 2, \dots$):

$$\mathbb{P}(X = k) = \frac{1}{k^4} - \frac{1}{(k+1)^4}.$$

Find the CDF of X. Compute $\mathbb{P}(X \ge k)$ for $k = 0, 1, 2, \dots$.

Solution. For $k = 1, 2, \dots$ we have $\mathbb{P}(X \le k) = \sum_{i=1}^k \mathbb{P}(X = i) = \sum_{i=1}^k \left(\frac{1}{i^4} - \frac{1}{(i+1)^4}\right) = \frac{1}{1^4} - \frac{1}{2^4} + \frac{1}{2^4} - \frac{1}{3^4} + \dots + \frac{1}{k^4}$ - $\frac{1}{(k+1)^4} = 1 - \frac{1}{(k+1)^4}$. Now, for any $t \in \mathbb{R}$ we have $\mathbb{P}(X \le t) = 0$ for $t < 1$. Also, $\mathbb{P}(X \le t) = \mathbb{P}(X \le \lfloor t \rfloor) = 1 - \frac{1}{(\lfloor t \rfloor + 1)^4}$ for $t \ge 1$. To calculate $\mathbb{P}(X \ge k)$ for $k = 1, 2, \dots$ we write $\mathbb{P}(X \ge k) = \sum_{i=k}^\infty \mathbb{P}(X = i) = \sum_{i=k}^\infty \left(\frac{1}{i^4} - \frac{1}{(i+1)^4}\right) = \frac{1}{k^4} - \frac{1}{(k+1)^4} + \frac{1}{(k+1)^4} - \frac{1}{(k+2)^4} + \dots = \frac{1}{k^4}$. ■

Problem 70. Let (X, Y) be a pair of continuous rv's whose joint density is

$$p_{XY}(x, y) = \frac{1}{2}\mathbf{1}_{[0,1]}(x)\mathbf{1}_{[0,2]}(y),$$

where $\mathbf{1}_A(x)$ is the indicator function of the set A, i.e.

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

Find the CDF of the vector (X, Y).

Solution. Let $s \ge 0$, $t \ge 0$. The CDF is:

$$\mathbb{P}(X < s, Y < t) = \frac{1}{2}\int_0^s \mathbf{1}_{[0,1]}(x)\mathrm{d}x \int_0^t \mathbf{1}_{[0,2]}(y)\mathrm{d}y = \frac{(s \wedge 1)(t \wedge 2)}{2}$$

where $u \wedge v$ is the minimum of u and v. ■

Problem 71. The joint PMF of (X, Y) is

	Y=1	2	3
X=0	0.2	0.1	0
1	0.1	0.3	0
2	0	0	0.3

Find the marginal probability mass functions of X and Y. Find the conditional probabilities $\mathbb{P}(X=0|Y=1)$, $\mathbb{P}(X=1|Y=1)$, $\mathbb{P}(X=2|Y=1)$, $\mathbb{P}(X=0|Y=2)$, $\mathbb{P}(X=1|Y=2)$, $\mathbb{P}(X=2|Y=2)$.

Solution. The definition of conditional probability is $\mathbb{P}(A|B)=\frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}$. To calculate

$$\mathbb{P}(X=0|Y=1)=\frac{\mathbb{P}(X=0,Y=1)}{\mathbb{P}(Y=1)}=\frac{0.2}{0.3}=\frac{2}{3},$$

where $\mathbb{P}(Y=1)=0.2+0.1+0=0.3$ and $\mathbb{P}(X=0,Y=1)=0.2$. Other conditional probabilities are calculated similarly. We find:

$$\mathbb{P}(X=1|Y=1)=\frac{0.1}{0.3}=\frac{1}{3}$$
$$\mathbb{P}(X=2|Y=1)=\frac{0}{0.3}=0$$
$$\mathbb{P}(X=0|Y=2)=\frac{0.1}{0.4}=\frac{1}{4}$$
$$\mathbb{P}(X=1|Y=2)=\frac{0.3}{0.4}=\frac{3}{4}$$
$$\mathbb{P}(X=2|Y=2)=\frac{0}{0.4}=0$$

■

Problem 72. The random vector (X,Y) is uniformly distributed over the following region in the 2D plane:

$$D=\{(x,y)\in\mathbb{R}^2 : x^2+y^2\le 2\}$$

i.e., the joint PDF is

$$p_{XY}(x,y)=\frac{1}{2\pi}\mathbf{1}_D(x,y)=\begin{cases}\frac{1}{2\pi} & \text{if } x^2+y^2\le 2;\\ 0 & \text{if } x^2+y^2>2.\end{cases}$$

Find the marginal densities of X and Y.

Solution. We apply the formula

$$p_X(x)=\int_{-\infty}^{\infty}p_{XY}(x,y)\mathrm{d}y=\frac{1}{2\pi}\int_{-\infty}^{\infty}\mathbf{1}_D(x,y)\mathrm{d}y$$

For $x<-\sqrt{2}$ and $x>\sqrt{2}$ we have $x^2+y^2>2$. Thus, $\mathbf{1}_D(x,y)=0$ and $p_X(x)=0$. Assume that $x\in[-\sqrt{2},\sqrt{2}]$. We have $\mathbf{1}_D(x,y)=1$ iff $y\in[-\sqrt{2-x^2},\sqrt{2-x^2}]$ and otherwise $\mathbf{1}_D(x,y)=0$. Then, $p_X(x)=$

$\frac{1}{2\pi}\int_{-\infty}^{\infty}\mathbf{1}_D(x,y)\mathrm{d}y = \frac{1}{2\pi}\int_{-\sqrt{2-x^2}}^{\sqrt{2-x^2}}1\mathrm{d}y = \frac{1}{\pi}\sqrt{2-x^2}$. Similarly, $p_Y(y)=0$ for $y<-\sqrt{2}$ and $y>\sqrt{2}$, and $p_Y(y)=\frac{1}{\pi}\sqrt{2-y^2}$ for $y\in[-\sqrt{2},\sqrt{2}]$. ■

Problem 73. Prove that X and Y, whose joint PDF is defined in Problem 72, are statistically independent. Calculate the covariance between X and Y.

Solution. Using the marginal densities obtained in Problem 72,
$p_X(x)=\frac{1}{\pi}\sqrt{2-x^2}$ for $x\in[-\sqrt{2},\sqrt{2}]$
$p_Y(y)=\frac{1}{\pi}\sqrt{2-y^2}$ for $y\in[-\sqrt{2},\sqrt{2}]$,
and

$$p_{XY}(x,y)=\frac{1}{2\pi}\mathbf{1}_D(x,y)=\begin{cases}\frac{1}{2\pi} & \text{if } x^2+y^2\le 2;\\ 0 & \text{if } x^2+y^2>2.\end{cases}$$

$$D=\{(x,y)\in\mathbb{R}^2 : x^2+y^2\le 2\}$$

we find that $p_{XY}(x,y)\neq p_X(x)p_Y(y)$. Thus X and Y are not statistically independent. The covariance is defined as

$$cov(X,Y)=\mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)]=\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]$$

Via direct computation:

$$\begin{aligned}\mathbb{E}[XY]=&\frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{d}x\int_{-\infty}^{\infty}xy\,\mathbf{1}_D(x,y)\,\mathrm{d}y=\frac{1}{2\pi}\iint_D xy\,\mathrm{d}x\mathrm{d}y\\ =&\frac{1}{2\pi}\int_{-2}^{2}\mathrm{d}x\int_{-\sqrt{2-x^2}}^{\sqrt{2-x^2}}xy\,\mathrm{d}y\\ =&\frac{1}{2\pi}\int_{-2}^{2}x\frac{1}{2}\left[(2-x^2)-(2-x^2)\right]\mathrm{d}x=0\end{aligned}$$

Also, we have that $\mathbb{E}[X]=0$ and $\mathbb{E}[Y]=0$ since their marginal densities are symmetric. Therefore, $cov(X,Y)=0$. This is an instance of two random variables that are statistically independent but uncorrelated. ■

Problem 74. Let X and Y be rv's whose joint PMF is given by:

	Y=1	2	3
X=0	0.2	0.1	0
1	0.1	0.3	0
2	0	0	0.3

Compute the covariance and correlation matrix of the random vector (X,Y).

Solution. Let $\mathbf{X} = (X, Y)$. The covariance matrix is:

$$cov(\mathbf{X}) = \begin{bmatrix} cov(X,X) & cov(X,Y) \\ cov(Y,X) & cov(Y,Y) \end{bmatrix} = \begin{bmatrix} \mathbb{E}(X-\mu_X)^2 & \mathbb{E}(X-\mu_X)(Y-\mu_Y) \\ \mathbb{E}(Y-\mu_Y)(X-\mu_X) & \mathbb{E}(Y-\mu_Y)^2 \end{bmatrix}$$

The correlation matrix is the covariance matrix whose entries are normalized (see correlation coefficient):

$$corr(\mathbf{X}) = \begin{bmatrix} \frac{\mathbb{E}(X-\mu_X)^2}{\sigma_X^2} & \frac{\mathbb{E}(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y} \\ \frac{\mathbb{E}(Y-\mu_Y)(X-\mu_X)}{\sigma_X\sigma_Y} & \frac{\mathbb{E}(Y-\mu_Y)^2}{\sigma_Y^2} \end{bmatrix} = \begin{bmatrix} 1 & \frac{\mathbb{E}(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y} \\ \frac{\mathbb{E}(Y-\mu_Y)(X-\mu_X)}{\sigma_X\sigma_Y} & 1 \end{bmatrix}$$

Computing the matrix elements:

$$\mu_X = \mathbb{E}X = 0.4 + 2 * 0.3 = 1$$

$$\mu_Y = \mathbb{E}Y = 1 * (0.2 + 0.1) + 2 * (0.1 + 0.3) + 3 * (0.3) = 2$$

$\sigma_X^2 = \mathbb{E}(X-\mu_X)^2 = (0.2+0.1)*(0-1)^2+(0.1+0.3)*(1-1)^2+0.3*(2-1)^2 = 0.6$

$\sigma_Y^2 = \mathbb{E}(Y-\mu_Y)^2 = (0.2+0.1)*(1-2)^2+(0.1+0.3)*(2-2)^2+0.3*(3-2)^2 = 0.6$

The off diagonal element is:

$$\begin{aligned}\mathbb{E}(X-\mu_X)(Y-\mu_Y) =& 0.2*(0-1)(1-2) + 0.1*(0-1)(2-2) \\ &+ 0.1*(1-1)(1-2) + 0.3*(1-1)(2-2) + 0.3*(2-1)(3-2) = 0.5\end{aligned}$$

Thus, we arrive at:

$$cov(\mathbf{X}) = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}, \qquad corr(\mathbf{X}) = \begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix}$$

■

Problem 75. Let rv X and Y have a joint PDF

$$p_{XY}(x,y) = \begin{cases} \frac{1}{2} & \text{if } 0 \le x \le y \le 2; \\ 0 & \text{otherwise} \end{cases}.$$

Are X and Y statistically independent? Compute the correlation matrix of the random vector (X, Y).

Solution.

$$\begin{aligned}\mathbb{E}XY =& \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} xy\, p_{XY}(x,y)\mathrm{d}y\right)\mathrm{d}x = \int_0^2\left(\int_x^2 \frac{1}{2}\mathrm{d}y\right)\mathrm{d}x = \frac{1}{2}\int_0^2 x\left(\int_x^2 y\mathrm{d}y\right)\mathrm{d}x \\ =& \frac{1}{2}\int_0^2 x\frac{1}{2}\, y^2\Big|_{y=x}^{y=2}\mathrm{d}x = \frac{1}{4}\int_0^2 x(4-x^2)\mathrm{d}x = \frac{1}{4}\int_0^2 (4x-x^3)\mathrm{d}x = \frac{1}{4}\,(4\frac{1}{2}x^2 - \frac{1}{4}x^4)\Big|_{x=0}^{x=2} \\ =& \frac{1}{4}(2\cdot 2^2 - \frac{1}{4}2^4) = 1.\end{aligned}$$

Next,

$$\begin{aligned}\mathbb{E}X &= \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} p_{XY}(x,y)\mathrm{d}y \right) \mathrm{d}x = \int_0^2 x \left(\int_x^2 \frac{1}{2}\mathrm{d}y \right) \mathrm{d}x = \frac{1}{2}\int_0^2 x \left(\int_x^2 1\mathrm{d}y \right) \mathrm{d}x \\ &= \frac{1}{2}\int_0^2 x \cdot y|_{y=x}^{y=2}\,\mathrm{d}x = \frac{1}{2}\int_0^2 x(2-x)\mathrm{d}x = \frac{1}{2}\int_0^2 (2x-x^2)\mathrm{d}x = \frac{1}{2}(2\frac{1}{2}x^2 - \frac{1}{3}x^3)\Big|_{x=0}^{x=2} \\ &= \frac{1}{2}(2^2 - \frac{1}{3}2^3) = \frac{2}{3}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}X^2 &= \int_{-\infty}^{\infty} x^2 \left(\int_{-\infty}^{\infty} p_{XY}(x,y)\mathrm{d}y \right) \mathrm{d}x = \int_0^2 x^2 \left(\int_x^2 \frac{1}{2}\mathrm{d}y \right) \mathrm{d}x = \frac{1}{2}\int_0^2 x^2 \left(\int_x^2 1\mathrm{d}y \right) \mathrm{d}x \\ &= \frac{1}{2}\int_0^2 x^2 \cdot y|_{y=x}^{y=2}\,\mathrm{d}x = \frac{1}{2}\int_0^2 x^2(2-x)\mathrm{d}x = \frac{1}{2}\int_0^2 (2x^2-x^3)\mathrm{d}x \\ &= \frac{1}{2}\left(2\frac{1}{3}x^3 - \frac{1}{4}x^4\right)\Big|_{x=0}^{x=2} = \frac{1}{2}\left(2\frac{1}{3}2^3 - \frac{1}{4}2^4\right) = \frac{2}{3}.\end{aligned}$$

We also have

$$\begin{aligned}\mathbb{E}Y &= \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} p_{XY}(x,y)\mathrm{d}x \right) \mathrm{d}y = \int_0^2 y \left(\int_0^y \frac{1}{2}\mathrm{d}x \right) dy = \frac{1}{2}\int_0^2 y \left(\int_0^y 1\mathrm{d}x \right) \mathrm{d}y \\ &= \frac{1}{2}\int_0^2 y \cdot x|_{x=0}^{x=y}\,\mathrm{d}y = \frac{1}{2}\int_0^2 y^2\mathrm{d}y = \frac{1}{2}\frac{1}{3}\,y^3\big|_{y=0}^{y=2} = \frac{1}{6}2^3 = \frac{4}{3}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}Y^2 &= \int_{-\infty}^{\infty} y^2 \left(\int_{-\infty}^{\infty} p_{XY}(x,y)\mathrm{d}x \right) \mathrm{d}y = \int_0^2 y^2 \left(\int_0^y \frac{1}{2}\mathrm{d}x \right) \mathrm{d}y = \frac{1}{2}\int_0^2 y^2 \left(\int_0^y 1\mathrm{d}x \right) \mathrm{d}y \\ &= \frac{1}{2}\int_0^2 y^2 \cdot x|_{x=0}^{x=y}\,\mathrm{d}y = \frac{1}{2}\int_0^2 y^3\mathrm{d}y = \frac{1}{2}\frac{1}{4}\,y^4\big|_{y=0}^{y=2} = \frac{1}{8}2^4 = 2.\end{aligned}$$

Then,

$$\begin{aligned}\rho(X,Y) &= \frac{cov(X,Y)}{\sigma_X\sigma_Y} = \frac{\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y}{\sqrt{\mathbb{E}X^2 - (\mathbb{E}X)^2}\sqrt{\mathbb{E}Y^2 - (\mathbb{E}Y)^2}} \\ &= \frac{1 - \frac{2}{3}\frac{4}{3}}{\sqrt{\frac{2}{3} - (\frac{2}{3})^2}\sqrt{2 - (\frac{4}{3})^2}} = \frac{\frac{1}{9}}{\sqrt{\frac{2}{9}}\sqrt{\frac{2}{9}}} = \frac{1}{2}\end{aligned}$$

and the correlation matrix of (X,Y) reads

$$\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}.$$

Since $\rho(X,Y) \neq 0$ the variables X and Y are dependent. ■

Problem 76. There is a bridge in Durham, NC nicknamed the "can opener" bridge. Watch this 10-minutes long compilation:

https://www.youtube.com/watch?v=USu8vT_tfdw

The meaning of the bridge's name should be apparent from this video. Consider all oversized trucks shown in the video. The trucks either get through with significant damage (can opener) or with minimal damage. We consider 2 different scenarios:

(A) While the truck is significantly oversized, the truck driver goes through anyways, causing the truck to undergo carnage and decapitation.

(B) Truck either follows the sign and turns away, or goes through anyways and the truck suffers minimal damage (small bump, then backing out) or barely scraping under (lucky driver).

Count the number of times you observe scenarios A and B. From this data, assign probabilities for events A and B. Suppose that type A events are associated with a low IQ truck driver (IQ=60), whereas type B events are associated with a higher IQ driver (IQ=140). Compute the average IQ of a truck driver in Durham, NC. (Note: This problem is a joke; we are not implying that truck drivers from anywhere are idiots.)

Solution. Suppose we count 15 severely damaged trucks and 5 mildly damaged ones. The probability of A is

$$\mathbb{P}(A) = \frac{15}{20} = 0.75$$

The probability of B is:

$$\mathbb{P}(B) = \frac{5}{20} = 0.25$$

The average IQ is:

$$\mathbb{E}(IQ) = IQ(A) \cdot \mathbb{P}(A) + IQ(B) \cdot \mathbb{P}(B) = 60 \cdot 0.75 + 140 \cdot 0.25 = 80.$$

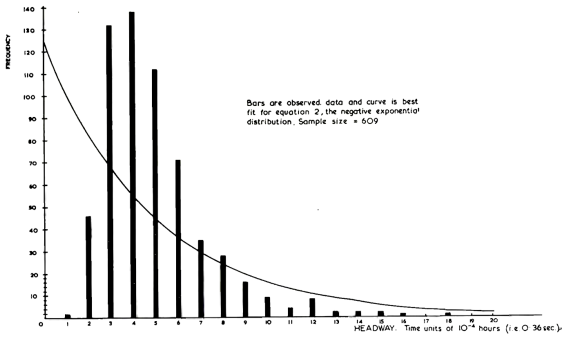
■

Problem 77. Watch 10 minutes of traffic video (preferably traffic that is not too dense, so you are able to count events). This webcam appears suitable:

https://www.youtube.com/watch?v=5_XSYlAfJZM

Choose a landmark such as a line on the road. Pick a lane of traffic. Count the time interval τ between consecutive vehicles crossing that lane. Plot a histogram of the time intervals. Compute the average $\langle \tau \rangle$. What distribution does τ follow? Fit the histogram to a suitable distribution. Obtain the parameters of the distribution.

Solution. An example data set is:



By inspection of this graph, $\langle \tau \rangle \approx 4 \cdot 10^{-4}$ hours. The distribution is called "headway distribution" or gap distribution. The commonly used distributions include the "displaced exponential" (for low-medium flows) and "Schuhl's composite exponential" (for normal-heavy flows) distributions. ■

Problem 78. For the traffic problem (#2) pick a time interval, say 4 minutes. Count the number of cars, n, that pass through the intersection/line (in a given lane) during that time interval. Plot of histogram of n. Find the distribution of n. Obtain the parameters of the distribution.

Solution. Suppose that we have 180 time windows (each lasting 4 seconds) and record the following observations (x: number of vehicles arriving per 4 second interval):

x	Obs. freq.	Total vehicles	Probability $P(x)$	Theoretical freq.
0	94	0	0.539	97.0
1	63	63	0.333	59.9
2	21	42	0.103	18.5
3	2	6	0.021	3.8
> 3	0	0	0.004	0.8
Total	180	111	1.000	180.0

To get the histogram, we plot the vector of observed frequencies vs x. In MATLAB, we could type

```
plot([94 63 21 2 0],'o');
```

The graph doesn't quite look like an exponential decay. On the other hand, a Poisson distribution seems suitable. The probability distribution function for Poisson takes the form:

$$P(k) = \frac{m^k e^{-m}}{k!}$$

where λ is a parameter to be derived from the data. Its physical interpretation is the average number of cars per 4-second time interval. Since there are 180 time intervals in our experiment, and the total number of vehicles

observed is 111:

$$m = \frac{\text{total vehicles}}{\text{total periods}} = \frac{111}{180} = 0.617; e^{-.617} = .539$$

$$P(x) = \frac{(0.617)^x}{x!} e^{-.617} = \frac{(.617)^x (.539)}{x!}$$

In the above table the column $P(x)$ is the probability calculated using the Poisson formula. The calculated "theoretical frequency" is equal to 180 $P(x)$. ■

Problem 79. For problems 76, 77 and 78 describe the probability space, the set of elementary outcomes, the random variable and the random events considered.

Solution. For problem 76, the set of possible outcomes, Ω, is the set of all possible trajectories $\omega \in \Omega$ that a given truck can take (this is best left as abstract). There are two events considered here: $A(\omega)$ (high impact), $B(\omega)$ (low or no impact). The random variable considered here is the IQ of a driver: $IQ(\omega)$, where ω refers to a particular truck/driver trajectory.

For problem 77, the set Ω of possible outcomes ($\omega \in \Omega$) is the traffic flow, i.e. all traffic scenarios giving rise to all possible gaps between consecutive cars (or some similar idea). We may consider events of the type $\{\tau = t\}$. Each of these events has probability zero (since the time intervals/bins have zero duration), however, for purposes of plotting a histogram we need to consider finite intervals of the form $\{t_1 \leq \tau \leq t_2\}$. The random variable is $\tau(\omega)$.

For problem 78, the set of outcomes is the same as in Problem 2, since the physical random experiment is the same (traffic flow). The random variable is $n(\omega)$, the number of cars in a given time interval. The events are of the form $\{n = x\}$, where x is an integer value (0, 1, 2, 3, ...). ■

Problem 80. In probability theory we often use integrals over sets. This is the same integral as you are used to, but written differently. For example, the integral of the exponential distribution, e^{-x}, over the set $[0, 1]$ is:

$$\int_{[1,3]} e^{-x} \mathrm{d}x = \int_1^3 e^{-x} \mathrm{d}x = -\, e^{-x}\big|_1^3 = e^{-1} - e^{-3} = 0.318$$

Let A be a set over the positive real numbers. Denote:

$$Q(A) = \int_A e^{-x} \mathrm{d}x$$

Compute:

(a) $Q([5,\infty))$

(b) $Q([1,3]\cup[3,5])$

(c) $Q([0,\infty))$

Solution. (a)

$$\int_5^\infty e^{-x}\mathrm{d}x = -\,e^{-x}\big|_5^\infty = e^{-5} \approx 0.007$$

(b)

$$\int_1^3 e^{-x}\mathrm{d}x + \int_3^5 e^{-x}\mathrm{d}x = \int_1^5 e^{-x}\mathrm{d}x \approx 0.36114$$

(c)

$$\int_0^\infty e^{-x}\mathrm{d}x = 1$$

■

Problem 81. The same can be done for multiple variables. For $A \subset \mathbb{R}^n$, define the set function:

$$Q(A) = \int\cdots\int_A \mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_n,$$

provided the integral exists. For example, if $A = \{(x_1, x_2, \ldots, x_n) : 0 \le x_1 \le x_2, 0 \le x_i \le 1, \text{ for } i = 2, 3, \ldots, n\}$, then

$$Q(A) = \int_0^1 \left[\int_0^{x_2} \mathrm{d}x_1\right]\mathrm{d}x_2 \cdot \prod_{i=3}^{n}\left[\int_0^1 \mathrm{d}x_i\right] = \frac{x_2^2}{2}\bigg|_0^1 \cdot 1 = \frac{1}{2}.$$

Let $B = \{(x_1, x_2, \ldots, x_n) : 0 \le x_1 \le x_2 \le \cdots \le x_n \le 1\}$. Calculate the numerical value of $Q(B)$.

Solution.

$$Q(B) = \int_0^1 \left[\int_0^{x_n} \ldots \left[\int_0^{x_3}\left[\int_0^{x_2} \mathrm{d}x_1\right]\mathrm{d}x_2\right]\ldots \mathrm{d}x_{n-1}\right]\mathrm{d}x_n = \frac{1}{n!}$$

where $n! = n(n-1)\ldots 3\cdot 2\cdot 1$. ■

Problem 82. Solve the following problems using set theory:

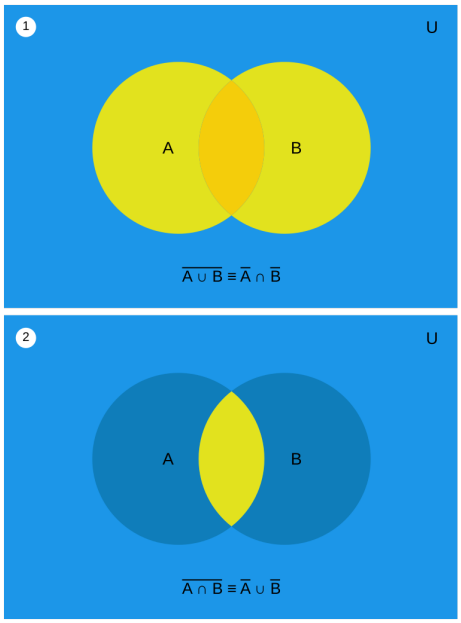
(a) Find the union $C_1 \cup C_2$ and the intersection $C_1 \cap C_2$ of the two sets C_1 and C_2, where $C_1 = \{(x, y) : 0 < x < 1, 0 < y < 3\}$, $C_2 = \{(x, y) : 0 < x < 2, 2 \le y < 3\}$.

(b) Find the complement C^c of the set C with respect to the space $\mathcal{C}$ if $\mathcal{C} = \{(x, y) : x^2 + y^2 \le 1\}$, $C = \{(x, y) : |x| + |y| < 1\}$.

(c) Prove, using Venn diagrams, that the following statements are true:

$$(A \cap B)^c = A^c \cup B^c$$
$$(A \cup B)^c = A^c \cap B^c$$

Illustrate with an example. Generalize these statements to countable unions and intersections.



(d) Consider the space $\mathcal{C}$ to be the set of points enclosed by a rectangle containing the circles C_1, C_2 and C_3. Use Venn diagrams to compare the following sets:

$$C_1 \cup (C_2 \cap C_3) \text{ and } (C_1 \cup C_2) \cap (C_1 \cup C_3)$$

(e) Show that the following sequences of sets, $\{C_k\}$, are nondecreasing (nested upwards), i.e. $C_k \subset C_{k+1}$ for $k = 1, 2, 3, \ldots$. For such a sequence, define

$$\lim_{k\to\infty} C_k = \cup_{k=1}^{\infty} C_k.$$

Take the following sequence:

$$C_k = \{(x, y) : 1/k \leq x^2 + y^2 \leq 4 - 1/k\}, \ k = 1, 2, 3, \ldots$$

Find the limit $\lim_{k\to\infty} C_k$.

(f) Show that the following sequence of sets, $\{C_k\}$, where

$$C_k = \{x : 2 < x \leq 2 + 1/k\},\ k = 1, 2, 3, \ldots,$$

is nonincreasing. A sequence of sets $\{A_n\}$ is said to be **nonincreasing** if $A_n \supset A_{n+1}$ for $n = 1, 2, 3, \ldots$. In this case, we define

$$\lim_{n\to\infty} A_n = \cap_{n=1}^{\infty} A_n.$$

Find $\lim_{k\to\infty} C_k$.

(g) For every two-dimensional set $C \subset \mathbb{R}^2$, let $Q(C) = \int\int_C (x^2 + y^2)dxdy$. If $C_1 = \{(x, y) : -1 \leq x \leq 1, -1 \leq y \leq 1\}$, $C_2 = \{(x, y) : -1 \leq x = y \leq 1\}$, and $C_3 = \{(x, y) : x^2 + y^2 \leq 1\}$, find $Q(C_1)$, $Q(C_2)$ and $Q(C_3)$.

(h) To join a club, a person must be either an idiot or a truck driver, or both. Of the 35 members in this club, 25 are idiots and 17 are truck drivers. How many persons in the club are both an idiot and a truck driver? How will these people fare when they encounter the "can opener" bridge? (Note: this problem is a joke; we are not implying that truck drivers are idiots.)

Solution. Union is a L-shaped region in the 2D plane defined by the coordinates:

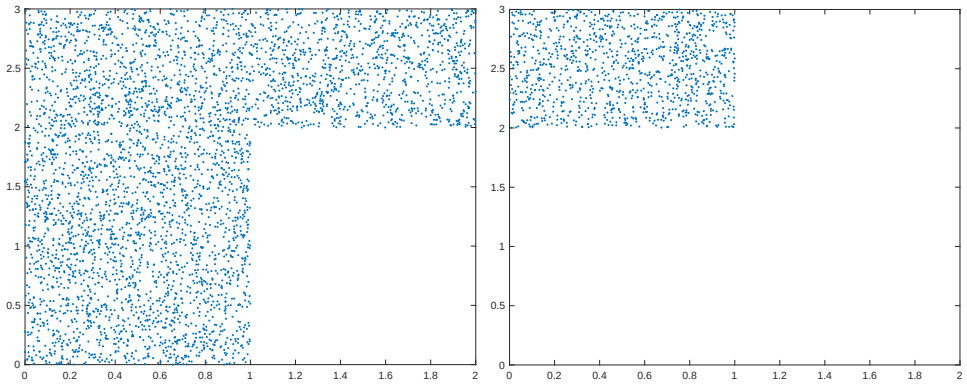
$$C_1 \cup C_2 = \{(x, y) : 0 < x < 1, 0 < y < 3 \text{ or } 0 < x < 2, 2 \leq y < 3\}$$

Intersection is a small square

$$C_1 \cap C_2 = \{(x, y) : 0 < x < 1, 2 \leq y \leq 3\}$$

(Notice the equality signs.)
The following MATLAB code can be used to plot the region

```
x=3*rand([1 10000]);
y=3*rand([1 10000]);
ll1=find(x>0 & x<1 & y>0 & y<3);
ll2=find(x>0 & x<2 & y>2 & y<3);
ll3=intersect(ll1,ll2);
ll3=union(ll1,ll2);
figure;
plot(x(ll3),y(ll3),'.');
axis([0 2 0 3]);
```

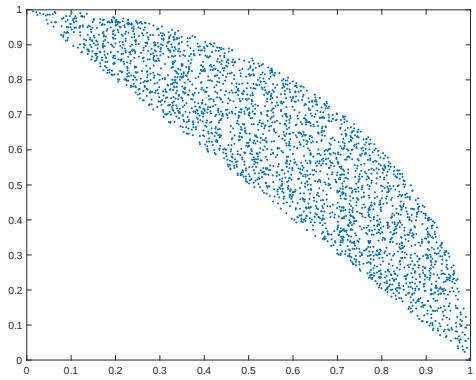


(b)

$$C^c = \{(x, y) : |x| + |y| \geq 1 \text{ and } x^2 + y^2 \leq 1\}$$

The following MATLAB code can be used to plot the region

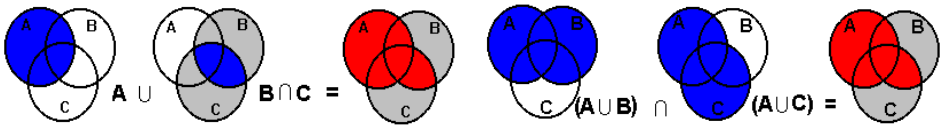
```
x=rand([1 10000]);
y=rand([1 10000]);
ll1=find(x.^2 + y.^2 < 1);
ll2=find(abs(x) + abs(y) > 1);
ll3=intersect(ll1,ll2);
figure;
plot(x(ll3),y(ll3),'.');
```



(c) Generalization is:

$$\overline{\cap_{i \in I} A_i} \equiv \cup_{i \in I} \overline{A_i}$$
$$\overline{\cup_{i \in I} A_i} \equiv \cap_{i \in I} \overline{A_i}$$

where I is some, possibly uncountable, indexing set.

(d) 

(e) The sequence is nondecreasing since

$\{(x,y) : 1/k \le x^2+y^2 \le 4-1/k\} \subset \{(x,y) : 1/(k+1) \le x^2+y^2 \le 4-1/(k+1)\}$

for all k. The limit is

$$\{(x,y) : 0 < x^2 + y^2 < 4\}$$

Note: the equality signs are gone because the end points are not part of the infinite union.

(f) The sequence is nonincreasing since

$$\{x : 2 < x \le 2 + 1/k\} \supset \{x : 2 < x \le 2 + 1/(k+1)\}$$

for all k. The limit set is:

$$\{x : 2 < x \le 2\}$$

Note: the equality sign remains because the term $1/k > 0$ for all k (even in the limit $k \to \infty$).

(g)

$$Q(C_1) = \int_{-1}^{1} \mathrm{d}x \int_{-1}^{1} (x^2+y^2)\mathrm{d}y = \frac{8}{3} \approx 2.66667$$

$Q(C_2) = \iint\limits_{C} (x^2+y^2)\mathrm{d}x\mathrm{d}y = 0$ since the set C is a thin line with zero area

$$Q(C_3) = \iint\limits_{\{(x,y)|x^2+y^2<R^2\}} (x^2+y^2)\mathrm{d}x\mathrm{d}y = \int_0^R r\mathrm{d}r \int_0^{2\pi} \mathrm{d}\theta r^2 = \frac{\pi R^4}{2} = \frac{\pi}{2}$$

(h) 25+17=42. 42-35=7. In all likelihood, the bridge shall open 7 cans of sardines. ■

Problem 83. Let Ω be the set of elementary outcomes and E a subset of Ω, called "event". Denote $\mathcal{F}$ the collection of all possible events. Technically, $\mathcal{F}$ is called a "σ-field of subsets". Let $\mathbb{P}$ be a real-valued function defined on $\mathcal{F}$. $\mathbb{P}$ is a probability set function of it satisfies the following three conditions:

(1) $\mathbb{P}(A) \ge 0$, for all $A \in \mathcal{F}$.

(2) $\mathbb{P}(\Omega) = 1$.

(3) If $\{A_n\}$ is a sequence of events in $\mathcal{F}$ and $A_m \cap A_n = \emptyset$ for all $m \neq n$, then

$$\mathbb{P}\left(\cup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

A collection of events whose members are pairwise disjoint is said to be a **mutually exclusive** collection and its union is often referred to as a **disjoint union**. The collection is further said to be **exhaustive** if the union of its events is the sample space, in which case $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = 1$. We say that a mutually exclusive and exhaustive collection of events forms a **partition** of Ω.

Using the above definition of probability:

(a) Prove that for each event $A \in \mathcal{F}$, $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$.

(b) Prove that the probability of a null set is zero, i.e. $\mathbb{P}(\emptyset) = 0$.

(c) Prove that if A and B are events such that $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

(d) Prove that for each $A \in \mathcal{F}$, $0 \leq \mathbb{P}(A) \leq 1$.

(e) Prove that if A and B are events in Ω, then

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

(f) For a finite sample space $\Omega = \{x_1, x_2, \ldots, x_m\}$ with m elements, let $p_1, p_2, \ldots, p_m$ be such that $0 \leq p_i \leq 1$ for $i = 1, 2, \ldots, m$ and $\sum_{i=1}^{m} p_i = 1$. Construct a probability set function $\mathbb{P}(A)$ on $\mathcal{F}$ (for all subsets $A \in \mathcal{F}$) such that all 3 above conditions are satisfied.

(g) Let $\Omega = \{x_1, x_2, \ldots, x_m\}$ be a finite sample space. Find the set of elementary probabilities p_i for all $i = 1, 2, \ldots, m$ such that $\mathbb{P}(A) = \#(A)/m$, where $\#(A)$ denotes the number of elements in A. Prove that $\mathbb{P}$ is a probability on Ω.

(h) Let $\Omega = \{x : 0 < x < \infty\}$. Let $C \subset \Omega$ be defined by $C = \{x : 0 < x < 10\}$. Define the function $\mathbb{P}(A) = \int_A \frac{1}{2} e^{-x/2} \mathrm{d}x$ for any event $A \subset \Omega$. Show that $\mathbb{P}(\Omega) = 1$. Evaluate $\mathbb{P}(C)$, $\mathbb{P}(C^c)$ and $\mathbb{P}(C \cap C^c)$.

Solution. (a) We have $\Omega = A \cup A^c$ and $A \cap A^c = \emptyset$. Thus from conditions 2 and 3 it follows that

$$1 = \mathbb{P}(A) + \mathbb{P}(A^c)$$

(b) Take $A = \emptyset$ so that $A^c = \Omega$. Using the result from (a),

$$\mathbb{P}(\emptyset) = 1 - \mathbb{P}(\Omega) = 1 - 1 = 0.$$

(c) Writing $B = A \cup (A^c \cap B)$ and $A \cap (A^c \cap B) = \emptyset$, condition 3 gives

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B)$$

From condition 1, $\mathbb{P}(A^c \cap B) \geq 0$. Hence, $\mathbb{P}(B) \geq \mathbb{P}(A)$.

(d) Since $\emptyset \subset A \subset \Omega$, we have by the results of part (c) that

$$\mathbb{P}(\emptyset) \leq \mathbb{P}(A) \leq \mathbb{P}(\Omega)$$

or $0 \leq \mathbb{P}(A) \leq 1$, the desired result.

(e) Each of the sets $A \cup B$ and B can be represented, respectively, as a union of nonintersecting sets as follows:

$$A \cup B = A \cup (A^c \cap B) \text{ and } B = (A \cap B) \cup (A^c \cap B).$$

These identities hold for all sets A and B, according to set theory. (You can also verify them using Venn diagrams.) From condition 3 we have

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B)$$

and

$$\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B).$$

If the second of these quantities is solved for $\mathbb{P}(A^c \cap B)$ and this result is substituted in the first equation, we obtain

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

(f) We can take $p_i = 1/m$ and $\mathbb{P}(A) = \#(A)/m$. See (g).

(g) Take the equilikely distribution $p_i = 1/m$. Define:

$$\mathbb{P}(A) = \sum_{x_i \in A} \frac{1}{m} = \frac{\#(A)}{m}.$$

Then, $\mathbb{P}$ is a probability on Ω. It is trivial to check that all 3 conditions are satisfied: $\mathbb{P}(A) \geq 0$, $\mathbb{P}(\Omega) = m/m = 1$, and for disjoint sets $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

(h)

$$\mathbb{P}(\Omega) = \int_0^\infty \tfrac{1}{2} e^{-x/2} dx = \left[-e^{-x/2}\right]_0^\infty = 0 - (-1) = 1$$

■

Problem 84. You write 3 letters and in a rush, put a random letter in each envelope. (There are 3 envelopes, 3 letters, 1 letter per envelope.). What is the probability that at least one letter is in the correct envelope?

Solution. Let C_i be the event that the i-th letter is in the correct envelope. Expand $\mathbb{P}(C_1 \cup C_2 \cup C_3)$ to determine the probability:

$$\mathbb{P}(C_1 \cup C_2 \cup C_3) = \mathbb{P}(C_1) + \mathbb{P}(C_2) + \mathbb{P}(C_3) - \mathbb{P}(C_1 \cap C_2) - \mathbb{P}(C_1 \cap C_3) \\ -\mathbb{P}(C_2 \cap C_3) + \mathbb{P}(C_1 \cap C_2 \cap C_3)$$

All pairwise terms $\mathbb{P}(C_1 \cap C_2)$, $\mathbb{P}(C_1 \cap C_3)$ and $\mathbb{P}(C_2 \cap C_3)$ are zero because it's not possible to have only 2 letters in correct envelopes without have all 3. Then,

$$\mathbb{P}(C_1 \cup C_2 \cup C_3) = \mathbb{P}(C_1) + \mathbb{P}(C_2) + \mathbb{P}(C_3) + \mathbb{P}(C_1 \cap C_2 \cap C_3).$$

Now the probabilities: There are 3!=6 ways to place 3 letters in 3 envelopes (order matters). There is 1 way to place letter 1 in envelope 1 (and only 1 way to place envelopes 2 and 3 in the remaining incorrect envelopes). Therefore $\mathbb{P}(C_1) = 1/6$. Same for $\mathbb{P}(C_2)$ and $\mathbb{P}(C_3)$. For the last term, $\mathbb{P}(C_1 \cap C_2 \cap C_3)$, we need to know the number of ways we can place all 3 letters in the correct envelopes. There's only 1 way to do that. Hence, $\mathbb{P}(C_1 \cap C_2 \cap C_3) = 1/6$. Thus,

$$\mathbb{P}(C_1 \cup C_2 \cup C_3) = \frac{4}{6}.$$

■

Problem 85. A random experiment consists of choosing a random number in the interval $(0, 1)$. (This number can be rational or irrational.) For any interval $(a, b) \subset (0, 1)$ it seems reasonable to define the probability $\mathbb{P}((a, b)) = b - a$, i.e. as equal to the length of the interval. Choose an appropriate sequence of subsets of $(0, 1)$ and use the following result:

$$\lim_{n\to\infty} \mathbb{P}(C_n) = \mathbb{P}(\lim_{n\to\infty} C_n) = \mathbb{P}\left(\cap_{n=1}^\infty C_n\right)$$

where $\{C_n\}$ is a decreasing sequence of events (i.e. $C_{n+1} \subset C_n$), to show that $\mathbb{P}(\{a\}) = 0$, for all $a \in (0, 1)$.

Solution. Construct the following decreasing sequence of events:

$$C_k = \{x : a - 1/k < x < a + 1/k\}$$

You can check that these events are open intervals $(a-1/k, a+1/k)$. Their intersection/limit is the point $\{a\}$:

$$\lim_{k\to\infty} C_k \equiv \cap_{k=1}^{\infty} C_k = \{a\}.$$

Meanwhile,

$$\mathbb{P}((a-1/k, a+1/k)) = a + 1/k - (a-1/k) = 2/k.$$

Taking the limit $k \to \infty$, we see that (applying the above 'result')

$$\mathbb{P}(\lim_{k\to\infty} C_k) = \mathbb{P}(\{a\}) = \lim_{k\to\infty} \mathbb{P}((a-1/k, a+1/k)) = \lim_{k\to\infty} \frac{2}{k} = 0.$$

Therefore, $P(\{a\}) = 0$. ■

Problem 86. Calculate the following probabilities:

(a) Consider a probability space where the set of elementary outcomes is the interval $\Omega = (0, 1)$, i.e. a number X (random variable) is chosen at random within that interval. Define a probability measure over that interval as

$$\mathbb{P}(X \in (a, b)) = b - a, \quad \text{for } 0 < a < b < 1.$$

Find an expression for the CDF. Derive a PDF from $\mathbb{P}$ (or the CDF). Compute the probability that X is less than an eighth or greater than seven eights. Would it be possible to use a discrete probability model for this experiment?

(b) In a random experiment we will an unbiased die. The set of outcomes is $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let X be the random variable that indicates the result (top face of die) of the experiment. X can take the value $\in \{1, 2, 3, 4, 5, 6\}$. Its PMF is $p_i = 1/6$ for $i = 1, \ldots, 6$. Plot the CDF of X, i.e. $F(x)$ vs x. Recall that the CDF is defined as $F(x) \equiv \mathbb{P}(X \in (-\infty, x])$. For $x < 1$ define $F(x) = 0$. What is the limiting value of $F(x)$ (i.e. as $x \to \infty$)? Using the CDF, can you obtain the PMF? Explain.

(c) Let X be a random variable representing a real random number chosen between 0 and 1. Obtain the CDF of X. You may assume that $\mathbb{P}(X \in (a, b)) = b - a$ for $0 < a < b < 1$. Sketch the CDF. Obtain the PDF. State the connection between CDF and PDF.

(d) Let X be a random variable with the CDF $F(x)$. Then for $a < b$, prove that the probability $\mathbb{P}(a < X \leq b) = F(b) - F(a)$.

(e) If X is a random variable and $F(x)$ its CDF, then for all a and b, if $a < b$ then $F(a) \leq F(b)$ (F is nondecreasing). Also, it can be shown that

$\lim_{x\to-\infty} F(x) = 0$ (the lower limit of F is 0), $\lim_{x\to\infty} F(x) = 1$ (the upper limit of F is 1), $\lim_{x\downarrow x_0} F(x) = F(x_0)$ (F is right-continuous).

Let X be the half-life of a radioactive isotope. Assume that X has the CDF

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-x} & 0 \le x. \end{cases}$$

Obtain the PDF of X. Show that the derivative of the CDF does not exist at $x = 0$, but that does not affect our ability to compute probabilities. Compute the probability that the half-life is between 1 and 3 years.

(f) The conditions expressed at the beginning of problem (e) show that CDFs are right-continuous and monotone. Such functions can be shown to have at most a countable number of discontinuities. For any random variable, prove that $\mathbb{P}(X = x) = F(x) - F(x-)$, for all $x \in \mathbb{R}$, where $F(x-) = \lim_{z\uparrow x} F(z)$. This results is not a mere curiosity; it allows us to deal with discontinuities in the distribution. Recall that for $\{C_n\}$ a nondecreasing sequence of events,

$$\lim_{n\to\infty} \mathbb{P}(C_n) = \mathbb{P}(\lim_{n\to\infty} C_n) = \mathbb{P}\left(\cup_{n=1}^{\infty} C_n\right)$$

Similarly for a decreasing sequence of events,

$$\lim_{n\to\infty} \mathbb{P}(C_n) = \mathbb{P}(\lim_{n\to\infty} C_n) = \mathbb{P}\left(\cap_{n=1}^{\infty} C_n\right)$$

(g) Let X have a CDF:

$$F(x) = \begin{cases} 0 & x < 0 \\ x/2 & 0 \le x < 1 \\ 1 & 1 \le x. \end{cases}$$

Compute the value $\mathbb{P}(-1 < X \le 1/2)$ and $\mathbb{P}(X = 1)$ (the value is not zero!).

(h) Let X have the PMF

$$p(x) = \begin{cases} cx & x = 1, 2, \ldots, 10 \\ 0 & \text{elsewhere} \end{cases}$$

for an appropriate constant c. Find the value c.

(i) Let X have the PDF

$$f(x) = \begin{cases} cx^3 & 0 < x < 2 \\ 0 & \text{elsewhere} \end{cases}$$

for a constant c. Compute c. Compute the probability $\mathbb{P}(1/4 < X < 1)$.

(j) Let $\Omega = \{x : 1 < x < 2\}$ be the space of X. If $D_1 = \{x : 1 < x \leq 4/3\}$ and $D_2 = \{x : 4/3 < x < 2\}$, find $\mathbb{P}(D_2)$ if $\mathbb{P}(D_1) = 1/3$.

(k) Choose five cards at random and without replacement from a normal deck of playing cards. Find the PMF of X, the number of hearts in the five cards. Determine $\mathbb{P}(X \leq 1)$.

Solution. (a) PDF is obtained by differentiating the CDF. CDF is

$$F(x) \equiv \mathbb{P}(X \in (0, x]) = \int_0^x dx = x.$$

Then PDF is:

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

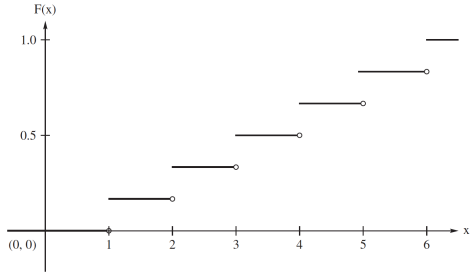
$$\mathbb{P}(\{X < 1/8\} \cup \{X > 7/8\}) = \int_0^{1/8} dx + \int_{7/8}^1 dx = 1/4.$$

Discrete: no, because the probability $\mathbb{P}(\{X = a\}) = 0$ for all $a \in (0, 1)$.

(b) The CDF is defined as the right-continuous function:

$$F(x) = \sum_{\{i : x_i \leq x\}} p_i$$

where the notation $\{i : x_i \leq x\}$ means "sum over all i such that $x_i \leq x$". Plotting this function gives a right-continuous function:



The PMF is given to us: $\{p_i\}$. From the CDF we can obtain the PDF by differentiating:

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x} = \sum_i p_i \delta(x - x_i),$$

where $\delta(x - x_i)$ are Dirac delta functions. What about the PMF? The PMF is defined as $\mathbb{P}(X = x)$, the probability that X takes a specific value x. Obviously this is zero unless $x = x_i$, the points where the CDF "jumps" (discontinuities of F). The size of the discontinuity gives p_i. Formally, the

probability of x_i is obtained by integrating the PDF:

$$\mathbb{P}(X = x_i) = \int_{\{x_i\}} f(x)\mathrm{d}x = \lim_{\epsilon\to 0}\int_{x_i-\epsilon}^{x_i+\epsilon} f(x)\mathrm{d}x = \lim_{\epsilon\to 0}\int_{x_i-\epsilon}^{x_i+\epsilon}\sum_j p_j\delta(x-x_j)\mathrm{d}x = p_i.$$

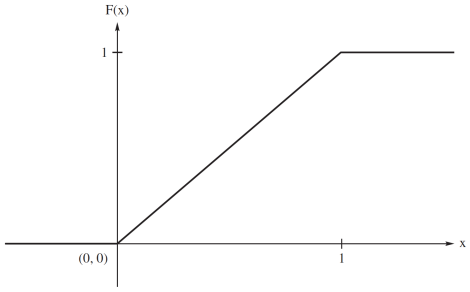
The PMF can also be obtained form the CDF as follows:

$$\mathbb{P}(X = x_i) = \lim_{\epsilon\to 0}\mathbb{P}(x_i-\epsilon < X \le x_i+\epsilon) = \lim_{\epsilon\to 0}\{F(x_i+\epsilon) - F(x_i-\epsilon)\} = p_i.$$

(c) Recall that the CDF is defined as $F(x) = \mathbb{P}(X \in (-\infty, x])$. Since the domain of definition of X is (0,1), we take 0 instead of $-\infty$ as the lower limit and we make sure that x does not exceed 1. We can take $\mathbb{P}(X \in (a, b)) = b - a$, replace a by 0 and b by x:

$$F(x) = \mathbb{P}(X \in (0, x]) = \begin{cases} 0 & x < 0 \\ x & x \in (0, 1) \\ 1 & x > 1 \end{cases}$$

The graph looks like:



The PDF is obtained from the CDF by differentiating:

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x} = 1$$

where $x \in (0, 1)$. This is the uniform distribution on the interval $(0, 1)$.

(d) Note that

$$\{-\infty < X \le b\} = \{-\infty < X \le a\} \cup \{a < X \le b\}.$$

The proof follows immediately because the union on the right side of this equation is a disjoint union.

(e) The PDF is obtained by differentiating

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x} = \begin{cases} e^{-x} & 0 < x < \infty \\ 0 & \text{elsewhere} \end{cases}$$

The derivative of a function $F(x)$ at the point a exists if the limit

$$\lim_{x\to a}\frac{F(x)-F(a)}{(x-a)}$$

exists.

That limit is also the slope of the tangent line to the curve $y = F(x)$ at $x = a$. That limit does not exist when the curve $y = F(x)$ does not have a tangent line at $x = a$ or when the curve does have a tangent line, but the tangent line has infinite slope. In the present case, there is no tangent line at $x = 0$ because this point is a sharp corner (plot the graph of $F(x)$ to see). This is of no consequence when computing probabilities involving X because $\mathbb{P}(X = 0) = 0$ (see problem f below). Therefore, we can assign $f(0) = 0$ without changing the probabilities involving X.

Finally,

$$\mathbb{P}(1 < X \leq 3) = F(3) - F(1) = \int_1^3 e^{-x}\mathrm{d}x.$$

(f) For any $x \in \mathbb{R}$, we have

$$\{x\} = \bigcap_{n=1}^{\infty}\left(x - \frac{1}{n}, x\right]$$

that is, $\{x\}$ is the limit of a decreasing sequence of sets. Hence,

$$\mathbb{P}(X = x) = \mathbb{P}\left(\bigcap_{n=1}^{\infty}\{x - \frac{1}{n} < X \leq x\}\right) = \lim_{n\to\infty}\mathbb{P}(x - 1/n < X \leq x)$$
$$= \lim_{n\to\infty}[F(x) - F(x - 1/n)] = F(x) - F(x-)$$

which is the desired result. The difference, $F(x) - F(x-)$ measures the discontinuity at x.

(g)

$$\mathbb{P}(-1 < X \leq 1/2) = F(1/2) - F(-1) = \frac{1}{4} - 0 = \frac{1}{4}.$$
$$\mathbb{P}(X = 1) = F(1) - F(1-) = 1 - \frac{1}{2} = \frac{1}{2}.$$

(h)

$$1 = \sum_{x=1}^{10} p(x) = \sum_{x=1}^{10} cx = c(1 + 2 + \cdots + 10) = 55c.$$

Hence, $c = 1/55$.

(i)

$$1 = \int_0^2 cx^3 \mathrm{d}x = x\left[\frac{x^4}{4}\right]_0^2 = 4c.$$

Hence, $c = 1/4$. Also,

$$\mathbb{P}(1/4 < X < 1) = \int_{1/4}^1 \frac{x^3}{4}\mathrm{d}x = \frac{255}{4096} = 0.06226.$$

(j) The two sets are disjoint, $D_1 \cap D_2 = \emptyset$. Since, $D_1 \cup D_2 = \Omega$, the collection of sets $\{D_1, D_2\}$ forms a partition of Ω. Then,

$$\mathbb{P}(D_1 \cup D_2) = \mathbb{P}(D_1) + \mathbb{P}(D_2) = \mathbb{P}(\Omega) = 1.$$

Hence,

$$\mathbb{P}(D_2) = 1 - \mathbb{P}(D_1) = 1 - 1/3 = 2/3.$$

(k) Let's assume a standard 52-card deck. There are 4 suits, 13 cards per suit. The number of ways to choose 5 cards without replacement, and without regard to order is:

$$\binom{52}{5} = \frac{52!}{(52-5)!5!} = 2,598,960.$$

First we consider the case $X = 1$. The number of ways to choose a heart is $\binom{13}{1}$. Cards 2-5: number of ways to choose 4 non-hearts is $\binom{51}{4}$. Number of ways to choose 1 heart, 4 non-hearts:

$$\binom{13}{1}\binom{39}{4}.$$

The probability of $X = 1$ is

$$\mathbb{P}(X = 1) = \frac{\binom{13}{1}\binom{39}{4}}{\binom{52}{5}} \approx 0.4114$$

Similarly,

$$\mathbb{P}(X = 2) = \frac{\binom{13}{2}\binom{39}{3}}{\binom{52}{5}} \approx 0.2743$$

$$\mathbb{P}(X = 3) = \frac{\binom{13}{3}\binom{39}{2}}{\binom{52}{5}} \approx 0.0815$$

$$\mathbb{P}(X = 4) = \frac{\binom{13}{4}\binom{39}{1}}{\binom{52}{5}} \approx 0.0107$$

$$\mathbb{P}(X = 5) = \frac{\binom{13}{5}\binom{39}{0}}{\binom{52}{5}} \approx 0.0005$$

For $\mathbb{P}(X \leq 1)$ we need to consider two disjoint events: $\{X = 0\}$ and $\{X = 1\}$. Then,

$$\mathbb{P}(X \leq 1) = \mathbb{P}(\{X = 0\}) \cup \{X = 1\}) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1)$$

■

Problem 87. Let X have the PMF

$$p_X(x) = \begin{cases} \frac{3!}{x!(3-x)!} \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{3-x} & x = 0, 1, 2, 3 \\ 0 & \text{elsewhere} \end{cases}$$

Find the PMF $p_Y(y)$ of the random variable $Y = X^2$.

Solution. The transformation $y = g(x) = x^2$ maps the set $\{x : x = 0, 1, 2, 3\}$ into $\{y : y = 0, 1, 4, 9\}$. In general, $y = x^2$ does not define a one-to-one transformation. Here, however, it does, as there are no negative values of x in the set (for x). That is, we have the single-valued inverse function $x = g^{-1}(y)$ (not $-\sqrt{y}$), and so

$$p_Y(y) = p_X(\sqrt{y}) = \frac{3!}{(\sqrt{y})!(3 - \sqrt{y})!} \left(\frac{2}{3}\right)^{\sqrt{y}} \left(\frac{1}{3}\right)^{3-\sqrt{y}}, \quad y = 0, 1, 4, 9$$

■

Problem 88. Consider a sequence of independent flips of a coin, each resulting in a head (H) or a tail (T). On each flip, we assume that H and T are equally likely. The sample space consists of sequences TTHTHTHTHT.... Let the random variable X equal to the number of flips needed to obtain the first head. For example, $X(\text{TTHTHHT} \cdots) = 3$. The space of X is $\Omega = \{1, 2, 3, 4, \dots\}$. We see what $X = 1$ when the sequence begins with an H and $\mathbb{P}(x = 1) = \frac{1}{2}$. Likewise, $X = 2$ when the sequence begins with TH, which has probability $\mathbb{P}(X = 2) = (\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$ (assuming statistical independence). More generally, if $X = x$, where $x = 1, 2, 3, 4, \dots$, there must be a string of $x - 1$ tails followed by a head. That is, $\text{TT}\cdots\text{TH}$, where there are $x - 1$ tails in $\text{TT}\cdots\text{T}$. Thus, from independence,

$$\mathbb{P}(X = x) = \left(\frac{1}{2}\right)^{x-1} \left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)^x, \quad x = 1, 2, 3, \dots$$

the space of which is countable. Calculate the probability of the event that the first head appears on an odd number of flips, i.e. $X \in \{1, 3, 5, \dots\}$. Let $Z = (X - 2)^2$. Compute the PMF of Z.

Solution. For the first part of the question,

$$\mathbb{P}\left(X \in \{1, 3, 5, \dots\}\right) = \sum_{x=1}^{\infty} \left(\frac{1}{2}\right)^{2x-1} = \frac{1}{2} \sum_{x=1}^{\infty} \left(\frac{1}{4}\right)^{x-1} = \frac{1/2}{1 - (1/4)} = \frac{2}{3}.$$

For $Z = (X-2)^2$, the space of Z is $\{0, 1, 4, 9, 16, \dots\}$. Note that $Z = 0$ if and only if $X = 2$. $Z = 1$ if and only if $X = 1$ or $X = 3$. For the other values of the space there is a one-to-one correspondence given by $x = \sqrt{z}+2$, for $z \in \{4, 9, 16, \dots\}$. Hence, the PMF of Z is

$$p_Z(z) = \begin{cases} p_X(2) = \frac{1}{4} & \text{for } z = 0 \\ p_X(1) + p_X(3) = \frac{5}{8} & \text{for } z = 1 \\ p_X(\sqrt{2}+2) = \frac{1}{4}\left(\frac{1}{2}\right)^{\sqrt{z}} & \text{for } z = 4, 9, 16, \dots \end{cases}$$

You can show that the PMF of Z sums to 1 over its space. ■

Problem 89. Suppose that we have a unit circle and select a point at random within the interior of the circle. Let X be the distance of the point to the origin (Euclidean distance). The sample space for the random point is $\Omega = \{(w, y) : w^2 + y^2 < 1\}$. If the points (chosen at random) have equal probability, write down a formula for the probability of the point landing within an area A contained within the interior of the circle. The event $\{X \leq x\}$ means the point lies in a circle of radius x. Compute the probability $\mathbb{P}(X \leq x)$. Write down the CDF of X. Obtain the PDF of X. Calculate the numerical value of $\mathbb{P}(1/4 < X \leq 1/2)$.

Solution.

$$\mathbb{P}(A) = \frac{\text{area of } A}{\pi}$$

For $0 < x < 1$, the event $\{X \leq x\}$ is equivalent to the point lying in a circle of radius x. By this probability rule, $\mathbb{P}(X \leq x) = \pi x^2/\pi = x^2$. Hence, the CDF of X is

$$F(x) = \begin{cases} 0 & x < 0 \\ x^2 & 0 \leq x < 1 \\ 1 & 1 \leq x \end{cases}$$

■

Problem 90. Suppose that a phone company operates a computerized switchboard designed to route phone calls across the busy telephone network. Let X be the random variable that is the time in seconds between (consecutive) incoming telephone calls. Suppose that the PDF of X is

$$f(x) = \begin{cases} \frac{1}{4}e^{-x/4} & 0 < x < \infty \\ 0 & \text{elsewhere} \end{cases}$$

Show that $f(x)$ is normalized (to 1) and that $f(x) \geq 0$. Calculate the probability that the time between successive phone calls exceeds 4 seconds, i.e. $\mathbb{P}(X > 4)$. Plot this PDF and illustrate the area under the graph that corresponds to this probability. Is this distribution skewed? Compute the skewness of the distribution and explain the value obtained.

Solution.

$$f(x) = \begin{cases} \frac{1}{4}e^{-x/4} & 0 < x < \infty \\ 0 & \text{elsewhere} \end{cases}$$

$$\mathbb{P}(X > 4) = \int_4^\infty \frac{1}{4}e^{-x/4}\mathrm{d}x = e^{-1} = 0.3679.$$

■

Problem 91. Obtain the distribution of $Y = X^2$, where the CDF of X is

$$F_X(x) = \begin{cases} 0 & x < 0 \\ x^2 & 0 \le x < 1 \\ 1 & 1 \le x \end{cases}$$

Both X and Y have the same support, i.e., the interval $(0, 1)$.

Solution. Let y be the support of Y, i.e., $0 < y < 1$. The CDF of Y is

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(X^2 \le y) = \mathbb{P}(X \le \sqrt{y}) = F_X(\sqrt{y}) = \sqrt{y}^2 = y.$$

It follows that the PDF of Y is

$$f_Y(y) = \begin{cases} 1 & 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

■

Problem 92. Let X be a continuous random variable with PDF

$$f(x) = \frac{e^x}{(1 + 5e^x)^{1.2}}, \quad -\infty < x < \infty.$$

Obtain the CDF of X. Plot the PDF. Compute the 3 quantiles: 0.25, 0.50 and 0.75 for X. Indicate on the graph of the PDF the position of the 3 quantiles. Definition (quantile): Let $0 < p < 1$. The quantile of order p of X is a value ξ_p such that $\mathbb{P}(X < \xi_p) \le p$ and $\mathbb{P}(X \le \xi_p) \ge p$. It is known as the $(100p)$th percentile of X.

Solution. The CDF of X is

$$F(x) = 1 + (1 + 5e^x)^{-.2} \qquad -\infty x < \infty,$$

which is confirmed by differentiation, $F'(x) = f(x)$. The quantiles are

$$q_1 = -0.4419242$$

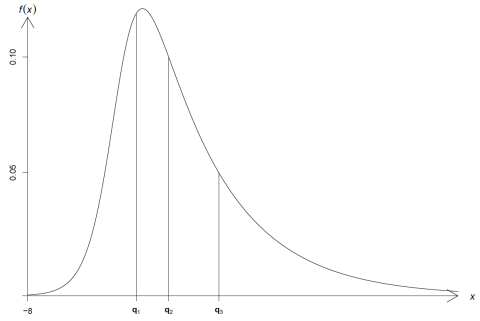
for 25%,

$$q_2 = 1.824549$$

for 50% and

$$q_3 = 5.321057$$

for 75%.



■

Problem 93. Let $f_X(x) = 1/2$, $-1 < x < 1$, zero elsewhere, be the PDF of X. X has uniform distribution within the interval of support $(-1, 1)$. Define $Y = X^2$. Find the PDF and CDF of Y.

Solution. If $y \geq 0$, the probability $\mathbb{P}(Y \leq y)$ is equivalent to

$$\mathbb{P}(X^2 \leq y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}).$$

Accordingly, the CDF of Y is given by

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{2} \mathrm{d}x = \sqrt{y} & 0 \leq y < 1 \\ 1 & 1 \leq y \end{cases}$$

■

Problem 94. Let X have a distribution

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x+1}{2} & 0 \leq x < 1 \\ 1 & 1 \leq x \end{cases}.$$

Calculate the value $\mathbb{P}(-3 < X \leq 1/2)$ and $\mathbb{P}(X = 0)$ (not zero!). Plot the graph of $F(x)$. Comment on any discontinuities and on the discrete (or non-discrete) nature of the distribution.

Solution.

$$\mathbb{P}(-3 < x \leq 1/2) = F(1/2) - F(-3) = \frac{3}{4} - 0 = \frac{3}{4}$$

$$\mathbb{P}(X = 0) = F(0) - F(0-) = \frac{1}{2} - 0 = \frac{1}{2}.$$

■

Problem 95. Compute the following expectation values of X:

(a) Let X have the PDF

$$f(x) = \begin{cases} 4x^3 & 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

(b) For $x = 1, 2, 3, 4$ the corresponding PMF is $p(x)$=4/10, 1/10, 3/10 and 2/10, respectively. Here, $p(x) = 0$ if x is not equal to one of the first four positive integers.

(c) Let X be continuous rv with PDF $f(x) = 2x$, which has support on the interval $(0, 1)$. Suppose $Y = 1/(1 + X)$. Find $\mathbb{E}(X)$ and $\mathbb{E}(Y)$.

(d) Let X have the PDF

$$f(x) = \begin{cases} 2(1 - x) & 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

Calculate $\mathbb{E}(X)$, $\mathbb{E}(X^2)$ and $\mathbb{E}(6X - 3X^2)$.

(e) Let X have the PMF

$$p(x) = \begin{cases} \frac{x}{6} & x = 1, 2, 3 \\ 0 & \text{elsewhere} \end{cases}$$

Compute $\mathbb{E}(6X^3 + X)$.

(f) Divide randomly a line segment of length 5 into two parts. If X is the length of the left-hand part, it is reasonable to assume that X has the PDF

$$f(x) = \begin{cases} \frac{1}{5} & 0 < x < 5 \\ 0 & \text{elsewhere} \end{cases}$$

Compute the expected value of the length $\mathbb{E}(X)$. Calculate $\mathbb{E}(5 - X)$. Calculate also $\mathbb{E}(X(5 - X))$ (expectation value of their product). Explain why $\mathbb{E}(X(5-X)) \neq \mathbb{E}(X)\cdot\mathbb{E}(5-X)$. In the physical sciences, we often encounter situations like this where the product of expectation values is not the same as the expectation value of the product. A famous example is the spatial dependence of the dipole-dipole interaction, which scales as $1/r^3$. In general, $\langle 1/r^3 \rangle \neq \langle 1/r \rangle^3$.

Solution. (a)

$$\mathbb{E}(X) = \int_0^1 x(4x^3)\mathrm{d}x = \int_0^1 4x^4\mathrm{d}x = \left.\frac{4x^5}{5}\right|_0^1 = \frac{4}{5}.$$

(b)

$$\mathbb{E}(X) = (1)\frac{4}{10} + (2)\frac{1}{10} + (3)\frac{3}{10} + (4)\frac{2}{10} = \frac{23}{10} = 2.3$$

(c)

$$\mathbb{E}(Y) = \int_0^1 \frac{2x}{1+x}\mathrm{d}x = \int_1^2 \frac{2u-2}{u}\mathrm{d}u = 2(1-\log 2).$$

(d)

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x = \int_0^1 (x)2(1-x)\mathrm{d}x = \frac{1}{3}$$
$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\mathrm{d}x = \int_0^1 (x^2)2(1-x)\mathrm{d}x = \frac{1}{6}$$
$$\mathbb{E}(6X + 3X^2) = 6(\frac{1}{3}) + 3(\frac{1}{6}) = \frac{5}{2}.$$

(e)

$$\mathbb{E}(6X^3 + X) = 6\mathbb{E}(X^3) + \mathbb{E}(X) = 6\sum_{x=1}^{3} x^3 p(x) + \sum_{x=1}^{3} xp(x) = \frac{301}{3}$$

(f)

$$\mathbb{E}(X(5-x)) = \int_0^5 x(5-x)(\frac{1}{5})\mathrm{d}x = \frac{25}{6} \neq (\frac{5}{2})^2.$$

■

Chapter 3

Propagation of Errors

In many experiments we are required to measure basic physical quantities and derive other quantities from these measurements. The derived quantities are often obtained from mathematical formulas. If we know the error in the basic measurements, what is the error in the derived quantities? This is the topic of error propagation. We shall denote errors in a quantity x as α_x, Δx or δx. The notation δ_x is preferred over Δx when we refer to small (e.g. infinitesimal) quantities. We will avoid the notation σ_x because σ refers to standard deviation. The error α_x could be taken as σ_x but it doesn't have to be; hence, the reason we avoid σ_x here.

Let's start with some simple examples. Suppose we measure acceleration and want to know the error in the force, as propagated through Newton's law, $F = ma$. The average acceleration obtained from repeat measurements is denoted $\overline{a}$ and its error bar is denoted Δa. The maximum value of a is $a_{max} = \overline{a} + \Delta a/2$ and its lowest value is denoted $a_{min} = \overline{a} - \Delta a/2$. The mass is assumed to be positive, $m > 0$ and so is $\Delta a > 0$. The change in F is:

$$\Delta F = F_{max} - F_{min} = m(a_{max} - a_{min}) = m\Delta a = \left(\frac{\partial F}{\partial a}\right)_{\overline{a}} \Delta a.$$

where the subscript $\overline{a}$ on the derivative indicates that the derivative is evaluated at the point $\overline{a}$. In a similar way, suppose that we have Hooke's law $F = kx^2$ and that we measure x. The error in F due to the error in x as propagated through this formula is:

$$\Delta F = F_{max} - F_{min} = k(x_{max}^2 - x_{min}^2)$$

to which we add and subtract $x_{max}x_{min}$:

$$\begin{aligned}\Delta F =& k(x_{max}^2 - x_{min}^2 + x_{max}x_{min} - x_{max}x_{min}) \\ =& k(x_{max} + x_{min})\Delta x = 2k\overline{x}\Delta x = \left(\frac{\partial F}{\partial x}\right)_{\overline{x}} \Delta x\end{aligned}$$

where $\Delta x = x_{max} - x_{min}$ and $\overline{x} = (x_{max} + x_{min})/2$. Similarly, one can show that regardless of the functional dependence on x, we will always have:

$$\Delta F = \left(\frac{\partial F}{\partial x}\right)_{\overline{x}} \Delta x.$$

We have only looked at $F(x)$ with first and second powers of x. If $F(x)$ is a smooth function, it can be expanded in a power series. Let $F(x) = a_n x^n$. Then:

$$\Delta F = F_{max} - F_{min} = a_n(x_{max}^n - x_{min}^n) = a_n[(\overline{x} + \Delta x/2)^n - (\overline{x} - \Delta x/2)^n]$$

Expanding each term using the binomial theorem:

$$(x+h)^n = x^n + nx^{n-1}h + \binom{n}{2}x^{n-2}h^2 + \ldots$$

and keeping only the terms that are first order in Δx:

$$= a_n[2n(\overline{x}/2)^{n-1}\Delta x + O(|\Delta x|^2)] = a_n n(\overline{x})^{n-1}\Delta x = a_n \left(\frac{\partial F}{\partial x}\right)_{\overline{x}} \Delta x.$$

Since this holds for a monomial $a_n x^n$ it holds for any linear combination of monomials (polynomials) and any smooth function F.

Now for the case of 2 variables, we can take $F(m, a) = ma$ and view it as a function of both m and a. The errors in m and a are denoted Δm and Δa, respectively. Then:

$$\Delta F = F_{max} - F_{min} = m_{max}a_{max} - m_{min}a_{min}.$$

Adding and subtracting the quantity $m_{max}a_{min}$, we have:

$$\begin{aligned}\Delta F =& m_{max}a_{max} - m_{min}a_{min} + m_{max}a_{min} - m_{max}a_{min} \\ =& m_{max}\Delta a + a_{min}\Delta m \\ =& (\overline{m} + \Delta m/2)\Delta a + (\overline{a} - \Delta a/2)\Delta m \\ =& \left(\frac{\partial F}{\partial a}\right)_{\overline{a},\overline{m}} \Delta a + \left(\frac{\partial F}{\partial m}\right)_{\overline{a},\overline{m}} \Delta m\end{aligned}$$

where we neglected the second-order small quantity $\Delta m \Delta a$ in the last step. Taking the limit of small Δm and Δa and writing δm and δa for the corresponding infinitesimal quantities, the formula for error propagation has the form: $\delta F(x_1, \ldots, x_n) = \sum_{i=1}^n \left(\frac{\partial F}{\partial x_i}\right)_{\overline{x}} \delta x_i$. Since δF and δx_i are required to

be positive, the partial derivatives must be taken positive:

$$\boxed{\delta F(x_1, \ldots, x_n) = \sum_{i=1}^{n} \left| \frac{\partial F}{\partial x_i} \right|_{\overline{x}} \delta x_i.}$$

This introduction is informal and does not tell the full story. The formula derived, for example, neglects possible correlations between the random variables. We will now derive formulae for error propagation slightly more carefully while explaining the assumptions made along the way. The method of error propagation based on partial derivatives predates the advent of modern computers and is limited to the use of first and second moments of the statistical distributions of the measured quantities. With modern computers we can utilize knowledge of the full distribution functions to propagate errors through formulae using Monte-Carlo methods (see Section 3.6). Monte-Carlo methods paint a more complete picture since all statistical moments are obtained.

3.1. Single Variable Case

It is best to illustrate the methods of error propagation by way of examples.

3.1.1. Entropy of a Gas. In statistical mechanics the entropy of an isolated system with energy U is given by the Boltzmann formula

$$S(U) = k_B \log W(U),$$

where W is the number of microstates whose energy equals the system's energy U. W is also a measure of the volume of phase space. It can be shown that the entropy of a monatomic ideal gas is:

$$S = k_B N \log \left[\frac{V}{N} \left(\frac{4\pi m}{3h^2} \frac{U}{N} \right)^{3/2} \right] + \frac{5}{2} k_B N,$$

where N is the number of particles, V is the gas volume, U is its internal energy and h is Planck's constant. This is called the Sackur-Tetrode equation. Saturn is a planet made of gas. It has the lowest density of gas (0.69 g cm^{-3}) of any gaseous planet (Jupiter, Neptune, Uranus, Neptune). Its equatorial radius is 54,445 $\pm$ 10 km. Since we know its density and volume (hence the value of N), in principle you can use the Sackur-Tetrode equation to obtain a numerical value for the entropy S. What is the uncertainty in its entropy? Taking $V = \frac{4}{3}\pi r^3$, we find that:

$$\delta S = 3 k_B N \log \left| \frac{r + \delta r}{r} \right|.$$

3.1.2. Black Hole Entropy. Black holes have entropy. In thermodynamics, $\frac{\partial S}{\partial E} = \frac{1}{T}$. The energy of the black hole is given by its rest mass $E = Mc^2$.

The entropy is obtained by integrating:

$$S = \int \frac{dE}{T} = \frac{8\pi G}{\hbar c} = \int M dM = \frac{4\pi G}{\hbar c} M^2 = \frac{c^3}{4\hbar G} A,$$

where A is the surface area of the black hole (size of the event horizon),

$$A = 4\pi r_s^2 = 16\pi \frac{M^2 G^2}{c^4},$$

with r_s the Schwarzschild radius ($r_s = 2MG/c^2$) and G is the gravitational constant. We used the Hawking formula for the temperature of a black hole:

$$T = \frac{\hbar c^3}{8\pi G M k_B}.$$

A solar mass black hole has temperature $T = 10^{-8}$ K. A supermassive black hole is about 1 million solar masses and has $T = 10^{-14}$ K. Suppose that we estimate the Schwarzschild radius, r_s, to be r, with an error δr. The corresponding error in S is:

$$S(r + \delta r) - S(r + \delta r) = \frac{c^3}{4\hbar G} 4\pi (r + \delta r)^2 - \frac{c^3}{4\hbar G} 4\pi (r)^2.$$

Expanding the square, $(r + \delta r)^2 = r^2 + 2r\delta r + (\delta r)^2$,

$$\begin{aligned} \delta S = S(r + \delta r) - S(r + \delta r) =& \frac{c^3}{4\hbar G} 4\pi \left(r^2 + 2r\delta r + (\delta r)^2\right) - \frac{c^3}{4\hbar G} 4\pi (r)^2 \\ =& \frac{c^3}{4\hbar G} 4\pi \left(2r\delta r + (\delta r)^2\right). \end{aligned}$$

Given an error δr in the value of r, this formula provides an estimate for the error in S.

3.1.3. Bragg's Law. Bragg's law gives the angles for coherent and incoherent scattering from a crystal lattice. When X-rays are incident on an atom, they induce the radiation of electromagnetic waves at the same frequency, but the angular distribution depends on the lattice parameters. Crystalline solids produce specific patterns of reflected X-rays. At certain specific wavelengths and incident angles, we get intense peaks of reflected radiation (Bragg peaks). This is explained by modeling the crystal as a set of discrete parallel planes separated by a constant parameter d (Fig. 3.1).

The incident X-ray radiation produces a Bragg peak if the reflections off the various planes interfered constructively (Fig. 3.2). The interference is constructive when the phase shift is a multiple of 2π (Fig. 3.2).

This condition can be expressed by Bragg's law,

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

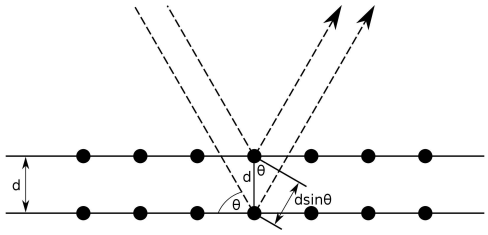
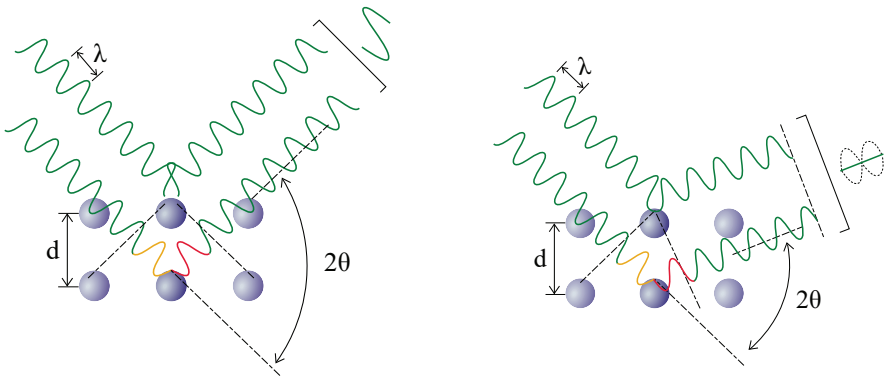


Figure 3.1. Bragg planes in a crystal.



(a) Constructive interference

(b) Destructive interference

Figure 3.2. Bragg's law. Two cases are shown: constructive vs destructive interference.

The case $n = 1$ is the first order peak. For structure determination, the value of λ (wavelength of incident radiation) would be known, and we would measure θ and solve for d the lattice constant.

Another possible experiment we could do is determine λ from a known value of d and measured θ. Suppose we measure θ and compute λ from the Bragg formula. This is shown graphically in Fig. 3.3, where the error in θ is denoted α_θ.

3.1.4. Linear Approximation Method. From this, we see that the error in λ is:

$$\alpha_\lambda = |f(\overline{\theta} + \alpha_\theta) - f(\overline{\theta})| \approx \alpha_\theta \left| \frac{\mathrm{d}f}{\mathrm{d}\theta}_{\theta=\overline{\theta}} \right|,$$

where in the last equality this error was obtained from calculus, by approximating the function f by its first derivative at the point $\overline{\theta}$.

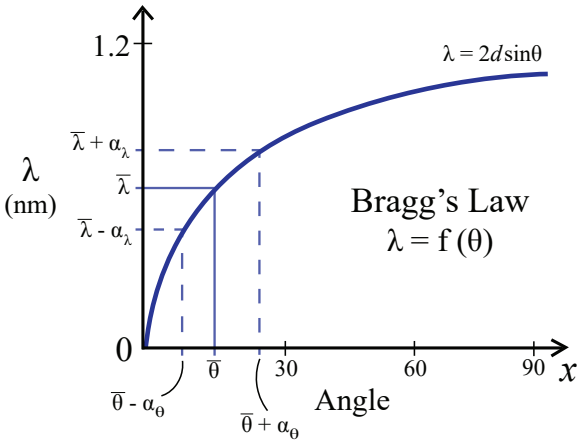


Figure 3.3. Error propagation for Bragg's law.

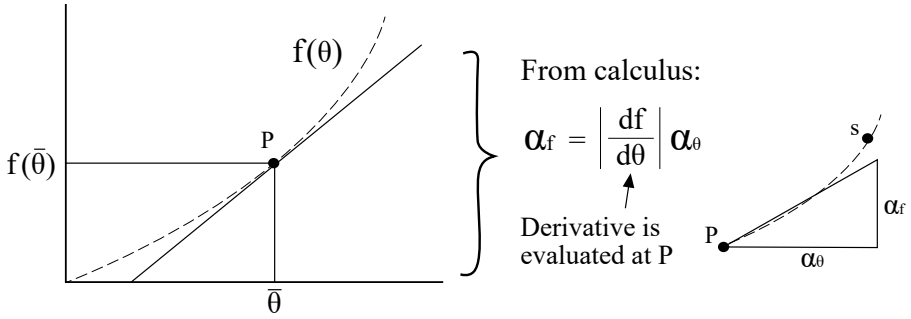


Figure 3.4. Method for error propagation in a single variable based on linear approximation.

This linear approximation is illustrated graphically in Fig. 3.4. This is the most common method that people use for propagating errors. Another way to think about it is to view θ as a random variable that is the sum of a mean value $\overline{\theta}$ and a deviation $\delta\theta$:

$$\theta = \overline{\theta} + \delta\theta.$$

With this decomposition, $\overline{\theta}$ is deterministic and $\delta\theta$ is random. Taylor expansion of f to first order gives:

$$\delta f = f(\overline{\theta} + \delta\theta) - f(\overline{\theta}) \approx \delta\theta \cdot \left(\frac{\mathrm{d}f}{\mathrm{d}\theta}\bigg|_{\theta=\overline{\theta}} \right).$$

Here δf is a random variable because it is equal to $f(\overline{\theta} + \delta\theta) - f(\overline{\theta})$, a function of a random variable $\delta\theta$. Because δf is random, and its value changes

$f(\theta)$	$\mathrm{d}f/\mathrm{d}\theta$	error (α_f)
$1/\theta$	$-1/\theta^2$	$\alpha_\theta/\theta^2 = f^2\alpha_\theta$ or $\lvert\alpha_f/f\rvert = \lvert\alpha_\theta/\theta\rvert$
$\exp(\theta)$	$\exp(\theta)$	$\exp(\theta)\alpha_\theta = f\alpha_\theta$
$\log\theta$	$1/\theta$	α_θ/θ
$\log_{10}\theta$	$\frac{1}{\log 10\cdot\theta}$	$\alpha_\theta/\log 10\cdot\theta$
θ^n	$n\theta^{n-1}$	$\lvert n\theta^{n-1}\rvert\alpha_\theta$ or $\lvert\alpha_f/f\rvert = \lvert n\alpha_\theta/\theta\rvert$
$\sin\theta$	$\cos\theta$	$\lvert\cos\theta\rvert\alpha_\theta$
$\cos\theta$	$-\sin\theta$	$\lvert\sin\theta\rvert\alpha_\theta$

Table 3.1. Formulae for propagation of errors in a single variable. This table can be found in Hughes & Hase's book.

every time an experiment is run, we can't immediately use this expression and call δf the "propagated error". We must take the additional step of computing its variance $var(\delta f)$, which can then be used to obtain the error (as say, the square root of the variance).

The Taylor approximation term, $\delta\theta\cdot\left.\frac{\mathrm{d}f}{\mathrm{d}\theta}\right|_{\theta=\overline{\theta}}$, is a random variable because $\delta\theta$ is a random variable. $\left.\frac{\mathrm{d}f}{\mathrm{d}\theta}\right|_{\theta=\overline{\theta}}$, on the other hand, is a deterministic quantity because it is the derivative of a deterministic function (f) evaluated at a deterministic argument ($\overline{\theta}$ is deterministic by definition, because the "mean" is just a number, hence deterministic).

Take the variance of δf and apply the property $var(aX) = a^2 var(X)$:

$$var(\delta f) = \left(\frac{\mathrm{d}f}{\mathrm{d}\theta}\right)^2 \cdot var(\delta\theta).$$

Because the square root of the variance is the standard deviation, let us take the error bar α_f to be the standard deviation $\sqrt{var(\delta f)}$ and similarly for $\alpha_\theta = \sqrt{var(\delta\theta)}$, writing:

$$\boxed{\alpha_f = \left|\frac{\mathrm{d}f}{\mathrm{d}\theta}\right|\alpha_\theta.}$$

This formula is identical to the one derived in the previous section. As an exercise, the reader should derive all the formulas in Table 3.1.

3.1.5. EXAMPLE: Error in Cosine (Single Variable). This example can be found in Taylor's book. Suppose we want to know the uncertainty in a cosine, i.e. $f(\theta) = \cos\theta$, where θ is a measured quantity. The measured angle is:

$$\theta = 20 \pm 3^\circ$$

Then,

$$(\cos\theta)_{best} = \cos 20^\circ = 0.94.$$

and

$$\alpha_{\cos\theta} = \left|\frac{\mathrm{d}\cos\theta}{\mathrm{d}\theta}\right| \alpha_\theta = |\sin\theta|\alpha_\theta \qquad \text{(angles in radians)}$$

with $\alpha_\theta = 3^\circ = 0.05$ rad. Note: the formula $\mathrm{d}\cos\theta/\mathrm{d}\theta = -\sin\theta$ only holds if θ is in radians. Then,

$$\alpha_{\cos\theta} = (\sin 20^\circ) \times 0.05 = 0.34 \times 0.05 = 0.02$$

so we report:

$$\cos\theta = 0.94 \pm 0.02$$

3.2. Multi-Variable Case

To avoid references to angles, let us switch notation from θ to x. Consider a function $f(\mathbf{x})$ which depends on several variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. The variables $\mathbf{x}$ refer to different physical quantities measured in the laboratory. For example, suppose we measure the length (L), width (W) and depth (D) of a rectangular box, each with their own error bars, and we want to know the volume of the box and its uncertainty. Since $V = L \times W \times D$, we will learn to derive the following result:

$$\left(\frac{\alpha_V}{V}\right) = \sqrt{\left(\frac{\alpha_L}{L}\right)^2 + \left(\frac{\alpha_W}{W}\right)^2 + \left(\frac{\alpha_D}{D}\right)^2},$$

where α_V is the error in V, and similarly for α_L, α_W and α_D. Another experiment could involve Ohm's law ($V = IR$; R, resistance; I, current; V, voltage). Suppose we measure voltage and current across a device and use this information to obtain the impedance (R) of the device. The error in R, α_R is related to the errors in I and V, α_V and α_I, as follows:

$$\left(\frac{\alpha_R}{R}\right) = \sqrt{\left(\frac{\alpha_I}{I}\right)^2 + \left(\frac{\alpha_V}{V}\right)^2}.$$

If the variables are statistically independent, the formula for error propagation takes a particularly simple form. The graphical method is easily generalized as follows:

$$|\delta f| = |f(\overline{\mathbf{x}} + \delta\mathbf{x}) - f(\overline{\mathbf{x}})|$$

and taking δ_{x_i} to be the error bars α_{x_i} in x_i (where $\alpha_{x_i} > 0$), α_f to be the error bar in f:

$$|\alpha_f| = |f(\overline{\mathbf{x}} + \vec{\alpha}_{\mathbf{x}}) - f(\overline{\mathbf{x}})|.$$

Example in 2D:

$$|\alpha_f| = |f(A + \alpha_A, B + \alpha_B) - f(A, B)|.$$

Thus, if you know f, A, B, α_A, α_B you can calculate the error bar in f. In general, this method should be sufficient.

In high dimensional spaces, it is impossible to visualize the function to be approximated. We can instead use derivatives:

$$|\delta f| = |f(\overline{\mathbf{x}} + \delta\mathbf{x}) - f(\overline{\mathbf{x}})| \approx \left| \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \delta x_i \right|.$$

If you know the derivatives of f, you can use this formula. Consider a function that adds several independent random variables:

$$f(x_1, x_2, \ldots, x_n) = x_1 + x_2 + \cdots + x_n.$$

In that case, all derivatives are equal to 1: $\partial_{x_i} f = 1$ for all $i = 1, \ldots, n$. Denoting the errors in the x_i as α_i and the error in δf as α_f, we find (since $\alpha_i > 0$):

$$|\alpha_f| = |\alpha_1 + \alpha_2 + \cdots + \alpha_n| = |\alpha_1| + |\alpha_2| + \cdots + |\alpha_n|.$$

We can also use the upper bound:

$$\left| \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \delta x_i \right| \leq \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| |\delta x_i| \,,$$

as our error bar in f. Then, denoting the error bars in x_i as α_{x_i} (instead of δx_i):

$$\boxed{|\alpha_f| = \sum_{i=1}^{n} |\partial_i f| \cdot |\alpha_{x_i}|.}$$

3.2.1. Remark: Adding Two Quantities. Consider the case $f(x_1, x_2) = x_1 + x_2$. One way to get the error in f in terms of the errors in x_1 and x_2 is to add the two error bars:

$$|\alpha_f| = |\alpha_{x_1}| + |\alpha_{x_2}|.$$

This formula is easily derived using the graphical method:

$$|\alpha_f| = |f(x_1+\alpha_{x_1}, x_2+\alpha_{x_2}) - f(x_1, x_2)| \approx \left| \frac{\partial f}{\partial x_1} \alpha_{x_1} + \frac{\partial f}{\partial x_2} \alpha_{x_2} \right| = |\alpha_{x_1}| + |\alpha_{x_2}|,$$

where $f = x_1 + x_2$, $\alpha_{x_1} > 0$ and $\alpha_{x_2} > 0$. (Note: we are treating all variables here as deterministic.)

Another method for obtaining the error bars in f when $f(x_1, x_2) = x_1 + x_2$ is to add error bars quadratically: $\alpha_f = \sqrt{(\alpha_{x_1})^2 + (\alpha_{x_2})^2}$. Both methods are valid; they simply report different information. The method of adding errors in quadrature yields a tighter error bound than adding the errors linearly, i.e.

$$|\alpha_{x_1}|^2 + |\alpha_{x_2}|^2 \leq |\alpha_{x_1}|^2 + |\alpha_{x_2}|^2 + 2|\alpha_{x_1}||\alpha_{x_2}| = (|\alpha_{x_1}| + |\alpha_{x_2}|)^2.$$

You can also see this graphically from the Pythagoras theorem: the hypothenuse $\sqrt{(\alpha_{x_1})^2 + (\alpha_{x_2})^2}$ is always shorter than (or equal to) the sum of the two remaining sides $|\alpha_{x_1}| + |\alpha_{x_2}|$.

3.2.2. Derivative Method: Case of Statistically Independent Variables. As in the 1D case, we view the experimental measurement as a random variable $\mathbf{x}$ which is the sum of a mean value $\overline{\mathbf{x}}$ (deterministic quantity) and a deviation $\delta\mathbf{x}$ (a random quantity):

$$\mathbf{x} = \overline{\mathbf{x}} + \delta\mathbf{x}.$$

The Taylor expansion of $f(\mathbf{x})$ about the point $\overline{\mathbf{x}}$ to first order is:

$$\delta f = f(\overline{\mathbf{x}} + \delta\mathbf{x}) - f(\overline{\mathbf{x}}) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\delta x_i + O(|\delta\mathbf{x}|^2), \tag{3.1}$$

where the derivatives are evaluated at the point $\overline{\mathbf{x}}$. Since $\delta\mathbf{x}$ are random variables, δf is also a random variable which depends on random variables $\delta\mathbf{x} = (\delta x_1, \ldots, \delta x_n)$. The δx_i are deviations from the point of expansion $\overline{\mathbf{x}}$. Again, we view the $\partial f/\partial x_i$ as deterministic quantities whereas the δx_i are random. Note: δf itself is not the error bar. It's a random variable. Its value changes every time a new experiment is done. However, an error can be obtained from δf by taking the square root of its variance.

If the random variables $\{\delta x_i\}$ are mutually independent, taking the variance of Eq. (3.1) gives:

$$var(\delta f) = \sum_{i=1}^{n} var(\partial_i f \cdot \delta x_i) = \sum_{i=1}^{n} (\partial_i f)^2 var(\delta x_i).$$

Noting that $\delta\mathbf{x} = \mathbf{x} - \bar{\mathbf{x}}$, and that $var(\delta x_i) = var(x_i)$ since constants such as $\bar{\mathbf{x}}$ don't affect the variance, we have:

$$\underbrace{var(\delta f)}_{(\alpha_f)^2} = \sum_{i=1}^{n} (\partial_i f)^2 \underbrace{var(x_i)}_{(\alpha_{x_i})^2},$$

or

$$\boxed{(\alpha_f)^2 = \sum_{i=1}^{n} (\partial_i f)^2 (\alpha_{x_i})^2.} \tag{3.2}$$

How is this related to the result of the previous section?

$$\underbrace{\sum_{i=1}^{n} |\partial_i f|^2 \cdot |\alpha_{x_i}|^2}_{l_2 \text{ norm squared}} \leq \underbrace{\left(\sum_{i=1}^{n} |\partial_i f| \cdot |\alpha_{x_i}| \right)^2}_{l_1 \text{ norm squared}}.$$

$f(x_1, x_2, \dots)$	error propagation
$x_1 \pm x_2$	$\alpha_f = \sqrt{(\alpha_{x_1})^2 + (\alpha_{x_2})^2}$
$x_1 \cdot x_2$ or x_1/x_2	$\frac{\alpha_f}{f} = \sqrt{\left(\frac{\alpha_{x_1}}{x_1}\right)^2 + \left(\frac{\alpha_{x_2}}{x_2}\right)^2}$
x_1^n	$\left\lvert\frac{\alpha_f}{f}\right\rvert = \left\lvert n\frac{\alpha_{x_1}}{x_1}\right\rvert$
$k\frac{x_1}{x_2}$	$\frac{\alpha_f}{f} = \sqrt{\left(\frac{\alpha_{x_1}}{x_1}\right)^2 + \left(\frac{\alpha_{x_2}}{x_2}\right)^2}$
$k\frac{x_1^n}{x_2^m}$	$\frac{\alpha_f}{f} = \sqrt{\left(n\frac{\alpha_{x_1}}{x_1}\right)^2 + \left(m\frac{\alpha_{x_2}}{x_2}\right)^2}$
$x_1 + x_2 - x_3 + x_4$	$\alpha_f = \sqrt{(\alpha_{x_1})^2 + (\alpha_{x_2})^2 + (\alpha_{x_3})^2 + (\alpha_{x_4})^2}$
$\frac{x_1 \cdot x_2}{x_3 \cdot x_4}$	$\frac{\alpha_f}{f} = \sqrt{\left(\frac{\alpha_{x_1}}{x_1}\right)^2 + \left(\frac{\alpha_{x_2}}{x_2}\right)^2 + \left(\frac{\alpha_{x_3}}{x_3}\right)^2 + \left(\frac{\alpha_{x_4}}{x_4}\right)^2}$

Table 3.2. Formulae for error propagation in several variables. This table can be found in Hughes & Hase's book.

The l_2-norm gives tighter bounds than the l_1 norm. Recall: the l_p-norm of a vector $\vec{x}$ is:

$$\|\vec{x}\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}\,. \tag{3.3}$$

The reader should derive all the formulas in Table 3.2.

Let us work out the first two lines of the table explicitly.

When $f(x_1, x_2) = x_1 \pm x_2$, the partials $\partial_{x_1} f = 1$ and $\partial_{x_2} f = \pm 1$. Thus,

$$\begin{aligned}(\alpha_f)^2 = &\left(\frac{\partial f}{\partial x_1}\right)^2 (\alpha_{x_1})^2 + \left(\frac{\partial f}{\partial x_2}\right)^2 (\alpha_{x_2})^2 \\ =&1 \cdot (\alpha_{x_1})^2 + (\pm 1)^2 \cdot (\alpha_{x_2})^2 = (\alpha_{x_1})^2 + (\alpha_{x_2})^2.\end{aligned}$$

Hence,

$$\alpha_f = \sqrt{(\alpha_{x_1})^2 + (\alpha_{x_2})^2}.$$

When $f = x_1 \cdot x_2$, we have $\partial_{x_1} f = x_2$ and $\partial_{x_2} f = x_1$. Then,

$$\begin{aligned}(\alpha_f)^2 = &\left(\frac{\partial f}{\partial x_1}\right)^2 (\alpha_{x_1})^2 + \left(\frac{\partial f}{\partial x_2}\right)^2 (\alpha_{x_2})^2 \\ =&x_2^2 \cdot (\alpha_{x_1})^2 + x_1^2 \cdot (\alpha_{x_2})^2\end{aligned}$$

Dividing both sides by $f^2 = x_1^2 x_2^2$ gives the result that fractional errors add up in quadrature:

$$\left(\frac{\alpha_f}{f}\right)^2 = \left(\frac{\alpha_{x_1}}{x_1}\right)^2 + \left(\frac{\alpha_{x_2}}{x_2}\right)^2.$$

We also get the same result for $f = x_1/x_2$.

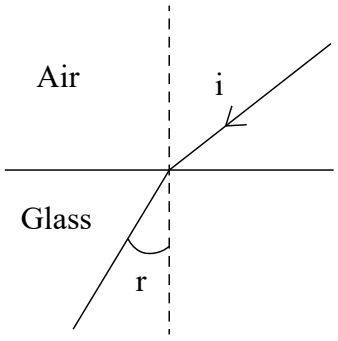


Figure 3.5. Snell's law. The angle of incidence (in radians) is denoted i; the angle of refraction is denoted r.

3.2.3. EXAMPLE: Snell's Law (Two Variables). This example can be found in Taylor's book. Consider light that enters a medium. Let (i) denote the angle of incidence and (r) the angle of refraction (Fig. 3.5). These two angles are related by Snell's law, through the index of refraction of the medium n (in this case, glass),

$$\sin i = n \sin r.$$

We ask what is the error in n given experimentally measured errors in i and r.

3.2.3.1. *Method 1, Using Propagation of Errors Formula.* From the previous lecture, $(\alpha_f)^2 = \sum_{i=1}^{n} (\partial_i f)^2 (\alpha_{x_i})^2$, and $n = \frac{\sin i}{\sin r}$, we have

$$(\alpha_n)^2 = \left(\frac{\cos i}{\sin r}\right)^2 (\alpha_i)^2 + \left(\frac{\sin i \cdot \cos r}{(\sin r)^2}\right)^2 (\alpha_r)^2.$$

Dividing throughout by n^2, or equivalently, multiplying by $1/n^2 = (\sin r / \sin i)^2$:

$$\left(\frac{\alpha_n}{n}\right)^2 = \left(\frac{\cos i}{\sin i}\right)^2 (\alpha_i)^2 + \left(\frac{\cos r}{\sin r}\right)^2 (\alpha_r)^2.$$

3.2.3.2. *Method 2, Using Table of Formulas.* Using the result for A/B in the table from the last page of the previous lecture, $\alpha_z/z = \sqrt{(\alpha_A/A)^2 + (\alpha_B/B)^2}$, and $n = \sin i / \sin r$ with $A = \sin i$ and $B = \sin r$, we have:

$$\frac{\alpha_n}{n} = \sqrt{\left(\frac{\alpha_{\sin i}}{\sin i}\right)^2 + \left(\frac{\alpha_{\sin r}}{\sin r}\right)^2},$$

where $\alpha_{\sin i} = |\cos i| \alpha_i$ and $\alpha_{\sin r} = |\cos r| \alpha_r$. Some example numbers are provided in Table 3.3, for the case where errors in angles are $\pm 1^\circ$, or 0.02 rad.

i (deg)	r (deg)	$\sin i$	$\sin r$	n	$\frac{\alpha_{\sin i}}{\lvert\sin i\rvert}$	$\frac{\alpha_{\sin r}}{\lvert\sin r\rvert}$	$\frac{\alpha_n}{n}$
20	13	0.342	0.225	1.52	5%	8%	9%
40	23.5	0.643	0.399	1.61	2%	4%	5%

Table 3.3. Error propagation through Snell's law.

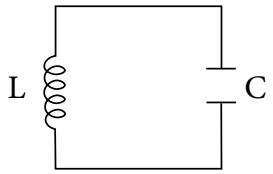


Figure 3.6. Resonant circuit.

3.2.4. EXAMPLE: LC Resonant Circuit (Two Variables). Suppose that we have an LC circuit (Fig. 3.6). Its resonance frequency can be shown to be[1]

$$f_0 = \frac{1}{2\pi}\frac{1}{\sqrt{LC}}.$$

It is generally difficult to measure L with good accuracy. A better method is to measure the resonant frequency f_0 (which can be done using a network analyzer) and the capacitance C. L is then given by

$$L = \frac{1}{(2\pi f_0)^2 C}.$$

What about the error in L?

$$(\alpha_L)^2 = \left|\frac{\partial L}{\partial f_0}\right|^2 (\alpha_{f_0})^2 + \left|\frac{\partial L}{\partial C}\right|^2 (\alpha_C)^2$$

where

$$\left|\frac{\partial L}{\partial f_0}\right| = \frac{1}{2\pi^2 f_0^3 C} = \frac{2L}{f_0}$$

and

$$\left|\frac{\partial L}{\partial C}\right| = \frac{1}{(2\pi f_0)^2 C^2} = \frac{L}{C}.$$

The result is:

$$\left(\frac{\alpha_L}{L}\right)^2 = 4\left(\frac{\alpha_{f_0}}{f_0}\right)^2 + \left(\frac{\alpha_C}{C}\right)^2,$$

[1]This is seen by equating the voltages across each lumped element, i.e. $j\omega L = \frac{1}{j\omega C}$, where $j = \sqrt{-1}$ denotes an imaginary number and $\omega = 2\pi f_0$.

which could also have been derived from the above look-up table formula with $z = kA^n/B^m$ for which

$$\frac{\alpha_z}{z} = \sqrt{\left(\frac{n\alpha_A}{A}\right)^2 + \left(\frac{m\alpha_B}{B}\right)^2}.$$

3.2.5. EXAMPLE: Generic (Two Variables). This example can be found in Taylor's book. Let's determine the error in $q = x^2y - xy^2$, where we have measured experimentally the quantities:

$$x = 3.0 \pm 0.1 \qquad \text{and} \qquad y = 2.0 \pm 0.1$$

The "x_{best}" value, as usual, is obtained from the reported values:

$$q_{best} = 3^2 \cdot 2 - 3 \cdot 2^2 = 6.0.$$

For the error, we need the following two terms:

$$\left|\frac{\partial q}{\partial x}\right| \alpha_x = |2xy - y^2|\alpha_x = |12 - 4| \cdot 0.1 = 0.8,$$

$$\left|\frac{\partial q}{\partial y}\right| \alpha_y = |x^2 - 2xy|\alpha_y = |9 - 12| \cdot 0.1 = 0.3.$$

Adding the errors in quadrature gives:

$$\alpha_q = \sqrt{(0.8)^2 + (0.3)^2} = 0.9.$$

Therefore, we report:

$$q = 6.0 \pm 0.9.$$

3.3. When Variables are Correlated

Previously, we found the following formula for propagation of errors

$$(\alpha_f)^2 = \sum_{i=1}^{n} (\partial_i f)^2 (\alpha_{x_i})^2, \tag{3.4}$$

where f is a function of n variables, i.e. $f = f(x_1, x_2, \ldots, x_n)$. $\partial_i f$ denotes the partial derivative of f with respect to x_i. This result relies on the assumption that the different random variables are statistically independent. However, if the random variables are not statistically independent, covariance

$$cov(X, Y) = \overline{XY} - \overline{X} \cdot \overline{Y}$$

enters the picture and an additional term is required to describe the interdependence between the variables. Before we begin, let us remark the following properties of the covariance, which easily follow from the definition of covariance:

$$cov(X, X) = var(X)$$

$$cov(X, Y) = cov(Y, X) \qquad \text{(symmetry)}$$

$cov(X+Y,Z) = cov(X,Z) + cov(Y,Z)$ (linearity in the first argument)

$cov(aX,Y) = a \cdot cov(X,Y)$ (linearity in the first argument)

from which you can derive

$cov(aX+bY,Z) = a{\cdot}cov(X,Z)+b{\cdot}cov(Y,Z)$ (linearity in the first argument)

$cov(Z,aX+bY) = a{\cdot}cov(Z,X)+b{\cdot}cov(Z,Y)$ (linearity in the second argument).

Also useful is the relationship:

$$cov(a+X,Y) = \overline{(a+X)Y} - \overline{(a+X)}\cdot\overline{Y} = \cancel{a\overline{Y}} + \overline{XY} - \cancel{a\overline{Y}} - \overline{X}\cdot\overline{Y} = cov(X,Y),$$

since additive constants don't change anything. Slightly more obvious is:

$$cov(a,Y) = \overline{aY} - \overline{a}\cdot\overline{Y} = \cancel{a\overline{Y}} - \cancel{a\overline{Y}} = 0.$$

Since *cov* is linear in both arguments, we say that it is *bilinear*. Let $\{X_1,\ldots,X_n\}$ be n random variables. Bilinearity gives:

$$var(\sum_{i=1}^{n} a_i X_i) = \sum_{i,j=1}^{n} a_i a_j\, cov(X_i,X_j) = \sum_{i=1}^{n} a_i^2 var(X_i) + 2\sum_{i<j} a_i a_j\, cov(X_i,X_j).$$

So taking the covariance of

$$\delta f = f(\overline{\mathbf{x}}+\delta\mathbf{x}) - f(\overline{\mathbf{x}}) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\delta x_i + O(|\delta\mathbf{x}|^2) \tag{3.5}$$

with itself to get the variance (the square root of which can be taken as the error bar):

$$\begin{aligned}\underbrace{var(\delta f)}_{(\alpha_f)^2} = cov(\delta f,\delta f) &= cov\left(\sum_{i=1}^{n}\frac{\partial f}{\partial x_i}\delta x_i, \sum_{j=1}^{n}\frac{\partial f}{\partial x_j}\delta x_j\right)\\ &= \sum_{i=1}^{n}(\partial_i f)^2 \underbrace{var(\delta x_i)}_{(\alpha_{x_i})^2} + 2\sum_{i<j}(\partial_i f)(\partial_j f) cov(\delta x_i,\delta x_j).\end{aligned}$$

Note: we neglected terms of order $O(|\delta\mathbf{x}|^2)$ and higher.

Recall that $\delta\mathbf{x} = \mathbf{x} - \bar{\mathbf{x}}$ and because $\bar{\mathbf{x}}$ is simply an additive constant, we have that:

$$\begin{aligned}cov(\delta x_i,\delta x_j) &= cov(x_i - \overline{x_i}, x_j - \overline{x_j})\\ &= cov(x_i,x_j) - cov(x_i,\overline{x_j}) - cov(\overline{x_i},x_j) + cov(\overline{x_i},\overline{x_j}).\end{aligned}$$

Observe that $cov(\overline{x_i},\overline{x_j}) = 0$ since both arguments are constant. Also, both $cov(x_i,\overline{x_j})$ and $cov(\overline{x_i},x_j)$ vanish since one of the arguments is a constant. This leaves:

$$cov(\delta x_i,\delta x_j) = cov(x_i,x_j).$$

Thus,

$$var(f) = \sum_{i=1}^{n} (\partial_i f)^2 var(x_i) + 2 \sum_{i<j} (\partial_i f)(\partial_j f) cov(x_i, x_j).$$

or,

$$\boxed{(\alpha_f)^2 = \underbrace{\sum_{i=1}^{n} (\partial_i f)^2 (\alpha_{x_i})^2}_{\text{diagonal term}} + \underbrace{2 \sum_{i<j} (\partial_i f)(\partial_j f) cov(x_i, x_j)..}_{\text{off-diagonal term}}} \tag{3.6}$$

Formula (3.6) differs from (3.4) by the emergence of a cross-correlation term,

$$2 \sum_{i<j} (\partial_i f)(\partial_j f) cov(x_i, x_j).$$

We note that if the random variables $\{x_i\}$ are statistically independent, $cov(x_i, x_j) = 0$, the second term vanishes and the formula reduces to (3.4).

3.3.1. EXAMPLE: case of two variables. Suppose that we measure mass and acceleration and compute the force according to $F = ma$. The above formula for the error in F gives:

$$var(F) = (\partial_m F)^2 var(m) + (\partial_a F)^2 var(a) + 2(\partial_m F)(\partial_a F) cov(a, m)$$

The partial derivatives of F are easily computed: $\partial_m(ma) = a$, $\partial_a(ma) = m$. Then,

$$var(F) = a^2 var(m) + m^2 var(a) + 2(ma) cov(a, m)$$

This can easily be computed from experimental data. Suppose we have random samples $\{(a_i, m_i)\}_{i=1}^n$. The sample means are:

$$\hat{\mu}_a = \frac{1}{n} \sum_{i=1}^{n} a_i, \qquad \hat{\mu}_m = \frac{1}{n} \sum_{i=1}^{n} m_i.$$

The sample variances are:

$$\hat{\sigma}^2_{n-1}(a) = \frac{1}{n-1} \sum_{i=1}^{n} (a_i - \hat{\mu}_a)^2, \qquad \hat{\sigma}^2_{n-1}(m) = \frac{1}{n-1} \sum_{i=1}^{n} (m_i - \hat{\mu}_m)^2.$$

The sample covariance is:

$$cov_{n-1}(a, m) = \frac{1}{n-1} \sum_{i=1}^{n} (a_i - \hat{\mu}_a)(m_i - \hat{\mu}_m).$$

We conclude that the formula for propagation of error can readily be used with experimental data. In the example $F = ma$ there is no *a priori* reason to assume a correlation between a and m. With other models, however, variables could be correlated.

3.4. Several Functions of Several Variables

Suppose that f is a vector-valued function, i.e.

$$\vec{f} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{pmatrix},$$

then, one such relationship holds for all m components f_i:

$$(\alpha_{f_i})^2 = \sum_{j=1}^{n} (\partial_j f_i)^2 (\alpha_{x_j})^2 + 2 \sum_{h<j} (\partial_h f_i)(\partial_j f_i) cov(x_h, x_j),$$

or

$$\boxed{var(f_i) = \sum_{j=1}^{n} (\partial_j f_i)^2 var(x_j) + 2 \sum_{h<j} (\partial_h f_i)(\partial_j f_i) cov(x_h, x_j).} \tag{3.7}$$

Of course, when $m = 1$ this expression reduces to Eq. (3.6).

Now suppose that we have m functions $f_1, f_2, \ldots, f_m$, each a function of n different random variables $x_1, x_2, \ldots, x_n$. In the general case, the different f_k will be correlated with one another, even if the $x_1, x_2, \ldots, x_n$ are uncorrelated. The variances of the f_k are given by Eq. (3.7), whereas the covariances are:

$$cov(f_k, f_l) = \overline{(f_k - \overline{f_k})(f_l - \overline{f_l})} = \overline{\delta f_k \delta f_l} = \overline{\left(\sum_{i=1}^{n} \frac{\partial f_k}{\partial x_i} \delta x_i \right) \left(\sum_{i=1}^{n} \frac{\partial f_l}{\partial x_i} \delta x_i \right)}$$
$$= \sum_{i,j=1}^{n} \frac{\partial f_k}{\partial x_i} \frac{\partial f_l}{\partial x_j} \overline{\delta x_i \delta x_j},$$

which leads to:

$$\boxed{cov(f_k, f_l) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{\partial f_k}{\partial x_i} \right) \left(\frac{\partial f_l}{\partial x_j} \right) cov(x_i, x_j).} \tag{3.8}$$

We made use of

$$\delta f_k = f_k(\overline{\mathbf{x}} + \delta \mathbf{x}) - f_k(\overline{\mathbf{x}}) = \sum_{i=1}^{n} \frac{\partial f_k}{\partial x_i} \delta x_i + O(|\delta \mathbf{x}|^2),$$

and neglected terms of order $O(|\delta \mathbf{x}|^2)$ and higher. If we denote $g_{kl} = \partial_l f_k$ the elements of a matrix G and F and X the covariance matrices of $\vec{f}$ and $\mathbf{x}$, respectively, then:

$$F = GXG^T,$$

which can also be written as:

$$\boxed{cov(\vec{f},\vec{f}) = G \cdot cov(\mathbf{x},\mathbf{x}) \cdot G^T,} \tag{3.9}$$

where $cov(\vec{f},\vec{f}) = F$ and $cov(\mathbf{x},\mathbf{x}) = X$. Note: the first derivatives in the G matrix are evaluated at the point $\overline{\mathbf{x}}$. This very simple formula, which encodes all there is to know about error propagation, should be the one taught rather than the one for uncorrelated variables, Eq. (3.4) (Section 3.2.2). Equation (3.4) is obtained from (3.9) by dropping the off-diagonal terms (covariances). The use of the covariance matrix should not scare students because the parameters themselves result from a least-squares analysis, and in general their covariances, which may be non-negligible, may be obtained as part of the analysis. We will see in Chapter 8 (specifically, in Sections 8.7 and 8.7.1) how to obtain the covariance matrix during non-linear least squares analysis.

3.4.1. Examples Using Matrix Method. We look at a few examples of how to apply Eq. (3.9). Suppose that $m = 1$ and $f : \mathbb{R}^3 \to \mathbb{R}$ depends on three variables, $f = f(x, y, z)$, assumed to be uncorrelated. The errors on x, y, z are $\sigma_x, \sigma_y, \sigma_z$, respectively. Then,

$$G = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right)$$

$$F = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right) \begin{pmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_y^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{pmatrix} \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{pmatrix}.$$

This gives:

$$F \equiv var(f) = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + \left(\frac{\partial f}{\partial z}\right)^2 \sigma_z^2.$$

Suppose instead that $f : \mathbb{R}^3 \to \mathbb{R}^3$ describes a change of coordinates from cylindrical to Cartesian, i.e. $x = r\cos\phi$, $y = r\sin\phi$ and $z = z$. Suppose that the random variables r, ϕ, z are uncorrelated. Then,

$$G = \begin{pmatrix} \cos\phi & -r\sin\phi & 0 \\ \sin\phi & r\cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$F = \begin{pmatrix} \cos\phi & -r\sin\phi & 0 \\ \sin\phi & r\cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \sigma_r^2 & 0 & 0 \\ 0 & \sigma_\phi^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{pmatrix} \begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -r\sin\phi & r\cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} \cos\phi & -r\sin\phi & 0 \\ \sin\phi & r\cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \sigma_r^2\cos\phi & \sigma_r^2\sin\phi & 0 \\ \sigma_\phi^2(-r\sin\phi) & \sigma_\phi^2 r\cos\phi & 0 \\ 0 & 0 & \sigma_z^2 \end{pmatrix}$$

Then, we conclude that:

$$\begin{pmatrix} var(x) & cov(x,y) & cov(x,z) \\ cov(y,x) & var(y) & cov(y,z) \\ cov(z,x) & cov(z,y) & var(z) \end{pmatrix} = \begin{pmatrix} \sigma_r^2\cos^2\phi + \sigma_\phi^2 r^2\sin^2\phi & (\sigma_r^2 - \sigma_\phi^2 r^2)\sin\phi\cos\phi & 0 \\ (\sigma_r^2 - \sigma_\phi^2 r^2)\sin\phi\cos\phi & \sigma_r^2\sin^2\phi + \sigma_\phi^2 r^2\cos^2\phi & 0 \\ 0 & 0 & \sigma_z^2 \end{pmatrix}$$

We see that the errors and correlations involving z are not affected by the transformation (since z is never transformed). The errors in x and y depend on where they are (r, ϕ) and on the variances of r, ϕ.

3.5. Additive And Multiplicative Systematic Errors

3.5.1. Additive Errors. Systematic errors often appear as an overall shift of the value we measure, i.e. instead of measuring x we measure x' which is the sum of x and an offset δ:

$$x' = x + \delta.$$

Another random variable y is subject to the same systematic drift:

$$y' = y + \delta.$$

Thus, x and y have a common systematic error. Assume that x and y are independent of each other and independent of δ. The error in x' is:

$$var(x') = var(x) + var(\delta)$$

and similarly for y. The covariance is:

$$cov(x', y') = cov(x + \delta, y + \delta) = var(\delta).$$

Thus, the covariance matrix for x', y' has the random and systematic error added in quadrature along the diagonal whereas off-diagonal elements are the square systematic errors:

$$\begin{pmatrix} var(x) & cov(x,y) \\ cov(y,x) & var(y) \end{pmatrix} = \begin{pmatrix} var(x) + var(\delta) & var(\delta) \\ var(\delta) & var(y) + var(\delta) \end{pmatrix}.$$

If the systematic error is constant (i.e. does not fluctuate), then $var(\delta) = 0$ and x', y' are uncorrelated.

3.5.2. Multiplicative Errors. Suppose instead that the systematic errors are multiplicative:

$$x' = \delta \cdot x.$$

Similarly for y:

$$y' = \delta \cdot y.$$

We assume that x and y are independent of each other. We also assume that δ is independent of x and y. Then,

$$var(x') = \overline{\delta}^2 var(x) + \overline{x}^2 var(\delta) + var(\delta)var(x)$$

and

$$var(y') = \overline{\delta}^2 var(y) + \overline{y}^2 var(\delta) + var(\delta)var(y).$$

The covariances are:

$$\begin{aligned} cov(x', y') =& \overline{(\delta \cdot x - \overline{\delta \cdot x})(\delta \cdot y - \overline{\delta \cdot y})} = \overline{\delta \cdot x \delta \cdot y} - \overline{\delta \cdot x} \cdot \overline{\delta \cdot y} \\ =& \overline{\delta^2 xy} - \overline{\delta x} \cdot \overline{\delta y} = \overline{\delta^2} \cdot \overline{x} \cdot \overline{y} - \overline{\delta} \cdot \overline{x} \cdot \overline{\delta} \cdot \overline{y} = var(\delta)\overline{x} \cdot \overline{y}. \end{aligned}$$

The covariance matrix of x', y' is therefore:

$$\begin{pmatrix} \overline{\delta}^2 var(x) + \overline{x}^2 var(\delta) + var(\delta)var(x) & var(\delta)\overline{x} \cdot \overline{y} \\ var(\delta)\overline{x} \cdot \overline{y} & \overline{\delta}^2 var(y) + \overline{y}^2 var(\delta) + var(\delta)var(y) \end{pmatrix}.$$

The amount of fluctuations in δ determine the magnitude of covariance between x' and y'. If δ does not fluctuate, $var(\delta) = 0$, δ is just a number and this reduces to:

$$\begin{pmatrix} \overline{\delta}^2 var(x) & 0 \\ 0 & \overline{\delta}^2 var(y) \end{pmatrix}.$$

There is no longer any covariance between x' and y'. As expected, the errors in x', y' depend on the mean value of the scaling factor δ. Since the multiplicative factor is just a constant, this result could also have been deduced from the theorem $var(aX) = a^2 var(X)$.

3.6. Monte-Carlo Method For Error Propagation

3.6.1. Toy Model. Let $X, Y \sim \mathcal{N}(0, \sigma^2)$ be iidrv's and x, y their corresponding values. Define the transformation:

$$r = \sqrt{x^2 + y^2}.$$

You may recall this mapping from Section 24 in the context of transformation to polar coordinates R, Θ from $X, Y \sim \mathcal{N}(0, \sigma^2)$, which yielded a Rayleigh distribution for R and a uniform distribution for Θ. For now, let us focus on error propagation through the function $r(x, y)$. Propagation of errors gives:

$$\sigma_r^2 = \left(\frac{\partial r}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial r}{\partial y}\right)^2 \sigma_y^2,$$

where

$$\frac{\partial r}{\partial x} = \frac{1}{2} \frac{2x}{\sqrt{x^2 + y^2}} = \frac{x}{\sqrt{x^2 + y^2}}$$

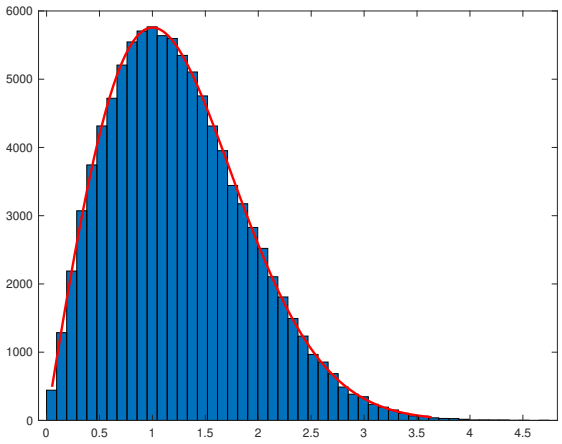


Figure 3.7. Histogram obtained via random number generation (Monte-Carlo) for the error propagation through $r = \sqrt{x^2 + y^2}$, where x, y are standard normals. The red line is a fit to the Rayleigh distribution.

$$\frac{\partial r}{\partial y} = \frac{1}{2}\frac{2y}{\sqrt{x^2+y^2}} = \frac{y}{\sqrt{x^2+y^2}}.$$

Substituting:

$$\sigma_r = \sqrt{\frac{x^2}{x^2+y^2}\sigma_x^2 + \frac{y^2}{x^2+y^2}\sigma_y^2}.$$

But since $X, Y \sim \mathcal{N}(0, \sigma^2)$, this reduces to:

$$\sigma_r = \sigma\sqrt{\frac{x^2}{x^2+y^2} + \frac{y^2}{x^2+y^2}} = \sigma. \tag{3.10}$$

We can check in MATLAB if the formula for propagation of errors is correct. Suppose that $\sigma = 1$, i.e. $X, Y \sim \mathcal{N}(0, 1)$. Then, according to Eq. (3.10), we should find $\sigma_r = 1$. However, we instead find `0.6548`:

```
>> x=randn([1 100000]);
>> y=randn([1 100000]);
>> r=sqrt(x.^2+y.^2);
>> figure; histfit(r,50,'rayleigh');
>> std(r)

ans =

    0.6548
```

The histogram is shown in Figure 3.7.

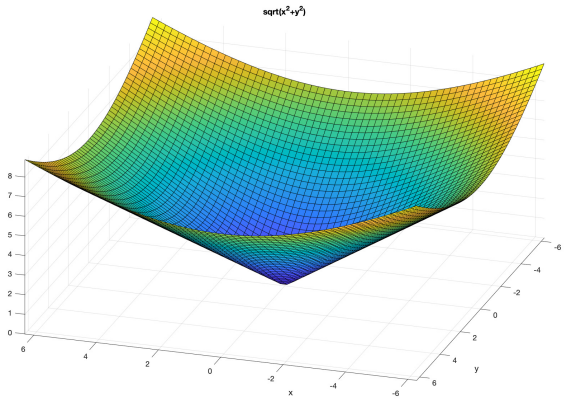


Figure 3.8. Plot of the function $r = \sqrt{x^2 + y^2}$.

How can this difference be explained? Let's begin by examining the origins of the error propagation formula. We started from the linear approximation of f:

$$\delta f = f(\overline{\mathbf{x}} + \delta\mathbf{x}) - f(\overline{\mathbf{x}}) = \sum_{i=1} \frac{\partial f}{\partial x_i} \delta x_i + O(|\delta\mathbf{x}|^2).$$

To first order, the variance of δf is:

$$var(\delta f) = \sum_{i=1} \left(\frac{\partial f}{\partial x_i} \right)^2 var(\delta x_i),$$

whose square root gives the error in f, i.e., $\sqrt{var(\delta f)} \equiv \sigma_f$. (In our case, f is $r = \sqrt{x^2 + y^2}$, so $\sigma_f \equiv \sigma_r$.) We are hoping that this linear approximation yields a good approximation to the true variance, $var(\delta f) = var(f(\overline{\mathbf{x}} + \delta\mathbf{x}) - f(\overline{\mathbf{x}}))$.We can check if the linear approximation is valid by estimating the magnitude of the remainder term, $O(|\delta\mathbf{x}|^2)$. A plot of the function $r(x, y) = \sqrt{x^2 + y^2}$ is shown in Figure 3.8.

The remainder term of the Taylor expansion is proportional to the second derivative. Consider a region centered on the origin with radius σ. If over this region the norm of the Hessian matrix (second derivative) of r is bounded by $q \leq \|\nabla\nabla r(x, y)\| \leq Q$, the remainder term satisfies the inequality $q\sigma^2/2 \leq O(|\delta\mathbf{x}|^2) \leq Q\sigma^2/2$. The Hessian matrix of $r = \sqrt{x^2 + y^2}$ is:

$$\nabla\nabla r(x, y) = \begin{bmatrix} \frac{1}{\sqrt{x^2+y^2}} - \frac{x^2}{(x^2+y^2)^{3/2}} & -\frac{xy}{(x^2+y^2)^{3/2}} \\ -\frac{xy}{(x^2+y^2)^{3/2}} & \frac{1}{\sqrt{x^2+y^2}} - \frac{x^2}{(x^2+y^2)^{3/2}} \end{bmatrix}$$

For the $y = 0$ and $x = 0$ directions the linear approximation is exact (Hessian vanishes). However, at 45° ($x = y$), the Hessian reaches a maximum:

$$\begin{bmatrix} \frac{1}{\sqrt{2x^2}} - \frac{x^2}{(2x^2)^{3/2}} & -\frac{x^2}{(2x^2)^{3/2}} \\ -\frac{x^2}{(2x^2)^{3/2}} & \frac{1}{\sqrt{2x^2}} - \frac{x^2}{(2x^2)^{3/2}} \end{bmatrix} = \begin{bmatrix} \frac{1}{2\sqrt{2}|x|} & -\frac{1}{2\sqrt{2}|x|} \\ -\frac{1}{2\sqrt{2}|x|} & \frac{1}{2\sqrt{2}|x|} \end{bmatrix} \sim \frac{1}{|x|},$$

whose eigenvalues are 0 and $1/(\sqrt{2}|x|)$ (i.e. the Frobenius norm is $1/(\sqrt{2}|x|)$). If the "small parameter" of the Taylor expansion is $\sim var(\delta x_i) \sim \sigma_x = \sigma_y \equiv \sigma$, then this Hessian has norm $\sim \frac{1}{\sqrt{2}\sigma}$. Over the circle with radius σ the second derivative is bounded by

$$\frac{1}{\sqrt{2}\sigma} \leq \|\nabla\nabla r(x, y)\| \leq \infty$$

Therefore, the Taylor remainder is bounded by

$$\frac{1}{\sqrt{2}\sigma}\frac{\sigma^2}{2}\left(\equiv \frac{\sigma}{2\sqrt{2}}\right) \leq O(|\delta\mathbf{x}|^2) \leq \infty,$$

with singular the upper limit corresponding to the limit of small $r = \sqrt{x^2 + y^2}$. In the Taylor expansion the first order term is $\sim (\partial_i f)\sigma$ ($i = x, y$), where the first derivatives themselves are $(x, y)/\sqrt{x^2 + y^2}$ and therefore, of order unity (magnitude is 1 when either $x = 0$ or y=0, and $1/\sqrt{2}$ at 45°, when $x = y$). The remainder is bounded from below by $\frac{\sigma}{2\sqrt{2}}$ (when $x = y$). Thus, the remainder is comparable in size to the first order term. The first-order Taylor expansion is therefore not a good approximation since the error term is just as large as the approximating term itself. This is an example of the breakdown of the linear approximation used in error propagation.

We have concluded that when the linear approximation breaks down, the correct error bar cannot be obtained by standard error propagation. We instead must know the distribution function of the new variable (r) and based on this distribution extract the error as a parameter of the distribution. We have already established numerically that the standard deviation should be `0.6548`. Let's check that this is consistent with the parameters of the Rayleigh distribution

$$p_R(r) = \frac{r}{b^2}e^{-r^2/(2b^2)},$$

whose mean is $b\sqrt{\pi/2}$ and variance is $b^2(4-\pi)/2$. Fitting the r-data to the distribution gives a value of `1.00083`, for the b-parameter:

```
>> fitdist(r','rayleigh')

ans =

  RayleighDistribution
```

```
  Rayleigh distribution
    B = 1.00083   [0.997739, 1.00394]

>> (1.00083^2)*sqrt((4-pi)/2)

ans =

    0.6562
```

Using the formula for the variance of the Rayleigh distribution we find that the standard deviation should be `0.6562`, which is in excellent agreement with our Monte-Carlo estimate of `0.6548`.

3.6.2. Monte-Carlo Method. This suggests a very simple but accurate and general numerical method for error propagation that does not rely on the linear approximation:

- For a function of n variables, $f = f(x_1, x_2, \ldots, x_n)$ generate N (N very large) random numbers for each rv $x_1, \ldots, x_n$ according to their joint distribution. Hopefully, these are normally distributed iidrv's, as most computer software can generate such random numbers.
- Propagate the error through the equation $f = f(x_1, x_2, \ldots, x_n)$ by feeding it the N random numbers corresponding to the N realizations of each of the n random variables ($N \times n$ random numbers total will have been generated for this purpose). This will result in N values of f, i.e., $f_1, \ldots, f_N$ (one value per realization of the rv's).
- Compute the histogram of the new random variable f, using the N values $f_1, \ldots, f_N$.
- Fit the histogram to an appropriate distribution. (If the distribution of f is known or can be derived, use it; otherwise pick a distribution that approximates the histogram well.) Obtain the parameters of the distribution.
- Alternatively (to fitting the distribution of f), you can instead numerically calculate the moments of f. For example, the variance may be sufficient if all you need is an error bar for f. However, make sure you inspect the distribution of f to make sure that the variance is a good measure of its spread.

3.6.3. Linear vs Non-Linear Propagation. The example covered in Section 3.6 illustrates an interesting potential weakness of the error propagation method you should be aware of. The error propagation method works well when f can be approximated by a linear map near $\overline{\mathbf{x}}$. Examples of linear maps were presented in Sections 2.24.2 and 2.24.1 where we covered

the the normal sum theorem,

$$\mathcal{N}(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)=\mathcal{N}_1(\mu_1,\sigma_1^2)+\mathcal{N}_2(\mu_2,\sigma_2^2),$$

and the normal linear transform theorem

$$\alpha+\beta\mathcal{N}(\mu,\sigma^2)=\mathcal{N}(\alpha+\beta\mu,\beta^2\sigma^2).$$

The first corresponds to a relationship of the form $f(X,Y)=X+Y$ (two variables), where X,Y are independent normal rv's. The second is a relationship of the form $f(X)=\alpha+\beta X$ (single variable), where X is normal. In both cases, the Taylor expansion stops after the linear term (the remainder term is zero). In the case $f(X,Y)=\sqrt{X^2+Y^2}$ the relationship is nonlinear and the Taylor expansion of f requires an infinite number of terms. In that case, we must be careful to assess the magnitude of the remainder term relative to the linear term. The error propagation formula based on Taylor's theorem is justified when $O(|\delta\mathbf{x}|^2)\ll|\sum_i(\partial_i f)\delta x_i|$.

3.7. Problems

Problem 96. The Richter magnitude of an earthquake is determined from the logarithm of the amplitude of waves recorded by seismographs (adjustments are included to compensate for the variation in the distance between the various seismographs and the epicenter of the earthquake). The formula is:

$$M_L=\log_{10}(A/A_0(\delta))$$

where A is the maximum excursion of the Wood-Anderson seismograph, the empirical function A_0 depends only on the epicentral distance of the station, δ. Both A and δ are prone to measurement error. Find the uncertainty in M_L due to errors in A and δ.

Solution. Propagation of error formula is:

$$\sigma_{M_L}^2=\left|\frac{\partial M_L}{\partial A}\right|^2\sigma_A^2+\left|\frac{\partial M_L}{\partial\delta}\right|^2\sigma_\delta^2,$$

where the derivatives are:

$$\frac{\partial M_L}{\partial A}=\frac{1}{\log(10)}\frac{A_0(\delta)}{A}\cdot\frac{1}{A_0(\delta)}=\frac{1}{A\cdot\log(10)}$$

$$\frac{\partial M_L}{\partial\delta}=-\frac{1}{\log(10)}\frac{A_0(\delta)}{A}\cdot\frac{1}{[A_0(\delta)]^2}\frac{\partial A_0(\delta)}{\partial\delta}=-\frac{\partial A_0(\delta)/\partial\delta}{A\cdot A_0(\delta)\cdot\log(10)}$$

■

Problem 97. You built a setup to detect light with an avalanche photodiode. In an avalanche photodiode the carrier production rate $\eta(P/h\nu)$ is increased by a factor of M because of the ionization by the drifting electrons and holes. Here, η is the detector quantum efficiency (number of

carriers generated divided by the number of photons absorbed), P is the power absorbed in the detector and $h\nu$ is the incident photon energy. The photocurrent is enhanced by the same value:

$$i = Me\left[\eta\left(\frac{P}{h\nu}\right) + g\right]$$

where g is a constant and e is the charge of the electron. Suppose that light incident on the photodetector is in a coherent state. A coherent state is a photon field that is a linear superposition of single-mode states. It is the closest approximation to a classical electromagnetic field, as far as quantum mechanics' Heisenberg uncertainty principle allows.

In a coherent state the fluctuations of the photon number follow Poisson statistics, i.e. for a measurement of the number of photons in the light field, the probability of the field containing n photons is

$$\mathbb{P}(n;\overline{n}) = \frac{e^{-\overline{n}}\overline{n}^n}{n!}$$

which is a Poisson distribution with mean $\overline{n}$. The light power absorbed in the detector (P) is proportional to n, i.e. $P = Cn$ where C is a constant.

(a) Compute the fluctuations in the photon count (n) of the coherent state in terms of the measured photocurrent (i). From the noise, estimate the actual photon count n in the electromagnetic field.

(b) Same as (a) but use the fluctuations in the measured electrical power instead of current (i). Electrical power is i^2R, where R is a constant (the load resistance).

Solution. (a)

$$i = Me\left[\eta\frac{Cn}{h\nu} + g\right] \quad \rightarrow \quad \delta i = \left|Me\eta\frac{c}{h\nu}\right|\delta n \quad \rightarrow \quad \delta n = \frac{\delta i}{\left|Me\eta\frac{c}{h\nu}\right|}$$

square δn to get n.
(b)

$$P = i^2R = R(Me)^2\left[\eta\frac{Cn}{h\nu} + g\right]^2$$

expanding the square bracket:

$$\eta^2\frac{c^2n^2}{(h\nu)^2} + 2\eta\frac{Cng}{h\nu} + g^2$$

solving for δP

$$\delta P = R(Me)^2 \left[\frac{2\eta^2 c^2 n}{(h\nu)^2} + \frac{2\eta C g}{h\nu}\right] \delta n$$

replace n by $(\delta n)^2$ and solve to get δn. Square to get n. ■

Problem 98. Use the method of error propagation to find the error in the work function W (due to a force $F(x)$ applied over a distance x, i.e. $W = \int F(x)\mathrm{d}x$ and you can assume Hooke's law for F) performed by dragging a light object of mass m across a distance x. The object is mounted at the end of a spring whose constant is k. The error in x is denoted δx.

(a) Find the error δW in terms of δx.

(b) What if the displacement x is measured using an interferometer whose output is designed to produce a voltage V (you may assume that V is linear in x).

Solution. (a) Substitute $F = kx$ and integrate to get $W = \frac{1}{2}kx^2$. Then $\delta W = |\partial W/\partial x|\delta x = |kx|\delta x$.

(b) Now we have $x = AV$, where A is some constant. Then, $W = \frac{1}{2}kA^2V^2$ and $\delta W = |kA^2V|\delta V$. ■

Problem 99. We have seen in class that when two random variables X and Y are added to form a new random variable, $Z = X + Y$, the errors add in quadrature:

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2.$$

However, in your chemistry labs, you may have been taught that the errors add in first power:

$$\sigma_Z = \sigma_X + \sigma_Y.$$

There are no restrictions on the distributions of X and Y apart from the obvious requirement of finite variances.

(a) Derive the second formula ($\sigma_Z = \sigma_X + \sigma_Y$, where $\sigma_X, \sigma_Y \geq 0$).

(b) Show the relationship between these two different methods for computing the error in Z, i.e., derive an inequality that relates these two different forms for σ_Z.

Solution. (a) The second formula follows from mapping the range of X and Y values onto the Z axis. $Z(X, Y) = X + Y$ is a function of two variables. As X and Y range over their allowed values (thanks to σ_X and σ_Y), Z

ranges from

$$Z_{min} = (X - \sigma_X) + (Y - \sigma_Y)$$

up to

$$Z_{max} = (X + \sigma_X) + (Y + \sigma_Y)$$

The difference between Z_{min} and Z_{max} is:

$$Z_{max} - Z_{min} = 2\sigma_x + 2\sigma_Y.$$

This is the total range. Error bar is half of this, since we write it symmetrically as $Z \pm \delta Z$:

$$\sigma_Z = \sigma_X + \sigma_Y.$$

(b) Observe that $(\sigma_X, \sigma_Y \geq 0)$:

$$|\sigma_X + \sigma_Y|^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_X\sigma_Y.$$

Hence,

$$|\sigma_X + \sigma_Y|^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_X\sigma_Y \geq \sigma_X^2 + \sigma_Y^2.$$

The left hand side is σ_Z^2, where $\sigma_Z = \sigma_X + \sigma_Y$. The right hand side is σ_Z^2, where $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$. Thus, the error bar obtained from $\sigma_Z = \sigma_X + \sigma_Y$ is larger than the error bars obtained by adding the errors in quadrature. ■

Problem 100. Use error propagation to show how an error in the measured radius leads to errors in the calculation of sphere volume (as calculated from the radius).

Solution. From $V = \frac{4}{3}\pi r^3$, $\delta V = 4\pi r^2 \delta r$. ■

Problem 101. Same question as the previous one, except that now consider the case where you only remember the value of π to a few digits (e.g. $\pi = 3.14$). How does this error in π (in addition to the error in r) contribute to the uncertainty in volume?

Solution. $(\delta V)^2 = |4\pi r^2|^2 |\delta r|^2 + \left|\frac{4}{3} r^3\right|^2 |\delta \pi|^2$. ■

Problem 102. Suppose you work at a particle accelerator and your job is to spend your life measuring two quantities in the laboratory, X and Y. X could be for example, a voltage, whereas Y could be a position. Both X and Y relate to some important physical quantity being measured. The two measurements are uncorrelated and statistically independent. It is known that the values of X range continuously from 0 to 1 and appear to be uniformly distributed whereas Y appears to be Gaussian distributed with mean 0 and variance 1.

(a) Calculate the value of the probability $\mathbb{P}(0.2 < X < 0.8, Y < 0)$.

(b) Calculate the probability $\mathbb{P}(0.2 < X < 0.8, Y \neq 0)$ (i.e. the joint probability that X lies in the range $[0.2, 0.8]$ and Y does not equal zero).

(c) A physical theory for the new particle predicts that its spin can be obtained from $Z = Xe^Y$, where X is related to the total orbital moment of the subatomic particles and Y is related to the total spin of these particles. Definite values of Z are hard to obtain due to experimental error, so the best you can do is *set an upper bound on its value*. Explain how you would set such an upper bound. Or equivalently, calculate the chance $\mathbb{P}(Z < z)$ that Z will take a value less than z.

Solution. (a) $\mathbb{P} = 0.6 * 0.5 = 0.30$

(b) $\mathbb{P} = 0.6 * 1 = 0.6$

(c) Method 1: since the distributions of X and Y are known:

$$\begin{aligned}\mathbb{P}(Z < z) =& \mathbb{P}(Xe^Y < z) = \int_{\{(x,y)|xe^y<z, 0\le x\le 1\}} p_{XY}(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_{\{(x,y)|xe^y<z, 0\le x\le 1\}} 1\cdot\frac{e^{-y^2/2}}{\sqrt{2\pi}}\mathrm{d}x\mathrm{d}y = \int_0^1 dx\int_{-\infty}^{\log(z/x)}\frac{e^{-y^2/2}}{\sqrt{2\pi}}\mathrm{d}y \\ &= \int_0^1 \Phi(\log(z/x))\mathrm{d}x.\end{aligned}$$

Method 2: using error propagation, which can be done since $z(x, y)$ is known and the errors in X and Y are also known. ■

Problem 103. From the coast of Normandy, you measure the position of an English flag along the coast of Great Britain, across the English channel, using a telescope. You use the angle the telescope makes relative to its base, and the known distance across the English channel, to convert the angle to lateral distance along the coast (using simple trigonometry). However, the intense winds make it difficult if not nearly impossible for you to readout the angle (and hence the distance) properly. The intense vibrations result in random fluctuations of this distance.

Fortunately, you know that averaging several measurements together can reduce the noise. Let $X_1, \ldots, X_n$ be n measurements of the flag's true position X (r.v. mean μ and variance σ^2; the variance is a measure of the area of the flag). What should the value of n be (n: no. of samples acquired in an experiment) needed to ensure that the probability that the position (calculated form the sample mean) does not deviate from the true position of the flag by more than $\sigma/10$ is at least 0.95.

Solution. Using the inequality on p.1 with $\epsilon = \sigma/10$

$$\mathbb{P}\left(|\overline{X}_n - \mu| \le \frac{\sigma}{10}\right) = 1 - \mathbb{P}\left(|\overline{X}_n - \mu| > \frac{\sigma}{10}\right) \ge 1 - \frac{var(\overline{X}_n)}{(\sigma/10)^2}$$

Where $\overline{X}_n = \dfrac{X_1 + \cdots + X_n}{n}$, $var(\overline{X}_n) = \dfrac{var(X)}{n} = \dfrac{\sigma^2}{n}$.

$$\mathbb{P}\left(|\overline{X}_n - \mu| \leq \frac{\sigma}{10}\right) \geq 1 - \frac{var(X)}{n(\sigma/10)^2} = 1 - \frac{100}{n}.$$

If we want this probability to be at least 0.95, we must have $100/n \leq 0.05$ or $n \geq 100/0.05 = 2000$. ■

Problem 104. Suppose we have Hooke's law $F = kx$, where k and x have uncertainties. Use error propagation to determine the error in F from the errors in k and x

Solution.

$$(\delta F)^2 = \left(\frac{\partial F}{\partial k}\right)^2 (\delta_k)^2 + \left(\frac{\partial F}{\partial x}\right)^2 (\delta_x)^2 = x^2(\delta_k)^2 + k^2(\delta_x)^2$$

$$(\delta F) = \sqrt{x^2(\delta_k)^2 + k^2(\delta_x)^2}.$$

■

Chapter 4

Statistical Parameter Estimation

Statistical distributions, such as the Gaussian distribution, contain parameters such as the mean and variance whose values may be unknown. For example, suppose that we want to estimate the mean of a distribution. We may have at our disposal a series of measurements of a random variable X (random sample) $\{X_1, \ldots, X_n\}$. Corresponding to a random experiment ω we denote their values by lowercase letters: $X_i(\omega) = x_i$. How should we calculate the average?

For example, should we use the arithmetic mean (sample mean),

$$\overline{X}_{arith.} \equiv \frac{1}{n} \sum_{i=1}^{n} x_i,$$

the harmonic mean

$$\overline{X}_{harm.} \equiv \left(\frac{1}{n} \sum_{i=1}^{n} x_i^{-1} \right)^{-1},$$

or the geometric mean

$$\overline{X}_{geom.} \equiv \left(\prod_{i=1}^{n} x_i \right)^{1/n} ?$$

Methods have been developed to estimate the parameters of statistical distributions. Estimation theory is a branch of statistics that deals with estimating the values of parameters based on measured empirical data. An

estimator is not an exact representation of the parameters, but instead attempts to approximate the parameters from the measurements using some underlying principle. There are several estimation methods that are commonly used, such as maximum likelihood estimation (MLE), Bayes estimation, the method of moments, the maximum a posteriori (MAP) method, etc. The simplest one is MLE; other methods require additional theory. We shall cover the MLE method.

4.1. Maximum Likelihood Estimation (MLE)

MLE estimates the parameters of a statistical model by finding the parameter values that maximize the likelihood of making the observations given the parameters.

4.1.1. Likelihood Function. Let us construct a so-called "likelihood function" $L(\boldsymbol{\theta})$, as a function of the unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, and conditioned by the observation of the random samples. We will set this likelihood function to be equal to the joint probability density of observing the values $X_1(\omega) = x_1, \ldots, X_n(\omega) = x_n$. Since the random samples are independent random variables, the joint density factorizes into a product of densities for each:

$$\boxed{L(\boldsymbol{\theta}|x_1, x_2, \ldots, x_n) \equiv p_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} p_X(x_i|\boldsymbol{\theta}).}$$

Here, $p_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n|\boldsymbol{\theta})$ is the joint PDF of $X_1, \ldots, X_n$ given $\boldsymbol{\theta}$. $p_X(x_i|\boldsymbol{\theta})$ is the PDF of X given $\boldsymbol{\theta}$. An interpretation of the likelihood function is as follows: given observed data $x_1, x_2, \ldots, x_n$ (corresponding to the realization ω of iidrv's $X_1, \ldots, X_n \sim X$), a distribution with parameters $\boldsymbol{\theta}$, then:

$$\text{probability that } X_i \text{ is in } [x_i, x_i + \mathrm{d}x_i] \text{ for all } i = \prod_{i=1}^{n} p_X(x_i|\boldsymbol{\theta})\mathrm{d}x_i,$$

where $\boldsymbol{\theta}$ represents one or more parameters of the distribution $p_X(x|\boldsymbol{\theta})$. The likelihood function, $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p_X(x_i|\boldsymbol{\theta})$, which is just the joint PDF of the x_i, is treated here as a function of the parameter(s), $\boldsymbol{\theta}$. The x_i, on the other hand, are treated as fixed (the experiment is over). The MLE method consists of solving the system of equations, $\nabla_\theta L = 0$, or $\partial L/\partial\theta_i = 0$, $i = 1, \ldots, p$ for the unknown parameters θ_i, as function of the data $(x_1, \ldots, x_n)$.

Next, we need to know or assume a distribution for the random variables $X_1, X_2, \ldots, X_n$. We did assume they were iidrv's, $X_i \sim X$. The distribution of X is arbitrary. By distribution we mean the CDF or PDF as well as its

parameters $\boldsymbol{\theta}$. As an example, if they are normally distributed, the parameters are $\boldsymbol{\theta} = (\sigma, \mu)$. Due to statistical independence, the joint probability distribution $p_{X_1,\ldots,X_n}(x_1,\ldots,x_n|\boldsymbol{\theta})$ for observing the whole set of n readouts $\{X_i(\omega) = x_i\}$ factorizes as a product:

$$p_X(x_1|\boldsymbol{\theta}) \cdot p_X(x_2|\boldsymbol{\theta}) \cdot \cdots \cdot p_X(x_n|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\sum_{i=1}^n (x_i-\mu)^2/2\sigma^2},$$

where $\boldsymbol{\theta} = (\mu, \sigma)$. The problem we want to solve is this: given the measurements $\{x_1,\ldots,x_n\}$ we want to estimate the unknown quantities μ and σ. In other words, we would like to find "best estimates" for these parameters in terms of the observations $x_1,\ldots,x_n$.

4.1.2. Principle of Maximum Likelihood. The *principle of maximum likelihood* states that the best estimates for $\boldsymbol{\theta}$ are those for which the observed data $x_1,\ldots,x_n$ are most likely to occur, i.e. for which the likelihood function $L(\boldsymbol{\theta}|x_1, x_2, \ldots, x_n)$ is a maximum with respect to $\boldsymbol{\theta}$. Suppose that $\boldsymbol{\theta} = (\mu, \sigma)$, maximization with respect to $\boldsymbol{\theta}$ means that we should enforce the following two conditions:[1]

$$\frac{\partial L}{\partial \mu} = 0 \qquad \text{and} \qquad \frac{\partial L}{\partial \sigma} = 0,$$

where the derivatives are evaluated at the point $(\mu, \sigma) = (\hat{\mu}, \hat{\sigma})$. The hat notation is used to denote the particular choice of θ values obtained by MLE.

Let us start with the first condition ($\frac{\partial L}{\partial \mu} = 0$). Inspection of the above "Gaussian" expression for $p_{\mu,\sigma}$ shows that μ only appears in the argument of exp. Finding the extremum of $\exp(f(\mu))$ with respect to μ is equivalent to finding the extremum of $f(\mu)$ with respect to μ, since:

$$\frac{\partial}{\partial \mu} e^{f(\mu)} = e^{f(\mu)} \frac{\partial f}{\partial \mu} = 0$$

[1] An extremum of L is found by setting $\mathrm{d}L = 0$. Take $\boldsymbol{\theta} = (\mu, \sigma)$, for example: since $\mathrm{d}L(\mu, \sigma) = \partial_\mu \mathrm{d}\mu + \partial_\sigma \mathrm{d}\sigma$ and $\mathrm{d}\mu, \mathrm{d}\sigma$ are arbitrary displacements, $\mathrm{d}L = 0$ implies that the gradient of L vanishes: $\nabla L(\mu, \sigma) = 0$. The vanishing gradient means that all partial derivatives vanish: $\partial_\mu = 0$ and $\partial_\sigma = 0$.

and dividing both sides by $e^{f(\mu)}$ implies that $\frac{\partial f}{\partial \mu} = 0$. In general, you can find the extremum of $\log(L)$ rather than L; this is sometimes easier.[2] Thus,

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{n} (x_i - \mu)^2 / 2\sigma^2 = 0 \qquad \text{or} \qquad \sum_{i=1}^{n} (x_i - \mu) = 0,$$

which leads to the formula for sample mean we have been using all along. We substitute $X_i(\omega)$ in place of x_i, suppress the ω notation in order to view $\hat{\mu}$ as a random variable:

$$\boxed{\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i.}$$

The hat notation $\hat{\mu}$ denotes that this particular value of μ corresponds to the MLE for the mean. This provides a justification for the use of $\hat{\mu}$ as the value of x_{best}.

For $\partial L / \partial \sigma$ the situation is different because σ also appears in the coefficient of exp. In this case, we look for the extremum of L instead of its log. Differentiating with respect to σ we get

$$(2\pi)^{n/2} \frac{\partial L}{\partial \sigma} = \frac{(-n)}{\sigma^{n+1}} e^{-\sum_i} + \frac{1}{\sigma^n} e^{-\sum_i} \left(- \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2} \right) \left(- \frac{2}{\sigma^3} \right) = 0.$$

Divide by $e^{-\sum_i}$, multiply by σ^{n+1}:

$$-n + \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} (x_i - \mu)^2 \right) = 0 \qquad \text{which gives} \qquad \boxed{\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2}$$

where $\hat{\mu}$ is the sample mean and Σ_i is shorthand for $\sum_{i=1}^{n} (x_i - \mu)^2 / 2\sigma^2$. In the last step we wrote $(\hat{\mu}, \hat{\sigma})$ to denote the fact that this particular choice of values for (μ, σ) corresponds to the MLE.

[2] We note that finding the extremum of L is equivalent to finding the extremum of $\log L$. This is because $\log(x)$ function is a monotonic function. Indeed, a necessary condition for the maximization of $\log L$ is:

$$\frac{\partial}{\partial x} \log L(x) = \frac{1}{L(x)} \frac{\partial}{\partial x} L(x) = 0,$$

where $L(x) \neq 0$. Multiplying both sides of the equation by $L(x)$ leads to $\frac{\partial}{\partial x} L(x) = 0$, which is a necessary condition for a maximum in $L(x)$. Thus, maximizing L is the same as maximizing $\log L$. (Exercise: our proof is not complete since we have only considered the *necessary* condition for a maximum. Can you complete the argument by analyzing the *sufficient* condition?)

In practice, we instead use the quantity with the coefficient $1/(n-1)$ instead of $1/n$:

$$\boxed{\hat{\sigma}^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2} \qquad \text{sample variance}$$

because the $n-1$ corrects for the underestimation of σ^2 when we use $\hat{\mu}$ (sample mean) as our estimator for μ. Recall that the choice of $\mu = \hat{\mu}$ (sample mean), by definition, minimizes the quantity $\sum_i (X_i - \mu)^2$ for $\mu = \hat{\mu}$. We say that σ^2 is a *biased* estimator of the variance whereas σ^2_{n-1} is an unbiased estimator. The proof for the $n/(n-1)$ correction factor is found in Section 4.2.4.

4.2. Estimator Bias

An estimator is said to be unbiased if its expectation value is equal to the true value of the parameter. For example, if our estimator for the mean is the arithmetic average $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (sample mean), the arithmetic average is said to be an unbiased estimator of the mean if $\mathbb{E}\hat{\mu}$ is equal to the true mean of X. The latter is defined as $\int x p_X(x) \mathrm{d}x$.

4.2.1. Random Sample. Suppose we have a random variable X. In the laboratory, we can measure X by acquiring several measurements of X. Denote these measurements of X by the set of values $(X_1, X_2, \ldots, X_n)$. Each X_i is iidrv with the same distribution of X. X_i are samples of X measured at different points in time. We say that the set $(X_1, X_2, \ldots, X_n)$ is a *random sample* of X.

4.2.2. MLE of the Mean: Is the Estimator Biased? Let X be a random variable with mean μ ($\mu = \int x p_X(x) \mathrm{d}x$ is the true mean). Let $(X_1, X_2, \ldots, X_n)$ be a random sample of X. Here we show that

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

is an unbiased estimator of μ. $\frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean. This statement follows from:

$$\mathbb{E}\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}(n\mu) = \mu.$$

To get the first equality, we used the linearity property of expectation value. In the second equality, we used the fact that $\{X_i\}$ is a random sample of X, whose mean is μ.

4.2.3. Variance of the Sample Mean Estimator. The sample mean

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is a rv because it is a sum of rv's $X_1, X_2, \ldots, X_n$. It has a mean equal to the true mean of X, namely μ, which we verified in the previous section, meaning that this particular expression for $\hat{\mu}$ is an unbiased estimator. By the LLN, it converges to the true mean as n increases. We can also check its variance:

$$var(\hat{\mu}) = \frac{1}{n^2}\sum_{i=1}^{n} var(X_i) = \frac{\sigma^2}{n},$$

meaning that it becomes "narrowly distributed" as n increases. This says that the estimator becomes more and more precise as n increases. This is a desirable feature of estimators. If you have an estimator whose variance increases with n, this is not a good estimator. The square root of $var(\hat{\mu})$ represents the error in the mean; you will recognize it as the standard error (aka the *standard deviation of the mean*).

4.2.4. Bias of the Variance Estimator. Let $(X_1, X_2, \ldots, X_n)$ be a random sample of X (with mean μ and variance σ^2). Here we show that

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2,$$

is a biased estimator of σ^2. By definition, the variance of X_i, which is also the variance of X (since $\{X_i\}$ is a random sample of X), is:

$$\sigma_{X_i}^2 = \sigma_X^2 = \mathbb{E}(X-\mu)^2 \equiv \int (x-\mu)^2 p_X(x)\mathrm{d}x.$$

Also,[3]

$$\begin{aligned}
\mathbb{E}\hat{\sigma}_n^2 =& \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2 = \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}[(X_i-\mu)-(\hat{\mu}-\mu)]^2 \\
=& \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}[(X_i-\mu)^2 - 2(X_i-\mu)(\hat{\mu}-\mu)+(\hat{\mu}-\mu)^2] \\
=& \mathbb{E}\frac{1}{n}\left[\left(\sum_{i=1}^{n}(X_i-\mu)^2\right) - n(\hat{\mu}-\mu)^2\right] \\
=& \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i-\mu)^2\right) - \mathbb{E}(\hat{\mu}-\mu)^2 \\
=& \frac{1}{n}\left(\sum_i \sigma^2\right) - var(\hat{\mu}) = \sigma^2 - var(\hat{\mu}),
\end{aligned}$$

and using the property,

$$var(\sum a_i X_i) = \sum a_i^2 var(X_i),$$

valid for independent random variables, we have:

$$var(\hat{\mu}) = var\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \sum_{i=1}\frac{1}{n^2}var(X_i) = \sum_{i=1}^{n}\frac{1}{n^2}\sigma^2 = \frac{1}{n}\sigma^2.$$

Thus,

$$\mathbb{E}\hat{\sigma}_n^2 = \sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2$$

which shows that $\hat{\sigma}_n^2$ is a biased estimator of σ^2. Thus, to get an estimator which, on the average, yields a result equal to the true value of σ^2 we should take instead $\frac{n}{n-1}\hat{\sigma}_n^2$ as the estimator of the variance. We denote this "bias-corrected" estimator of the variance by $\hat{\sigma}_{n-1}^2$:

$$\boxed{\hat{\sigma}_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\hat{\mu})^2} \qquad \text{sample variance}$$

The factor $\frac{n}{n-1}$ is called Bessel's correction.

[3]To get from the second to the third line, we take the summation inside the square bracket and apply it to the second and third terms. (After multiplying the second term by $n/n = 1$.) The third term inherits a coefficient of n because it is summed n times. In the second term, we replace $(1/n)\sum_i X_i$ by $\hat{\mu}$. The second term, which becomes $-2n(\hat{\mu}-\mu)^2$ is of the same form as the third term, $n(\hat{\mu}-\mu)^2$. Adding the two terms together gives $-n(\hat{\mu}-\mu)^2$. When going to the fourth line, the factor of n cancels the $1/n$ in front of the square bracket, leaving $-\mathbb{E}(\hat{\mu}-\mu)^2$, minus 1 times the variance of $\hat{\mu}$.

This result is the expression for the sample variance introduced earlier in the course (see, e.g., Eq. 1.1). *The sample variance is basically the MLE estimator for the variance, but corrected for bias.* Finally, we note that for large n, $1/n$ and $1/(n-1)$ are asymptotically equal. Thus, $\hat{\sigma}_n^2$ is asymptotically unbiased, i.e.

$$\lim_{n\to\infty} \hat{\sigma}_n^2 = \hat{\sigma}_{n-1}^2.$$

Finally, we would like to know how "sharp" this estimator is. Ideally, we would want an estimator whose uncertainty is small. We do this by computing its variance:

$$var(\hat{\sigma}_{n-1}^2) = \mathbb{E}(\hat{\sigma}_{n-1}^4) - (\mathbb{E}\hat{\sigma}_{n-1}^2)^2 = \mathbb{E}(\hat{\sigma}_{n-1}^4) - (\sigma^2)^2.$$

where

$$\begin{aligned}
\mathbb{E}\hat{\sigma}_n^4 =& \mathbb{E}\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\hat{\mu})^2\frac{1}{n-1}\sum_{j=1}^{n}(X_j-\hat{\mu})^2 \\
=& \frac{1}{(n-1)^2}\left(\sum_{i=1}^{n}\mathbb{E}(X_i-\hat{\mu})^4 + \sum_{i\neq j}^{n}\mathbb{E}(X_i-\hat{\mu})^2(X_j-\hat{\mu})^2\right) \\
=& \frac{1}{(n-1)^2}\left(\sum_{i=1}^{n}\mathbb{E}(X_i-\hat{\mu})^4 + \sum_{i\neq j}^{n}\mathbb{E}(X_i-\hat{\mu})^2\cdot\mathbb{E}(X_j-\hat{\mu})^2\right) \\
=& \frac{1}{(n-1)^2}\left(n\mu_4 + n(n-1)\sigma^4\right)
\end{aligned}$$

where $\mu_4 = \mathbb{E}(X_i - \hat{\mu})^4$ is the fourth central moment. In the limit of large n this quantity tends to:

$$\lim_{n\to\infty}\mathbb{E}\hat{\sigma}_n^4 = \lim_{n\to\infty}\frac{1}{(n-1)^2}\left(n\mu_4 + n(n-1)\sigma^4\right) = \sigma^4.$$

Substituting into the above expression for $var(\hat{\sigma}_{n-1}^2)$, we find:

$$\lim_{n\to\infty} var(\hat{\sigma}_{n-1}^2) = \lim_{n\to\infty}\mathbb{E}(\hat{\sigma}_{n-1}^4) - (\sigma^2)^2 = 0.$$

Therefore, the uncertainty of our variance estimator $\hat{\sigma}_{n-1}^2$ decreases as $n \to \infty$. In fact, $var(\hat{\sigma}_{n-1}^2) \propto \frac{n}{(n-1)^2}$ or $1/n$ for large n. The ability to reduce the uncertainty of our estimator by acquiring more data is a useful feature.

4.2.5. MLE Can Fail. The MLE method does not always work. In fact, it can fail even in the simplest cases. Suppose that the density is a mixture of two normal densities:

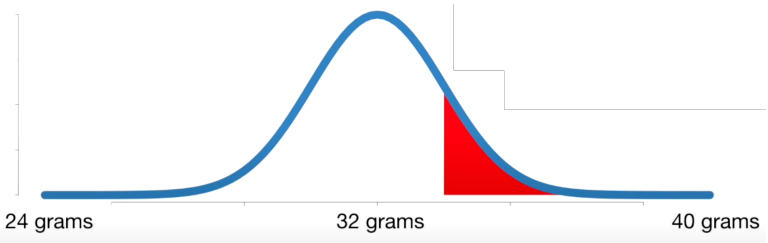
$$p(x|a,\sigma) = \frac{1}{2\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-a)^2}{2\sigma^2}\right] + \frac{1}{2\sqrt{2\pi}}\exp\left[-\frac{x^2}{2}\right],$$

where the parameters (a, σ) are unknown. Suppose that we measure iidrv $X_1, \ldots, X_n$ (whose values are denoted $x_1, \ldots, x_n$). For any constant $\delta > 0$, there exists small enough $\sigma = \sigma_0$ that for $a = x_1$,

$$\begin{aligned} L(x_1, \sigma_0) &= \sum_{i=1}^{n} \log p(x_i | x_1, \sigma_0) \\ &> \log\left(\frac{1}{2\sigma_0\sqrt{2\pi}}\right) + \sum_{i=2}^{n} \log\left(\frac{1}{2\sqrt{2\pi}} \exp\left[-\frac{x_i^2}{2}\right]\right) \\ &= -\log\sigma_0 - \left(\sum_{i=2}^{n} \frac{x_i^2}{2}\right) - n\log 2\sqrt{2\pi} > \delta \end{aligned}$$

From this inequality, we can conclude that the likelihood does not exist (i.e. it can always be made to "blow up" by taking the limit $\sigma_0 \to 0$). Therefore, MLE is unable to estimate the parameters a and σ. Thus, the range of applicability of MLE is limited.

4.2.6. Difference between probability vs likelihood. The difference between probability and likelihood is best illustrated with an example. Suppose that we have a distribution of animal weights with a mean of 32 grams and a standard deviation of 2.5. The probability that we will weigh a randomly selected animal between 32 and 34 grams is given by the area under the curve between 32 and 34 grams (left).



In this case, the area under the curve happens to be 0.29, meaning that there is a 29% chance a randomly selected animal will weigh between 32 and 34 grams. Mathematically we express this as follows:

$$\mathbb{P}(\text{weight between 32 and 34 grams} | \text{mean=32, standard deviation=2.5}) = 0.29$$

or

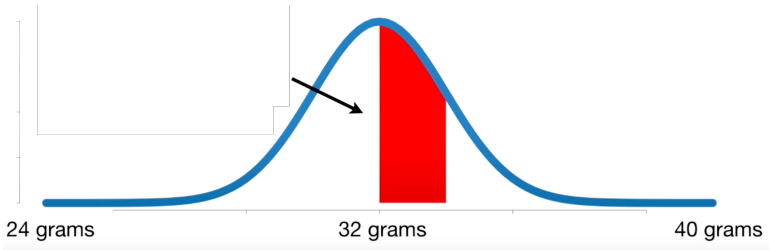
$$\mathbb{P}(32 \leq X \leq 34 | \mu = 32, \sigma = 2.5) = 0.29.$$

Another example is:

$$\mathbb{P}(\text{animal weighs} > 34 \text{ grams} | \text{mean=32, standard deviation=2.5}).$$

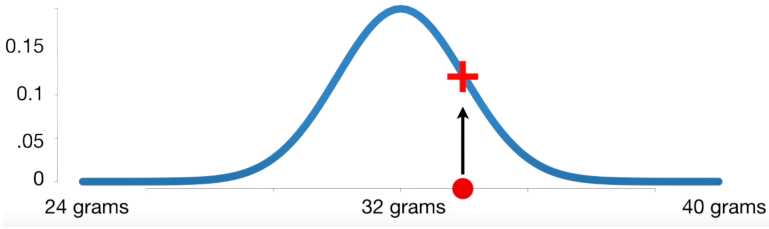
or

$$\mathbb{P}(X > 34|\mu = 32, \sigma = 2.5)$$



The area under the curve is described by the left hand side of the probabilistic expression $\mathbb{P}(X > 34|\mu = 32, \sigma = 2.5)$, i.e. $X > 34$. The right hand side, $\mu = 32, \sigma = 2.5$, describes the same distribution for both examples. This distribution is fixed.

Likelihood, on the other hand, deals with fixed data and variable distributions. Suppose that we have some data, i.e. we have an animal and weighed it (or more animals, and their weights). An animal weighs 34 grams. The likelihood of weighting a 34 gram animal is this point on the curve:



Projecting onto the vertical axis, that value is 0.12. Mathematically, we use the following notation:

$$L(\text{mean=32, standard deviation=2.5}|\text{animal weighs 34 grams}) = 0.12,$$

or

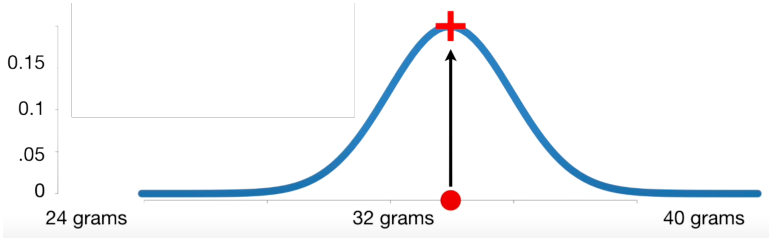
$$L(\mu = 32, \sigma = 2.5|\text{m=34 g}) = 0.12.$$

The likelihood of a distribution with mean 32 and standard deviation 2.5 given that we weighed a 34 gram animal is 0.12.

If we shifted the distribution such that

$$L(\text{mean=34, standard deviation=2.5}|\text{animal weighs 34 grams}),$$

the likelihood would be 0.21. So with likelihoods, the information on the right hand side (e.g. "animal weighs 34 grams") is fixed. We modify the shape and location of the distribution with the left hand side.



To summarize, probabilities are the areas under a fixed distribution. They answer questions such as what is the probability under conditions where the distribution is fixed:

$$\mathbb{P}(\text{data}|\text{distribution})$$

Likelihoods are the y-axis values for fixed data points with distributions that can be moved:

$$\mathbb{P}(\text{distribution}|\text{data}).$$

In this chapter, we "move" the distribution by adjusting its parameters by way of the maximum likelihood criteria.

4.2.7. Prior vs Posterior Distribution. In this section we will need two results from probability theory. The first is the *Bayes' theorem*:[4]

$$\mathbb{P}(\theta|X) = \frac{\mathbb{P}(X|\theta)\mathbb{P}(\theta)}{\mathbb{P}(X)}, \tag{4.1}$$

where $\mathbb{P}(X) \neq 0$. X is called the "evidence" or data. θ is sometimes called the parameters or the distribution. Often, the evidence is fixed and we write

$$\mathbb{P}(\theta|X) \propto \mathbb{P}(X|\theta)\mathbb{P}(\theta).$$

If the random events $\{X_j\}$ partition the sample space, we may write $\mathbb{P}(\theta)$ using the *law of total probability*:[5]

$$\mathbb{P}(X) = \sum_j \mathbb{P}(\theta|X_j)\mathbb{P}(X_j).$$

Then,

$$\boxed{\mathbb{P}(\theta|X) = \frac{\mathbb{P}(X|\theta)\mathbb{P}(\theta)}{\sum_j \mathbb{P}(\theta|X_j)\mathbb{P}(X_j)}.}$$

[4]The proof uses the definition of conditional probability. On one hand we have $\mathbb{P}(A|B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}$. On the other hand we have $\mathbb{P}(B|A) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(A)}$. Solving for $\mathbb{P}(A\cap B)$ in both equations and equating gives: $\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$. Rearranging gives $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.

[5]If the events $\{B_n\}$ form a partition of the probability space, i.e. $\cup_n B_n = \Omega$ and $B_i \cap B_j = \emptyset$ $(i \neq j)$, then

$$\mathbb{P}(A) = \sum_n \mathbb{P}(A \cap B_n) = \sum_n \mathbb{P}(A|B_n)\mathbb{P}(B_n).$$

This is another way to express the same result. The events can be of the type $\{X = x\}$, etc. For example:

$$\mathbb{P}(\theta = a|X = b) = \frac{\mathbb{P}(X = b|\theta = a)\mathbb{P}(\theta = a)}{\mathbb{P}(X = b)}.$$

They can also be intervals $\{a \leq X < b\}$ or infinitesimals, e.g.

$$\mathbb{P}(a \leq \theta < a + \mathrm{d}a|b \leq X < b + \mathrm{d}b) = \frac{\mathbb{P}(b \leq X < b + \mathrm{d}b|a \leq \theta < a + \mathrm{d}a)\mathbb{P}(a \leq \theta < a + \mathrm{d}a)}{\mathbb{P}(b \leq X < b + \mathrm{d}b)}, \tag{4.2}$$

which is the same as[6]

$$p_{\theta|b\leq X<b+\mathrm{d}b}(a)\mathrm{d}a = \frac{p_{X|a\leq\theta<a+\mathrm{d}a}(b)\mathrm{d}b\, p_\theta(a)\mathrm{d}a}{p_X(b)\mathrm{d}b}.$$

Canceling the $\mathrm{d}a$ and $\mathrm{d}b$'s,

$$p_{\theta|b\leq X<b+\mathrm{d}b}(a) = \frac{p_{X|a\leq\theta<a+\mathrm{d}a}(b)\, p_\theta(a)}{p_X(b)}.$$

Taking the limit $|\mathrm{d}b|, |\mathrm{d}a| \to 0$, we have for continuous random variables:

$$\boxed{p_{\theta|X=b}(a) = \frac{p_{X|\theta=a}(b)\, p_\theta(a)}{p_X(b)}.}$$

The posterior probability is the probability of the parameters θ given the data X. It is denoted as $p(\theta|X)$. Posterior, in this context, means after taking into account the relevant evidence (data) related to the particular case being examined. Posterior probability differs from the likelihood function, which is the probability of observing the data given some parameters (i.e. or a fixed distribution), $L(\theta|X) \equiv p(X|\theta)$.

Let's say our data is in the form of a dataset $X = \{X_i\}_{i=1}^N$ (independent rv's) where $X_i \cap X_j = \emptyset$ $(i \neq j)$. According to Eq. (4.1), the posterior distribution is

$$L \equiv p(\theta|X) \propto p(\theta)p(X|\theta) = p(\theta)\prod_{i=1}^{N} p(X_i|\theta),$$

where the equality follows from the statistical independence of the X_i's.

---

[6] This follows from the definition of conditional probability $p_{X|Y=y}(x) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$. For example, the left hand side

$$\mathbb{P}(a \leq \theta < a + \mathrm{d}a|b \leq X < b + \mathrm{d}b) = \frac{\mathbb{P}(a \leq \theta < a + \mathrm{d}a \cap b \leq X < b + \mathrm{d}b)}{\mathbb{P}(b \leq X < b + \mathrm{d}b)} = \frac{p_{\theta,X}(a,b)\mathrm{d}a\,\mathrm{d}b}{p_X(b)\mathrm{d}b} = p_{\theta|X=b}(a)\mathrm{d}a. \tag{4.3}$$

4.3. Method of Moments

Another method for deriving estimators is the method of moments. Recall the strong law (SLLN) of large numbers (Section 2.25.1):

Theorem 4.1. *Let $X_1, \ldots, X_n$ be iidrv with finite first absolute moment, i.e. $\mathbb{E}[|X_1|] < +\infty$. Then,*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}[X_1] \text{ almost surely as } n \to \infty.$$

Remark 4.2. "Almost Surely" means for all ω except for a set of measure 0, i.e.

$$\mathbb{P}\big(\lim_{n\to\infty}\big\{\omega\big|\frac{1}{n}\sum_{i=1}^{n} X_i(\omega) \nrightarrow \mathbb{E}(X_1)\big\}\big) = 0.$$

In other words, the set of outcomes for which the random variable $\frac{1}{n}\sum_{i=1}^{n} X_i$ does not converge to $\mathbb{E}[X_1]$ has probability 0 as $n \to \infty$.

A second theorem of probability theory (not proved here) that we will need is:

Theorem 4.3. *Let $Y_1, \ldots, Y_n$ be a sequence of rv's (not necessarily iidrv), such that $Y_n \to Y$ almost surely, and if h is a continuous function, then $h(Y_n) \to h(Y)$ almost surely as $n \to \infty$.*

Let $X_1, \ldots, X_n$ be iidrv whose PDF is the exponential probability

$$p_X(x|\theta) = \theta^{-1} e^{-x\theta^{-1}} \mathbf{1}_{\{x>0\}},$$

where $\mathbf{1}_{\{x>0\}}$ is the indicator function that equals 1 when $x > 0$ and 0 otherwise. Suppose we want to estimate $\theta = \mathbb{E}[X_1]$ from experimental data. We propose the following estimator for the first moment:

$$\hat{\theta}_n^{(1)} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

which is such that $\hat{\theta}_n^{(1)} \to \theta$ almost surely as $n \to \infty$. The SLLN guarantees that this estimator converges to $\mathbb{E}[X_1]$, which equals θ, the desired parameter. This gives us a possible estimator for θ.

On the other hand, $\mathbb{E}[X_1^2] = 2\theta^2$. According to the SLLN,

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \to \mathbb{E}[X_1^2] = 2\theta^2 \text{ almost surely as } n \to \infty.$$

Thus, if we divide by 2 and take the square root, we should get another estimate for θ. Let

$$Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2,$$

and take $h(x) = \sqrt{x/2}$. We know that the sequence of rv's Y_n converges to the rv $Y = \mathbb{E}[X_1^2] = 2\theta^2$ almost surely per the SLLN. According to the second theorem, $h(Y_n) \to h(Y) = \theta$ almost surely as $n \to \infty$. Thus,

$$\hat{\theta}^{(2)} = \sqrt{\frac{1}{2n}\sum_{i=1}^{n} X_i^2},$$

is another possible estimator for θ. So we have derived 2 estimators so far, $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$; you can also derive more. Which estimator you should use depends on your particular application. Factors that may influence your decision include bias, errors, rate of convergence, etc.

Based on the above theorems, a general recipe for the method of moments consists of:

(1) Calculate the first few moments of X using the known PDF of X. Since the PDF is a function of the parameters of the distribution, the result will also be in terms of those parameters.

(2) Estimate the moments using experimental data $X_1, \ldots, X_n$, according to the LLN.

(3) From (1), solve for the parameters of the distributions in terms of the moments, and express the moments in terms of estimated moments found in (2).

Let us work out an example to illustrate this. Let X be normally distributed with mean μ and variance σ^2. The two parameters of the distribution are μ and σ.

Step 1: Using the PDF of the normal distribution, one calculates the moments:

$$\mathbb{E}[X^n] = \int_{-\infty}^{\infty} x^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \mathrm{d}x.$$

The first few moments are found to be:

$$n = 1 \qquad \mathbb{E}[X^1] = \mu \tag{4.4}$$

$$n = 2 \qquad \mathbb{E}[X^2] = \mu^2 + \sigma^2 \tag{4.5}$$

$$n = 3 \qquad \mathbb{E}[X^3] = \mu^3 + 3\mu\sigma^2 \tag{4.6}$$

$$\vdots$$

The higher order moments are not needed because the above moments are already expressed in terms of both parameters μ and σ.

Step 2: Given a random sample $X_1, \ldots, X_n$ of X and its corresponding experimental measurement $\{X_i(\omega) = x_i\}$, the LLN enables us[7] to write:

$$n = 1 \qquad \mathbb{E}[X^1] \approx \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.7}$$

$$n = 2 \qquad \mathbb{E}[X^2] \approx \frac{1}{n}\sum_{i=1}^{n} x_i^2 \tag{4.8}$$

$$n = 3 \qquad \mathbb{E}[X^3] \approx \frac{1}{n}\sum_{i=1}^{n} x_i^3 \tag{4.9}$$

$$\vdots$$

Step 3: We solve the system of equations of Step 1 in terms of μ and σ, and express these in terms of the equations found in Step 2.

Adding a hat to μ, substituting $X_i(\omega) = x_i$ and omitting ω from the notation, Eqs. (4.4) and (4.7) give:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{4.10}$$

Next, we take Eq. (4.5), solve for σ and substitute Eq. (4.7) and (4.8):

$$\sigma^2 = \mathbb{E}[X^2] - \mu^2 \approx \frac{1}{n}\sum_i x_i^2 - \left(\frac{1}{n}\sum_i x_i\right)^2$$

[7] Recall that the LLN says that $\frac{1}{n}\sum_i X_i(\omega) \approx \mathbb{E}[X]$, where X is a random variable and $\{X_i\}_{i=1}^n$ is a random sample drawn according to the same distribution as X. $X_i(\omega)$ denotes the value of X_i after an experiment ω. Since functions of random variables are also random variables; in particular, the powers (moments) of X can be estimated from $\frac{1}{n}\sum_{i=1}^n [X_i(\omega)]^m \approx \mathbb{E}[X^m]$.

Next, we substitute $X_i(\omega) = x_i$, $\hat{\mu}(\omega) = (1/n)\sum_i X_i(\omega)$ and drop ω from the notation:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2\hat{\mu}X_i + \hat{\mu}^2) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2. \tag{4.11}$$

Here, $\hat{\mu}$ is calculated using Eq. (4.10) using the data $X_i(\omega) = x_i$. Our two "method of moments" estimators are given by Eq. (4.10) and (4.11).

As an exercise, you should work out the method of moments estimators for other distributions (Problem 105).

4.4. Problems

Problem 105. Use the method of moments to find parameter estimators for the mean and variance of other distributions. For example: use the method of moments to find estimate(s) of the θ parameter for the uniform distribution of a random variable X, with PDF $p_X(x|\theta) = \theta^{-1}\mathbf{1}_{[0,\theta]}(x)$, $\theta > 0$, where θ defines the upper bound of the support of $p_X(x|\theta)$, and $\mathbb{E}(X) = \theta/2$.

Problem 106. Let X be a rv and $X_1, \ldots, X_n$ ($X_i(\omega) \geq 0$) a random sample of X (iidrv). Denote $X_i(\omega) = x_i$ their respective values. The PDF for each of these rv's is:

$$p_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad i = 1, \ldots, n$$

where $\alpha > 0$ and $\beta > 0$ are parameters for the distribution and $\Gamma(\cdot)$ is the gamma function. Write down the likelihood function (joint PDF) for these n observations (the parameters α and β are the same for all n observations, since they are iidrv). Show that the maximum likelihood estimators for α and β are obtained by solving these two equations:

$$\hat{\beta} = \frac{\hat{\alpha}}{\frac{1}{n}\sum_{i=1}^{n} x_i}, \quad \text{and} \quad \log(\hat{\alpha}) - \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = \log\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - \frac{1}{n}\sum_{i=1}^{n}\log x_i.$$

Explain how you would calculate the values of $\hat{\alpha}$ and $\hat{\beta}$ in terms of the observations $x_1, x_2, \ldots, x_n$.

Solution. The likelihood function is

$$L \equiv p_{X_1,\ldots,X_n}(x_1, \ldots, x_n|\alpha, \beta) = \frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n}\prod_{i=1}^{n} x_i^{\alpha-1} e^{-\beta\sum_{i=1}^{n} x_i}$$

The log likelihood function is

$$l = \log L \equiv n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1)\Big(\sum_{i=1}^{n} \log x_i\Big) - \beta \sum_{i=1}^{n} x_i.$$

Differentiating this expression with respect to α and β we get

$$\frac{\partial l}{\partial \alpha} = n \log \beta - \frac{n}{\Gamma(\alpha)}\Gamma'(\alpha) + \sum_{i=1}^{n} \log x_i \Big|_{\alpha,\beta=\hat{\alpha},\hat{\beta}} = 0$$

$$\frac{\partial l}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i \Big|_{\alpha,\beta=\hat{\alpha},\hat{\beta}} = 0.$$

At the extremum ($\partial_\alpha l = 0$, $\partial_\beta l = 0$), we use the hat notation $\hat{\alpha}$ and $\hat{\beta}$. Viewing α and β as random variables (i.e. $\hat{\alpha} \equiv \hat{\alpha}(\omega)$ and $\hat{\beta} \equiv \hat{\beta}(\omega)$), substituting $X_i(\omega) = x_i$ and dropping ω from the notation:

$$\hat{\beta} = \frac{\hat{\alpha}}{\frac{1}{n}\sum_{i=1}^{n} X_i}, \quad \text{and} \quad \log(\hat{\alpha}) - \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = \log\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) - \frac{1}{n}\sum_{i=1}^{n} \log X_i.$$

We need $\hat{\alpha}$ to calculate $\hat{\beta}$, but in order to calculate $\hat{\alpha}$ you have a highly nonlinear equation in $\hat{\alpha}$. There is no way you can solve for α analytically. So given values for the rv's $X_1, \ldots, X_n$ you could use Newton Raphson (or similar method) to solve for $\hat{\alpha}$ numerically. ■

Problem 107. Suppose $X_1, \ldots, X_n$ are iidrv (random sample of X) with density $p_X(x|\theta) = \theta e^{-\theta x}$, $x > 0$, $\theta > 0$ and corresponding values $X_i(\omega) = x_i$. Write down the likelihood function for n observations. Find the maximum likelihood estimate for θ.

Solution. Likelihood function is

$$L = \theta e^{-\theta x_1} \cdot \theta e^{-\theta x_2} \ldots \theta e^{-\theta x_n} = \theta^n e^{-\theta \sum_{i=1}^{n} x_i}.$$

and using the log likelihood

$$\log L = n \log \theta - \theta \sum_{i=1}^{n} x_i$$

differentiating with respect to θ and setting derivative equal to zero we get

$$\frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0.$$

Using the hat notation $\hat{\theta}$ for this extremum, substituting $X_i(\omega) = x_i$ and dropping ω from the notation:

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} X_i}.$$

■

Problem 108. Let X be a random variable and $X_1, \ldots, X_n$ (iidrv) be a random sample of X. We denote the measurements as $X_i(\omega) = x_i$. Find the maximum likelihood estimate of the parameter θ for the density $p_X(x|\theta) = 1/\theta$, $0 \leq x \leq \theta$, based on the measurements $x_1, \ldots, x_n$.

Solution.

$$L(\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \left(\frac{1}{\theta}\right)^n.$$

This function will be maximized by choosing θ as small as possible, subject to the restriction $0 \leq x_i \leq \theta$, $i = 1, \ldots, n$. The smallest possible value of θ that satisfies these inequalities is clearly the largest value of the x_i. Thus, the MLE of θ is given by

$$\hat{\theta} = \max\{X_1, X_2, \ldots, X_n\},$$

where $X_i(\omega) = x_i$. ■

Problem 109. Let X be a random variable and $X_1, \ldots, X_n$ be a random sample of X. Denote their values as $X_i(\omega) = x_i$. If X follows a Lorentz distribution:

$$p_X(x|\theta) = \frac{1}{\pi} \frac{\epsilon}{(x-\theta)^2 + \epsilon^2}$$

with unknown median θ and known spread $\epsilon > 0$. Derive a maximum likelihood estimator for θ.

Solution. Differentiating the likelihood function gives

$$\frac{\partial p_{X_1,\ldots,X_n}(x_1, \ldots, x_n|\theta)}{\partial \theta} = \frac{2(x_1 - \theta)}{(x_1 - \theta)^2 + \epsilon^2} + \cdots + \frac{2(x_n - \theta)}{(x_n - \theta)^2 + \epsilon^2} = 0.$$

This is a nasty equation to solve for θ. We could do it by computer by implementing an iterative process for solving the nonlinear equations. Suppose that θ is a solution. Then the terms in the equation corresponding to data points x_i that are far from θ are close to zero. The terms in the equation corresponding to data points x_i that are closed to θ then each have magnitude about $(x_i - \theta)/\epsilon^2$. So the θ solution is, roughly speaking, a sample mean of part of the data, leaving out the more extreme values. While this estimator is more efficient than the sample median, the sample median begins to look attractive from the point of view of convenience. ■

Problem 110. Show that the following random variable

$$X = e^{\mu + \sigma Z}$$

where Z is a standard normal rv, i.e. $Z \sim N(0, 1)$, has the PDF:

$$p_X(x|\mu, \sigma) = \frac{e^{-(\log x - \mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma x}, \quad x > 0.$$

Find the MLE for the distribution parameters μ and σ, given a random sample $X_1, \ldots, X_n$ of X and their values $X_i(\omega) = x_i$. Each random variable X_i are iidrv with the above PDF for X.

Solution. We proceed as usual:

$$\begin{aligned}\mathbb{P}(X < x) =& \mathbb{P}(e^{\mu+\sigma Z} < x) = \mathbb{P}(\mu + \sigma Z < \log x) = \mathbb{P}\left(Z < \frac{\log x - \mu}{\sigma}\right) \\ =& \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\log x - \mu}{\sigma}} e^{-\xi^2/2} d\xi.\end{aligned}$$

which we recognize as being

$$\Phi\left(\frac{\log x - \mu}{\sigma}\right)$$

Differentiation with respect to x gives the PDF:

$$p(x) = \frac{e^{-(\log x - \mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma x}.$$

The likelihood function is

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{e^{-(\log x_i - \mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma x_i}.$$

The log likelihood is:

$$l(\mu, \sigma) = \sum_{i=1}^{n} \left\{ -\frac{1}{2}\log(2\pi) - \log(\sigma) - \log(x_i) - (\log x_i - \mu)^2/2\sigma^2 \right\}.$$

Differentiation with respect to σ and μ yields the following MLE:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \log X_i}{n} \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (\log X_i - \hat{\mu})^2}{n}.$$

■

Problem 111. Suppose that you make n measurements of the random variable X, knowing the X is Poisson distributed with Poisson parameter λ, and m measurements of another random variable Y, which is known to be exponentially distributed, i.e. its PDF is $p_Y(y) = a \cdot \exp(-ay)$, $a > 0$. Find estimators for the parameters λ and a in terms of the experimental data $X_1(\omega), \ldots, X_n(\omega)$ and $Y_1(\omega), \ldots, Y_m(\omega)$. (All these rv's can be assumed independent.)

Solution. Likelihood function:

$$L(\lambda, a) = \frac{e^{-\lambda n} \lambda^{x_1 + \cdots + x_n}}{x_1! x_2! \ldots x_n!} a^m e^{-a(y_1 + \cdots + y_m)}$$

Taking derivative of $\log L$ with respect to λ:

$$-n + (x_1 + \cdots + x_n)\frac{1}{\lambda} = 0$$

with respect to a:

$$m\frac{1}{a} - (y_1 + \cdots + y_m) = 0.$$

The first equation gives:

$$\hat{\lambda} = \frac{1}{n}\left(X_1 + \cdots + X_n\right).$$

The second gives:

$$\hat{a} = \frac{m}{(Y_1 + \cdots + Y_m)}.$$

Here $X_i(\omega) = x_i$ and $Y_i(\omega) = y_i$. ■

Problem 112. The maximum likelihood estimator (MLE) is the set of parameters denoted $\hat{\theta}$ for which the likelihood function is maximized (provided that one or more maxima exists):

$$\hat{\theta} \in \left\{ \arg\max_{\theta} L(\theta|x_1, \ldots, x_n) \right\}$$

Alternatively, we can also maximize the log-likelihood function

$$\hat{\theta} \in \left\{ \arg\max_{\theta} l(\theta|x_1, \ldots, x_n) \right\}$$

(a) A MLE estimate is the same regardless of whether we maximize the likelihood or the log-likelihood function. Why?

(b) In the case of a Gaussian distribution write down the log likelihood function.

(c) Carry out the maximization of the likelihood function (L) and the log-likelihood function (l) in the case of a Gaussian distribution. Do you get the same result? (Recall from (a) that you should get the same result.) Which procedure is simpler?

(d) In the case of Poisson distribution distributed random variables (the case of discrete random variables, but whose parameter(s) are continuous variables!), carry out the same MLE procedure and determine the maximum likelihood estimator for the mean and variance. For simplicity, assume that the random variables measured (let's call them $n_1, n_2, \ldots, n_N$) are iidrv. This exercise teaches how to compute "sample mean" and "sample variance" when performing repeated measurements of a Poisson-distributed random

variable. Does the ML prescription require you to compute an arithmetic mean?

Solution. (a) Because log is a monotonically increasing function. Consequently, a maximum of the likelihood function is also a maximum of the log-likelihood function.

(b) The joint Gaussian for iidrv is

$$L(\overline{X}, \sigma^2 | x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\chi^2/2}, \qquad \text{where } \chi^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{X})^2}{\sigma^2}$$

taking the log we get

$$l = -(n/2)\log 2\pi - n\log\sigma - \chi^2/2$$

Here maximization of L corresponds to the minimization of χ^2 because of the negative sign in front of χ^2.

(c) We have done L in class already. The case of l gives:

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma}\chi^2 = 0$$

which yields the same result as for L:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

where $X_i(\omega) = x_i$.

And likewise, the partial with respect to $\overline{X}$ yields the arithmetic mean, as can be seen by inspection. In this case, using the log likelihood is the simpler method to use.

(d) RVs are iidrv, so the joint density is a product of individual Poisson's. Denoting their mean as θ (in class, we used $\overline{n}$). ML method applied to the log likelihood:

$$\frac{\partial}{\partial\theta}\log\prod_{i=1}^{N}\left(e^{-\theta}\frac{\theta^{n_i}}{n_i!}\right) = \frac{\partial}{\partial\theta}\sum_{i=1}^{N}\left[-\theta + n_i\log\theta - \log n_i!\right] = \sum_{i=1}^{N}\left(-1 + \frac{n_i}{\theta}\right)$$

where in this expression n_i is shorthand notation for the value $n_i(\omega)$. We set this derivative equal to zero. This yields the arithmetic mean:

$$\hat{\theta}(\omega) = \frac{1}{N}\sum_{i=1}^{N} n_i(\omega).$$

(And for Poisson distributions the variance equals the mean, so there is nothing else we need to calculate.) ■

Problem 113. Let X be distributed according to the distribution: $\mathbb{P}(X = 2) = \theta$, $\mathbb{P}(X = 3) = 2\theta$ and $\mathbb{P}(X = 1) = 1 - 3\theta$. (The state space is discrete:

$X \in \{1, 2, 3\}$.) This distribution has a single parameter θ.

(a) What is the range of allowed values for θ? Find the mean and variance of X (in terms of θ).

(b) Suppose you want to estimate the parameter θ using n random samples of X. Let n_1, n_2 and n_3 be the number of times X equals 1, 2 or 3, respectively, within this random sample. What is the probability law for n_1?

(c) Find the maximum likelihood estimator $\hat{\theta}$ for θ. Calculate the mean and variance of $\hat{\theta}$. Is the estimator biased?

Solution. (a) θ can take values in the interval $[0, \frac{1}{3}]$, so that these probabilities must remain between 0 and 1. The mean is

$$\mathbb{E}[X] = 1 - 3\theta + 2\theta + 3 \times 2\theta = 1 + 5\theta.$$

variance is

$$var(X) = 1 - 3\theta + 4\theta + 9 \times 2\theta - (1 + 5\theta)^2 = \theta(9 - 25\theta).$$

(b) n_1 follows a binomial law, with parameters n and $p = 1 - 3\theta$.

(c) The likelihood function is

$$L(\vec{X}|\theta) = \prod_{i=1}^{n} \mathbb{P}(X = x_i) = (1 - 3\theta)^{n_1} \times \theta^{n_2} (2\theta)^{n_3}$$

Minimizing the log of L with respect to θ,

$$\frac{\partial}{\partial \theta} \log L = -\frac{3n_1}{1 - 3\theta} + \frac{n_2}{\theta} + \frac{2n_3}{2\theta} = \frac{-3n_1}{1 - 3\theta} + \frac{n - n_1}{\theta} = 0,$$

where n_i stands for the value $n_i(\omega)$. Thus, we conclude that

$$\hat{\theta}(\omega) = \frac{n - n_1(\omega)}{3n}.$$

Its mean is

$$\mathbb{E}[\hat{\theta}] = \frac{1}{3} - \frac{1}{3n}(1 - 3\theta)n = \theta$$

and the variance is

$$var(\hat{\theta}) = \frac{1}{9n^2} n(1 - 3\theta)3\theta = \frac{\theta(1 - 3\theta)}{3n}.$$

Thus, $\hat{\theta}$ is an unbiased estimator. Its variance decreases with n. ■

Problem 114. Tesla Motors has a plant in Nevada that manufactures car batteries. If a battery has length greater than a, it will not fit into the car, and this is a big problem because the cost of fixing such defective batteries

is high. Tesla has been plagued recently regarding problems in its assembly line, related to poor quality control. You were tasked to visit the plan in your capacity as inspector, and do a statistical analysis of the assembly line. The first question that comes to mind is what is the probability that a battery (chosen at random in the assembly line) has length greater than a, i.e. $\mathbb{P}(X \geq a)$. (Let's assume a 1D problem.) Show that a times this probability can never be larger than the mean of X (i.e., $\mathbb{E}X$). What is the significance of this statement?

Solution.

$$\mathbb{E}[X] = \int_0^\infty xp(x)\mathrm{d}x \geq \int_a^\infty xp(x)\mathrm{d}x \geq \int_a^\infty ap(x)\mathrm{d}x = a\mathbb{P}(X \geq a)$$

■

Problem 115. Suppose that you make n measurements of the random variable X, knowing the X is Poisson distributed with Poisson parameter λ, and m measurements of another random variable Y, which is known to be exponentially distributed, i.e. its PDF is $p(y) = a \cdot \exp(-ay)$, $a > 0$. Find estimators for the parameters λ and a in terms of the experimental data $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$. (All these rv's can be assumed independent.)

Solution. Likelihood function:

$$L(\lambda, a) = \frac{e^{-\lambda n}\lambda^{x_1+\cdots+x_n}}{x_1!x_2!\ldots x_n!}a^m e^{-a(y_1+\cdots+y_m)}$$

Taking derivative of $\log L$ with respect to λ:

$$-n + (x_1 + \cdots + x_n)\frac{1}{\lambda} = 0$$

with respect to a:

$$m\frac{1}{a} - (y_1 + \cdots + y_m) = 0.$$

The first equation gives:

$$\hat{\lambda} = \frac{1}{n}\left(X_1 + \cdots + X_n\right).$$

The second gives:

$$\hat{a} = \frac{m}{(Y_1 + \cdots + Y_m)}.$$

Here, $X_i(\omega) = x_i$ and $Y_i(\omega) = y_i$. ■

Problem 116. What is the meaning of "bias" in the context of a statistical estimator? How does one determine the amount of bias in an estimator? What desirable properties should the estimator possess? Give an example of a biased estimator and explain what makes it biased.

Solution. Biased, since the coefficient is $1/n$ instead of $1/(n-1)$. ■

Problem 117. The Richter magnitude of an earthquake is determined from the logarithm of the amplitude of waves recorded by seismographs (adjustments are included to compensate for the variation in the distance between the various seismographs and the epicenter of the earthquake). The formula is:

$$M_L = \log_{10}(A/A_0(\delta))$$

where A is the maximum excursion of the Wood-Anderson seismograph, the empirical function A_0 depends only on the epicentral distance of the station, δ. Both A and δ are prone to measurement error. Find the uncertainty in M_L due to errors in A and δ.

Solution. Propagation of error formula is:

$$\sigma_{M_L}^2 = \left|\frac{\partial M_L}{\partial A}\right|^2 \sigma_A^2 + \left|\frac{\partial M_L}{\partial \delta}\right|^2 \sigma_\delta^2,$$

where the derivatives are:

$$\frac{\partial M_L}{\partial A} = \frac{1}{\log(10)}\frac{A_0(\delta)}{A}\cdot\frac{1}{A_0(\delta)} = \frac{1}{A\cdot\log(10)}$$

$$\frac{\partial M_L}{\partial \delta} = -\frac{1}{\log(10)}\frac{A_0(\delta)}{A}\cdot\frac{1}{[A_0(\delta)]^2}\frac{\partial A_0(\delta)}{\partial \delta} = -\frac{\partial A_0(\delta)/\partial\delta}{A\cdot A_0(\delta)\cdot\log(10)}$$

■

Problem 118. The fluorescence decay of a fluorophore molecule is modeled by:

$$I(t) = \tau^{-1}e^{-t/\tau},$$

where τ is the fluorescence decay lifetime, or equivalently, τ^{-1} is the decay rate. (You can think of this model as giving the probability of a molecule chosen randomly from an ensemble to be found in the excited state.) To some extent, the decay lifetime is indicative of the immediate chemical environment of the molecule. Consequently, you decide to investigate its potential use as a chemical sensor by measuring lifetime in the presence of different types of solvents and solutes. Your experiments returns values of t (i.e., $t_1, t_2, \ldots, t_n$) at which the fluorescence intensity drops to $1/e$ of its initial value; i.e., the point where $t = \tau$. Such measurements of $t_1, t_2, \ldots, t_n$ are subject to considerable errors. To reduce variability you repeat the measurement many times and derive a suitable average.

(a) Explain how you would estimate an average lifetime in terms of the experimentally measured lifetimes $t_1, t_2, \ldots, t_n$ which maximizes the chances of observing this experimentally measured data.

(b) If the scientific journal you are trying to publish your results in asks you to report average decay rates instead of lifetimes, how would you estimate average decay rates from the data?

Solution. (a) For simplicity, we shall use the same notation t_i for the random variable as well as its value $t_i(\omega)$. Since $t_1, t_2, \ldots, t_n$ are iidrv, the likelihood function is a product:

$$L(t_1, t_2, \ldots, t_n) = \tau^{-1} e^{-t_1/\tau} \ldots \tau^{-1} e^{-t_n/\tau} = \tau^{-n} \exp\left(-\tau^{-1} \sum t_i\right).$$

The log likelihood is

$$\log L = -n \log \tau - \tau^{-1} \sum t_i.$$

Differentiating with respect to τ and setting equal to zero:

$$-\frac{n}{\tau} + \frac{1}{\tau^2} \sum t_i = 0.$$

Solving for τ we get the MLE which is the arithmetic average of the measured lifetimes:

$$\boxed{\hat{\tau} = \frac{1}{n} \sum t_i.}$$

(b) Decay rate is inverse of lifetime:

$$\hat{\tau}^{-1} = \frac{n}{\sum t_i}.$$

■

Problem 119. Suppose you record the number of cars crossing some intersection going northbound at regular intervals of 5 minutes after midnight, i.e. when there is very little traffic, so you can assume that the events of cars crossing the intersection are Poisson distributed. Let n_i the number of cars recorded in the ith time interval. Your job description requires you to produce a daily report to the city by 10:00 am which includes the data recorded along with some basic statistical analysis. In particular, the city wants you to state on the cover page the mean number of vehicles and the variance in a 5 minute time interval. However, because the numbers are Poisson distributed you are unsure if you are allowed to use the formula for sample mean, since you may recall it was derived using the assumption of Gaussian statistics — not Poisson. Derive the correct formula for the sample mean and variance in the case of Poisson statistics.

Solution. See Problem 112. ■

Problem 120. You measure a signal, y, that is the sum of a function f contaminated by additive noise ξ. In discrete form:

$$y_i = f(x_i, \beta) + \xi_i$$

where β is a parameter for the function f that we wish to obtain. (The form of the function f is known, but we must determine this parameter from the data.) The density of the noise ξ_i is known, $p(\xi)$. Estimate the function

$f(x, \beta_0)$ from the set of functions $f(x, \beta)$ (i.e. determine the value of β), using the data obtained by measurements y of the function f corrupted by noise ξ.

Solution. The data is given by the pairs:

$$(x_1, y_1), \ldots, (x_n, y_n)$$

We estimate β_0 using the ML method by maximizing the log likelihood:

$$l(\beta) = \sum_{i=1}^{n} \log p(y_i - f(x_i, \beta))$$

Where $p(\xi)$ is a known function and $\xi = y - f(x, \beta)$. To make further progress we must know the form of $p(\xi)$. If p is Gaussian,

$$p(\xi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\xi^2}{2\sigma^2}\right)$$

(zero mean, known variance), we then obtain the least squares method:

$$l(\beta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2 - n \log(\sqrt{2\pi}\sigma)$$

Maximizing $l(\beta)$ over the parameters β is the same as minimizing the least squares functional:

$$\sum_{i=1}^{n} (y_i - f(x_i, \beta))^2$$

The particular value of β that minimizes this functional is called β_0. For other distributions, $p(\xi)$, a different functional may be obtained. ■

Chapter 5

Data Fitting

5.1. Linear Least Squares

An important aspect of experimental research is the measurement of physical quantities. From these physical quantities we seek to confirm or disprove certain hypotheses. This could be, for example, verification that a theory holds in a certain regime. We then need to "fit" the data to an equation and determine the unknown coefficients in the equation. This is the topic of data fitting.

5.1.1. Least Squares Method. An old and trusted method to data fitting is the method of least squares. If the distance between the experimental data and the model is normally distributed with mean 0 and finite variance, the joint pdf describing the measurement of the data points $\{y_1, \ldots, y_n\}$ is such that the application of the principle of maximum likelihood estimation (MLE) for the model parameters yields the method of least squares.

In contrast, if the measurements are not normally distributed about the model, MLE does not yield least squares. (Exercise: can you demonstrate this?) In this course, we will limit our discussion to data points which are normally distributed about the model. Linear least squares is a special case of least squares when the model is linear in the fitting parameters. Non-linear least squares problems are more complicated and generally cannot be solved analytically. Later in the course, we will solve non-linear least squares problems using computer-based methods.

5.1.2. Straight line. Let us begin with the method of linear least squares. And for simplicity we shall consider the problem of fitting data to a straight

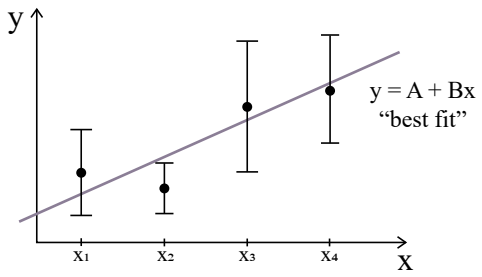


Figure 5.1. Straight line fit.

line model.[1] Suppose we have the data shown in Fig. 5.1 and we would like to fit these points to a straight line

$$y(x) = A + Bx.$$

The measured data is the set of pairs

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n).$$

If we assume that

- The uncertainty in $\{x_i\}$ is negligible.
- Each y_i is Gaussian-distributed with the same width σ_y.
- All measurements $\{y_i\}$ are statistically independent.

The hope is that the model (if correct) gives the true value of y:

$$y(x_i) \equiv (\text{true value of } y_i) = A + Bx_i.$$

From the above assumptions, it follows that the probability of a single measurement y_i is

$$p(y_i) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left[-\frac{(y_i - A - Bx_i)^2}{2\sigma_y^2}\right].$$

Because of *statistical independence* the joint probability density, $L(A, B) \equiv p(y_1, \ldots, y_n)$, of obtaining a complete set of measurements $y_1, \ldots, y_n$ is the product of probabilities of individual measurements:

$$\underbrace{p(y_1) \cdot p(y_2) \cdot \cdots \cdot p(y_n)}_{\text{by independence}} = \underbrace{\frac{1}{(2\pi)^{n/2}\sigma_y^n} \exp\left(-\sum_{i=1}^{n} \frac{(y_i - A - Bx_i)^2}{2\sigma_y^2}\right)}_{\text{by assumption of independent } \textit{Gaussians}}.$$

[1] Linear least squares and the straight line model are two different things. Fitting a polynomial is still a linear least squares problem because linear means linear in the model parameters. It is not a statement about the degree of the polynomial being fitted.

We shall rewrite this expression in slightly more convenient form:

$$\boxed{L(A,B|y_1,\ldots,y_n) \equiv p(y_1,\ldots,y_n) = \frac{1}{(2\pi)^{n/2}\sigma_y^n} e^{-\chi^2/2},}$$

where

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - A - Bx_i)^2}{\sigma_y^2} = \sum_{i=1}^{n} \tilde{R}_i^2.$$

Here, $\tilde{R}_i \equiv (y_i - y(x_i))/\sigma_i \sim \mathcal{N}(0,1)$.

According to the **principle of maximum likelihood**, the best estimates for the unknown constants A and B are those for which $p(y_1,\ldots,y_n)$ is maximized. Or equivalently, for which χ^2 is a minimum. This strategy is the basis for the method of **least squares fitting**.[2]

$$\frac{\partial \chi^2}{\partial A} = -\frac{2}{\sigma_y^2}\sum_{i=1}^{n}(y_i - A - Bx_i) = 0$$

$$\frac{\partial \chi^2}{\partial B} = -\frac{2}{\sigma_y^2}\sum_{i=1}^{n} x_i(y_i - A - Bx_i) = 0$$

which we rewrite as

$$An + B\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$A\sum_{i=1}^{n} x_i + B\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

These 2 equations and 2 unknowns are easily solved (see Section 12.1) to yield:

$$A = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{\Delta}, \tag{5.1}$$

$$B = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\Delta}, \tag{5.2}$$

where

$$\Delta = n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2.$$

The resulting line $y = A + Bx$ is called the least squares fit to the data, or the line of regression of y on x.

[2]The method is called least squares because it involves minimizing chi square. This is a consequence of our assumption that the data is normally distributed about the model. Because of the particular form of L obtained, maximizing L is equivalent to minimizing χ^2.

However, A and B are derived from experimental data. Thus, there is uncertainty associated with A and B. Another question of interest is how do we obtain an estimate for σ_y based on experimental data. σ_y is the parameter of the (assumed) Gaussian distribution that describes the fluctuations of the experimental data about the model.

5.1.3. Estimating σ_y for a Straight Line Model. The deviations $y_i - A - Bx_i$ are normally distributed, all with mean 0 and width σ_y. We can derive an expression for σ_y by viewing L as a function of σ_y and use maximum likelihood. Thus, we set:

$$\frac{\partial}{\partial \sigma_y}\left(\frac{1}{\sigma_y^n} e^{-\chi^2/2}\right) = 0, \qquad \text{where} \qquad \chi^2 = \sum_{i=1}^{n} \frac{(y_i - A - Bx_i)^2}{\sigma_y^2}.$$

Carrying out the differentiation we get

$$\frac{-n}{\sigma_y^{n+1}} e^{-\chi^2/2} + \frac{1}{\sigma_y^n} e^{-\chi^2/2} (\frac{-1}{2}) \left(\frac{-2}{\sigma_y^3}\right) \sum_{i=1}^{n} (y_i - A - Bx_i)^2 = 0$$

from which we conclude that $\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - A - Bx_i)^2$. In practice we take the prefactor to be $1/(n-2)$ instead of $1/n$ because A and B are computed from the data and this results in a bias of the estimator σ_y^2 (verify this!). A prefactor of $1/(n-2)$ corrects this bias. This "fix" makes sense if we imagine fitting a straight line to only $n = 2$ data points. The fit would always be perfect ($\sigma_y = 0$). The prefactor $1/(n-2)$ ensures that we get a division by zero: $\sigma_y = 0/0$ (undefined), indicating that such a situation should be avoided.

$$\boxed{\sigma_{y,n-2}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - A - Bx_i)^2.}$$

5.1.4. Estimating the Magnitude of σ_A and σ_B. For simplicity let us assume that the errors in $\{y_i\}$ are all identical, i.e. $\sigma_{y_i} = \sigma_y$. Here we cannot use maximum likelihood because $L(A, B, \sigma_y)$ does not depend on σ_A or σ_B. However, since A and B are well-defined functions of $y_1, y_2, \ldots, y_n$, we can find σ_A and σ_B by simple error propagation:[3]

$$A = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{\Delta}, \qquad \Delta = n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2$$

[3]It is worth reminding ourselves of our assumption that the error in $\{x_i\}$ is negligible. If this error is not negligible, we need to modify our strategy.

$$\begin{aligned}
\sigma_A^2 =& \sum_{j=1} \Big| \frac{\partial A}{\partial y_j} \Big|^2 \sigma_{y_j}^2 = \sum_j \left| \frac{\sum\limits_{i=1}^{n} x_i^2 - x_j \sum\limits_{i=1}^{n} x_i}{\Delta} \right|^2 \sigma_y^2 \\
=& \sum_j \Big[(\sum_{i=1}^{n} x_i^2)^2 - 2 \sum_{i=1}^{n} x_i^2 (x_j) \cdot \sum_{i=1}^{n} x_i + x_j^2 (\sum_{i=1}^{n} x_i)^2 \Big] \sigma_y^2 / \Delta^2 \\
=& \Big[n (\sum_{i=1}^{n} x_i^2)^2 - 2 \sum_{i=1}^{n} x_i^2 (\sum_{i=1}^{n} x_i)^2 + (\sum_{i=1}^{n} x_j^2)(\sum_{i=1}^{n} x_i)^2 \Big] \sigma_y^2 / \Delta^2 \\
=& \Big[n (\sum_{i=1}^{n} x_i^2)^2 - \sum_{i=1}^{n} x_i^2 (\sum_{i=1}^{n} x_i)^2 \Big] \sigma_y^2 / \Delta^2 \\
=& \Big\{ \sum_{i=1}^{n} x_i^2 \underbrace{\Big[n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2 \Big]}_{\Delta} \Big\} \sigma_y^2 / \Delta^2
\end{aligned}$$

from which we get

$$\boxed{\sigma_A = \sigma_y \sqrt{\frac{\sum\limits_{i=1}^{n} x_i^2}{\Delta}}.}$$

We proceed the same manner for σ_B.

Given:

$$B = \frac{n \sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{\Delta}, \; \Delta = n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2$$

we apply the method of error propagation to derive the error in B:

$$\begin{aligned}
\sigma_B^2 =& \sum_{j=1}^{n} \left| \frac{\partial B}{\partial y_j} \right|^2 \sigma_y^2 = \sum_{j=1}^{n} \left| \frac{n x_j - \sum\limits_{i=1}^{n} x_i}{\Delta} \right|^2 \sigma_y^2 \\
=& \sigma_y^2 \sum_{j=1}^{n} \frac{n^2 x_j^2 - 2 n x_j \sum\limits_{i=1}^{n} x_i + (\sum\limits_{i=1}^{n} x_i)^2}{\Delta^2} \\
=& \underbrace{n \Big[n \sum_{j=1}^{n} x_j^2 - (\sum_{i=1}^{n} x_i)^2 \Big]}_{\Delta} \frac{\sigma_y^2}{\Delta^2} = \frac{n \sigma_y^2}{\Delta}
\end{aligned}$$

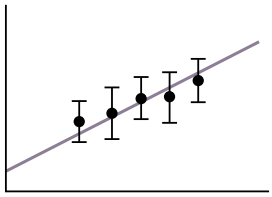


Figure 5.2. Straight line model and the factors affecting the quality of the fit. To get more precise slope and intercept, it is desirable to have small error bars near the end points of the data set.

from which we conclude that

$$\boxed{\sigma_B = \sigma_y \sqrt{\frac{n}{\Delta}}.}$$

5.1.5. Factors that Affect a Straight Line Fit. Suppose that we fit data to the straight line:

$$y(x_i|A, B) = A + Bx_i$$

A number of factors will affect the results (slope and intercept).

- **Intercept.** To increase the precision in the intercept, we must reduce the error bars on points that are close to the y-axis.
- **Slope.** To increase the precision of the slope, we must reduce the error bars on the points located at extrema of the data set (first and last points).

This is illustrated in Figure 5.2.

5.1.5.1. *Linear Least Squares: Summary of Assumptions.* It is important to remember the assumptions behind the method of linear least squares. First of all, we assumed Gaussian-distributed errors. By error we mean deviations of the data from the model, $y_i - y(x_i)$. This is reflected in the use of Gaussian PDFs that describe the probability of a measurement y_i being found near the model $y(x_i)$:

$$\mathbb{P}(y_i \leq Y_i \leq y_i + \mathrm{d}y_i) = p_{Y_i}(y_i)\mathrm{d}y_i = \frac{1}{\sqrt{2\pi\sigma_{y_i}^2}} e^{-(y_i - y(x_i))^2/2\sigma_{y_i}^2} \mathrm{d}y_i.$$

i.e., each Y_i obeys a normal law with mean $y(x_i)$ and variance $\sigma_{y_i}^2$, or $Y_i \sim \mathcal{N}(y(x_i), \sigma_{y_i}^2)$. We have also assumed that the measured data points $\{y_1, y_2, \ldots, y_n\}$ are statistically independent of each other. This assumption of statistical independence allowed us to write the likelihood function as a product of Gaussians:

$$p_{Y_1}(y_1) \cdot p_{Y_2}(y_2) \cdot \cdots \cdot p_{Y_n}(y_n) = \frac{1}{(2\pi)^{n/2} \prod_{i=1}^{n} \sigma_{y_i}} e^{-\chi^2/2}, \quad \chi^2 = \sum_{i=1}^{n} \tilde{R}_i^2.$$

Finally, the last assumption is our choice of the method of MLE as a way to derive equations for the fitting parameter. MLE involves choosing the fitting parameters that lead to the maximum likelihood of observing the experimental data $\{(x_i, y_i)\}$.

5.1.6. Linear Least Squares: Geometric Interpretation of A, B. For the straight line model,

$$y(x_i|A, B) = A + Bx_i,$$

we obtained the result for the linear least squares method as:

$$A = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{\Delta},$$
$$B = \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\Delta},$$
$$\Delta = n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2.$$

This result is rather difficult to memorize. I will show you how you can remember it using a simple geometric argument. We note that A and B can be expressed in terms of sample averages. In the equation for A, we divide the numerator and denominator by n^2, and write:

$$\boxed{A = \frac{\langle x^2\rangle\langle y\rangle - \langle x\rangle\langle xy\rangle}{\langle x^2\rangle - \langle x\rangle^2},}$$

where the angle brackets denote *sample means*, i.e., $\langle x\rangle = \frac{1}{n}\sum_{i=1}^n x_i$, $\langle x^2\rangle = \frac{1}{n}\sum_{i=1}^n x_i^2$ and $\langle xy\rangle = \frac{1}{n}\sum_{i=1}^n x_i y_i$. We note that $\langle xy\rangle$ is simply an inner product of two vectors, $\vec{x} = (x_1, \ldots, x_n)$ and $\vec{y} = (y_1, \ldots, y_n)$, sometimes written as $\langle \vec{x}, \vec{y}\rangle$ in a linear algebra course.

In the expression for B, if we divide the numerator and denominator by n^2, we get:

$$\boxed{B = \frac{\langle xy\rangle - \langle x\rangle\langle y\rangle}{\langle x^2\rangle - \langle x\rangle^2} = \frac{cov(x, y)}{var(x)}.}$$

The factor $\frac{cov(x,y)}{var(x)}$ represents an orthogonal projection of the vector $\vec{y}$ onto $\vec{x}$. In linear algebra such orthogonal projections are accomplished with the use of projection operators:

$$P_{\vec{x}}(\vec{y}) = \frac{\langle \vec{y}, \vec{x}\rangle}{\langle \vec{x}, \vec{x}\rangle}\vec{x} = \frac{\langle \vec{y}, \vec{x}\rangle}{\|\vec{x}\|^2}\vec{x} = \langle \vec{y}, \hat{e}_x\rangle \hat{e}_x = \|\vec{y}\| \cos\theta\, \hat{e}_x, \quad \hat{e}_x = \frac{\vec{x}}{\|\vec{x}\|},$$

where θ is the angle between the vectors $\vec{y}$ and $\vec{x}$.

Furthermore,

$$\langle y\rangle - B\langle x\rangle = \langle y\rangle \left[\frac{\langle x^2\rangle - \cancel{\langle x\rangle^2}}{\langle x^2\rangle - \langle x\rangle^2}\right] - \langle x\rangle \left[\frac{\langle xy\rangle - \cancel{\langle x\rangle\langle y\rangle}}{\langle x^2\rangle - \langle x\rangle^2}\right] = A.$$

Thus, the intercept is:

$$\boxed{A = \langle y\rangle - B\langle x\rangle = \langle y\rangle - \langle x\rangle \frac{cov(x,y)}{var(x)}.}$$

Now let's look at our model again, $y = A + Bx$, and view x and y as random variables and A and B as constants. Taking its average, $\langle y\rangle = A + B\langle x\rangle$, yields the relationship between A and B, $A = \langle y\rangle - B\langle x\rangle$. The equation $B = \frac{cov(x,y)}{var(x)}$ is obtained from $y = A + Bx$ by simply "projecting" it onto x using the projection operator $P_{\vec{x}}(\cdot) = \frac{cov(x,\cdot)}{var(x)}$, i.e.

$$P_{\vec{x}}(\vec{y}) \equiv \frac{cov(x,y)}{var(x)} = \frac{cov(x, A + Bx)}{var(x)} = B\frac{cov(x,x)}{var(x)} = B.$$

5.1.7. Maximizing the Likelihood Function. The MLE method requires us to maximize the likelihood function $L(\boldsymbol{\theta}|y_1, y_2, \ldots, y_n)$ with respect to the fitting parameters $\boldsymbol{\theta} = (A, B, C, \ldots)^T$. There are at least 3 ways to do this:

- Maximize L by inspection. It is sometimes possible to find the maximum by inspection. For example: maximize functions of the form $(1/2 - \eta)^2$, $(1/2 + \eta)$ and $(1/2 - \eta)(1/2 + \eta)$ with respect to η, for $-1/2 \leq \eta \leq 1/2$. This can be done by inspection: in the first case, $\eta = -1/2$; in the second case, $\eta = 1/2$ and in the third case, $\eta = 0$.
- You can maximize L using calculus. In one variable the necessary condition for an extremum is $\mathrm{d}L/\mathrm{d}A = 0$ whereas the second derivative test for a maximum is $\mathrm{d}^2L/\mathrm{d}A^2 < 0$. For example, in the case of two variables $\boldsymbol{\theta} = (A, B)^T$, the necessary condition for an extremum is $L \equiv L(A, B)$ is $\mathrm{d}L = (\partial_A L)\mathrm{d}A + (\partial_B L)\mathrm{d}B = 0$, which implies that $\partial_A L = \partial_B L = 0$. This gives 2 equations and two unknowns, allowing us to solve for A and B. Checking for a maximum requires a second (or higher-order) derivative test.
- You can maximize the log of L, $l = \log(L)$, also known as the "log-likelihood" function. Since log is a monotonic function, maximizing L is the same as maximizing $\log(L)$. The advantage of working with the log of L is because log converts products into sums, i.e. $\log(AB) = \log(A) + \log(B)$.

5.1.8. Weighted Average. Consider the experiment where we measure a random variable X using different methods. For example, X could be the

weight of an object and the weight can be measured using different types of balances, where each balance has its own uncertainty. Let $X_1, \ldots, X_n$ be a random sample of X. The measurements are denoted by lowercase variables $X_i(\omega) = x_i$. These rv's are independent but not necessarily identically distributed. Thus, we have the measurements $x_1, x_2, \ldots, x_n$, each with uncertainty $\sigma_1, \sigma_2, \ldots, \sigma_n$. What value should we report for X, in terms of the data $x_1, \ldots, x_n$ given that the uncertainties are different? Naturally, you may expect that the measurements with smaller uncertainty should carry more weight. It does not make sense to use the sample mean where all readouts are weighted equally, since some of those readouts may carry very large error bars.

Let us assume that $X_1, \ldots, X_n$ are Gaussian-distributed[4] iidrv with the same mean (μ) but different variances σ_i^2. We will then find the ML estimator for the mean, $\hat{\mu}$, and an estimator for the *variance of the mean*, i.e. σ_μ^2, where $\mu = \hat{\mu}$. In other words, assume that the σ_i^2 are given to you (known values). The problem consists of finding σ_μ^2 in terms of the known values x_i and σ_i^2.

The likelihood function is:

$$L(\mu, \sigma | x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-(x_i-\mu)^2/2\sigma_i^2}.$$

The log-likelihood is:[5]

$$\log L = -\frac{1}{2} n \log(2\pi) - \sum_{i=1}^{n} \log \sigma_i - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma_i^2}.$$

Extrema are found from:

$$\frac{\partial(\log L)}{\partial \mu} = \sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma_i^2} = \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} - \mu \sum_{i=1}^{n} \frac{1}{\sigma_i^2} = 0.$$

Substituting $X_i(\omega) = x_i$ and dropping the ω notation, we get the following estimator for the mean:

$$\boxed{\hat{\mu} = \frac{\sum_{i=1}^{n} \frac{X_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}.} \qquad \text{(5.3)} \qquad \text{"weighted mean"}$$

This estimator should be viewed as a random variable, i.e. $\hat{\mu} \equiv \hat{\mu}(\omega)$ is a function of the $X_i(\omega)$'s.

Note: if σ_i are all identical (i.e. $\sigma_i = \sigma$ for all i) then this reduces to the simple *arithmetic average*, $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$. We can get a partial check that

[4]The assumption of Gaussian distribution is made in order to simplify the math.

[5]Notice how L is a product of functions, whereas $l = \log L$ is a summation. Thus, working with the log-likelihood converts products into sums, which are easier to handle.

this is a maximum of L:

$$\frac{\partial^2(\log L)}{\partial \mu^2} = -\sum_{i=1}^{n} \frac{1}{\sigma_i^2} < 0.$$

(The "proof" that this is a maximum also requires additional considerations.)

We can check for bias:

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\frac{\sum_{i=1}^n \frac{X_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2}} = \frac{\sum_{i=1}^n \frac{\mathbb{E}[X_i]}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2}} = \frac{\sum_{i=1}^n \frac{\mu}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2}} = \mu. \quad \text{(unbiased)}$$

Given a random sample, we substitute $X_i(\omega) = x_i$ into Eq. (5.3), and obtain the variance of the mean by error propagation

$$\sigma_{\hat{\mu}}^2 = \sum_{i=1}^{n} \sigma_i^2 \left(\frac{\partial \mu}{\partial x_i}\right)^2.$$

where $\mu = \hat{\mu}(\omega)$ and

$$\frac{\partial \mu}{\partial x_i} = \frac{\partial}{\partial x_i} \frac{\sum_{i=1}^n (x_i/\sigma_i^2)}{\sum_{i=1}^n (1/\sigma_i^2)} = \frac{1/\sigma_i^2}{\sum_{i=1}^n (1/\sigma_i^2)},$$

so

$$\sigma_{\hat{\mu}}^2 = \sum_{i=1}^{n} \sigma_i^2 \left(\frac{1/\sigma_i^2}{\sum_{j=1}^n (1/\sigma_j^2)}\right)^2 = \sum_{i=1}^{n} \frac{1/\sigma_i^2}{[\sum_{j=1}^n (1/\sigma_j^2)]^2} = \frac{1}{\sum_{j=1}^n (1/\sigma_j^2)}.$$

Thus, $\sigma_{\hat{\mu}}^2$ is equal to the harmonic mean[6] of the variances of each measurement, σ_j^2, divided by n:

$$\boxed{\sigma_{\hat{\mu}}^2 = \frac{1}{n}\left(\frac{\sum_{j=1}^n (1/\sigma_j^2)}{n}\right)^{-1}.}$$

Note: if σ_i are all identical (i.e. $\sigma_i = \sigma$ for all i) then this reduces to $\sigma_{\hat{\mu}}^2 = \sigma^2/n$, the *standard error* (or *standard deviation of the mean*).

5.1.9. Weighted Least Squares. If the measured data $\{y_i\}$ have different uncertainties $\{\sigma_{y_i}\}$ then we need to account for their relative "weights" when fitting the curve (Fig. 5.3), i.e. points with exceedingly large error bars should not play a dominant role in the fitting results.

We define the "weight" of the i-th measurement as $w_i = \frac{1}{\sigma_{y_i}^2}$. We can apply the principle of maximum likelihood. We first write down the formula for

[6] The harmonic mean $H(x_1, \ldots, x_n)$ of $x_1, \ldots, x_n$ is:

$$\frac{1}{H(x_1, \ldots, x_n)} = \frac{1}{n}\left(\sum_{i=1}^{n} \frac{1}{x_i}\right).$$

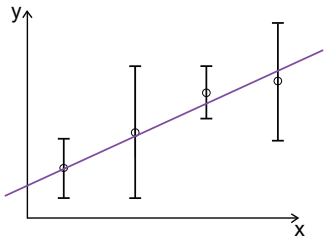


Figure 5.3. Weighted least squares fit accounts for the variations in the error bars from point to point.

chi-square:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - A - Bx_i)^2}{\sigma_{y_i}^2}.$$

Differentiating,

$$\begin{aligned} \frac{\partial \chi^2}{\partial A} &= -2 \sum_{i=1}^{n} \sigma_{y_i}^{-2}(y_i - A - Bx_i) = 0 \\ \frac{\partial \chi^2}{\partial B} &= -2 \sum_{i=1}^{n} \sigma_{y_i}^{-2} x_i (y_i - A - Bx_i) = 0, \end{aligned} \tag{5.4}$$

which we can rewrite as:

$$\begin{aligned} A \sum_{i=1}^{n} w_i + B \sum_{i=1}^{n} w_i x_i - \sum_{i=1}^{n} w_i y_i &= 0 \\ A \sum_{i=1}^{n} w_i x_i + B \sum_{i=1}^{n} w_i x_i^2 - \sum_{i=1}^{n} w_i x_i y_i &= 0. \end{aligned}$$

These 2 equations in 2 unknowns can be solved to yield (see next section):

$$A = \frac{(\sum w_i x_i^2)(\sum w_i y_i) - (\sum w_i x_i)(\sum w_i x_i y_i)}{\Delta},$$

$$B = \frac{(\sum w_i)(\sum w_i x_i y_i) - (\sum w_i x_i)(\sum w_i y_i)}{\Delta}$$

where

$$\Delta = (\sum_{i=1}^{n} w_i)(\sum_{i=1}^{n} w_i x_i^2) - (\sum_{i=1}^{n} w_i x_i)^2.$$

We can also easily show that:[7]

$$\sigma_A = \sqrt{\frac{\sum_{i=1}^{n} w_i x_i^2}{\Delta}} \quad \text{and} \quad \sigma_B = \sqrt{\frac{\sum_{i=1}^{n} w_i}{\Delta}}.$$

5.1.10. Solving for A, B using Matrix Inverse. For the system of equations (5.4) in the previous section, we would first write it in matrix form:

$$\begin{pmatrix} \sum w_i & \sum w_i x_i \\ \sum w_i x_i & \sum w_i x_i^2 \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} \sum w_i y_i \\ \sum w_i x_i y_i \end{pmatrix},$$

then solve for $(A\,B)$ with the matrix inverse:

$$\begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} \sum w_i & \sum w_i x_i \\ \sum w_i x_i & \sum w_i x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum w_i y_i \\ \sum w_i x_i y_i \end{pmatrix}. \tag{5.5}$$

The matrix inverse is easily found:

$$\begin{pmatrix} \sum w_i & \sum w_i x_i \\ \sum w_i x_i & \sum w_i x_i^2 \end{pmatrix}^{-1} = \frac{1}{\sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2} \begin{pmatrix} \sum w_i x_i^2 & -\sum w_i x_i \\ -\sum w_i x_i & \sum w_i \end{pmatrix}.$$

Denoting the determinant by Δ and carrying out the multiplication of the inverse with the column vector $\begin{pmatrix} \sum w_i y_i \\ \sum w_i x_i y_i \end{pmatrix}$, as required by Equation (5.5), we obtain $(A\,B)$:

$$\begin{pmatrix} A \\ B \end{pmatrix} = \frac{1}{\Delta} \begin{pmatrix} \sum w_i x_i^2 \sum w_i y_i - \sum w_i x_i \sum w_i x_i y_i \\ \sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i \end{pmatrix}$$

where Δ is defined as before. You can check by setting $w_i = 1$ that you recover the results of the unweighted least squares method previously covered. For the weighted least squares, we conclude that the coefficients A and B from the fitting procedures should be calculated according to the formulae:

$$A = \frac{\sum w_i x_i^2 \sum w_i y_i - \sum w_i x_i \sum w_i x_i y_i}{\Delta},$$

$$B = \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\Delta}.$$

5.1.11. Fitting Exponentials using a Straight Line Model. The exponential function

$$y(x) = Ae^{Bx}$$

does not appear to be amenable to the technique of linear least squares because it is not linear in the parameters A and B. Such exponential functions are ubiquitous in nature. They arise as solutions of differential equations of the type:

$$\frac{dy}{dt} = \lambda y \qquad \Longrightarrow \qquad y(t) = y(0)e^{\lambda(t-t_0)},$$

[7] If you had difficulty with this type of calculation when we previously dealt with the case of identical error bars, here is a good opportunity to gain some more practice with error propagation.

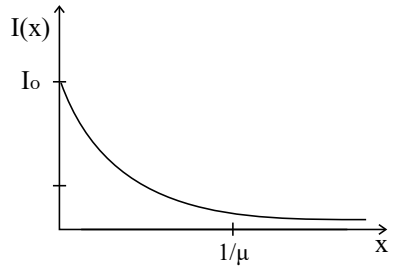


Figure 5.4. Decay of beam intensity as function of depth of penetration into a material. Attenuation is usually exponential with distance.

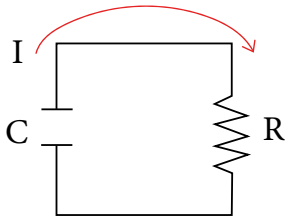


Figure 5.5. *RC* circuit.

as you can easily check by substitution of $y(t)$ into the differential equation. As an example, the intensity I of a radiation beam penetrating a distance x through a shield obeys the equation (see Fig. 5.4)

$$I(x) = I_0 e^{-\mu x}$$

I_0: intensity of the incident beam
μ: absorption coefficient (a property of the shield material)

Another example is an RC circuit (Fig. 5.5). The charge Q accumulated on capacitor C drains away exponentially fast when the capacitor is connected to a resistor. The time-dependence of the charge is described by:

$$Q(t) = Q_0 e^{-\lambda t}$$

Q_0: initial charge (at time $t = 0$)
$\lambda = 1/RC$: inverse time constant
R: resistance
C: capacitance

Fortunately, these models can be treated by linear least squares if we take the log of the equation:

$$\log y = \log A + Bx$$

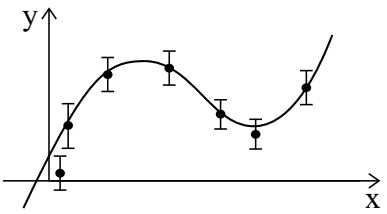


Figure 5.6. In general, a polynomial can be fit to most smooth curves. The challenge is to choose a fitting model that has the lowest number of parameters possible while fitting the data well.

and fit a straight line to the equation ($z_i = \log y_i$):

$$z_i = \log A + Bx_i$$

If the errors σ_y in y are identical at each point, then for a function of the form $z = \log y$, σ_z depends on y according to $\sigma_z = \left|\frac{dz}{dy}\right| \sigma_y = \frac{\sigma_y}{y}$. We can then use the weighted least squares method.

5.1.12. Fitting a Polynomial. We can derive least squares formulae for higher order polynomial models (see Fig. 5.6)

$$y(x|A, B, C, \ldots, H) = A + Bx + Cx^2 + \cdots + Hx^{p-1}$$

For example, the height of a falling body should obey the equation

$$y(t|y_0, v_0, g) = y_0 + v_0 t - \frac{1}{2} g t^2$$

where
y_0, v_0: initial height and velocity, respectively
g: acceleration due to gravity.
Consider the model:

$$y(x|A, B, C) = A + Bx + Cx^2,$$

which is still considered a "linear model" because linear refers to linearity in the fitting parameters A, B and C, which this function fulfills. The corresponding chi-square function is:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - A - Bx_i - Cx_i^2)^2}{\sigma_y^2}.$$

To obtain the coefficients A, B and B we invoke the principle of maximum likelihood. The likelihood function

$$L(A, B, C|y_1, \ldots, y_n) = p(y_1, \ldots, y_n) \propto e^{-\chi^2/2},$$

is viewed as a function of the fitting parameters A, B, C. Minimization is performed with respect to these parameters. Setting $\partial\chi^2/\partial A = 0$, $\partial\chi^2/\partial B =$

0 and $\partial\chi^2/\partial C = 0$:

$$An + B\sum_{i=1}^{n} x_i + C\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i$$
$$A\sum_{i=1}^{n} x_i + B\sum_{i=1}^{n} x_i^2 + C\sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i$$
$$A\sum_{i=1}^{n} x_i^2 + B\sum_{i=1}^{n} x_i^3 + C\sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i$$

These 3 equations in 3 unknowns can be solved to yield values of A, B and C in terms of the data.

5.1.13. General Linear Function of the Parameters. The method works for functions $y = f(x)$ which depend linearly on the coefficients A, B, ... For example,

$$y(x|A, B) = A\sin(x) + B\cos(x)$$

depends linearly on A and B. Likewise, any function of the form:

$$y(x|A, B, \ldots, H) = Af(x) + Bg(x) + \cdots + Hk(x)$$

where $f, g, \ldots, k$ are known functions.

5.1.14. Multiple Regression. Many problems require 2 or more variables. An example is the ideal gas law $PV = Nk_BT$ for fixed N, which expresses the relationship between pressure, volume and temperature, i.e. $P = f(V, T)$.

Suppose that

$$z(x, y|A, B, C) = A + Bx + Cy$$

and we measure the data points (set of triples)

$$\{(x_i, y_i, z_i)\}, \qquad i = 1, \ldots, n$$

z_i: all have the same uncertainty (σ_z)
x_i, y_i: are assumed to be "exact" (negligible uncertainty)

Maximum likelihood yields:

$$An + B\sum_{i=1}^{n} x_i + C\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} z_i$$
$$A\sum_{i=1}^{n} x_i + B\sum_{i=1}^{n} x_i^2 + C\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i z_i$$
$$A\sum_{i=1}^{n} y_i + B\sum_{i=1}^{n} x_i y_i + C\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} y_i z_i$$

and we solve for A, B and C to get the best fit (in the least squares sense).

5.1.15. Linear vs Nonlinear Dependence on Parameters. So far we have discussed the case of "linear least squares". By linear, we mean that the model is a linear function of the model's parameters. For example:

$$y(x|A,B) = A + Bx$$
$$y(x|A,B,C) = A + Bx + Cx^2$$
$$y(x|A,B,C,D,E) = A + Bx + Cz + Dz^2 + Exz$$

are all linear models. In general, a linear model is of the form:

$$y(x|\boldsymbol{\theta}) = \sum_{i=1}^{p} \theta_i f_i(x)$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ and each term contains the first power of the θ_i's. The functions $f_i(x)$ can be nonlinear in x. In matrix form, $y(x|\boldsymbol{\theta}) = \boldsymbol{\theta}^T \vec{f}(x)$, where and $\vec{f}$ is the column vector $(f_1, \ldots, f_p)^T$ (the subscript T denotes "transpose"). For example,

$$y(x|A,B) = A\cos(x) + B\log(x)$$

is still considered a "linear least squares" problem because A and B show up in first power. On the other hand, $y(x|A,B) = A\cos(Bx)$ is not linear in B because B shows up in even powers up to infinity. Similarly, $y(x|A) = A + A^2x$ is nonlinear because of the A^2. If the dependence on the parameters is linear, we can use the maximum likelihood technique to obtain estimates of the parameters in terms of the data. This is called "linear least squares".

5.2. How to Determine if a Fit is Good

5.2.1. Inspect the residuals and look for possible trends. The residuals measure the distance between the model, $\{y(x_i)\}$, and the data, $\{y_i\}$:

$$R_i = y_i - y(x_i|\boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ denotes the set of fitting parameters, e.g. $\boldsymbol{\theta} = (A,B)^T$ in the case of a straight line model. Plotting the residuals (see Fig. 5.8) yields a useful

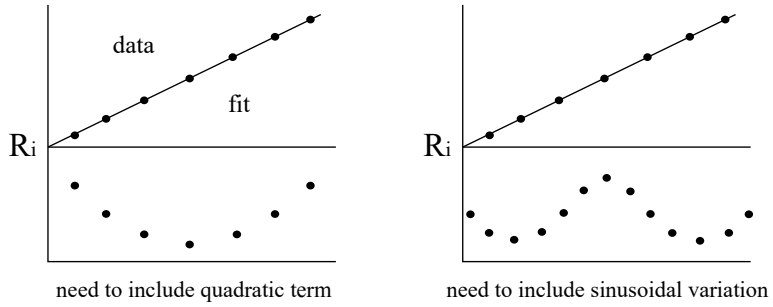


Figure 5.7. Residuals can be used to assess the adequacy of fitting model or to propose changes to the model.

tool to judge if a fitting model is adequate. In the figure below we have illustrated two examples where the fitting model is inadequate because the residuals show some clear (non-random) trends. In the first case, a quadratic dependence should be added to the model. In the second case, the model fails to capture a sinusoid dependence.

Suppose that $y(x)$ is a model for the experimental data. The set of ordered pairs, $\{(x_i, y_i)\}$, is the experimental data. For simplicity, we still assume that the errors in x_i are negligible. So far we have assumed that the following random variable is normally distributed about the model's trendline (with parameters: mean, $y(x_i)$ and variance, σ_y^2):

$$y_i \sim \mathcal{N}(y(x_i), \sigma_{y_i}^2).$$

An equivalent statement is:

$$R_i = y_i - y(x_i) \sim \mathcal{N}(0, \sigma_{y_i}^2).$$

Invoking $var(aX) = a^2 \cdot var(X)$, another equivalent statement is:

$$\boxed{\tilde{R}_i = \frac{y_i - y(x_i)}{\sigma_{y_i}} \sim \mathcal{N}(0, 1).}$$

where σ_{y_i} is the error in y_i. The basic *assumption* of the least squares method that data points are Gaussian-distributed about the model's trendline is a good assumption in most cases. This is a consequence of the central limit theorem and the fact that most physical measurements of macroscopic properties are the result (sum) of a very large quantity of smaller microscopic processes, such as molecular collisions and other scattering events. There are exceptions to this, where in some cases the distribution is not Gaussian. $\tilde{R}_i$ is known as the *normalized residuals*. In the sketch below, the plot on the left shows hypothetical experimental data whose error bars are larger on the right than on the left. This leads to the residuals $\{R_i\}$ shown in the middle plot. The error grows from left to right, as seen in the residuals. However, if

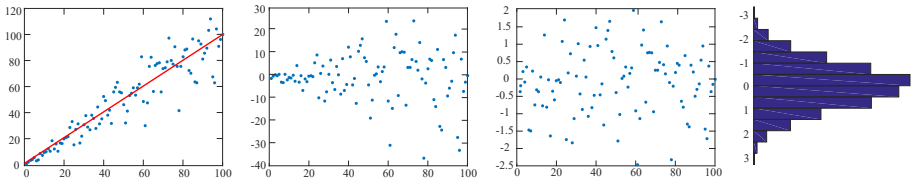


Figure 5.8. Using (normalized) residuals to assess the quality of a fit. From left to right: (1) Experimental data with straight line fit. (2) Residuals. (3) Normalized residuals. (4) Histogram of normalized residuals follows a Gaussian distribution with mean 0 and variance 1.

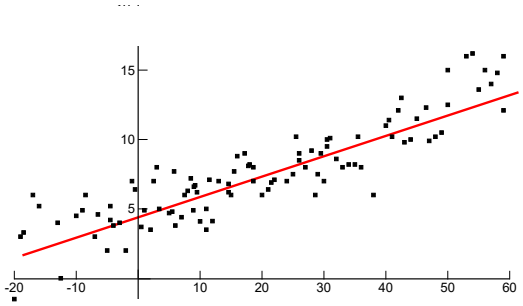


Figure 5.9. Linear regression.

we normalize the residuals, $\{\tilde{R}_i\}$, the results show a uniform error bar since $\tilde{R}_i \sim \mathcal{N}(0,1)$. This is shown on the plot to the right, where I have sketched a bell-shaped histogram to illustrate the standard Gaussian distribution of the normalized residuals (Fig. 5.8).

An adequate model which captures all the needed trends should leave normalized residuals that are normally distributed with mean 0 and width 1. A good fit will yield:

68% of data scattered within ± 1 of 0.
95% of data scattered within ± 2 of 0.
99.7% of data scattered within ± 3 of 0.

5.2.2. Chi-Square as a Goodness-of-Fit Parameter. For convenience, we shall denote the parameters of the model $(A, B, C, \dots)$ by the vector $\boldsymbol{\theta} = (A, B, C, \dots)^T$. Consider a set of data points $\{(x_i, y_i)\}$ and a model $y(x_i|\boldsymbol{\theta})$ for the data (Fig. 5.9).

A measure of how good the fit is should be a distance metric that measures how far the data points lie from the curve. One possible such measure can

be obtained by summing all differences (squared) between the data points and the model. The chi-square

$$\chi^2 = \sum_{i=1}^{n} \frac{[y_i - y(x_i|\boldsymbol{\theta})]^2}{\sigma_{y_i}^2}$$

is a measure of the distance between the model and the data, i.e. the "goodness-of-fit". Some remarks:

- Fitting the data corresponds to minimizing χ^2. Doing so ensures that the model closely represents the measured data. The best fit parameters are found when χ^2 is minimized.
- χ^2 can also be viewed as a random variable since it is a function of y_i which are themselves random variables. Thus, it has its own distribution. (Exercise: can you derive its probability distribution?)
- It can lead to analytical expressions for the coefficients in some cases. For example, we have seen that for linear models, analytical expressions can be obtained.
- When analytical expressions are not possible, we can always use computer-based minimization. (Computer-based optimization is the topic of subsequent lectures.)

5.2.2.1. *Chi-Square is (l_2-norm)^2 of the Normalized Residuals.* The formula for chi-square

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{y_i - y(x_i|\boldsymbol{\theta})}{\sigma_{y_i}}\right]^2 = \sum_{i=1}^{n} \tilde{R}_i^2$$

where

$$\tilde{R}_i = \frac{y_i - y(x_i|\boldsymbol{\theta})}{\sigma_{y_i}}$$

is the *normalized residual*, can be viewed also known as the l_2-norm of a n-dimension vector whose components are the normalized residuals:

$$\vec{R} = (\tilde{R}_1, \tilde{R}_2, \ldots, \tilde{R}_n).$$

The l_2 norm of a vector $\vec{x}$ is also known as the Euclidean norm:

$$\|\vec{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}.$$

where $\vec{x} = (x_1, x_2, \ldots, x_n)$. The Euclidean norm is frequently used to measure distances. The Euclidean distance between two vectors $\vec{x}$ and $\vec{y}$ is,

$$\|\vec{x} - \vec{y}\| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}.$$

The l_1 norm of a vector $\vec{x}$ is defined as:

$$\|\vec{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|.$$

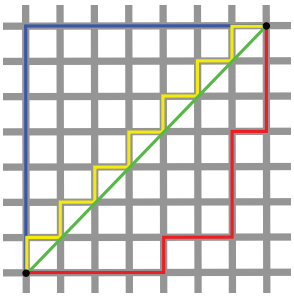


Figure 5.10. l_1 vs l_2 distance metrics.

The l_1 norm is also known as the *Manhattan metric* or the *taxicab norm*. The taxicab distance is thus:

$$\|\vec{x} - \vec{y}\|_1 = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|.$$

In Fig. 5.10, the taxicab distance in 2D ($n = 2$) is shown by the red line and the Euclidean distance is shown by the green curve.

Generally, the l_p-norm of a vector $\vec{x}$ is:

$$\|\vec{x}\|_p = \left(|x_1|^p + \cdots + |x_n|^p\right)^{1/p}.$$

As you can see there are many possible choices for the distance metric in order to measure the distance between the data and the model. With the chi-square function, as constructed, we have a n-dimensional space and least squares fitting consists of minimizing χ^2, the l_2-norm between the data points $\{y_i\}$ and the model $\{y(x_i|\boldsymbol{\theta})\}$. The l_2 norm is easier to work with from a mathematical standpoint than other distance metrics. For example, it is generally difficult to work with absolute values.

5.2.3. Analysis of Variance.

5.2.3.1. *Conditional Variance.* The conditional variance of a random variable Y given another random variable X is defined as:

$$var(Y|X) = \mathbb{E}\Big(\big(Y - \mathbb{E}(Y \mid X)\big)^2 \mid X\Big).$$

The conditional variance tells us how much variance is left if we use $\mathbb{E}(Y \mid X)$ to "predict" Y. Here, $\mathbb{E}(Y \mid X)$ stands for the conditional expectation of Y given X, which we may recall, is a random variable itself (a function of X, determined up to probability one). As a result, $var(Y|X)$ itself is a random variable (and is a function of X).

5.2.3.2. *Law of Total Variance.* The law of total variance states that if X and Y are random variables on the same probability space, and the variance of Y is finite, then

$$var(Y) = \mathbb{E}[var(Y \mid X)] + var(\mathbb{E}[Y \mid X]).$$

The two terms are called the "unexplained" and the "explained" components of the variance, respectively.

Proof. The law of total variance can be proved using the law of total expectation. First,

$$var[Y] = \mathbb{E}\left[Y^2\right] - \mathbb{E}[Y]^2$$

from the definition of variance. Again, from the definition of variance, and applying the law of total expectation, we have

$$\mathbb{E}\left[Y^2\right] = \mathbb{E}\left[\mathbb{E}[Y^2 \mid X]\right] = \mathbb{E}\left[var[Y \mid X] + [\mathbb{E}[Y \mid X]]^2\right].$$

Now we rewrite the conditional second moment of Y in terms of its variance and first moment, and apply the law of total expectation on the right hand side:

$$\mathbb{E}\left[Y^2\right] - \mathbb{E}[Y]^2 = \mathbb{E}\left[var[Y \mid X] + [\mathbb{E}[Y \mid X]]^2\right] - [\mathbb{E}[\mathbb{E}[Y \mid X]]]^2.$$

Since the expectation of a sum is the sum of expectations, the terms can now be regrouped:

$$= (\mathbb{E}[var[Y \mid X]]) + \left(\mathbb{E}\left[\mathbb{E}[Y \mid X]^2\right] - [\mathbb{E}[\mathbb{E}[Y \mid X]]]^2\right).$$

Finally, we recognize the terms in the second set of parentheses as the variance of the conditional expectation $\mathbb{E}[Y \mid X]$:

$$= \mathbb{E}[var[Y \mid X]] + var[\mathbb{E}[Y \mid X]].$$

□

5.2.3.3. *Explained and Unexplained Variation.* We are interested in two measures used in correlation and regression studies: the coefficient of determination and the standard error of estimate. In doing so, we must also learn how to construct a prediction interval for y using a regression line and a given value of x. To study these concepts, we need to understand and calculate the total variation, explained deviation, and the unexplained deviation for each ordered pair in a data set.

Assume that we have a collection of paired data $\{(x_i, y_i)\}_{i=1}^n$. Together with a model $y(x)$ that predicts the value of y. The sample mean will be denoted $\overline{y}$. The total variation about a regression line is the sum of the squares of the differences between the y-value of each ordered pair and the mean of y:

$$\text{total variation} = SS_{tot} = \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

The explained variation is the sum of the squared of the differences between each predicted y-value and the mean of y:

$$\text{explained variation} = SS_{ex} = \sum_{i=1}^{n} (y(x_i) - \overline{y})^2.$$

The unexplained variation is the sum of the squared differences between the y-value of each ordered pair and each corresponding predicted y-value:

$$\text{unexplained variation} = SS_{res} = \sum_{i=1}^{n} (y(x_i) - y_i)^2.$$

Since

$$(y_i - \overline{y})^2 = (y_i - y(x_i) + y(x_i) - \overline{y})^2 = (y_i - y(x_i))^2 + (y(x_i) - \overline{y})^2 + 2(y_i - y(x_i))(y(x_i) - \overline{y})$$

Summation over the last term, $\sum_i (y_i - y(x_i))(y(x_i) - \overline{y})$, yields the covariance of $\epsilon_i = y_i - y(x_i)$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $y(x_i)$ (with mean $\overline{y}$). This covariance vanishes since both ϵ_i and $y(x_i)$ are independent and have mean zero:

$$\overline{y} = \frac{1}{n} \sum_i y_i, \quad \sum_i \frac{1}{n} (y(x_i) - \overline{y}) = \frac{1}{n} \sum_i y(x_i) - \overline{y} = -\frac{1}{n} \sum_i \epsilon_i \approx 0.$$

We conclude that the sum of the explained and unexplained variations is equal to the total variation:

$$\text{total variation} = \text{explained variation} + \text{unexplained variation}.$$

As its name implies, the explained variation can be explained by the relationship between x and y. The unexplained variation cannot be explained by the relationship between x and y and is due to chance or other variables.

In the previous section we have seen the law for total variance:

$$var(Y) = \underbrace{\mathbb{E}[var(Y \mid X)]}_{unexplained} + \underbrace{var(\mathbb{E}[Y \mid X])}_{explained}.$$

where the two terms are "unexplained" and the "explained" components of the variance, respectively. We can check that this formula is equivalent to the above result

$$\underbrace{\sum_{i=1}^{n} (y_i - \overline{y})^2}_{total} = \underbrace{\sum_{i=1}^{n} (y(x_i) - y_i)^2}_{unexplained} + \underbrace{\sum_{i=1}^{n} (y(x_i) - \overline{y})^2}_{explained}$$

if we replace variance by sample variance, and expectation by sample mean:

$$\begin{aligned} var(Y) =& \mathbb{E}[var(Y \mid X)] + var(\mathbb{E}[Y \mid X]) \\ =& \mathbb{E}\left[\mathbb{E}\Big(\big(Y - \mathbb{E}(Y \mid X)\big)^2 \mid X\Big)\right] + var(\mathbb{E}[Y \mid X]) \\ =& \frac{1}{n} \sum_{i=1}^{n} (y_i - y(x_i))^2 + \frac{1}{n} \sum_{i=1}^{n} (y(x_i) - \overline{y})^2 \end{aligned}$$

where we used $y(X) \equiv \mathbb{E}[Y \mid X]$ and

$$\mathbb{E}_X\Big[\mathbb{E}\Big(\big(Y - \mathbb{E}[Y \mid X]\big)^2 \mid X\Big)\Big] = \mathbb{E}_X\big(Y - y(X)\big)^2 \approx \frac{1}{n}\sum_{i=1}^{n}(y_i - y(x_i))^2.$$

$$var(\mathbb{E}_X[Y \mid X]) = \mathbb{E}_X\left(\mathbb{E}[Y \mid X]\right) - \mathbb{E}[Y])^2 = \mathbb{E}_X(y(X) - \overline{y})^2 \approx \frac{1}{n}\sum_{i=1}^{n}(y(x_i) - \overline{y})^2.$$

If the model is linear (see Section 5.2.3.5 below),

$$Y = a + bX + \epsilon, \qquad y(X) \equiv \mathbb{E}[Y \mid X] = \mathbb{E}[a + bX + \epsilon \mid X] = a + bX.$$

5.2.3.4. *Adam's Law.* For any rv's X and Y,

$$\mathbb{E}\big[\mathbb{E}[Y \mid X]\big] = \mathbb{E}[Y].$$

Proof. We consider the case where X and Y are both discrete (the proofs for other cases are analogous). Let $\mathbb{E}[Y \mid X] = g(X)$. Then,

$$\begin{aligned}
\mathbb{E}[g(X)] &= \sum_x g(X)\mathbb{P}(X = x) \\
&= \sum_x \left(\sum_y y\mathbb{P}(Y = y \mid X = x)\right)\mathbb{P}(X = x) \\
&= \sum_x \sum_y y\mathbb{P}(X = x)\mathbb{P}(Y = y \mid X = x) \\
&= \sum_y y \sum_x \mathbb{P}(X = x, Y = y) \\
&= \sum_y y\mathbb{P}(Y = y) = \mathbb{E}[Y].
\end{aligned}$$

□

5.2.3.5. *Linear Regression.* In its most basic form, the linear regression model uses a single explanatory variable X to predict a response variable Y, and it assumes that the conditional expectation of Y is linear in X:

$$\mathbb{E}[Y \mid X] = a + bX.$$

An equivalent way to express this is to write

$$Y = a + bX + \epsilon,$$

where ϵ is a rv (called the error) with $\mathbb{E}[\epsilon \mid X] = 0$. This can be proven by taking $Y = a + bX + \epsilon$, with $\mathbb{E}[\epsilon \mid X] = 0$. By linearity:

$$\mathbb{E}[Y \mid X] = \mathbb{E}[a \mid X] + \mathbb{E}[bX \mid X] + \mathbb{E}[\epsilon \mid X] = a + bX.$$

Conversely, suppose that $\mathbb{E}[Y \mid X] = a + bX$, define

$$\epsilon = Y - (a + bX).$$

Then, $Y = a + bX + \epsilon$ with

$$\mathbb{E}[\epsilon \mid X] = \mathbb{E}[Y \mid X] - \mathbb{E}[a + bX \mid X] = \mathbb{E}[Y \mid X] - (a + bX) = 0.$$

We can also solve for the constants a and b in terms of $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $cov(X, Y)$ and $var(X)$. This is done by invoking Adam's law, $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$, and taking the expectation of both sides gives

$$\mathbb{E}[Y] = a + b\mathbb{E}[X].$$

Note that ϵ has mean 0 and X and ϵ are uncorrelated, since

$$\mathbb{E}[\epsilon] = \mathbb{E}[\mathbb{E}[\epsilon \mid X]] = \mathbb{E}[0] = 0$$

and

$$\mathbb{E}[\epsilon X] = \mathbb{E}[\mathbb{E}[\epsilon X \mid X]] = \mathbb{E}[X\mathbb{E}[\epsilon \mid X]] = \mathbb{E}[0] = 0.$$

Taking the covariance with X of both sides in $Y = a + bX + \epsilon$, we have

$$cov(X, Y) = cov(X, a) + b \cdot cov(X, X) + cov(X, \epsilon) = b \cdot var[X].$$

Thus, we have the two results:

$$b = \frac{cov(X, Y)}{var(X)},$$

$$a = \mathbb{E}[Y] - b\mathbb{E}[X] = \mathbb{E}[Y] - \frac{cov(X, Y)}{var(X)} \cdot \mathbb{E}[X].$$

Numerical values of a and b can then be obtained from experimental data $\{(x_i, y_i)\}_{i=1}^n$ by substituting the corresponding formulae for sample means, variance and covariance in lieu of $\mathbb{E}[Y]$, $\mathbb{E}[X]$, $var(X)$ and $cov(X, Y)$, respectively.

5.2.3.6. *R-Squared Value: The Coefficient of Determination.* The coefficient of determination, R^2, is defined as:

$$\boxed{R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\text{unexplained variance}}{\text{total variance}}}$$

where

$$SS_{res} = \sum_{i=1}^{n} (y_i - y(x_i))^2, \quad SS_{tot} = \sum_{i=1}^{n} (y_i - \overline{y})^2, \quad \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

$\{y_i\}$ are the observed data, $\{y(x_i)\}$ are the model predictions. R^2 is method that can tell us how well observed outcomes are replicated by a model. Like χ^2, it also uses the sum of square differences (L^2 distance) between the data and the model. The better the fit, the closer the coefficient of determination gets to $R^2 = 1$.

Example 5.1. The correlation coefficient for Twitter's advertising expenses and company sales data is 0.90. Find the coefficient of determination. What

does this tell you about the explained variation of the data about the regression line? About the unexplained variation? ($\rho = 0.90$ suggests a strong positive linear correlation). Thus, $R^2 = 0.81$. This means that about 81.0% of the variation in the company sales can be explained by the variation in the advertising expenditures. About 19.0% of the variation is unexplained and is due to chance or other variables.

5.2.3.7. *Correlation Coefficient vs Coefficient of Determination.* Let X and Y be two random variables. X is the independent variable and Y is the dependent variable. We would like to know the type of relationship that exists between X and Y (if any). The coefficient of correlation (ρ) and coefficient of determination (R^2) provide useful information. As an example, X and Y could be related linearly. We would describe the linear dependence by a model, i.e. $y(x) = \hat{a} + \hat{b}x$, where $\hat{a}$ and $\hat{b}$ are estimators for the true coefficients a and b. The random variables themselves are related by $Y = a + bX + \epsilon$, where ϵ is needed to describe the noise (i.e. imagine the special case where X is noiseless; ϵ is needed to explain the variability in Y).

Coefficient of determination (R^2):

(1) The square root of R^2 is equal to the correlation coefficient (ρ). See Section 5.2.3.8 below for proof.

(2) It provides percentage variation in Y which is explained by all the Y variables together.

(3) R^2 value is (usually) between 0 and 1 and indicates strength of Linear Regression model.

(4) The higher the R^2 value, the less scattered the data points are (i.e. it is a good model). The lesser the R^2 value is the more scattered the data points are.

Coefficient of Correlation (ρ):

(1) It measures the strength and the direction of a linear relationship between two variables (x and y) with possible values between -1 and 1.

(2) Positive correlation ($\rho > 0$) indicates that two variables rise and fall together. $\rho = 1$ means perfect positive correlation, i.e. $Y = a + bX + \epsilon$, where $b > 0$.

(3) Negative correlation ($\rho < 0$) indicates that two variables are perfect opposites: when one goes up the other goes down (and vice versa). $\rho = -1$ means perfect negative correlation (anti-correlation), i.e. , i.e. $Y = a + bX + \epsilon$, where $b < 0$.

(4) No correlation when ρ is close to 0. This means the correlation between X and Y is weak or non-existent. It could be due to X

and Y being statistically independent, although $\rho = 0$ is not a sufficient condition to prove independence.

5.2.3.8. *Relationship between ρ and R^2 in Linear regression.*

Theorem 5.2. *Assume a simple linear regression model with independent observations*

$$Y = a + bX + \varepsilon,\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2),\ i = 1, \ldots, n \tag{5.6}$$

and consider estimation using ordinary least squares. Then, the coefficient of determination is equal to the squared correlation coefficient between X and Y:

$$R^2 = \rho(X, Y)^2. \tag{5.7}$$

Proof. The ordinary least squares estimates for simple linear regression are

$$\hat{a} = \bar{y} - \hat{b}\bar{x}, \quad \hat{b} = \frac{s_{xy}}{s_x^2}, \tag{5.8}$$

where

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}), \quad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The coefficient of determination R^2 is defined as the proportion of the variance explained by the independent variables, relative to the total variance in the data. This can be quantified as the ratio of explained sum of squares to total sum of squares:

$$R^2 = \frac{SS_{ex}}{SS_{tot}} = \frac{\sum_{i=1}^{n}(y(x_i) - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

Using the explained and total sum of squares for simple linear regression, we have:

$$R^2 = \frac{\sum_{i=1}^{n}(y(x_i) - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n}(\hat{a} + \hat{b}x_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

By applying (5.8), we can further develop the coefficient of determination:

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n}(\bar{x} - \hat{b}\bar{x} + \hat{b}x_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^{n}\left(\hat{b}(x_i - \bar{x})\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= \hat{b}^2\,\frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= \hat{b}^2\,\frac{s_x^2}{s_y^2} = \left(\frac{s_x}{s_y}\,\hat{b}\right)^2 = \rho(X,Y)^2, \end{aligned}$$

where $s_y^2 = SS_{tot}/(n-1)$ and in the last step we used the relationship (5.8) between correlation coefficient and slope estimate (slope $= \rho \cdot s_y/s_x$). □

Chapter 6

Non-Linear Least Squares Optimization

If the model is nonlinear in the fitting parameters we must use iterative techniques. In many cases, the best we can do is "guess" the initial parameter values and evolve them over time until we reach a satisfactory solution. Our focus is on the function $\chi^2(\boldsymbol{\theta})$ because χ^2 measures the difference between data and a model, data fitting amounts to minimizing χ^2:

$$\min_{w.r.t.(\boldsymbol{\theta})} \chi^2(\boldsymbol{\theta}) = \chi^2(\overline{\boldsymbol{\theta}})$$

with respect to the parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$. These are the fitting parameters we previously denoted A, B, ... These parameters span a multidimensional space. The parameters of χ^2 are sometimes subjected to constraints.

If the data is fitted to a function $f(x|\boldsymbol{\theta})$ that is linear in the parameters, we can use linear least squares. In other cases (f is nonlinear in the θ_i's) we must resort to iterative techniques. The iterative approach consists of:

- Start with initial guess $\{\theta_i^{(0)}\}_{i=1}^p$. We denote the iteration index by a superscript. The component of the vector $\boldsymbol{\theta}$ is denoted by a subscript.
- Iterate until χ^2 is minimized, i.e. obtain a new set of coefficients $\{\theta_i^{(s)}\}_{i=1}^p$ from the previous ones $\{\theta_i^{(s-1)}\}_{i=1}^p$ using some suitable rule. (Very often, the iterations stop when χ^2 no longer changes appreciably.)
- Final $\{\theta_i^{final}\}_{i=1}^p$'s yield a global minimum of χ^2. At least, we hope that the minimum we reached is a *global* minimum.

6.1. Newton-Raphson method

Newton-Raphson is a method for finding successively better approximations to the roots (or zeros) of a real-valued function $f(x)$. These zeros are solutions to the equation $f(x) = 0$. When applied to the function $f'(x)$, it can be used to find extrema of f.

6.1.1. Finding Zeros of a Function. We can find the zeros of a function, $\{x|f(x) = 0\}$, even when x is defined implicitly. For example, suppose you are asked to find the maximum likelihood estimator of α, a parameter of the gamma distribution. The likelihood function for a set of gamma distributed random variables is:

$$L(\alpha, \beta) \equiv p(x_1, \ldots, x_n; \alpha, \beta) = \frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n} \prod_{i=1}^{n} x_i^{\alpha-1} e^{-\beta \sum_{i=1}^{n} x_i}.$$

where $X_i(\omega) = x_i$ are the experimentally measured data and $\Gamma(\cdot)$ is the gamma function. Taking the log of L and differentiating with respect to β gives the maximum likelihood estimator:

$$\hat{\beta} = \frac{\hat{\alpha}}{\frac{1}{n}\sum_{i=1}^{n} X_i}.$$

However, the derivative with respect to α gives

$$\log(\hat{\alpha}) - \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = \log\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \frac{1}{n}\sum_{i=1}^{n} \log X_i,$$

an expression that is highly nonlinear in $\hat{\alpha}$. We cannot solve for α by writing this equation in the form $\hat{\alpha} = h(\{X_i\})$, for some function h of the data points $\{X_i(\omega) = x_i\}$. Instead, we can write this equation in the form $f(\hat{\alpha}|\{x_i\}) = 0$, where

$$f(\hat{\alpha}) = \log(\hat{\alpha}) - \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} - \log\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + \frac{1}{n}\sum_{i=1}^{n} \log X_i,$$

and use the Newton-Raphson method to solve for $\hat{\alpha}$ in terms of the data $\{x_i\}$. As an exercise, you should write a MATLAB program to solve for the zeros of this function for given data $x_1, \ldots, x_n$.

The Newton-Raphson algorithm consists of:

- Choose a starting point, $x^{(1)}$.
- Approximate $f(x^{(1)})$ near $x^{(1)}$.

$$f(x^{(1)} + h) \equiv 0 \approx \underbrace{f(x^{(1)}) + f'(x^{(1)})h}_{\text{Taylor expansion}}.$$

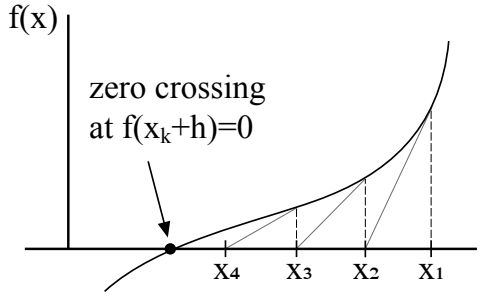


Figure 6.1. The Newton-Raphson method is used to iteratively find the zero-crossing of a function. For $k > 0$ sufficiently large, a zero of f is eventually found (if one exists). In the text I changed the notation for iteration to a superscript, e.g. $x^{(4)}$ instead of x_4.

so that

$$f(x^{(1)}) + f'(x^{(1)})h \approx 0$$

and therefore

$$h \approx -\frac{f(x^{(1)})}{f'(x^{(1)})}.$$

- This gives an approximation that takes us closer to the zero crossing

$$x^{(2)} = x^{(1)} - \frac{f(x^{(1)})}{f'(x^{(1)})}.$$

- In general (i.e. after the k-th step) the update rule is:

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}. \quad \text{(to find zeros of } f\text{)}$$

The working principle is illustrated in Fig. 6.1.

6.1.2. Finding Extrema of a Function. If we want to find extrema of a function $f(x)$, we can apply the Newton-Raphson method to the function $f'(x)$ instead of $f(x)$, as shown in Fig. 6.2. The update rule is:

$$x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})}. \quad \text{(to find extremum of } f\text{)}$$

In other words, Taylor expanding the gradient of f around the point $x^{(k)}$ and setting equal to zero gives the update rule for extrema of a function:

$$f'(x^{(k)} + h) = f'(x^{(k)}) + h \cdot f''(x^{(k)}) = 0 \quad \Rightarrow \quad h = -\frac{f'(x^{(k)})}{f''(x^{(k)})}, \tag{6.1}$$

which leads to

$$\boxed{x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})}.}$$

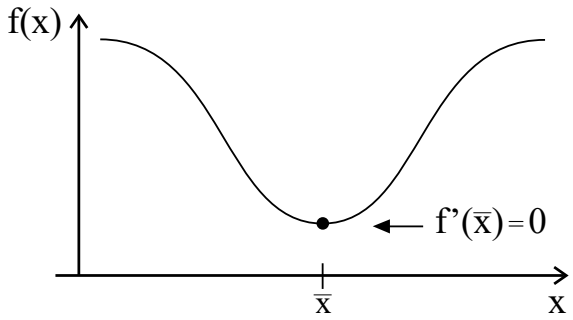


Figure 6.2. The Newton-Raphson method can also be used to find extrema of a function.

6.1.3. Alternative derivation. We can obtain this update rule directly for $f : \mathbb{R} \to \mathbb{R}$ because we know that $f'(x) = 0$ at the minimum of $f(x)$. Approximating f with a Taylor expansion about some point $x^{(k)}$:

$$f(x^{(k)} + h) \approx f(x^{(k)}) + h \cdot f'(x^{(k)}) + \frac{h^2}{2} f''(x^{(k)}).$$

We want to choose h so that $f(x^k+h)$ is a minimum. The necessary condition for an extremum is:

$$\frac{\mathrm{d}}{\mathrm{d}h}\left(f(x^{(k)}) + h \cdot f'(x^{(k)}) + \frac{h^2}{2} f''(x^{(k)})\right) = f'(x^{(k)}) + f''(x^{(k)})h = 0,$$

which leads to

$$h = -\frac{f'(x^{(k)})}{f''(x^{(k)})},$$

and the update rule

$$x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})} = x^{(k)} - (f''(x^{(k)}))^{-1} f'(x^{(k)}).$$

Its generalization to multiple dimensions, $f : \mathbb{R}^n \to \mathbb{R}$, involves replacing derivatives $f'(x) \to \nabla f(x)$ (gradient vector) and $f''(x) \to \nabla\nabla f(x)$ (Hessian matrix) to obtain the update rule $x^{(k+1)} = x^{(k)} - (\nabla\nabla f(x^{(k)}))^{-1} \nabla f(x^{(k)})$. This will be discussed below in Section 6.12.

6.2. Gradient (Steepest) Descent Method

$\nabla\chi^2$ lies along the direction where χ changes most rapidly (Fig. 6.3). Consequently, we may construct an update rule as

$$\boxed{\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \lambda \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)}).}$$

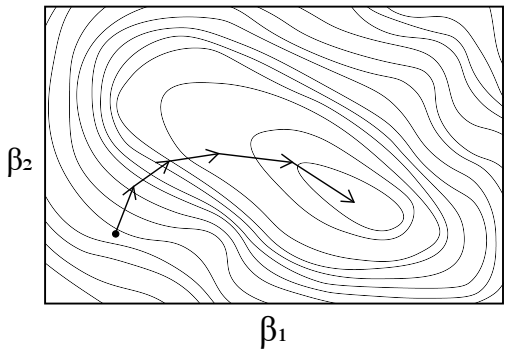


Figure 6.3. For a height function of the type $z = f(x, y)$, the gradient descent method moves in the xy plane along steps parallel to the local direction of steepest descent. Note: in this figure β should be replaced by θ.

where $\lambda > 0$ is called the "learning rate" because it controls the speed at which the model parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ are learned from the data (χ^2 is a function of the data).

The learning rate λ cannot be too large (to avoid overshoot) and must not be too small (so it converges reasonably fast) while $\boldsymbol{\theta}$ is a column vector containing all p parameters at the k-th iteration. (The negative sign is used because $\nabla\chi^2$ points in the direction of steepest increase of χ^2 and we want the direction of steepest decrease.) $\nabla\chi^2$ can be computed analytically if the functional form of χ^2 (hence, the model) is known.

If we don't have an analytical formula available for the gradient, we can approximate $(\nabla\chi^2)_j = \frac{\partial\chi^2}{\partial\theta_j}$ numerically by finite differences:

$$\frac{\chi^2(\theta_1, \ldots, \theta_{j-1}, \theta_j + \delta\theta_j, \theta_{j+1}, \ldots, \theta_p) - \chi^2(\theta_1, \ldots, \theta_{j-1}, \theta_j, \theta_{j+1}, \ldots, \theta_p)}{\delta\theta_j},$$

where $\delta\theta_j$ is a small step along the j-th direction.

This method moves toward the minimum because the directional derivative of f along $\hat{y}$ evaluated at the point $\vec{a}$

$$\frac{\mathrm{d}f}{\mathrm{d}y}(\vec{a}) \equiv \nabla f(\vec{a}) \cdot \hat{y} = \|\nabla f(\vec{a})\| \cos\phi$$

is largest when $\hat{y}$ points along (parallel to) the vector $\nabla f(\vec{a})$ (i.e. when $\phi = 0$). The function f undergoes its maximum rate of change in the direction of $\nabla f(\vec{a})$ (Fig. 6.4). Thus, the gradient $\nabla f(\vec{a})$ is a vector that points along the direction of steepest increase of f (at the point $\vec{a}$).

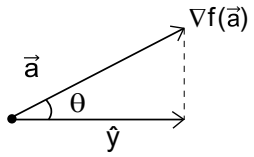


Figure 6.4. The gradient of f is a vector that points in the direction of maximum rate of change in f. Note: θ has been changed to ϕ in the main text.

6.2.1. Interpretation as Gradient Flow. Let X be a vector space (e.g. $\mathbb{R}^n$) and $f : X \to \mathbb{R}$ a smooth mapping. The gradient flow (or steepest descent curve) is a smooth curve $x : \mathbb{R}_+ \to X$ ($\mathbb{R}_+ \equiv \{t \in \mathbb{R} | t \geq 0\}$) such that

$$x'(t) = -\nabla f(x(t)),$$

where $x'(t) = \mathrm{d}x(t)/\mathrm{d}t$. Let f be a convex function. The equilibrium points ($x'(t) = 0$) of the gradient flow are the zeros of ∇f, which are also the minimizers of f.

The solution $x(t)$ of this differential equation is obtained as function of t subject to some initial condition $x(0)$. The forward Euler discretization of the gradient flow with step size $\lambda > 0$ leads to

$$\frac{x^{(k+1)} - x^{(k)}}{\lambda} = -\nabla f(x^{(k)})$$

Solving for the next iterate $x^{(k+1)}$ gives the scheme

$$x^{(k+1)} = x^{(k)} - \lambda \nabla f(x^{(k)})$$

known as the standard gradient descent iteration with step size λ. Thus, the gradient descent method can be interpreted as the forward Euler method for numerical integration applied to the gradient flow.

6.2.2. Proximal Point Method. Convergence of the forward Euler method depends on the proper selection of the step size λ. In order to get rid of the ill-conditioning of the forward step method, an alternative is the backward Euler approximation which may be done by a slight change of the above equation, i.e., by writing

$$\frac{x^{(k+1)} - x^{(k)}}{\lambda} = -\nabla f(x^{(k+1)}).$$

This method is known to have better approximation properties than forward Euler, especially for differential equations that converge, as the gradient flow does. Its main disadvantage is that it cannot be rewritten as an iteration

that gives $x^{(k+1)}$ explicitly in terms of $x^{(k)}$. For this reason, it is called an implicit method, in contrast to explicit methods like forward Euler.

To find $x^{(k+1)}$ we solve the equation

$$x^{(k+1)} + \lambda \nabla f(x^{(k+1)}) = x^{(k)}.$$

To solve it, one writes

$$x^{(k+1)} = (I + \lambda_k \nabla f)^{-1} x^{(k)},$$

and we replace ∇f by an operator F

$$x^{(k+1)} = (I + \lambda_k F)^{-1} x^{(k)},$$

where $\{\lambda_k\}$ is a sequence of positive real numbers. This is known as the *proximal point method*. The difficulty associated with the proximal point algorithm is due to the inverse operation $(I + \lambda F)^{-1}$. A common approach is to split F into two operators A and B such that $F = A + B$ and $(I + \lambda A)$ and $(I + \lambda B)$ are easily inverted.

For more information about proximal methods, see:

`https://web.stanford.edu/~boyd/papers/pdf/prox_algs.pdf`

6.3. Stochastic Gradient Descent (SGD) Method

A variant of gradient descent is the stochastic gradient method. In the update rule for gradient descent, $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \lambda \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)})$, we recall that $\chi^2(\boldsymbol{\theta}^{(k)})$ is a norm that measures the difference between data and model:

$$\chi^2(\boldsymbol{\theta}^{(k)}) = \sum_{i=1}^{n} \frac{\|y_i - y(x_i|\boldsymbol{\theta}^{(k)})\|^2}{\sigma_i^2} = \sum_{i=1}^{n} (\tilde{r}_i^{(k)})^2.$$

Substituting into the update rule we have

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \lambda \sum_{i=1}^{n} \nabla (\tilde{r}_i^{(k)})^2 = \boldsymbol{\theta}^{(k)} - \lambda \left(\nabla_{\boldsymbol{\theta}} (\tilde{r}_1^{(k)})^2 + \cdots + \nabla_{\boldsymbol{\theta}} (\tilde{r}_n^{(k)})^2 \right).$$

As you can see from the linearity of the gradient, the term $\nabla_{\boldsymbol{\theta}} \chi^2(\theta^{(k)})$ is the same as repeated (sequential) additions of $\nabla_{\boldsymbol{\theta}} (\tilde{r}_i^{(k)})^2$. We can take the extreme case and add them one at a time:

`for` $i = 1$ `to` n `do:`

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \lambda \nabla_{\boldsymbol{\theta}} (\tilde{r}_i^{(k)})^2.$$

`end for`

(where the samples are shuffled randomly prior to the `for` loop)

It can be shown that this method also converges to a local extremum, if it exists. Stochastic gradient descent is a popular algorithm for training a wide range of models in machine learning (ML). In ML λ is called the *learning*

rate. When using during backpropagation, it is one of the best algorithms for training artificial neural networks.

Normally, one uses a compromise between computing the true gradient and the gradient in a single sample to compute the gradient against more than one training samples (called a "mini-batch") at each step. For example:

`for` $i = 1$ `to` $[n/m]$ `do:`

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \lambda \sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} (\tilde{r}_i^{(k)})^2.$$

`end for`

(where the samples are shuffled randomly prior to the `for` loop)

This can perform significantly better than the stochastic gradient descent whose steps are computed one sample at a time. It may also result in smoother convergence, as the gradient computed at each step is averaged over more training examples.

More generally speaking, one has a *loss function* $l(x)$, $x \in \mathbb{R}^n$ (denotes model parameters), which is not required to be of the same form as χ^2. The simplest algorithm to solve the smooth problem

$$\min_{x \in \mathbb{R}^n} l(x)$$

is the gradient descent method $x^{(k+1)} = x^{(k)} - h\nabla l(x^{(k)})$, where $h > 0$ is the step size and $k = 0, 1, \ldots$ is the iteration number. The gradient descent is an explicit Euler discretization of the gradient flow $\dot{x} = -\nabla l(x)$, where $x = x(t)$. This deterministic minimization problem is replaced by the stochastic counterpart:

$$\min_{x \in \mathbb{R}^n} \mathbb{E}[l(x; \omega)],$$

where $\omega \in \Omega$ denotes the realization of a random process. We may view the ω's training data, $\{\omega_1, \ldots, \omega_N\}$ so that $l_i(x) \equiv l(x; \omega_i)$ is a random variable. We may invoke the LLN to approximate the above expectation value by the empirical (arithmetic) average:

$$\bar{l}(x) \equiv \frac{1}{N} \sum_{i=1}^{N} l_i(x),$$

which is exact when $N \to \infty$. Thus, instead of computing

$$\overline{\nabla l}(x) = \frac{1}{N} \sum_{i=1}^{N} \nabla l_i(x),$$

which may not be feasible, at each iteration of the algorithm we sample a "minibatch" $\mathcal{B}$ of size S, drawn uniformly at random (without replacement) from an index set $\{1, \ldots, N\}$ and compute the so-called *stochastic gradient*

given by

$$\boxed{\tilde{\nabla} l(x) \equiv \frac{1}{S} \sum_{i \in \mathcal{B}} \nabla l_i(x).}$$

Note that when $S = N$ the stochastic gradient becomes the true gradient of the empirical loss. Importantly, when the dataset is very large, i.e. $S \ll N$ and $N \to \infty$, the central limit theorem states that

$$\tilde{\nabla} l(x) = \overline{\nabla l}(x) + \xi(x)$$

where $\xi(x) \sim \mathcal{N}(0, \Sigma(x))$. Thus, the stochastic gradient is an unbiased estimator of the true gradient of the empirical loss.

6.3.1. Line search Method. In the update rule

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \lambda \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)}),$$

there is no general prescription for the value of λ, the learning rate. $\mathbf{d}_k = \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)})$ is called the descent direction. For a given descent direction, λ is chosen such that $\chi^2(\boldsymbol{\theta}^{(k)} + \lambda \mathbf{d}_k)$ is minimized over some range $\lambda \in [\lambda_{min}, \lambda_{max}]$. Using this optimal value for λ, we compute the update rule

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \hat{\lambda} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)}),$$

and keep iterating until convergence. $\hat{\lambda}$ denotes the optimal value of λ obtained from the line search. This line search is repeated at each iteration (for all k).

6.4. Random Search Method

The random optimization approach, as applied to the problem

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^n} f(\boldsymbol{\theta})$$

where f is a differentiable function involves sampling a point $\mathbf{y}$ randomly around the current position $\boldsymbol{\theta}$ (in accordance to a Gaussian distribution) and move to $\mathbf{y}$ if $f(\mathbf{y}) < f(\boldsymbol{\theta})$. This method is discussed in the following publications:

- C. Dorea, Expected number of steps of a random optimization method, JOTA, 39(1983), pp.165–171.
- J. Matyas, Random optimization. Automation and Remote Control, 26 (1965), pp. 246–253
- M. Sarma, On the convergence of the Baba and Dorea random optimization methods, JOTA, 66 (1990), pp. 337–343.

An improvement of this method is discussed in the following publication:

- B. Polyak, Introduction to Optimization. Optimization Software - Inc., Publications Division, New York, 1987.

In particular, it was mentioned that the scheme

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - h_k \frac{f(\boldsymbol{\theta}^{(k)} + \mu_k u) - f(\boldsymbol{\theta}^{(k)})}{\mu_k} u,$$

where u is a random vector distributed uniformly over the unit sphere and converges under assumption $\mu_k \to 0$. However, no explicit rules for choosing the parameters were given, and no particular rate of convergence was established. It appears that the most powerful version of this scheme corresponds to $\mu_k \to 0$. Then we get the following process:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - h_k f'(\boldsymbol{\theta}^{(k)}, u) u,$$

where $f'(\boldsymbol{\theta}, u)$ is a directional derivative of the function $f(\boldsymbol{\theta})$ along $u \in \mathbb{R}^n$. As compared with the gradient, the directional derivative is a much simpler object. Its value can be easily computed even for non-convex non-smooth functions by a forward differentiation. Or it can be approximated very well by finite differences.

6.5. Classical Momentum (CM) Method

Given an objective function $f(\boldsymbol{\theta})$ to be minimized, classical momentum (CM) method are of the form:

$$v_{k+1} = \mu v_k - \epsilon \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)})$$
$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + v_{k+1}$$

where $\epsilon > 0$ is the learning rate, $\mu \in [0, 1]$ is the momentum coefficient and $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)})$ is the gradient with respect to $\boldsymbol{\theta}^{(k)}$.

6.6. Nesterov Momentum Method

The Nesterov accelerated gradient (NAG) method converges faster than the CM method. The idea of the NAG method is that in principle, we can get a superior step direction by applying the momentum vector to the parameters before computing the gradient.

The NAG update rule is:

$$v_{k+1} = \mu v_k - \epsilon \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)} + \mu v_k)$$
$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + v_{k+1}$$

For more details on CM and NAG see:

`http://www.cs.toronto.edu/~fritz/absps/momentum.pdf`

6.7. Adaptive Gradient (AdaGrad) Method

AdaGrad is a modified stochastic gradient descent algorithm that increases the learning rate for sparser parameters and decreases the learning rate for ones that are less sparse. This strategy often improves convergence performance over standard stochastic gradient descent in settings where data is sparse and sparse parameters are more informative.

The update rule is:

$$\begin{aligned}\mathbf{g}_k &= \nabla_{\boldsymbol{\theta}^{(k-1)}} f(\boldsymbol{\theta}^{(k-1)}) \\ \mathbf{n}_k &= \mathbf{n}_{k-1} + \mathbf{g}_k^2 \\ \boldsymbol{\theta}^{(k)} &= \boldsymbol{\theta}^{(k-1)} - \lambda \frac{\mathbf{g}_k}{\sqrt{\mathbf{n}_k} + \epsilon}\end{aligned}$$

where ϵ is a fudge factor, $\boldsymbol{\theta}^{(k-1)}$ denote the parameters vector at step $k-1$, $\nabla_{\boldsymbol{\theta}^{(k-1)}}$ denote the gradient with respect to the parameters vector at the previous step and $\mathbf{n}_k$ is a norm vector.

Since the square root of a vector is not defined, the last equation is best understood in component form:

$$\boldsymbol{\theta}_i^{(k)} = \boldsymbol{\theta}_i^{(k-1)} - \lambda \frac{(\mathbf{g}_k)_i}{\sqrt{(\mathbf{n}_k)_i} + \epsilon}$$

This algorithm divides the learning rate of every step by the L_2 norm of all previous gradients ($\mathbf{g}_k^2$). This slows down learning along dimensions that have already changed significantly and speeds up learning along dimensions that have only changed slightly, stabilizing the model's representation of common features and allowing it to rapidly "catch up" its representation of rare features.

For more details, see:

`http://cs229.stanford.edu/proj2015/054_report.pdf`

6.8. RMSProp Method

One notable problem with AdaGrad is that the norm vector $\mathbf{n}$ eventually becomes so large that training slows to a halt, preventing the model from reaching the local minimum; [16] go on to motivate RMSProp, an alternative to AdaGrad that replaces the sum in $\mathbf{n}_k$ with a decaying mean parameterized here by ν. This allows the model to continue to learn indefinitely.

$$\begin{aligned}\mathbf{g}_k &= \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k-1)}) \\ \mathbf{n}_k &= \nu \mathbf{n}_{k-1} + (1-\nu)\mathbf{g}_k^2 \\ \boldsymbol{\theta}^{(k)} &= \boldsymbol{\theta}^{(k-1)} - \lambda \frac{\mathbf{g}_k}{\sqrt{\mathbf{n}_k} + \epsilon}\end{aligned}$$

6.9. Adaptive Moment Estimation (Adam) Method

Adam combines the momentum-based and norm-based (AdaGrad, RMSProp) methods to provide the advantages of both. More specifically, Adam combines CM (using a decaying mean instead of decaying sum) with RMSProp.

The steps of an Adam iteration are:

(1) $\mathbf{g}_k = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k-1)})$
(2) $\mathbf{m}_k = \mu \mathbf{m}_{k-1} + (1-\mu)\mathbf{g}_k$
(3) $\hat{\mathbf{m}}_k = \frac{\mathbf{m}_k}{1-\mu^k}$
(4) $\mathbf{n}_k = \nu \mathbf{n}_{k-1} + (1-\nu)\mathbf{g}_k^2$
(5) $\hat{\mathbf{n}}_k = \frac{\mathbf{n}_k}{1-\nu^k}$
(6) $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(k-1)} - \lambda \frac{\hat{\mathbf{m}}_k}{\sqrt{\hat{\mathbf{n}}_k}+\epsilon}$

6.10. AdaMax

The L_2 norm can be replaced by the L_∞ norm, eliminating the need for $\hat{\mathbf{n}}_k$. The updates are:

(1) $\mathbf{g}_k = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k-1)})$
(2) $\mathbf{m}_k = \mu \mathbf{m}_{k-1} + (1-\mu)\mathbf{g}_k$
(3) $\hat{\mathbf{m}}_k = \frac{\mathbf{m}_k}{1-\mu^k}$
(4) $\mathbf{n}_k = \max(\nu \mathbf{n}_{k-1}, |\mathbf{g}_k|)$
(5) $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(k-1)} - \lambda \frac{\hat{\mathbf{m}}_k}{\mathbf{n}_k+\epsilon}$

6.11. Non-Linear Conjugate Gradient (NCG) Method

Let $\mathbf{A}$ be a symmetric $n \times n$ matrix (i.e. $\mathbf{A}^T = \mathbf{A}$). We define an inner product of two vectors $\mathbf{d}_0, \mathbf{d}_1 \in \mathbb{R}^n$ with respect to $\mathbf{A}$ as follows:

$$\langle \mathbf{d}_0, \mathbf{d}_1 \rangle_A \equiv \mathbf{d}_0^T \mathbf{A} \mathbf{d}_1 = \mathbf{d}_1^T \mathbf{A} \mathbf{d}_0,$$

where the last equality follows by symmetry. Let $f(\boldsymbol{\theta})$ be a function of n variables to minimize. Its gradient $\nabla f(\boldsymbol{\theta})$ is the direction of maximum increase. Let $\boldsymbol{\theta}^{(0)}$ be the starting position. We take the first step in the opposite (steepest descent) direction:

$$\mathbf{d}_0 = -\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(0)}),$$

with a step size (learning rate) λ that is chosen by performing a line search in this direction until it reaches a minimum of f:

$$\lambda_0 = \arg\min_{\alpha} f(\boldsymbol{\theta}^{(0)} + \lambda \mathbf{d}_0).$$

This gives the new position:

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} + \lambda_0 \mathbf{d}_0.$$

After this first iteration in the direction $\mathbf{d}_0$, we perform the following steps along a conjugate direction $\mathbf{d}_k$,

(1) Calculate the steepest descent direction

$$\mathbf{h}_i = -\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)})$$

(2) Compute ν_k, the step size using one of the following formulas:

$$\nu_k = \frac{\mathbf{h}_k^T \mathbf{h}_k}{\mathbf{h}_{k-1}^T \mathbf{h}_{k-1}}, \quad \text{or} \quad \frac{\mathbf{h}_k^T (\mathbf{h}_k - \mathbf{h}_{k-1})}{\mathbf{h}_{k-1}^T \mathbf{h}_{k-1}}, \quad \text{or} \quad \frac{\mathbf{h}_k^T (\mathbf{h}_k - \mathbf{h}_{k-1})}{-\mathbf{d}_{k-1}^T (\mathbf{h}_k - \mathbf{h}_{k-1})},$$

$$\text{or} \quad \frac{\mathbf{h}_k^T \mathbf{h}_k}{-\mathbf{d}_{k-1}^T (\mathbf{h}_k - \mathbf{h}_{k-1})}.$$

(These formulas were proposed by different authors.)

(3) Obtain the conjugate direction

$$\mathbf{d}_k = \mathbf{h}_k + \nu_k \mathbf{d}_{k-1}$$

(4) Perform a line search

$$\lambda_k = \arg\min_{\lambda} f(\boldsymbol{\theta}^{(k)} + \lambda \mathbf{d}_k)$$

(5) Update the position

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \lambda_i \mathbf{d}_k.$$

This sequence of steps is repeated (iterated) until convergence. Search directions lose conjugacy over time, requiring the search direction to be reset to the steepest descent direction after some time. However, resetting too often would turn the method into steepest descent. The algorithm stops when it finds the minimum, determined when no progress is made after a direction reset (i.e. in the steepest descent direction), or when some tolerance criterion is reached.

6.12. Newton Method

Near a minimum $\boldsymbol{\theta}^{(k)}$, χ^2 looks parabolic whereas *at* the minimum we have $\nabla_{\boldsymbol{\theta}} \chi^2 = 0$. Let us proceed as we did for the Newton-Raphson method (see

equation 6.1) by Taylor expanding[1] the gradient of χ^2 near the minimum:

$$\nabla_{\boldsymbol{\theta}}\chi^2(\boldsymbol{\theta}^{(k)}+\mathbf{h}_k) = \nabla_{\boldsymbol{\theta}}\chi^2(\boldsymbol{\theta}^{(k)}) + \mathbf{h}_k \cdot \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\chi^2(\boldsymbol{\theta}^{(k)}) = 0,$$

where $\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\chi^2(\boldsymbol{\theta}^{(k)})$ is the $p \times p$ Hessian matrix. We abbreviate it here as $\mathbf{H}_k$. Its matrix elements are:

$$\mathbf{H}_k \equiv \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\chi^2(\boldsymbol{\theta}^{(k)}) = \begin{bmatrix} \frac{\partial^2\chi^2}{\partial\theta_1^2} & \cdots & \frac{\partial^2\chi^2}{\partial\theta_1\partial\theta_p} \\ \vdots & & \vdots \\ \frac{\partial^2\chi^2}{\partial\theta_p\partial\theta_1} & \cdots & \frac{\partial^2\chi^2}{\partial\theta_p^2} \end{bmatrix},$$

where $\chi^2 = \chi^2(\boldsymbol{\theta}^{(k)})$. We must therefore solve the system of equations

$$\mathbf{H}_k\mathbf{h}_k = -\nabla_{\boldsymbol{\theta}}\chi^2(\boldsymbol{\theta}^{(k)}),$$

for $\mathbf{h}_k$, for example, using the method of LU factorization. Here, both $\mathbf{h}_k$ and $-\nabla_{\boldsymbol{\theta}}\chi^2(\boldsymbol{\theta}^{(k)})$ are column vectors with p rows. If the Hessian matrix is invertible,

$$\mathbf{h}_k = -\mathbf{H}_k^{-1}\nabla_{\boldsymbol{\theta}}\chi^2(\boldsymbol{\theta}^{(k)}).$$

The update rule $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}_k + \mathbf{h}_k$ is:

$$\boxed{\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \mathbf{H}_k^{-1}\nabla_{\boldsymbol{\theta}}\chi^2(\boldsymbol{\theta}^{(k)}).} \qquad \text{(Newton)}$$

This requires calculating and inverting a Hessian matrix. If $\mathbf{H}_k = \mathbf{1}/\lambda$ ($\mathbf{1}$ is a $p \times p$ unit matrix) this method is equivalent to the steepest descent. $\mathbf{H}_k$ is also called the curvature matrix because second partial derivatives, $\partial_i^2\chi^2$, are indicative of the curvature of the function (χ^2 here) along each direction i.

6.13. Gauss-Newton Method

In the Newton method the update rule is:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \mathbf{H}_k^{-1}\nabla\chi^2(\boldsymbol{\theta}^{(k)}) \qquad \text{(Newton)}$$

It is a second-order method because it involves computing the Hessian (matrix of second partial derivatives). We must also invert the Hessian matrix. The Gauss-Newton method is also a "second order" method, but a simpler

[1] We made use of $f(x+h) = f(x) + h \cdot \nabla f(x) + O(|h|^2)$, where $f(x) = \nabla\chi^2(x)$ and $x, h \in \mathbb{R}^p$. We also note that

$$f(x+h) = f(x) + h \cdot \nabla f(x) + \frac{1}{2}h \cdot \nabla\nabla f(x) \cdot h + O(|h|^3),$$

The necessary condition for a minimum is $\nabla_h f(x+h) = 0$. This gives:

$$\nabla_h(f(x) + h \cdot \nabla f(x) + \frac{1}{2}h \cdot \nabla\nabla f(x) \cdot h) = \nabla f(x) + \nabla\nabla f(x) \cdot h = 0.$$

Solving for h gives $h = -(\nabla\nabla f(x))^{-1}\nabla f(x)$.

form of the Hessian matrix is used. Let

$$\chi^2 = \sum_{i=1}^{n} \tilde{R}_i^2, \qquad \tilde{R}_i = \frac{y_i - y(x_i|\boldsymbol{\theta})}{\alpha_i} \sim \mathcal{N}(0,1)$$

where n is the total number of data points. Define the $n \times p$ Jacobian matrix:

$$\mathbf{J}_k \equiv \mathbf{J}(\boldsymbol{\theta}^{(k)}) = \begin{bmatrix} \frac{\partial \tilde{R}_1}{\partial \theta_1} & \cdots & \frac{\partial \tilde{R}_1}{\partial \theta_p} \\ \vdots & & \vdots \\ \frac{\partial \tilde{R}_n}{\partial \theta_1} & \cdots & \frac{\partial \tilde{R}_n}{\partial \theta_p} \end{bmatrix}.$$

Now, the gradient of χ^2 is

$$(\nabla_{\boldsymbol{\theta}} \chi^2)_j \equiv \frac{\partial \chi^2}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \tilde{R}_i^2 = 2 \sum \frac{\partial \tilde{R}_i}{\partial \theta_j} \tilde{R}_i,$$

which can be written compactly as $\nabla_{\boldsymbol{\theta}} \chi^2 = \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{n} \tilde{R}_i^2 = 2\mathbf{J}^T \vec{\tilde{R}}$, where $\vec{\tilde{R}}$ is the vector of normalized residuals (one entry for each of the n data points):

$$\vec{\tilde{R}} \equiv \begin{bmatrix} \tilde{R}_1 \\ \tilde{R}_2 \\ \vdots \\ \tilde{R}_n \end{bmatrix}.$$

The Hessian matrix is therefore:

$$(\mathbf{H}_k)_{jm} \equiv \frac{\partial^2 \chi^2}{\partial \theta_j \partial \theta_m} = \frac{\partial^2}{\partial \theta_j \partial \theta_m} \sum_i \tilde{R}_i^2 = 2 \sum_i \frac{\partial \tilde{R}_i}{\partial \theta_j} \frac{\partial \tilde{R}_i}{\partial \theta_m} + \underbrace{2 \sum_i \tilde{R}_i \frac{\partial^2 \tilde{R}_i}{\partial \theta_j \partial \theta_m}}_{\text{neglect}}.$$

where the last term is neglected since $\tilde{R}_i$ is small ($O(1)$) and normally distributed with mean 0 and variance 1. This yields $\mathbf{H}_k \approx 2\mathbf{J}_k^T \mathbf{J}_k$, which is easier to evaluate than the full Hessian. This has the advantage of requiring fewer steps to compute. The final update rule is:

$$\boxed{\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - (2\mathbf{J}_k^T \mathbf{J}_k)^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)})} \qquad \text{(Gauss-Newton)}$$

The disadvantage of second-order methods is they don't work well outside parabolic surfaces. This led to the development of the Levenberg method and Marquardt-Levenberg methods.

6.14. Hessian-Free (HF) Method

The Hessian-Free (HF) method is a 2nd order optimization method that does not require calculation of the Hessian matrix. Instead one computes

so-called "Hessian-vector" products. The latter can be computed accurately by the finite differences method, or other algorithms. HF differs from Newton's method only because it is performing an incomplete optimization (via un-converged conjugate-gradient, CG) in lieu of doing a full matrix inversion. The linear CG method (as opposed to NCG) is powerful because it approximates the optimization problem by a quadratic form. The quadratic nature of the optimization problem it solves is used to iteratively generate a set of "conjugate directions" and optimize along these directions independently and exactly.

The updates of the gradient descent are not optimal for two reasons: 1) the learning rate is unspecified; 2) the directions chosen do not lead to the shortest path to the nearest extremum, as the directions may undo the work of previous iterations. The conjugate gradient method requires that our directions be conjugate to one another.

Consider the quadratic form

$$f(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{b} + c, \qquad \boldsymbol{\theta}, \mathbf{b} \in \mathbb{R}^n, c \in \mathbb{R}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric ($\mathbf{A} = \mathbf{A}^T$) and positive definite ($\mathbf{y}^T\mathbf{A}\mathbf{y} > 0$ for any $\mathbf{y} \in \mathbb{R}^n$), $\boldsymbol{\theta}^T$ denotes the transpose of the column vector θ. Here, $\boldsymbol{\theta}^T\mathbf{b}$ is the dot product $\sum_{i=1}^n \theta_i b_i$. $\boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta}$ is the quadratic form $\sum_{i,j=1}^n \theta_i A_{ij}\theta_j$. The second derivative of f, also known as the Hessian matrix, is

$$\nabla\nabla f(\boldsymbol{\theta}) = \mathbf{A}.$$

This can be seen in component form:

$$\begin{aligned}
[\nabla\nabla f(\boldsymbol{\theta})]_{ij} =& \frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\sum_{\alpha,\beta=1}^{n}\frac{1}{2}\theta_\alpha A_{\alpha\beta}\theta_\beta \\
=& \frac{\partial}{\partial\theta_i}\sum_{\alpha,\beta=1}^{n}\frac{1}{2}\left(\delta_{\alpha j}A_{\alpha\beta}\theta_\beta + \theta_\alpha A_{\alpha\beta}\delta_{\beta j}\right) \\
=& \sum_{\beta=1}^{n}\frac{1}{2}\left(A_{j\beta}\delta_{\beta i} + \delta_{\alpha i}A_{\alpha j}\right) = \frac{1}{2}(A_{ji} + A_{ij}) = A_{ij}.
\end{aligned}$$

The last equality follows because $\mathbf{A}$ is symmetric. (The second term $\boldsymbol{\theta}^T\mathbf{b}$ was not included since its second derivative with respect to $\boldsymbol{\theta}$ is zero.) The necessary condition for an extremum is $\nabla f(\boldsymbol{\theta}) = 0$:

$$\nabla f(\boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\theta} - \mathbf{b} = 0.$$

This can be verified using components:

$$\frac{\partial}{\partial \theta_i}\left(\frac{1}{2}\sum_{\alpha,\beta=1}^{n}\theta_\alpha A_{\alpha\beta}\theta_\beta - \sum_{j=1}^{n}\theta_j b_j\right) = \frac{1}{2}\sum_{\alpha,\beta=1}^{n}(\delta_{i\alpha}A_{\alpha\beta}\theta_\beta + \theta_\alpha A_{\alpha\beta}\delta_{i\beta}) - \sum_{j=1}^{n}\delta_{ij}b_j$$
$$= \frac{1}{2}(A_{i\beta}\theta_\beta + \theta_\alpha A_{\alpha i}) - b_i,$$

which leads to $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\theta} - \mathbf{b}$ because $\mathbf{A}$ is symmetric. Thus, the condition for an extremum of the quadratic form $f(\boldsymbol{\theta})$ is equivalent to solving the linear system of equations $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$.

Conjugate gradient is an iterative method whereby iterations are chosen conjugate to the previous direction. Suppose we want to find an extremum of a function f. We approximate f the quadratic form shown above. Let θ_0 be the initial position and $\mathbf{d}_0$ initial direction. The issue of the learning rate λ can be resolved if we choose λ such that

$$f(\boldsymbol{\theta}^{(0)} + \lambda\mathbf{d}_0) = \frac{1}{2}(\boldsymbol{\theta}^{(0)} + \lambda\mathbf{d}_0)^T\mathbf{A}(\boldsymbol{\theta}^{(0)} + \lambda\mathbf{d}_0) - (\boldsymbol{\theta}^{(0)} + \lambda\mathbf{d}_0)^T\mathbf{b} + c$$
$$= \frac{1}{2}\lambda^2\mathbf{d}_0^T\mathbf{A}\mathbf{d}_0 + \mathbf{d}_0^T(\mathbf{A}\boldsymbol{\theta}^{(0)} - \mathbf{b})\lambda + (\frac{1}{2}\boldsymbol{\theta}^{(0)T}\mathbf{A}\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^{(0)T}\mathbf{b} + c)$$

is an extremum. (Since f here is assumed to be a quadratic form, it has either a minimum or a maximum, and that extremum is unique.) Taking the derivative with respect to λ, setting equal to zero and solving for λ gives

$$\lambda = -\frac{\mathbf{d}_0^T(\mathbf{A}\boldsymbol{\theta}^{(0)} - \mathbf{b})}{\mathbf{d}_0^T\mathbf{A}\mathbf{d}_0}.$$

From this result, we start at $\boldsymbol{\theta}^{(0)}$ and iterate to get our first point $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} - \lambda\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(0)})$. So far, this is identical to gradient descent except that λ has been explicitly derived for the case of a quadratic form.

We have already moved in the $\mathbf{d}_0 = -\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(0)})$ direction. In the conjugate gradient method we choose the next direction to be conjugate to the previous direction. This is done by starting with the gradient of θ_1 and subtracting off anything that is related to the previous direction:

$$\mathbf{d}_1 = -\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(1)}) + \nu_0\mathbf{d}_0,$$

where the amount ν_0 is derived by requiring that $\mathbf{d}_0$ and $\mathbf{d}_1$ be conjugate, i.e. $\mathbf{d}_1^T\mathbf{A}\mathbf{d}_0 = 0$. The definition of conjugacy can be viewed as an orthogonality condition between $\mathbf{d}_0$ and $\mathbf{d}_1$, where the inner product is $\langle \mathbf{d}_0, \mathbf{d}_1\rangle_A \equiv \mathbf{d}_0^T\mathbf{A}\mathbf{d}_1 = \mathbf{d}_1^T\mathbf{A}\mathbf{d}_0$ (symmetry). Expanding $\mathbf{d}_1$ gives

$$\langle \mathbf{d}_0, \mathbf{d}_1\rangle_A = \mathbf{d}_1^T\mathbf{A}\mathbf{d}_0 = -\nabla_{\boldsymbol{\theta}} f(\theta_1)^T\mathbf{A}\mathbf{d}_0 + \nu_0\mathbf{d}_0^T\mathbf{A}\mathbf{d}_0 = 0,$$

which leads to

$$\nu_0 = \frac{\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(1)})^T\mathbf{A}\mathbf{d}_0}{\mathbf{d}_0^T\mathbf{A}\mathbf{d}_0}.$$

This procedure is done iteratively. Each next move is conjugate to the previous ones. At each iteration we choose the learning rate. The algorithm for quadratic functions $f(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta} - \mathbf{b}^T\boldsymbol{\theta} + c$ is

(1) Let $\boldsymbol{\theta}^{(0)}$ be the initial guess. Compute the initial direction as $\mathbf{d}_0 = -\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(0)})$.

(2) Find the learning rate (step size) using the equation

$$\lambda = -\frac{\mathbf{d}_i^T(\mathbf{A}\boldsymbol{\theta}^{(i)} - \mathbf{b})}{\mathbf{d_i}^T\mathbf{A}\mathbf{d}_i}.$$

This is the direction that minimizes the function $f(\boldsymbol{\theta}^{(i)} + \lambda\mathbf{d}_i)$.

(3) Update the position:

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} + \lambda\mathbf{d}_i.$$

(4) Update the direction:

$$\mathbf{d}_{i+1} = -\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(i+1)}) + \nu_i\mathbf{d}_i$$

where ν_i is given by:

$$\nu_i = \frac{\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(i+1)})^T\mathbf{A}\mathbf{d}_i}{\mathbf{d}_i^T\mathbf{A}\mathbf{d}_i}.$$

(5) Repeat steps 2-4 until n directions have been found.

The conjugate gradient method can be used to find extrema of general functions if we consider a Taylor expansion[2] of f around $\boldsymbol{\theta}$:

$$f(\boldsymbol{\theta} + \mathbf{h}) \approx f(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})^T\mathbf{h} + \mathbf{h}^T\mathbf{H}\mathbf{h},$$

where $\mathbf{H}$ is the Hessian of f evaluated at the point $\boldsymbol{\theta}$. This is a quadratic form and we can apply the algorithm as many times as needed until convergence. We note that the Hessian matrix is not needed. The quantities that are needed are Hessian-vector products, $\mathbf{H}\mathbf{v}$ where $\mathbf{v}$ is a vector. Notice that $\mathbf{H}\mathbf{v} = \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} f \cdot \mathbf{v}$:

$$[\mathbf{H}\mathbf{v}]_i = \frac{\partial}{\partial\theta_i}\sum_{j=1}^{n}\frac{\partial f}{\partial\theta_j}(\boldsymbol{\theta})v_j = [\nabla_{\boldsymbol{\theta}} D_{\mathbf{v}} f(\boldsymbol{\theta})]_i$$

where $D_{\mathbf{v}} f(\boldsymbol{\theta})$ is the directional derivative of f along $\mathbf{v}$. Thus

$$\mathbf{H}\mathbf{v} = \nabla_{\boldsymbol{\theta}}\lim_{\epsilon\to 0}\frac{f(\boldsymbol{\theta} + \epsilon\mathbf{v}) - f(\boldsymbol{\theta})}{\epsilon} = \lim_{\epsilon\to 0}\frac{\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta} + \epsilon\mathbf{v}) - \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})}{\epsilon}.$$

---

[2] Comparing

$$f(\boldsymbol{\theta} + \mathbf{h}) \approx f(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})^T\mathbf{h} + \mathbf{h}^T\mathbf{H}\mathbf{h},$$

with

$$g(\mathbf{h}) = \frac{1}{2}\mathbf{h}^T\mathbf{A}\mathbf{h} - \mathbf{h}^T\mathbf{b} + c, \qquad \mathbf{h}, \mathbf{b} \in \mathbb{R}^n, c \in \mathbb{R}$$

We have the correspondence $\mathbf{H} = \frac{1}{2}\mathbf{A}$, $\mathbf{b} = -\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$ and $f(\boldsymbol{\theta}) = c$.

This latest result is important because evaluation of the Hessian matrix is avoided and reduced to the calculation of Hessian-vector product, which involves two gradient calculations, $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta} + \epsilon \mathbf{v})$ and $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$. This method, called Hessian-Free (HF), and was proposed by James Martens for use in deep learning.

`https://www.cs.toronto.edu/~jmartens/docs/Deep_HessianFree.pdf`

6.15. Quasi-Newton Methods, incl. BFGS

Quasi-Newton methods can be used if the Jacobian or Hessian is unavailable or is too expensive to compute at every iteration. Suppose we want to find a minimum of a function $f(\boldsymbol{\theta})$. Taylor expansion around a point $\boldsymbol{\theta}^{(k)}$:

$$f(\boldsymbol{\theta}^{(k)} + \mathbf{h}) \approx f(\boldsymbol{\theta}^{(k)}) + \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)})^T \mathbf{h} + \frac{1}{2} \mathbf{h}^T \mathbf{B} \mathbf{h},$$

where $\mathbf{B}^T = \mathbf{B}$ (symmetric) is an *approximation* to the Hessian matrix. Taking the derivative with respect to $\mathbf{h}$,

$$\frac{\partial}{\partial \mathbf{h}} f(\boldsymbol{\theta}^{(k)} + \mathbf{h}) \approx \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)}) + \mathbf{B}\mathbf{h},$$

which, when setting this equal to zero gives $\mathbf{h} = -\mathbf{B}^{-1} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)})$. The Hessian approximation is chosen to satisfy $\frac{\partial}{\partial \mathbf{h}} f(\boldsymbol{\theta}^{(k)} + \mathbf{h}) = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)}) + \mathbf{B}\mathbf{h}$. It is customary to start with $\mathbf{B}_0 = const \times I$ (I: unit matrix). Updates $\mathbf{B}_{k+1}$ are chosen close to $\mathbf{B}_k$ in some norm, $\mathbf{B}_{k+1} = \arg\min_{\mathbf{B}} \|\mathbf{B} - \mathbf{B}_k\|_{\mathbf{A}}$, where $\mathbf{A}$ is a positive definite matrix that defines the norm.

Starting from a point $\boldsymbol{\theta}^{(0)}$ the following steps are taken:

(1) Compute the step update

$$\mathbf{h}_k = -\lambda_k \mathbf{B}_k^{-1} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)}),$$

where λ_k is a step size. There is no rule for choosing λ_k. You can fix the step size or do a line search.

(2) Compute the next position

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \mathbf{h}_k$$

(3) Compute the gradient at the new position $\nabla f(\boldsymbol{\theta}^{(k+1)})$ and

$$\mathbf{y}_k = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k+1)}) - \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)})$$

(4) Update the the approximate Hessian $\mathbf{B}_{k+1}$ or directly its inverse $\mathbf{B}_{k+1}^{-1}$ using the Sherman-Morrison formula (see table below).

The most popular update formulas are:

Method	$\mathbf{B}_{k+1}$	$\mathbf{A}_{k+1} = \mathbf{B}_{k+1}^{-1}$
BFGS	$\mathbf{B}_k + \frac{\mathbf{y}_k\mathbf{y}_k^T}{\mathbf{y}_k^T h_k} - \frac{\mathbf{B}_k\mathbf{h}_k(\mathbf{B}_k\mathbf{h}_k)^T}{\mathbf{h}_k^T\mathbf{B}_k\mathbf{h}_k}$	$\left(I - \frac{\mathbf{h}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{h}_k}\right)\mathbf{A}_k\left(I - \frac{\mathbf{y}_k\mathbf{h}_k^T}{\mathbf{y}_k^T\mathbf{h}_k}\right) + \frac{\mathbf{h}_k\mathbf{h}_k^T}{\mathbf{y}_k^T\mathbf{h}_k}$
Broyden	$\mathbf{B}_k + \frac{\mathbf{y}_k - \mathbf{B}_k\mathbf{h}_k}{\mathbf{h}_k^T\mathbf{h}_k}\mathbf{h}_k^T$	$\mathbf{A}_k + \frac{(\mathbf{h}_k - \mathbf{A}_k\mathbf{y}_k)\mathbf{h}_k^T\mathbf{A}_k}{\mathbf{h}_k^T\mathbf{A}_k\mathbf{y}_k}$
BFGS/DFP	$(1-\varphi_k)\mathbf{B}_{k+1}^{\text{BFGS}} + \varphi_k\mathbf{B}_{k+1}^{\text{DFP}}, \quad \varphi \in [0,1]$	
DFP	$\left(I - \frac{y_k\mathbf{h}_k^T}{\mathbf{y}_k^T\mathbf{h}_k}\right)\mathbf{B}_k\left(I - \frac{\mathbf{h}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{h}_k}\right) + \frac{\mathbf{y}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{h}_k}$	$\mathbf{A}_k + \frac{\mathbf{h}_k\mathbf{h}_k^T}{\mathbf{h}_k^T\mathbf{y}_k} - \frac{\mathbf{A}_k\mathbf{y}_k\mathbf{y}_k^T\mathbf{A}_k}{\mathbf{y}_k^T\mathbf{A}_k\mathbf{y}_k}$
SR1	$\mathbf{B}_k + \frac{(\mathbf{y}_k - \mathbf{B}_k\mathbf{h}_k)(\mathbf{y}_k - \mathbf{B}_k\mathbf{h}_k)^T}{(\mathbf{y}_k - \mathbf{B}_k\mathbf{h}_k)^T\mathbf{h}_k}$	$\mathbf{A}_k + \frac{(\mathbf{h}_k - \mathbf{A}_k\mathbf{y}_k)(\mathbf{h}_k - \mathbf{A}_k\mathbf{y}_k)^T}{(\mathbf{h}_k - \mathbf{A}_k\mathbf{y}_k)^T\mathbf{y}_k}$

6.15.1. Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. The most popular quasi-Newton method is BFGS. As mentioned earlier, Newton methods obtain the search direction at stage k by solving the equation

$$\mathbf{H}_k\mathbf{d}_k = -\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)}).$$

Quasi-Newton methods avoid working with the Hessian matrix $\mathbf{H}_k$ directly. Instead one uses $\mathbf{B}_k$, an approximation to the Hessian matrix:

$$\mathbf{B}_k\mathbf{d}_k = -\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)}).$$

A line search is used to obtain the next point $\boldsymbol{\theta}^{(k+1)}$ by minimizing $f(\boldsymbol{\theta}^{(k)} + \lambda\mathbf{d}_k)$ over the scalar $\lambda > 0$. The quasi-Newton condition is:

$$\mathbf{B}_k(\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}) = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k+1)}) - \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)}).$$

If we set $\mathbf{y}_k = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k+1)}) - \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)})$ and $\mathbf{h}_k = \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}$, then $\mathbf{B}_{k+1}\mathbf{h}_k = \mathbf{y}_k$. For $\mathbf{B}_{k+1}$ to be positive definite we need $\mathbf{h}_k^T\mathbf{B}_{k+1}\mathbf{h}_k = \mathbf{h}_k\mathbf{y}_k > 0$. This condition on $\mathbf{B}_{k+1}$ is called the convexity condition, since the Hessian matrix deals with curvature.

The approximate Hessian is updated by adding two matrices: $\mathbf{B}_{k+1} = \mathbf{B}_k + U_k + V_k$. To maintain symmetry and positive definitiveness of $\mathbf{B}_{k+1}$ we choose $\mathbf{B}_{k+1} = \mathbf{B}_k + \alpha\mathbf{u}\mathbf{u}^T + \beta\mathbf{v}\mathbf{v}^T$. Imposing the condition $\mathbf{B}_{k+1}\mathbf{h}_k = \mathbf{y}_k$ (ensuring that $\mathbf{B}_{k+1}$ is positive definite) and choosing $\mathbf{u} = \mathbf{y}_k$ and $\mathbf{v} = \mathbf{B}_k\mathbf{h}_k$, i.e.,

$$\mathbf{B}_{k+1}\mathbf{h}_k = \mathbf{B}_k\mathbf{h}_k + \alpha\mathbf{y}_k\mathbf{y}_k^T\mathbf{h}_k + \beta\mathbf{B}_k\mathbf{h}_k\mathbf{h}_k^T\mathbf{B}_k^T\mathbf{h}_k = \mathbf{y}_k,$$

we find $\alpha = \frac{1}{\mathbf{y}_k^T\mathbf{h}_k}$ and $\beta = -\frac{1}{\mathbf{h}_k^T\mathbf{B}\mathbf{h}_k}$. Substituting these into the equation for $\mathbf{B}_{k+1}$ we get the BFGS update rule:

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{\mathbf{y}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{h}_k} - \frac{\mathbf{B}_k\mathbf{s}_k\mathbf{h}_k^T\mathbf{B}_k^T}{\mathbf{h}_k^T\mathbf{B}_k\mathbf{h}_k}. \tag{6.2}$$

Starting from an initial position $\boldsymbol{\theta}^{(0)}$ and approximate Hessian $\mathbf{B}_0$ that is positive definite (can be the unit matrix) the algorithm is:

(1) Choose the k-th direction by solving

$$\mathbf{B}_k\mathbf{d}_k = -\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)})$$

(2) Do a line search to get the step size

$$\lambda_k = \arg\min_{\lambda} f(\boldsymbol{\theta}^{(k)} + \lambda \mathbf{d}_k)$$

(3) Set $\mathbf{h}_k = \lambda_k \mathbf{d}_k$ and obtain the next position:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \mathbf{h}_k$$

(4) Compute the update to the approximate Hessian using the BFGS rule (Equation 6.2), with $\mathbf{y}_k = \nabla_{\boldsymbol{\theta}} f(\theta_{k+1}) - \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)})$.

In the first step we solve the system of equations $\mathbf{B}_k \mathbf{d}_k = -\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)})$ for $\mathbf{d}_k$. Alternatively we can use the inverse, $\mathbf{B}_k^{-1}$, and compute $\mathbf{d}_k = -\mathbf{B}_k^{-1} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)})$ direct. In that case, we update the inverse of the approximate Hessian. The rule for this can be derived using the Sherman-Morrison formula. The end result is:

$$\mathbf{B}_{k+1}^{-1} = \left(I - \frac{\mathbf{s}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right) \mathbf{B}_k^{-1} \left(I - \frac{\mathbf{y}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right) + \frac{\mathbf{s}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}.$$

6.16. Levenberg Method

The steepest descent method works best on flat terrain (away from the minimum) whereas the Newton method works best near the minimum, which is well approximated by a parabola. Levenberg proposed an improved update rule which interpolates between the two methods:

$$\boxed{\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - (\mathbf{H}_k + \mu I)^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)}),} \qquad \text{(Levenberg)}$$

which is a modified Newton method. I is a $p \times p$ unit matrix. The term μI is a regularization term and μ is called "trust-region parameter". When μ is large (denoting $\mu^{-1} = \tilde{\mu}$), $\|\mathbf{H}_k\| \ll \|\mu I\|$,[3]

$$\begin{aligned} \boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - (\mathbf{H}_k + \mu I)^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)}) &\approx \boldsymbol{\theta}^{(k)} - (\mu I)^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)}) \\ &= \boldsymbol{\theta}^{(k)} - \tilde{\mu} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)}), \end{aligned} \tag{6.3}$$

we recover the steepest descent method. On the other hand, when μ is small, $\|\mathbf{H}_k\| \gg \|\mu I\|$, and

$$\begin{aligned} \boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - (\mathbf{H}_k + \mu I)^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)}) &\approx \boldsymbol{\theta}^{(k)} - (\mathbf{H}_k)^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)}) \\ &= \boldsymbol{\theta}_k - \mathbf{H}_k^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)}), \end{aligned} \tag{6.4}$$

we recover the Newton method. Far from the minimum, we want to use large values of μ whereas close to the minimum we want to use small values of μ. The parameter μ is adjusted at each iteration. We stop iterating when χ^2 does not change appreciably.

[3]Here $\|\mathbf{A}\|$ denotes the norm of the matrix $\mathbf{A}$. Any suitable norm can be used. For example, it can be the largest of all matrix entires: $\|\mathbf{A}\| = \max_{ij} |A_{ij}|$.

6.17. Marquardt-Levenberg Method

We recall the Levenberg update rule:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - (\mathbf{H}_k + \mu I)^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)}), \qquad \text{(Levenberg)}$$

which interpolates between steepest descent and the Newton method by way of the regularization term, μI. Indeed, for large λ we have the steepest descent method and for small μ we recover the Newton method. Far from the minimum, we want to use large values of μ whereas close to the minimum we want to use small values of μ.

An improvement over the Levenberg method was proposed by Marquardt in the form of a modified update rule:

$$\boxed{\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - (\mathbf{H}_k + \mu \,\mathrm{diag}[\mathbf{H}_k])^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)}).} \qquad \text{(Marquardt-Levenberg)}$$

where $\mathrm{diag}[\mathbf{H}_k]$ is the matrix $\mathbf{H}_k$ where all entries have been zeroed out except those along the diagonal:

$$\mathrm{diag}[\mathbf{H}_k] = \mathrm{diag} \begin{bmatrix} H_{11} & H_{12} & \dots & H_{1p} \\ H_{21} & H_{22} & \dots & H_{2p} \\ \vdots & \dots & \ddots & \vdots \\ H_{p1} & \dots & \dots & H_{pp} \end{bmatrix} = \begin{bmatrix} H_{11} & 0 & \dots & 0 \\ 0 & H_{22} & \dots & 0 \\ \vdots & \dots & \ddots & \vdots \\ 0 & \dots & \dots & H_{pp} \end{bmatrix}$$

where H_{ij} is the ij-th element of the matrix $\mathbf{H}_k$. The Marquardt-Levenberg method is equivalent to modifying the Hessian to $H_{jj} \to H_{jj}(1+\mu)$ and $H_{ij} \to H_{ij}$ $(i \neq j)$. For large μ, the matrix $\mathbf{H}_k + \mu \,\mathrm{diag}[\mathbf{H}_k]$ is said to be "diagonally dominant", i.e. it has the form

$$\begin{aligned} \mathbf{H}_k + \mu \,\mathrm{diag}[\mathbf{H}_k] &= \begin{bmatrix} H_{11}(1+\mu) & H_{12} & \dots & H_{1p} \\ H_{21} & H_{22}(1+\mu) & \dots & H_{2p} \\ \vdots & & \ddots & \vdots \\ H_{p1} & \dots & \dots & H_{pp}(1+\mu) \end{bmatrix} \\ &\approx (1+\mu) \begin{bmatrix} H_{11} & 0 & \dots & 0 \\ 0 & H_{22} & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & H_{pp} \end{bmatrix} \quad \text{(large } \mu\text{)} \end{aligned}$$

where $H_{ii} = \frac{\partial^2 \chi^2}{\partial \theta_{ii}^2}$ is the curvature of χ^2 surface along i-th direction. The inverse of a diagonal matrix involves taking the inverse of every diagonal

element:

$$(\mathbf{H}_k + \mu \operatorname{diag}[\mathbf{H}_k])^{-1} \approx \frac{1}{1+\mu} \begin{bmatrix} \frac{1}{H_{11}} & 0 & \dots & 0 \\ 0 & \frac{1}{H_{22}} & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & \frac{1}{H_{pp}} \end{bmatrix}.$$

This matrix then multiplies the column vector $\nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)})$ to yield the update rule:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - (\mathbf{H}_k + \mu \operatorname{diag}[\mathbf{H}_k])^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)}),$$

where for large μ, this approximates to:

$$\approx \boldsymbol{\theta}^{(k)} - (1+\mu)^{-1} \begin{bmatrix} \frac{(\vec{g}_k)_1}{H_{11}} \\ \frac{(\vec{g}_k)_2}{H_{22}} \\ \vdots \\ \frac{(\vec{g}_k)_p}{H_{pp}} \end{bmatrix} = \boldsymbol{\theta}^{(k)} - (1+\mu)^{-1} \begin{bmatrix} \frac{\partial_1 \chi^2}{\partial_{11}^2 \chi^2} \\ \frac{\partial_2 \chi^2}{\partial_{22}^2 \chi^2} \\ \vdots \\ \frac{\partial_p \chi^2}{\partial_{pp}^2 \chi^2} \end{bmatrix}$$

where we used the shorthand notations $\vec{g}_k = \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)})$, $\partial_i \chi^2 \equiv \frac{\partial \chi^2}{\partial \theta_i}$ and $\partial_{ii}^2 \chi^2 \equiv \frac{\partial^2 \chi^2}{\partial \theta_{ii}^2}$. This update rule looks like the gradient (steepest) descent method, except that each entry has been scaled in such a way that large steps are made in the direction of low curvature (flat terrain) and small steps in the direction with high curvature (steep incline). Levenberg-Marquardt is considered one of the best "local optimization" algorithms and is widely used in applications. In MATLAB, it is implemented in the function `lsqnonlin`.

6.17.1. Adjusting Trust-Region Parameter for Levenberg and Marquardt-Levenberg Methods. The parameter μ (learning rate) is adjusted at each iteration. We stop iterating when χ^2 does not change appreciably. $\mathbf{H}_k$ is called the curvature matrix. Here is a possible implementation of the Levenberg-Marquardt method (or Levenberg method):

- Pick initial guess for set of fitted parameters $\boldsymbol{\theta}^{(0)}$.
- Compute $\chi^2(\boldsymbol{\theta}^{(0)})$.
- Pick a value for μ, say $\mu = 0.1$.
- (*) Denote the current step by $k = 0, 1, \dots$. Solve for $\delta\boldsymbol{\theta}^{(k)} = -(\mathbf{H}_k + \mu \cdot \operatorname{diag}[\mathbf{H}_k])^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)})$ (Marquardt-Levenberg update rule; for the Levenberg method, use the Levenberg update rule instead) and evaluate $\chi^2(\boldsymbol{\theta}^{(k)} + \delta\boldsymbol{\theta}^{(k)})$.

- If $\chi^2(\boldsymbol{\theta}^{(k)} + \delta\boldsymbol{\theta}^{(k)}) \geq \chi^2(\boldsymbol{\theta}^{(k)})$, reject the move. Increase μ by a factor of 10 and go back to (*).
- If $\chi^2(\boldsymbol{\theta}^{(k)} + \delta\boldsymbol{\theta}^{(k)}) < \chi^2(\boldsymbol{\theta}^{(k)})$, accept the move. Update the trial solution $\boldsymbol{\theta}^{(k+1)} \leftarrow \boldsymbol{\theta}^{(k)} + \delta\boldsymbol{\theta}^{(k)}$. Increment $k \leftarrow k + 1$. Decrease μ by a factor of 10. Go back to (*). [In other words, if the total distance error of the updated parameters is less than the previous one, the updated parameters are taken to be the current parameters and μ is decreased.]
- Stopping criterion: changes in parameters that yield changes in χ^2 by an amount $\ll 1$ are not statistically meaningful.
- When finished, use the Hessian to compute the errors in the fitted parameters (see section 6.18).

The parameter μ can be initialized to be large so that first updates are small steps in the steepest-descent direction. If an iteration happens to result in a worse approximation, μ is increased. As the solution improves, μ is decreased, the Levenberg-Marquardt method approaches the Gauss-Newton method, and the solution typically accelerates to the local minimum.

For more information about this algorithm see:

- M.I.A. Lourakis. A brief description of the Levenberg-Marquardt algorithm implemented by `levmar`, Technical Report, Institute of Computer Science, Foundation for Research and Technology - Hellas, 2005.
- K. Madsen, N.B. Nielsen, and O. Tingleff. Methods for nonlinear least squares problems. Technical Report. Informatics and Mathematical Modeling, Technical University of Denmark, 2004.
- D.W. Marquardt. "An algorithm for least-squares estimation of nonlinear parameters," Journal of the Society for Industrial and Applied Mathematics, 11(2):431-441, 1963.

6.17.2. Confidence regions. If we plot contour lines of equal χ^2 (see Fig. 6.5), the interior region of these contours can be associated with the likelihood of a set of fitted parameters (random vector) lies within that contour.

6.17.3. Local optimization techniques: summary of update rules. The algorithms covered so far for nonlinear optimization are called "local optimization" techniques, because they are designed to search for the nearest minimum of χ^2. This minimum may not necessarily be the global minimum of the χ^2 surface. In subsequent lectures we will look at global optimization techniques. We summarize in Table 6.1 some of the most important update rules derived so far, for these local optimization techniques.

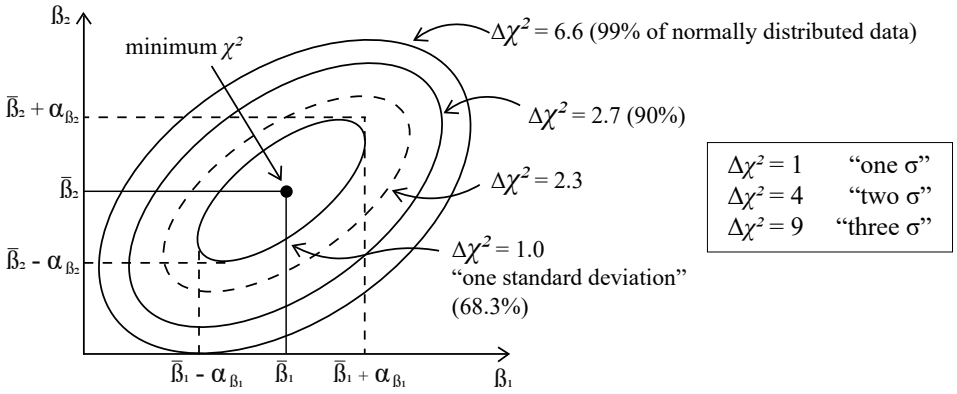


Figure 6.5. Confidence regions. Note: in this figure β should be replaced by θ.

Method	Update rule (displacement at k-th iteration)
gradient descent	$\mathbf{h}_k = -\lambda \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)})$
Newton method	$\mathbf{h}_k = -(\mathbf{H}_k)^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)})$
Gauss-Newton	$\mathbf{h}_k = -\frac{1}{2}(\mathbf{J}_k^T \mathbf{J}_k)^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)})$
Levenberg	$\mathbf{h}_k = -(\mathbf{H}_k + \mu I)^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)})$
Levenberg-Marquardt	$\mathbf{h}_k = -(\mathbf{H}_k + \mu \, \mathrm{diag}[\mathbf{H}_k])^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)})$

Table 6.1. Various update rules for non-linear local optimization.

Note: the Newton, Gauss-Newton, Levenberg and Levenberg-Marquardt methods do not have a learning rate as stated in the table. In practice, we often use a learning rate, i.e., $\mathbf{h}_k = -\lambda(\mathbf{H}_k + \mu I)^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)})$ for Levenberg and similarly for others.

6.18. Fitting Parameter Errors from Covariance Matrix

The errors in the fitted parameters $\boldsymbol{\theta}$ can be extracted from the main diagonal of the covariance matrix (omitting the superscript k in $\boldsymbol{\theta}^{(k)}$ momentarily to avoid cluttering the notation):

$$cov(\theta, \theta) \equiv \begin{bmatrix} \boxed{var(\theta_1)} & cov(\theta_1, \theta_2) & \dots & cov(\theta_1, \theta_p) \\ cov(\theta_2, \theta_1) & \boxed{var(\theta_2)} & \dots & cov(\theta_1, \theta_p) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\theta_p, \theta_1) & cov(\theta_p, \theta_2) & \dots & \boxed{var(\theta_p)} \end{bmatrix}$$

whereas the off-diagonal elements describe possible relationships (e.g. such as redundancy) among the fitting parameters. But how do we obtain the

matrix $cov(\boldsymbol{\theta}, \boldsymbol{\theta})$? It can be shown (see Chapter 8 for proof) that the covariance matrix can be obtained from the Hessian:

$$\boxed{cov(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}) = 2\mathbf{H}_k^{-1}}$$

For this to work, of course, the matrix $\mathbf{H}_k$ needs to be invertible (non-singular). This may not always be the case. (The subscript s here indicates that covariances can be monitored in real-time, at every time step of the iterative optimization process.)

6.19. Constrained Optimization

6.19.1. Method of Lagrange Multipliers.

$$df(x, y) = 0$$
$$df(x, y) \equiv \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy$$
$$\nabla f(x, y) = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y})$$

$$d\vec{r} = (dx, dy)$$
$$df = \nabla f \cdot d\vec{r} = 0$$
$$\nabla f = 0$$

For the Lagrangian

$$dL(x, \lambda) = 0$$
$$\nabla L = 0$$
$$\nabla(f - \lambda g) = \nabla f - \lambda \nabla g = 0$$
$$\nabla f = \lambda \nabla g$$

6.19.2. Inequality Constraints: The Karush-Kuhn-Tucker Conditions. In mathematical optimization, the Karush-Kuhn-Tucker (KKT) conditions, also known as the Kuhn–Tucker conditions, are first derivative tests (sometimes called first-order necessary conditions) for a solution in nonlinear programming to be optimal, provided that some regularity conditions are satisfied. Allowing inequality constraints, the KKT approach to nonlinear programming generalizes the method of Lagrange multipliers, which allows only equality constraints. Similar to the Lagrange approach, the constrained maximization (minimization) problem is rewritten as a Lagrange function whose optimal point is a saddle point, i.e. a global maximum (minimum) over the domain of the choice variables and a global minimum (maximum) over the multipliers, which is why the Karush-Kuhn-Tucker theorem is sometimes referred to as the saddle-point theorem. The KKT conditions were originally named after Harold W. Kuhn and Albert W. Tucker, who first

published the conditions in 1951. Later scholars discovered that the necessary conditions for this problem had been stated by William Karush in his master's thesis in 1939.

== Nonlinear optimization problem ==

Consider the following nonlinear [[optimization problem—minimization or maximization problem]]:

$$\text{Optimize } f(\mathbf{x})$$

subject to

$$g_i(\mathbf{x}) \leq 0,$$
$$h_j(\mathbf{x}) = 0.$$

where $\mathbf{x} \in \mathbf{X}$ is the optimization variable chosen from a convex subset of $\mathbb{R}^n$, f is the objective or utility function, g_i $(i = 1, \ldots, m)$ are the inequality constraint functions and h_j $(j = 1, \ldots, \ell)$ are the equality constraint functions. The numbers of inequalities and equalities are denoted by m and ℓ respectively. Corresponding to the constraint optimization problem one can form the Lagrangian function

$$L(\mathbf{x}, \mu, \lambda) = f(\mathbf{x}) + \mu^T \mathbf{g}(\mathbf{x}) + \lambda^T \mathbf{h}(\mathbf{x})$$

where $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_m(\mathbf{x}))^T$, $\mathbf{h}(\mathbf{x}) = (h_1(\mathbf{x}), \ldots, h_\ell(\mathbf{x}))^T$. The "Karush-Kuhn-Tucker theorem" then states the following.

Theorem. If $(\mathbf{x}^*, \mu^*)$ is a saddle point of $L(\mathbf{x}, \mu)$ in $\mathbf{x} \in \mathbf{X}$, $\mu \geq \mathbf{0}$, then $\mathbf{x}^*$ is an optimal vector for the above optimization problem. Suppose that $f(\mathbf{x})$ and $g_i(\mathbf{x})$, $i = 1, \ldots, m$, are convex in $\mathbf{x}$ and that there exists $\mathbf{x}_0 \in \mathbf{X}$ such that $\mathbf{g}(\mathbf{x}_0) < 0$. Then with an optimal vector $\mathbf{x}^*$ for the above optimization problem there is associated a non-negative vector μ^* such that $L(\mathbf{x}^*, \mu^*)$ is a saddle point of $L(\mathbf{x}, \mu)$.

Since the idea of this approach is to find a supporting hyperplane on the feasible set $\mathbf{\Gamma} = \{\mathbf{x} \in \mathbf{X} : g_i(\mathbf{x}) \leq 0, i = 1, \ldots, m\}$, the proof of the Karush-Kuhn-Tucker theorem makes use of the hyperplane separation theorem.

The system of equations and inequalities corresponding to the KKT conditions is usually not solved directly, except in the few special cases where a closed-form solution can be derived analytically. In general, many optimization algorithms can be interpreted as methods for numerically solving the KKT system of equations and inequalities.

== Necessary conditions ==

Suppose that the objective function $f : \mathbb{R}^n \to \mathbb{R}$ and the constraint functions $g_i : \mathbb{R}^n \to \mathbb{R}$ and $h_j : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable at a point $x^* \in \mathbb{R}^n$. If x^* is a local optimum and the optimization problem satisfies some regularity conditions (see below), then there exist constants μ_i $(i = 1, \ldots, m)$

and λ_j $(j = 1, \ldots, \ell)$, called KKT multipliers, such that the following four groups of conditions hold:

Stationarity

For maximizing $f(x)$:

$$\nabla f(x^*) - \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) - \sum_{j=1}^{\ell} \lambda_j \nabla h_j(x^*) = \mathbf{0}$$

For minimizing $f(x)$:

$$\nabla f(x^*) + \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) + \sum_{j=1}^{\ell} \lambda_j \nabla h_j(x^*) = \mathbf{0}$$

Primal feasibility

$$g_i(x^*) \leq 0, \text{ for } i = 1, \ldots, m$$
$$h_j(x^*) = 0, \text{ for } j = 1, \ldots, \ell$$

Dual feasibility

$$\mu_i \geq 0, \text{ for } i = 1, \ldots, m$$

Complementary slackness

$$\sum_{i=1}^{m} \mu_i g_i(x^*) = 0.$$

The last condition is sometimes written in the equivalent form:

$$\mu_i g_i(x^*) = 0, \text{ for } i = 1, \ldots, m.$$

In the particular case $m = 0$, i.e., when there are no inequality constraints, the KKT conditions turn into the Lagrange conditions, and the KKT multipliers are called Lagrange multipliers.

If some of the functions are non-differentiable, subdifferential versions of KKT conditions are available.

=== Matrix representation ===

The necessary conditions can be written with Jacobian matrices of the constraint functions. Let $\mathbf{g}(x) : \mathbb{R}^n \to \mathbb{R}^m$ be defined as $\mathbf{g}(x) = (g_1(x), \ldots, g_m(x))^T$ and let $\mathbf{h}(x) : \mathbb{R}^n \to \mathbb{R}^\ell$ be defined as $\mathbf{h}(x) = (h_1(x), \ldots, h_\ell(x))^T$. Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^T$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_\ell)^T$. Then the necessary conditions can be written as:

Stationarity

For maximizing $f(x)$:

$$\nabla f(x^*) - D\mathbf{g}(x^*)^T \boldsymbol{\mu} - D\mathbf{h}(x^*)^T \boldsymbol{\lambda} = \mathbf{0}$$

For minimizing $f(x)$:

$$\nabla f(x^*) + D\mathbf{g}(x^*)^T \boldsymbol{\mu} + D\mathbf{h}(x^*)^T \boldsymbol{\lambda} = \mathbf{0}$$

Primal feasibility

$$\mathbf{g}(x^*) \leq \mathbf{0}$$
$$\mathbf{h}(x^*) = \mathbf{0}$$

Dual feasibility

$$\boldsymbol{\mu} \geq \mathbf{0}$$

Complementary slackness

$$\boldsymbol{\mu}^T \mathbf{g}(x^*) = 0.$$

== Sufficient conditions ==

In some cases, the necessary conditions are also sufficient for optimality. In general, the necessary conditions are not sufficient for optimality and additional information is required, such as the Second Order Sufficient Conditions (SOSC). For smooth functions, SOSC involve the second derivatives, which explains its name.

The necessary conditions are sufficient for optimality if the objective function f of a maximization problem is a concave function, the inequality constraints g_j are continuously differentiable convex functions and the equality constraints h_i are affine functions. Similarly, if the objective function f of a minimization problem is a convex function, the necessary conditions are also sufficient for optimality.

It was shown by Martin in 1985 that the broader class of functions in which KKT conditions guarantees global optimality are the so-called Type 1 invex functions.

=== Second-order sufficient conditions ===

For smooth, non-linear optimization problems, a second order sufficient condition is given as follows.

The solution x^*, λ^*, μ^* found in the above section is a constrained local minimum if for the Lagrangian,

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \mu_i g_i(x) + \sum_{j=1}^{\ell} \lambda_j h_j(x)$$

then,

$$s^T \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) s \geq 0$$

where $s \neq 0$ is a vector satisfying the following,

$$[\nabla_x g_i(x^*), \nabla_x h_j(x^*)]^T \, s = 0$$

where only those active inequality constraints $g_i(x)$ corresponding to strict complementarity (i.e. where $\mu_i > 0$) are applied. The solution is a strict constrained local minimum in the case the inequality is also strict.

If $s^T \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) s = 0$, the third order Taylor expansion of the Lagrangian should be used to verify if x^* is a local minimum. The minimization of $f(x_1, x_2) = (x_2 - x_1^2)(x_2 - 3x_1^2)$ is a good counter-example, see also Peano surface.

Generalizations

With an extra multiplier $\mu_0 \geq 0$, which may be zero (as long as $(\mu_0, \mu, \lambda) \neq 0$), in front of $\nabla f(x^*)$ the KKT stationarity conditions turn into

$$\mu_0 \, \nabla f(x^*) + \sum_{i=1}^{m} \mu_i \, \nabla g_i(x^*) + \sum_{j=1}^{\ell} \lambda_j \, \nabla h_j(x^*) = 0, \tag{6.5}$$

$$\mu_j g_i(x^*) = 0, \quad i = 1, \ldots, m, \tag{6.6}$$

which are called the Fritz John conditions. This optimality conditions holds without constraint qualifications and it is equivalent to the optimality condition KKT or (not-MFCQ).

The KKT conditions belong to a wider class of the first-order necessary conditions (FONC), which allow for non-smooth functions using subderivatives.

Example 6.1. Find the maximum of $f(x, y) = xy$ subject to the constraint $g(x, y) = x^2 + y^2 = 2$. Lagrange's method instructs us to solve the system $\nabla f = \lambda \nabla g$ and $g(x, y) = 2$. The solutions are found to be $(x, y, \lambda) = (1, 1, \frac{1}{2})$, $(-1, -1, \frac{1}{2})$, $(1, -1, -\frac{1}{2})$, and $(-1, 1, -\frac{1}{2})$. The critical points are $(\pm 1, \pm 1)$ and the maximum value of $f(x, y)$ is $f(1, 1) = 1$.

This problem is simpler if we use polar coordinates $x = r \cos\theta$, $y = r \sin\theta$. The expression for f and g are $f(r, \theta) = (r \cos\theta)(r \sin\theta) = \frac{1}{2} r^2 \sin 2\theta$ and $g(r, \theta) = (r \cos\theta)^2 + (r \sin\theta)^2 = r^2$. Thus the problem reduces to finding the

maximum of $f(r,\theta) = \frac{1}{2}r^2 \sin 2\theta$, subject to the constraint $r^2 = 2$. We can do that by solving the equation $\frac{\partial f}{\partial \theta}(\sqrt{2},\theta) = 0$; that is, $2\cos 2\theta = 0$. The critical points are $(r,\theta) = (\sqrt{2}, \pi/4 + (n\pi)/2)$, where n is any integer – the same four points previously obtained.

Note that in the second solution the Lagrange multiplier λ did not appear. In polar coordinates the variable r was eliminated because it was constant on the constrained curve, reducing the problem to the unconstrained maximum problem in the remaining variable θ. This example demonstrates the advantage of thinking of the coordinate system not as an immutable quantity but as something that can be adapted to the problem.

6.19.3. Method of Differential Forms. The algebraic machinery of differential forms[4] also allows us to solve constrained optimization problems. For the special case of optimizing a function f of three variables and a single constraint, the Lagrange condition $\nabla f = \lambda \nabla g$ can be reformulated as $\nabla f \times \nabla g = 0$. The multiplier λ is avoided because two vectors in space are parallel if and only if their cross product is zero. In this special case, the components of $\nabla f \times \nabla g$, together with the constraint, provide four equations in three unknowns instead of Lagrange's four equations in four unknowns. In fact, the condition $\nabla f \times \nabla g = 0$ is slightly better than $\nabla f = \lambda \nabla g$, because points where $\nabla g = 0$ are critical points that must be included as candidates for the extremum.

The traditional statement of the Lagrange multiplier theorem considers only points where the gradients of the constraints are linearly independent, so it skirts an important consideration. The condition $\nabla f \times \nabla g = 0$ is better because it handles all possibilities simultaneously. Unfortunately, the cross product is defined only for 3D vectors, so this approach is limited in scope. However, there is a vector multiplication operation, similar to the vector cross product, called the wedge product, that removes the dimensional restriction at little cost.

If u and v are vectors, their wedge product $u \wedge v$ is a new object (sometimes called a *bivector*). The set of all linear combinations of bivectors is a vector space, and the wedge product operation has the following two properties:

(1) The wedge product is linear in each variable separately. That is, if α and β are scalars then

$$(\alpha u + \beta v) \wedge w = \alpha(u \wedge w) + \beta(v \wedge w)$$
$$u \wedge (\alpha v + \beta w) = \alpha(u \wedge v) + \beta(u \wedge w).$$

(2) The wedge product is anti-commutative: $u \wedge v = -v \wedge u$.

[4]This section is based on the paper: Zizza, F., 1998. Differential forms for constrained max-min problems: eliminating Lagrange multipliers. The College Mathematics Journal, 29(5), pp.387-396.

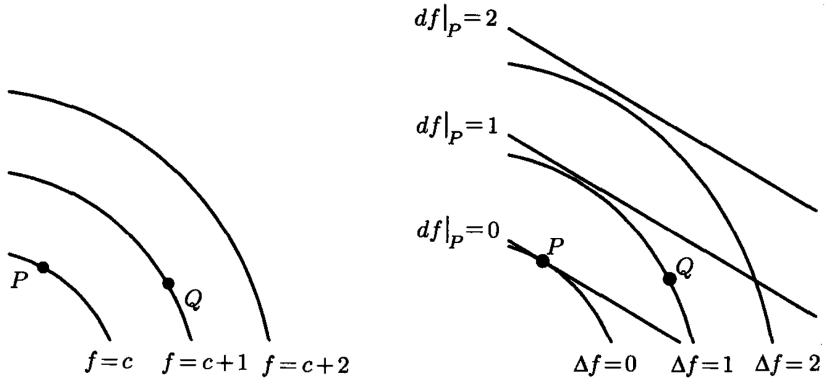
It follows that $u \wedge u = 0$ for any vector u. A crucial fact for our purpose is that $u \wedge v = 0$ if and only if the pair $\{u, v\}$ is linearly independent. These properties are satisfied by the cross product, but the bivectors are new objects, distinct from their vector factors and subject only to the conditions above. The advantage gained is that the wedge product makes sense for vectors of any dimension, not just vectors in 3D space.

Example 6.2. Let $v = \alpha\hat{x} + \beta\hat{y}$ and $w = \gamma\hat{x} + \delta\hat{y}$. Show that $u \wedge w = \begin{vmatrix} \alpha & \beta \\ \gamma & \delta \end{vmatrix} \hat{x} \wedge \hat{y}$ and interpret this result geometrically in terms of (signed) area.

The other ingredient in our plan of eliminating the multipliers from Lagrange's method is to replace gradients by differentials. Differentials of functions are better behaved than gradients. If f is a function and $(x_1, x_2, \ldots, x_n)$ is a coordinate system valid in the domain of f, then the differential of f expressed in this coordinate system is

$$df = \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2 + \cdots + \frac{\partial f}{\partial x_x} dx_n.$$

A geometric interpretation of the differential df can be obtained from a contour diagram of the function f.



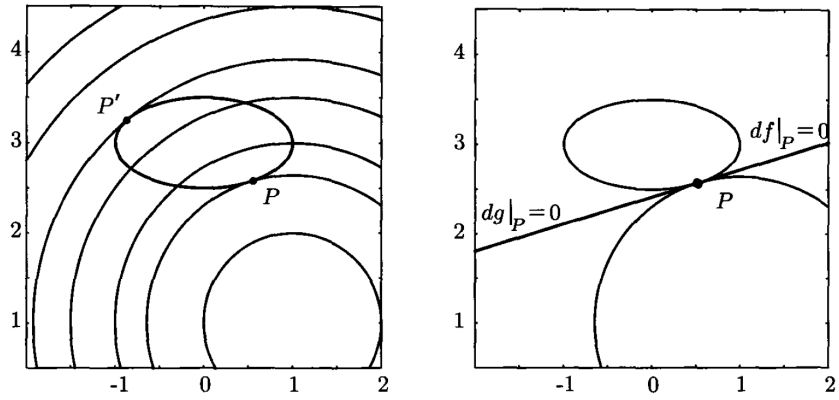
In the Figure (left), the contour at height c (the set of points $f^{-1}(c)$) is labeled as $f = c$. Consider the point P on the contour of height c as fixed and the points Q as a variable. The change in f from P to Q is defined as $\Delta f = f(Q) - f(P)$, which we think of as a function of Q. Using this definition, the contour of f at height c is exactly the same as the contour of Δf at height 0. Therefore, we can re-label the contours $f = c$, $f = c+1$ and $f = c+2$ as $\Delta f = 0$, $\Delta f = 1$ and $\Delta f = 2$, respectively. When evaluated at P, the differential $df|_P = \left.\frac{\partial f}{\partial x}\right|_P dx + \left.\frac{\partial f}{\partial y}\right|_P dy$ becomes a linear function of the variables dx and dy, which represent arbitrary changes in x and y from

their values at P. An equation of the tangent line to the contour of $f = c$ at P is $\frac{\partial f}{\partial x}\Big|_P (x - x_P) + \frac{\partial f}{\partial y}\Big|_P (y - y_P) = 0$, or $\frac{\partial f}{\partial x}\Big|_P dy + \frac{\partial f}{\partial y}\Big|_P dy = 0$.

Thus, in the (dx, dy) coordinate system, whose origin corresponds to the point P the linear equation $df|_P = 0$ gives the tangent at P to the contour labeled $\Delta f = 0$. Furthermore, the solutions of $df|_P = 1$ and $df|_P = 2$ form linear approximations to the contours $\Delta f = 1$ and $\Delta f = 2$ as in the Figure (right). This interpretation of df extends to functions whose domain is of arbitrary dimension.

The gradient of the function f evaluated at a point P is a vector perpendicular to the contour curve of f through P. The formula for the gradient expresses this vector as a linear combination of the unit vectors perpendicular to the coordinate contours through P, so it changes with the coordinate system. For example, in rectangular coordinates $\nabla f = \frac{\partial f}{\partial x}\hat{x} + \frac{\partial f}{\partial y}\hat{y}$, but in polar coordinates $\nabla f = \frac{\partial f}{\partial r}\hat{r} + (1/r)\frac{\partial f}{\partial \theta}\hat{\theta}$. The expression for the differential is the same no matter what coordinate system is used. In polar coordinates, for instance, $df = \frac{\partial f}{\partial r}dr + \frac{\partial f}{\partial \theta}d\theta$.

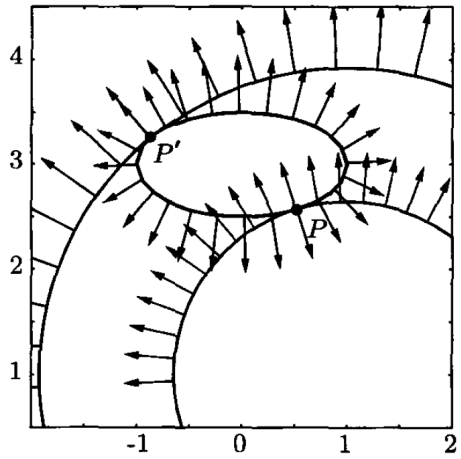
The traditional explanation of Lagrange's condition for maximizing or minimizing f, subject to one constraint of the form $g = c$, hinges on the observation that at a critical point the contours of f and g are tangent. For example, the problem of finding the point on the ellipse $x^2 + 4(y-3)^2 = 1$ that is closest to the point $(1, 1)$. We can view this as the problem of minimizing the squared distance function $f(x, y) = (x-1)^2 + (y-1)^2$, subject to the constraint $g(x, y) = x^2 + 4(y-3)^2 = 1$.



This Figure (left) shows several contours of $f(x, y)$ – points at a fixed distance from $(1, 1)$. The minimum distance occurs at the point P where the contours of f first make contact with the constraint ellipse, and the maximum distance occurs at P', where the last contour of f touches the ellipse.

At these critical points the two curves are tangent; for example, the tangent lines $df|_P = 0$ and $dg|_P = 0$ are the same, as shown in the Figure (right).

But the lines $df|_P = 0$ and $dg|_P = 0$ are the same when $df|_P$ is a multiple of $df|_P$. For example, the lines $2dx - 3dy = 0$ and $-4dx + 6dy = 0$ are identical because $-4dx + 6dy$ is a multiple ($\lambda = -2$) of $2dx - 3dy$. In other words, the equation $df|_P = \lambda\, dg|_P$ between differentials is equivalent to the traditional Lagrange condition $\nabla f|_P = \lambda\, \nabla g|_P$; but geometrically the differential condition says the tangent lines to the contour and the constraint curves are identical, while the gradient condition says the normal vectors to these lines are parallel.



Normal vectors are parallel at critical points.

This Figure shows the (normalized) gradient vectors at several points along the constraint ellipse and the contours of f. The gradient vectors are clearly parallel at the critical points P and P'.

Example 6.3. Geometric interpretation of $df \wedge dg$. As c varies, the equations $df|_P = c$ are the equations of all lines parallel to the tangent line $df|_P = 0$. If the lines tangent to the contours of f and g through P are not parallel, then the points between the four lines whose equations are $df|_P = 0$, $df|_P = 1$, $dg|_P = 0$ and $dg|_P = 1$ form a parallelogram. Let A denote the area of this parallelogram and show that $df \wedge dg|_P = \pm 1(1/A)dx \wedge dy$ (in the xy-coordinate system).

How can we eliminate the multiplier λ from Lagrange's optimization criterion $\nabla|_P = \lambda\, \nabla g|_P$? Let's return to a previous example and set up the Lagrange multiplier equations using differentials, so we have $df = ydx + xdy$ and $dg = 2xdx + 2ydy$. The equations of the tangent lines are $df = 0$ and $dg = 0$; that is, $ydx + xdy = 0$ and $2xdx + 2ydy = 0$. The equation $df = \lambda dg$ together with the constraint equation give the same system of

three equations in x, y and λ that we found earlier in solving Example 1 by the traditional gradient version of Lagrange's method. But now recall that $\{df, dg\}$ is linearly dependent precisely when $df \wedge dg = 0$. (In the expressions for df and dg, dx and dy are variables that we will consider as vector quantities; x and y are the coordinates of the points we are seeking and will be considered as unknown scalars.) Calculating this wedge product, we get

$$\begin{aligned} 0 = df \wedge dg =& (ydx + xdy) \wedge (2xdx + 2ydy) \\ =& 2xydx \wedge dx + 2y^2 dx \wedge dy + 2x^2 dy \wedge dx + 2xydy \wedge dy \\ =& 0 + 2y^2 dx \wedge dy - 2x^2 dx \wedge dy + 0 = (2y^2 - 2x^2) dx \wedge dy. \end{aligned}$$

Thus the critical points (x, y) satisfy the two equations $2y^2 - 2x^2 = 0$, $x^2 + y^2 = 2$. Solving this system yields the same four critical points as before. In fact, since $\frac{\partial f}{\partial \theta} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial \theta} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial \theta} = (y)(-y) + (x)(x) = x^2 - y^2$, we see that $\frac{\partial f}{\partial \theta} = 0$ and $df \wedge dg = 0$ are equivalent, though not identical, conditions. By expressing the Lagrange's optimality condition in the form $df \wedge dg = 0$, we have succeeded in finding a system of equations equivalent to that obtained by changing to polar coordinates.

To generalize the above differential method to handle more than one constraint, we need to extend the wedge product to an arbitrary number of factors. That can be done in a unique way, such that the wedge product is multilinear (linear in each variable) and is an alternating function (the sign changes whenever two terms are transposed). Consider the case of three vectors:

$$u \wedge v \wedge w = -u \wedge v \wedge w = v \wedge w \wedge u = -w \wedge v \wedge u = w \wedge u \wedge v = -u \wedge w \wedge v.$$

In particular, if $u = v$ then $u \wedge v \wedge w = 0$, since the only vector that equals its opposite is the zero vector. The properties of the wedge product lead to a simple way to recover the coefficients from a linear combination $z = \alpha u + \beta v + \delta w$, we simply take the wedge product of z with $v \wedge w$, the product of all the vectors other than u, whose coefficient we wish to find:

$$z \wedge v \wedge w = (\alpha u + \beta v + \delta w) \wedge (v \wedge w) = \alpha u \wedge v \wedge w + 0 + 0 = \alpha u \wedge v \wedge w.$$

Similarly,

$$z \wedge u \wedge w = \beta v \wedge u \wedge w = -\beta u \wedge v \wedge w$$

and

$$z \wedge u \wedge v = \delta w \wedge u \wedge v = \delta u \wedge v \wedge w.$$

So all three coefficients can be found as multipliers of $u \wedge v \wedge w$. (This device is reminiscent of the way we recover the Fourier coefficients of a vector with respect to an orthonormal basis, by taking dot products with the basis vectors.)

Finally, the wedge product provides a simple test for linear independence: The vectors $\{w_1, w_2, \ldots, w_k\}$ are linearly dependent if and only if $w_1 \wedge w_2 \wedge \cdots \wedge w_k = 0$.

Example 6.4. Find the extrema of the function $f(x, y, z) = xyz$ subject to the constraints $x^2 + y^2 + z^2 = 4$ and $x^2 + y^2 - z^2 = 0$.

Lagrange's method uses the constraint functions $g(x, y, z) = x^2 + y^2 + z^2$ and $h(x, y, z) = x^2 + y^2 - z^2$ and directs us to solve the system of equations $df = \lambda dg + \mu h$, $g(x, y, z) = 4$, and $h(x, y, z) = 0$. This approach produces the following system of equations:

$$\begin{aligned} yz &= \lambda(2x) + \mu(2x), & x^2 + y^2 + z^2 &= 4 \\ xz &= \lambda(2y) + \mu(2y), & x^2 + y^2 - z^2 &= 0 \\ xy &= \lambda(2z) + \mu(-2z). \end{aligned}$$

The wedge product reformulation replaces Lagrange's (linear combination) condition $df = \lambda dg + \mu dh$; instead, it finds where df is a linear combination of dg and dh by considering the condition $df \wedge dg \wedge dh = 0$. Calculating, we find

$$df \wedge dg \wedge dh = (yzdx + xzdy + xydz) \wedge (2xdx + 2ydy + 2zdz) \wedge (2xdx + 2ydy - 2zdz).$$

After some manipulations (distributing, interchanging of factors and simplifying), this reduces to $8(x^2 - y^2)z^2 dx \wedge dy \wedge dz$. So the new system of equations for the critical points is

$$\begin{aligned} 8(x^2 - y^2)z^2 &= 0, \\ x^2 + y^2 + z^2 &= 4, \\ x^2 + y^2 - z^2 &= 0. \end{aligned}$$

We now have three equations and three unknowns instead of the Lagrange method's five equations and five unknowns. The solutions of both sides of equations correspond to the critical points $(\pm 1, \pm 1, \pm\sqrt{2})$. Not surprisingly, computer algebra systems solve the second set of equations about 60% faster than they solve the Lagrange set.

6.20. KFAC paper

Suppose we have data drawn from a distribution

$$x_1, \ldots, x_N \sim q(x|\boldsymbol{\theta})$$

The likelihood function is

$$\prod_{i=1}^{N} q(x_i|\boldsymbol{\theta})$$

Then take -log

$$F(\boldsymbol{\theta}) \equiv -\log \prod_{i=1}^{N} q(x_i|\boldsymbol{\theta}) = -\sum_{i=1}^{N} \log q(x_i|\boldsymbol{\theta})$$

Then minimize this with respect to $\boldsymbol{\theta}$. We may do gradient descent:

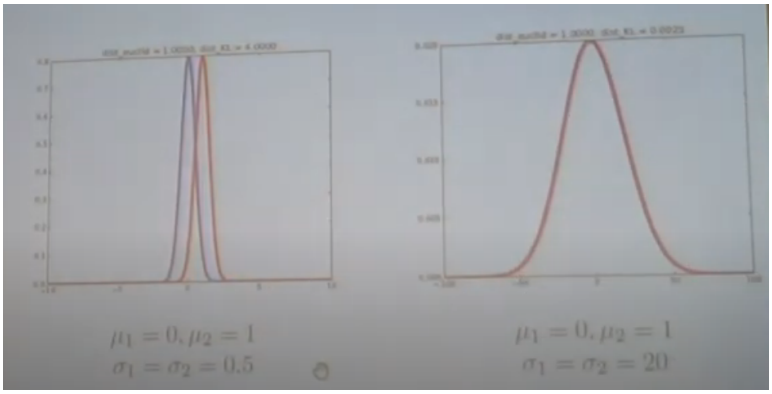
$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha_k \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}_k)$$

In a sense, F depends on the distribution q, i.e., $F(\boldsymbol{\theta}) = F(q)$ (by abuse of notation), and we need to minimize F with respect to the distribution q, which is a parametric family:

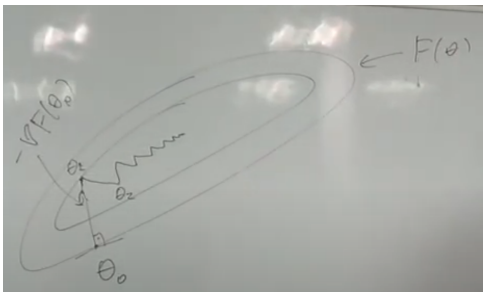
$$\min_{q \in \{q(x|\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}} F(q)$$

Thus, maximum likelihood problem is actually an optimization problem with respect to distribution q that we choose to model our data.

However, distance in parametric space may not be a problem for optimization. For example, in both cases below the Euclidean distance in parametric space is 1. However, on the left the two distributions are completely different. On the right, they are the same:



Consider level sets of function we optimize:



A standard gradient descent will go in a zig-zag trajectory because it ignores the curvature of the surface. We instead should choose the natural gradient. Let's take the directional derivative of our function F of $\boldsymbol{\theta}$ in some direction d:

$$F'(\boldsymbol{\theta}; d) = \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})^T d$$

and minimize with respect to d, within some small ball $\|d\|^2 \leq \epsilon$.

Minimizing

$$L(d, \lambda) = \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})^T d + \lambda(d^T d - \epsilon)$$

with respect to direction (d)

$$\nabla_d L(d, \lambda) = \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) + 2\lambda f = 0$$

gives

$$d = -\frac{1}{2\lambda} \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$$

The optimal λ should be taken to satisfy the condition $\|d\|^2 \leq \epsilon$:

$$\|d\|^2 = \frac{1}{4\lambda^2} \|\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})\|^2 \leq \epsilon$$

which implies $\lambda \geq \|\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})\|(1/2\sqrt{\epsilon})$. For λ we can take the equality:

$$d_{opt} = -\frac{\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})\sqrt{\epsilon}}{\|\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})\|}.$$

Let us replace the optimization problem with the following more general formulation:

$$\min_d \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})^T d$$

subject to the condition

$$d^T G(\boldsymbol{\theta}) d \leq \epsilon$$

involving a quadratic form. $G(\boldsymbol{\theta})$ is some positive-definite matrix, i.e. $G(\boldsymbol{\theta}) \succ 0$. Repeating the optimization under these new conditions we get:

$$d_{opt} \propto -(G(\boldsymbol{\theta}))^{-1} \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$$

This is called a natural gradient. Compare with Newton method, where $G(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$ (Hessian).

Let us introduce the natural distance in the space of distributions. Distances in the space of distributions are usually measured using the KL divergence. KL divergence is not symmetric, but can be symmetrized:

$$\rho(\boldsymbol{\theta}, \boldsymbol{\theta} + \delta\boldsymbol{\theta}) \equiv D_{KL}(q(x|\boldsymbol{\theta}) : q(x|\boldsymbol{\theta} + \delta\boldsymbol{\theta})) + D_{KL}(q(x|\boldsymbol{\theta} + \delta\boldsymbol{\theta}) : q(x|\boldsymbol{\theta}))$$

Using the definition of KL divergence:

$$= \int q(x|\boldsymbol{\theta}) \log \frac{q(x|\boldsymbol{\theta})}{q(x|\boldsymbol{\theta} + \delta\boldsymbol{\theta})} dx + \int q(x|\boldsymbol{\theta} + \delta\boldsymbol{\theta}) \log \frac{q(x|\boldsymbol{\theta} + \delta\boldsymbol{\theta})}{q(x|\boldsymbol{\theta})} dx$$

$$= \int (q(x|\boldsymbol{\theta} + \delta\boldsymbol{\theta}) - q(x|\boldsymbol{\theta}))(\log q(x|\boldsymbol{\theta} + \delta\boldsymbol{\theta}) - \log q(x|\boldsymbol{\theta})) dx$$

Taylor expand,

$$= \int (\nabla_{\boldsymbol{\theta}} q(x|\boldsymbol{\theta})^T \delta\boldsymbol{\theta} + O(\|\delta\boldsymbol{\theta}\|^2))(\nabla_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta})^T \delta\boldsymbol{\theta} + O(\|\delta\boldsymbol{\theta}\|^2)) dx$$

Then apply $\{\cdot\}$ to get:

$$\left\{ \nabla_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta}) = \frac{\nabla_{\boldsymbol{\theta}} q(x|\boldsymbol{\theta})}{q(x|\boldsymbol{\theta})} \right\} = \delta\boldsymbol{\theta}^T \mathbb{E}_{q(x|\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta})^T \delta\boldsymbol{\theta} + O(\|\delta\boldsymbol{\theta}\|^3)$$

$$= \rho(\boldsymbol{\theta}, \boldsymbol{\theta} + \delta\boldsymbol{\theta}).$$

This is valid for small δ. The first term gives $G(\boldsymbol{\theta})$, the Fisher matrix:

$$G(\boldsymbol{\theta}) = \mathbb{E}_{q(x|\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta})^T$$

The natural gradient distance approach:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha_k (G(\boldsymbol{\theta}_k))^{-1} \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}_k)$$

Then, it is possible to show that G is related to the Hessian:[5]

$$G(\boldsymbol{\theta}) = -\mathbb{E}_{q(x|\boldsymbol{\theta})} \nabla^2_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta})$$

In our optimization problem we have (let's add a factor of $1/N$, which doesn't change the result of optimization)

$$F(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{i=1}^{N} \log q(x_i|\boldsymbol{\theta})$$

which should be minimized with respect to $\boldsymbol{\theta}$. From this, let's estimate the Hessian of F:

$$\nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) = -\mathbb{E}_{\hat{q}(x)} \nabla^2_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta})$$

[5] The matrix

$$\nabla\nabla \log q = \nabla(\nabla \log q) = \nabla(\frac{\nabla q}{q}) = \frac{\nabla\nabla q}{q} - \frac{\nabla q \nabla q}{q^2} = \frac{\nabla\nabla q}{q} - \nabla \log q \nabla \log q$$

does not equal to $-\nabla \log q \nabla \log q$ unless the first term, $\frac{\nabla\nabla q}{q}$, vanishes. That term is nonzero but its expectation value is zero (see text).

where $\hat{q}(x)$ is our empirical distribution (based on our data),

$$\hat{q}(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x - x_i).$$

This is almost the same as the Fisher matrix $G(\boldsymbol{\theta})$, which depends on our chosen model $q(x|\boldsymbol{\theta})$.

Derivation:

$$\nabla_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta}) = \frac{\nabla_{\boldsymbol{\theta}} q(x|\boldsymbol{\theta})}{q(x|\boldsymbol{\theta})}$$

$$\nabla^2_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta}) = \frac{\nabla^2_{\boldsymbol{\theta}} q(x|\boldsymbol{\theta}) q(x|\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} q(x|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} q(x|\boldsymbol{\theta})^T}{(q(x|\boldsymbol{\theta}))^2}$$

$$= \frac{\nabla^2_{\boldsymbol{\theta}} q(x|\boldsymbol{\theta})}{q(x|\boldsymbol{\theta})} - \nabla_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta})^T$$

Then we need to take expectation of $\nabla^2_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta})$ with respect to $q(x|\boldsymbol{\theta})$, i.e. $\mathbb{E}_{q(x|\boldsymbol{\theta})}(\dots)$. The first term is:

$$\mathbb{E}_{q(x|\boldsymbol{\theta})} \frac{\nabla^2_{\theta} q(x|\theta)}{q(x|\boldsymbol{\theta})} = \int \nabla^2_{\boldsymbol{\theta}} q(x|\boldsymbol{\theta}) dx = \nabla^2_{\boldsymbol{\theta}} \int q(x|\boldsymbol{\theta}) dx = \nabla^2_{\boldsymbol{\theta}} 1 = 0.$$

Let's consider the simplest possible ML estimation problem, where data is approximated by 1-D normal distribution

$$x_1, \dots, x_N \overset{i.i.d}{\sim} \mathcal{N}(x|\mu, \sigma^2)$$

The values $x_i \in \mathbb{R}$. Taking the log of the density:

$$\log \mathcal{N}(x|\mu, \sigma^2) = -\frac{1}{2}\log 2\pi - \log \sigma - \frac{1}{2\sigma^2}(x - \mu)^2$$

$$\frac{\partial}{\partial \mu} \log \mathcal{N} = \frac{1}{\sigma^2}(x - \mu)$$

$$\frac{\partial}{\partial \sigma} \log \mathcal{N} = -\frac{1}{\sigma} + \frac{(x - \mu)^2}{\sigma^3}$$

$$\mathbb{E}_{q(x|\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log q(x|\boldsymbol{\theta})^T \approx \frac{1}{N}\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \log q(x_i|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log q(x_i|\boldsymbol{\theta})^T$$

where $x_i \sim q(x|\boldsymbol{\theta})$.

Levenberg-Marquardt

$$d = -(G(\boldsymbol{\theta}^{(k)}) + \tau_k I)^{-1} \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^{(k)}), \qquad \tau_k > 0$$

If we choose d from this expression, it is the same thing as saying that our function F is modeled by a quadratic function like this:

$$F(\boldsymbol{\theta}^{(k)} + d) \approx m_k(d) = F(\boldsymbol{\theta}^{(k)}) + \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^{(k)})^T d + \frac{1}{2} d^T (G(\boldsymbol{\theta}^{(k)}) + \tau_k I) d$$

Minimizing this expression for F with respect to d gives the above result for d.

Since $G(\theta)$ is positive semi-definite matrix we have the constrained minimization problem:

$$F(\boldsymbol{\theta}^{(k)}) + \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^{(k)})^T d + \frac{1}{2} d^T G(\boldsymbol{\theta}^{(k)}) d$$
$$\|d\|^2 \leq \Delta k$$

This is a convex minimization problem because this expression for F is convex quadratic and the constraint set is also convex. Since the set is convex, we can formulate the first order necessary condition using Lagrange function. There is a 1-to-1 correspondence between τ_k and Δ_k. The smaller Δ_k is, the larger τ_k is. The expression $F(\boldsymbol{\theta}^{(k)}) + \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^{(k)})^T d + \frac{1}{2} d^T G(\boldsymbol{\theta}^{(k)}) d$ without the constraint $\|d\|^2 \leq \Delta k$ is a plain natural gradient method.

By adjusting τ_k, we have a simple trust-region approach for our natural gradient, e.g.

$$\tau_0 = 1$$
$$\text{for } k = 0, 1, 2, \ldots$$
$$\text{Estimate } G(\boldsymbol{\theta}^{(k)})$$
$$d = -(G(\boldsymbol{\theta}^{(k)}) + \tau_k I)^{-1} \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^{(k)}) \qquad (*)$$
$$\text{if } F(\boldsymbol{\theta}^{(k)} + d) > F(\boldsymbol{\theta}^{(k)})$$

this means our model is "untrustworthy" and we need to narrow the region (lower Δ_k, which means increase τ_k). Go back to $(*)$.

Then we compare the change in our function (numerator) to the forecast of this difference taken from our model:

$$\rho = \frac{F(\boldsymbol{\theta}^{(k)} + d) - F(\boldsymbol{\theta}^{(k)})}{m_k(d) - m_k(0)}$$

This quantity is positive. If it is close to 1, then our model forecast is very good. If it is close to 0, our model is not good.

If $\rho < \frac{1}{4}$ (our quadratic model is less trustworthy), then $\tau_k \leftarrow 2\tau_k$ (switch towards gradient descent).

If $\rho > \frac{3}{4}$ (our model is good, we should trust it more), then $\tau_k \leftarrow \frac{\tau_k}{2}$ (we decrease τ_k, moving towards plain natural gradient).

In all cases, this is followed by

$$\tau_{k+1} = \tau_k + d,$$

i.e., we do not reject the steps. Only the size of the trust-region is varied.

For neural networks, our G is estimated using our current mini-batch:

$$\hat{G}(\boldsymbol{\theta}^{(k)}) = \frac{1}{|I_k|} \sum_{i \in I_k} \nabla_{\boldsymbol{\theta}} \log q(x_i|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log q(x_i|\boldsymbol{\theta})^T$$

$$G(\boldsymbol{\theta}^{(k)}) = (1 - \epsilon_k)\hat{G}(\boldsymbol{\theta}^{(k)}) + \epsilon_k G(\boldsymbol{\theta}^{(k-1)})$$

A simple heuristic rule is

$$\epsilon_k = \min(1 - 1/k, 0.95),$$

which says that initially, we use our mini-batches, but as time goes by we use G from the past.

KFAC. Neural networks has layers:

$$a_0 = x$$

$$s_i = W_i \overline{a_{i-1}}, \qquad \overline{a_{i-1}} = \begin{bmatrix} a_{i-1} \\ 1 \end{bmatrix}$$

$$a_i = \varphi(s_i), \qquad i = 1, \ldots, l$$

$$z(x, \boldsymbol{\theta}) = a_l$$

Our optimization problem is:

$$F(\boldsymbol{\theta}) \equiv \frac{1}{N} \sum_{i=1}^{N} L(y_i, z(x_i, \boldsymbol{\theta}))$$

which is minimized with respect to $\boldsymbol{\theta}$. The parameters are stored in the following column vector:

$$\boldsymbol{\theta} = [vec(W_1)^T, \ldots, vec(W_l)^T]^T$$

Here we don't have any statistical model, so we cannot apply natural gradient method. However, we can require that this loss function is not arbitrary, but is taken as the log of some real distribution:

$$-\frac{1}{N} \sum_{=1}^{N} \log q(y_i|z(x_i, \boldsymbol{\theta}))$$

which is minimized with respect to $\boldsymbol{\theta}$. So we have an optimization with respect to distribution q:

$$\min_{q(y,x,\boldsymbol{\theta})\hat{q}(x)} F(q)$$

i.e. conditional distribution of target variables given the inputs, $q(y|x, \boldsymbol{\theta})$, times empirical distribution of Fisher matrix $\hat{q}(x)$. Note: there are two distributions shown here, $q(y, x, \boldsymbol{\theta})\hat{q}(x)$, because our data is in the form $(y_1, x_1), \ldots, (y_N, x_N)$.

Define:

$$\mathcal{D}v = -\frac{\partial}{\partial v} \log q(y|z(x, \boldsymbol{\theta}))$$

In this notation our Fisher matrix is denoted:

$$G(\boldsymbol{\theta}) = \mathbb{E}_{q(y|z(x,\boldsymbol{\theta}))\hat{q}(x)} \mathcal{D}\boldsymbol{\theta} \mathcal{D}\boldsymbol{\theta}^T$$

where G is a $l \times l$ block matrix:

$$G_{ij}(\boldsymbol{\theta}) = \mathbb{E} d_i d_j^T, \qquad d_i = vec(\mathcal{D}W_i)$$

The derivative of our output function, according to the chain rule, is:

$$\mathcal{D}W_i = \mathcal{D}s_i \cdot \overline{a_{i-1}}^T$$

where $\overline{a_{i-1}}^T$ is the derivative of s_i with respect to W_i.

Then, using the relationship $vec(uv^T) = v \otimes u$,

$$vec(\mathcal{D}W_i) = vec(\mathcal{D}s_i \overline{a}_{i-1}^T) = \overline{a}_{i-1} \otimes \mathcal{D}s_i$$

Therefore, our Fisher matrix is

$$G_{ij}(\boldsymbol{\theta}) = \mathbb{E} d_i d_j^T = \mathbb{E} vec(\mathcal{D}W_i) vec(\mathcal{D}W_j)^T = \mathbb{E}[\overline{a}_{i-1} \otimes \mathcal{D}s_i][\overline{a}_{j-1} \otimes \mathcal{D}s_j]$$
$$= \mathbb{E}[\overline{a}_{i-1}\overline{a}_{j-1}^T \otimes \mathcal{D}s_i \mathcal{D}s_j^T] \approx \mathbb{E}[\overline{a}_{i-1}\overline{a}_{j-1}^T] \otimes \mathbb{E}[\mathcal{D}s_i \mathcal{D}s_j^T]$$

The last step is called the KFAC approximation. It is a product of two terms, the first is obtained by forward propagation. The second is obtained by backpropagation. It is of the form of a Khatri-Rao product.

Block diagonal approximation:

$$\tilde{G}_{ij} = 0 \qquad \forall i \neq j$$
$$\tilde{G}_{ii} = \mathbb{E}\overline{a}_{i-1}\overline{a}_{i-1}^T \otimes \mathbb{E}\mathcal{D}s_i \mathcal{D}s_i^T$$

We need to invert the matrix:

$$\tilde{G}_{ii}^{-1} = (\mathbb{E}\overline{a}_{i-1}\overline{a}_{i-1}^T)^{-1} \otimes (\mathbb{E}\mathcal{D}s_i \mathcal{D}s_i^T)^{-1}$$

If we have:

$$n_i \text{ inputs}$$
$$m_i \text{ outputs}$$

then $\tilde{G}_{ii}$ is a $n_i m_i \times n_i m_i$ matrix, whereas $\mathbb{E}\overline{a}_{i-1}\overline{a}_{i-1}^T$ is $n_i \times n_i$ and $\mathbb{E}\mathcal{D}s_i \mathcal{D}s_i^T$ is $m_i \times m_i$. Thus, KFAC provides a clear computational advantage. In KFAC each block corresponds to one layer.

Then to compute $u = \tilde{G}^{-1} v$, we can make use of the identity $(A \otimes B) vec(X) = vec(BXA^T)$ to get

$$U_i = (\mathbb{E}\mathcal{D}s_i \mathcal{D}s_i^T)^{-1} V_i (\mathbb{E}\overline{a}_{i-1}\overline{a}_{i-1}^T)^{-1}$$

where v maps to $(V_1, V_2, \ldots, V_l)$ and u maps to $(U_1, U_2, \ldots, U_l)$ in an analogous way[6] to how θ maps to $(W_1, W_2, \ldots, W_l)$.

[6] Recall that $\boldsymbol{\theta} = [vec(W_1)^T vec(W_2)^T \ldots vec(W_l)^T]^T$.

For Levenberg-Marquardt what we need to invert is $(\tilde{G}_{ii}+\tau_k I)^{-1}$. Consider the following approximation based on the Kronecker product trick:

$$(\tilde{G}_{ii}+\tau_k I) \approx (\mathbb{E}\overline{a}_{i-1}\overline{a}_{i-1}^T + \pi_i\sqrt{\tau_k}I) \otimes (\mathbb{E}\mathcal{D}s_i\mathcal{D}s_i^T + \frac{1}{\pi_i}\sqrt{\tau_k}I)$$

$$= \mathbb{E}\overline{a}_{i-1}\overline{a}_{i-1}^T \otimes \mathbb{E}\mathcal{D}s_i\mathcal{D}s_i^T + \underbrace{\pi_i\sqrt{\tau_k}I \otimes \mathbb{E}\mathcal{D}s_i\mathcal{D}s_i^T + \frac{1}{\pi_i}\sqrt{\tau_k}\mathbb{E}\overline{a}_{i-1}\overline{a}_{i-1}^T \otimes I} + \tau_k I \otimes I$$

which differs from the matrix $(\tilde{G}_{ii}+\tau_k I)$ by the two terms denoted by the underbrace. These two terms can be minimized with respect to π_i

$$\pi_i\sqrt{\tau_k}\|I \otimes \mathbb{E}\mathcal{D}s_i\mathcal{D}s_i^T\| + \frac{1}{\pi_i}\sqrt{\tau_k}\|\mathbb{E}\overline{a}_{i-1}\overline{a}_{i-1}^T \otimes I\|.$$

Then,

$$\pi_{i,opt}^2 = \frac{\|\mathbb{E}\overline{a}_{i-1}\overline{a}_{i-1}^T \otimes I\|_2}{\|I \otimes \mathbb{E}\mathcal{D}s_i \cdot \mathcal{D}s_i^T\|_2} = \frac{\|\mathbb{E}\overline{a}_{i-1}\overline{a}_{i-1}^T\|}{\|\mathbb{E}\mathcal{D}s_i\mathcal{D}s_i^T\|}$$

6.20.1. Line search vs Trust region. These are two different philosophies. We can choose to do line search and pick the best learning rate that lowers F the most. Or we can vary the size of the trust region. If our trust region is good, d (direction) will vary accordingly.

6.21. Problems

Problem 121. Regarding the problem of data fitting, where we need gradients, Jacobians, etc. (review lecture notes for the definitions of $\nabla\chi^2(\vec{\beta}_k)$, $\mathbf{J}_k$, etc.). (a) Take $y(x) = Ax + B$ as the fitting model. Calculate $\nabla\chi^2(\vec{\beta}_k)$, $\mathbf{J}_k$, $\mathbf{H}_k$ and the covariance matrix.

(b) Let the fitting model be given by $y(x) = A + Bx + C\cos(Dx)$. Calculate $\nabla\chi^2(\vec{\beta}_k)$ and $\mathbf{J}_k$.

(c) The model is $y(x) = A\exp(-Bx) + C$. Calculate $\nabla\chi^2(\vec{\beta}_k)$, $\mathbf{J}_k$ and $\mathbf{H}_k$.

Solution. Let's do the Jacobian here for (i), the linear model. From the definition of the Jacobian, this is a $n \times p$ matrix of first partial derivatives of the normalized residuals. The rows are the data points and the columns are the fitting parameters (here, A and B):

$$\mathbf{J}_k = \begin{bmatrix} \frac{\partial\tilde{R}_1}{\partial A} & \frac{\partial\tilde{R}_1}{\partial B} \\ \vdots & \vdots \\ \frac{\partial\tilde{R}_n}{\partial A} & \frac{\partial\tilde{R}_n}{\partial B} \end{bmatrix}$$

where $\tilde{R}_i = \frac{y_i - y(x_i;A,B)}{\alpha_i}$ and $y(x_i; A, B) = Ax_i + B$. Thus,

$$\mathbf{J}_k = \begin{bmatrix} \frac{x_1}{\alpha_1} & \frac{1}{\alpha_1} \\ \vdots & \vdots \\ \frac{x_n}{\alpha_n} & \frac{1}{\alpha_n} \end{bmatrix}$$

■

Problem 122. The following model is to be used for data fitting

$$y(x) = 40B \log x^2 + 4A \cos x + 2A$$

where A and B are parameters to be determined from the experimental data. In terms of this model, find the Jacobian, Hessian and covariance matrices and explain their role/purpose. Indicate which entries of the covariance matrix measure the redundancy between the fitting parameters.

Problem 123. Write MATLAB code to implement the Levenberg algorithm (not Levenberg-Marquardt) to minimize the following functions:

(a) $f(x, y) = \sin(5y)\sin^{-1}(x) - \sin(5x)\sin^{-1}(y)$ on the domain $[-1, 1]^2$.

(b) $f(x, y) = -\left|\sin(x)\cos(y)e^{(1/2)|1-\sqrt{|x|+|y|}|}\right|$ on the domain $[-10, 10]^2$.

Instead of using a `for` loop for the iterations, use a `while` loop instead. (No credits will be awarded if you use a `for` loop.) Monitor the running average of $f(x, y)$ (say, using the last 5 iterations) and if changes in $f(x, y)$ from iteration to iteration differ by less than 10^{-6}, stop the loop. For $f(x, y)$, make sure to use the normalized $f(x, y)$, i.e. divide $f(x, y)$ by the number of points in the summation.

Chapter 7

Global Optimization

The algorithms discussed so far can only take us to the nearest minimum. Once a minimum has been reached, there is no built-in mechanism to escape the minimum. If this minimum is not a global minimum, the solution obtained is not optimal. A general $\chi^2(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^p$ surface:

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{(y_i - y(x_i|\boldsymbol{\theta}))^2}{\sigma_i^2}. \qquad \text{(textbook writes } \alpha_i \text{ instead of } \sigma_i\text{)}$$

could potentially have several minima, as shown in Fig. 7.1 for the 2D case $\boldsymbol{\theta} = (\theta_1, \theta_2)$.

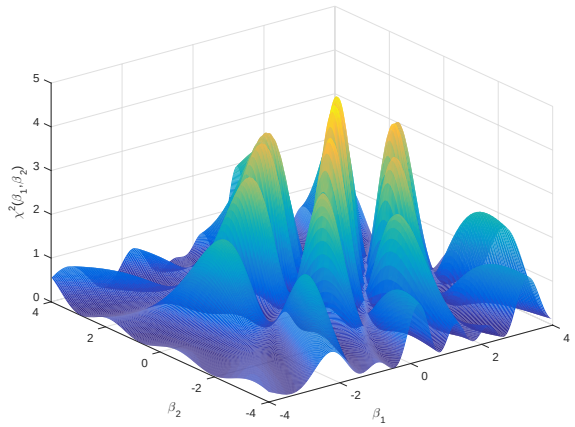


Figure 7.1. Global optimization aims to find the global minimum of the χ^2 surface. Note: in this figure β should be replaced by $\boldsymbol{\theta}$.

The reason that gradient (or Hessian) based methods failed to see the global minimum is because they only possess knowledge of the first and second derivative of χ^2 near the point $\boldsymbol{\theta}$. This local neighborhood of $\boldsymbol{\theta}$ does not extend very far into the parameter space.

Global optimization algorithms tend to rely on the use of multiple random initial guesses, or on the use of evolution steps that may include some element of randomness, or a combination of both. These random elements are essential in order to "push" the search toward regions of the parameter space ($\boldsymbol{\theta}$) that gradient-based techniques would otherwise not be able to reach. While global optimization schemes may or may not also use gradient-based searches as part of the optimization strategy, the "global" nature of the algorithm generally owes to the addition of some element of randomness.

Here we will cover two such global methods: the simulated annealing (Metropolis) method and the genetic algorithm. You should, however, be aware that there are many additional methods in use: quantum annealing, stochastic tunneling, tabu search, reactive search optimization, stochastic gradient descent, graduated optimization, ant colony optimization, cross-entropy method, harmony search, particle swarm optimization, intelligent water drops and parallel tempering. Each method has its advantages and disadvantages. Another ideal setting for these global optimization methods is in cases where derivative information is not available.

7.1. The Metropolis Algorithm (Simulated Annealing)

The Metropolis algorithm (Kirkpatrick, 1983) - also known as Metropolis Monte-Carlo - is a version of simulated annealing (SA) that utilizes the so-called Metropolis criteria [due to Metropolis (1953) and Hastings (1970)]. It is inspired by the annealing process in metallurgy. Annealing involves heating and controlled cooling of a metal or alloy to produce a high quality crystalline lattice. Heat increases thermal motion of the atoms and eases diffusional motion. The subsequent cooling causes the atoms to migrate to sites of local minimum energy, which usually translates into a lower amount of defects and a global energy minimum.

The algorithm can be summarized as follows:

(1) Choose an initial configuration $\boldsymbol{\theta} \in \mathbb{R}^p$ for the fitting parameters. The starting parameters can be selected randomly.

(2) Choose a starting temperature $T > 0$. Temperature should be high enough, as the goal of this algorithm is to cool (by annealing) over a long period of time.

(3) Calculate $\chi^2(\boldsymbol{\theta})$ for this configuration.

(4) Let $\delta\boldsymbol{\theta}$ by a random change to this configuration. The new configuration is $\boldsymbol{\theta} + \delta\boldsymbol{\theta}$.

(5) Calculate the new energy $\chi^2(\boldsymbol{\theta} + \delta\boldsymbol{\theta})$

(6) Accept the move with probability

$$\min\left(1, \exp(-(\chi^2(\boldsymbol{\theta} + \delta\boldsymbol{\theta}) - \chi^2(\boldsymbol{\theta}))/T)\right).$$

This step is called the Metropolis criterion. If the move is accepted, the new configuration $\boldsymbol{\theta} + \delta\boldsymbol{\theta}$ is taken to be the current configuration, $\boldsymbol{\theta}$. If the move is rejected, the old configuration $\boldsymbol{\theta}$ is kept unchanged.

(7) Repeat steps 3-6 until convergence while lowering the temperature T.

A few words are in order. First, the temperature T is dimensionless and is not a real temperature but a parameter that simulates the effects of temperature. The higher the temperature, the closer to 1 is the probability of acceptance, meaning that almost every state is accessible. At low temperatures, the $\exp(\cdot)$ (Boltzmann) factor is smaller and acceptance of the moves is less likely unless the energy of the system (as measured by χ^2) decreases as a result of the new move.

Finite temperatures correspond to "thermal energy" supplied by a reservoir to the thermodynamic system. If the temperature is not lowered, the system will remain with the same average energy per unit volume and moves will continue to be accepted at the same rate. In order to find the "ground state" of the system the goal is to reach a point where the majority of moves are rejected. This can only happen at or near $T = 0$.

If the Boltzmann factor

$$\exp\left(-\left[\chi^2(\boldsymbol{\theta}') - \chi^2(\boldsymbol{\theta})\right]/T\right) = e^{-\Delta E/T}$$

has $\Delta E = \chi^2(\boldsymbol{\theta}') - \chi^2(\boldsymbol{\theta}) > 0$, the new move results in a higher energy configuration. In this case, rather than rejecting it, the move is accepted with probability $e^{-\Delta E/T} < 1$. This allows for the possibility of random jumps that could lift the system out of a local minimum. On the other hand, $\Delta E < 0$ corresponds to a move that lowers the energy of the system. In this case, $e^{-\Delta E/T} > 1$ and the move is accepted with probability 1 (always).

7.2. Accepting a Move With Probability P

What does it mean when we are asked to accept a move with probability P? For example,

$$P = e^{-\Delta E/T}.$$

Given a probability $P \in [0, 1]$, regardless of the probability distribution it originates from, we may decide whether or not a random experiment occurs with probability P as follows:

- Generate a uniformly distributed random number R in the interval
$$R \in [0, 1]$$
- If $P > R$, accept the move (the event has occurred).
- If $P \leq R$, reject the move (the event did not occur).

The rationale for this reasoning is explained in the section below.

7.3. Sampling From a Distribution

In the Metropolis scheme and many other algorithms, we are asked to sample from a distribution, such as the exponential distribution. However, your random number generator may not be able to generate samples from the distribution of your choice. It is likely able to generate uniformly distributed samples $U([0, 1])$ in the interval $[0, 1]$. For example, the command `rand` in MATLAB will generate uniformly distributed random numbers:

```
>> help rand
 rand Uniformly distributed pseudorandom numbers.
    R = rand(N) returns an N-by-N matrix containing pseudorandom
    values drawn from the standard uniform distribution on the open
    interval(0,1).  rand(M,N) or rand([M,N]) returns an M-by-N matrix.
    rand(M,N,P,...) or rand([M,N,P,...]) returns an M-by-N-by-P-by-...
    array.  rand returns a scalar.  rand(SIZE(A)) returns an array
    the same size as A.

...
```

Fortunately, MATLAB has another command, `randn` which can generate random numbers sampled from a Gaussian distribution:

```
>> help randn
 randn Normally distributed pseudorandom numbers.
    R = randn(N) returns an N-by-N matrix containing
    pseudorandom values drawn from the standard
    normal distribution.  randn(M,N) or randn([M,N]) returns
    an M-by-N matrix. randn(M,N,P,...) or randn([M,N,P,...])
    returns an M-by-N-by-P-by-... array. randn returns a
    scalar.  randn(SIZE(A)) returns an array the same size as A.

...
```

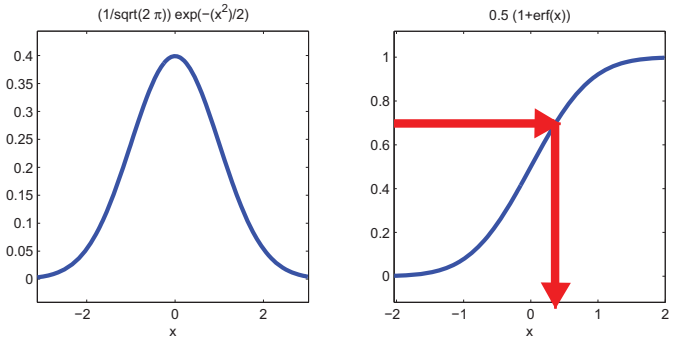


Figure 7.2. Sampling from an arbitrary distribution with the help of its CDF and the uniform distribution.

But to my knowledge in the basic version of MATLAB there are no random number generators for other distribution functions. The problem also exists with many programming languages such as C or FORTRAN, where random number generators exist but only for uniform distributions.

Luckily, a random number sampled from the uniform distribution, $U([0,1])$, can be used to generate random samples from any distribution function. Suppose that we want to generate Gaussian random numbers $X \sim \mathcal{N}(0,1)$ for a given random variable X. The probability density is $p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. In the figure below, we plot the CDF as the graph $(x, y(x))$:

$$y(x) = \mathbb{P}(X < x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2} \mathrm{d}x = \int_{-\infty}^{x} p_X(x) \mathrm{d}x$$

in the case of a Gaussian density. The left-most plot shows a PDF $p_X(x)$ corresponding to some random variable X. On the right, we have its cumulative distribution function $y(x) = \int_{-\infty}^{x} p_X(x)\mathrm{d}x$. We generate uniformly distributed samples y in the interval $[0,1]$ and invert $y = y(x)$ to produce a sample x according to the desired distribution $p_X(x)$. This is illustrated in Figure 7.2.

If we consider the y axis as a random variable $Y \sim U([0,1])$ which is sampled uniformly on the interval $[0,1]$, let us check that a uniform distribution for Y gives rise to the desired distribution for X.

$$\begin{aligned}\mathbb{P}(Y < y(x)) =& \mathbb{P}\left(Y < \int_{-\infty}^{x} p_X(x)\mathrm{d}x\right) = \int_0^{\int_{-\infty}^{x} p_X(x)\mathrm{d}x} 1 \cdot \mathrm{d}y \\ =& \int_{-\infty}^{x} p_X(x)\mathrm{d}x = \mathbb{P}(X < x).\end{aligned}$$

Therefore, a uniform distribution for Y implies picking X according to the density $p_X(x)$.

7.4. Genetic Algorithms

Genetic algorithms (Holland, 1975) enable us to to find global minima. In the early days, GAs have been applied to least squares curve fitting (Karr, 1991; Rogers 1991). Nowadays, GAs are applied to many different optimization problems in the physical and social sciences. Software packages such as MATLAB, IDL, Mathematica or Octave feature implementations of genetic algorithms (GA) that are relatively easy to use. One advantage of GA codes is they are inherently parallel and easily implemented on parallel hardware. Another advantage is they can handle large data sets. Their main disadvantage is they are slow. The general idea consists of generating candidate solutions and sending them to an evaluator for testing. If a candidate solution is not optimal, then the procedure is repeated. In genetic algorithms, the procedure is repeated iteratively over a large set of candidate solutions. Because this set can be large, a significant number of possible solutions can be tested simultaneously.

They are a type of parallel heuristic search method inspired by the laws of nature (genetics) that govern evolution of biological organisms. Each candidate solution is called an organism. A chromosome is a list of elements called genes. In the simplest case, an organism consists of a single chromosome (haploid), although there are cases when the organism consists of dual-strand chromosomes (diploid). Chromosomes usually consist of linear lists of genes. A gene can assume any of a number of values called alleles, which are taken from the base set. Generally, problem solutions are encoded as strings of alleles (most commonly, strings of 0's and 1's).

7.4.1. General idea. Let us look at an example strategy for a possible genetic algorithm. The algorithm below is purposely left vague so you can see the main steps involved and the parallel with evolution. In real implementations, the specifics of the algorithm strongly depend on the application. We will look at a specific example later.

(1) Generate a large population of random chromosomes $\{\boldsymbol{\theta}^{(i)}\}$, where $\boldsymbol{\theta}^{(i)} \in \mathbb{R}^p$ and $i = 1, \ldots, M$.

(2) Each chromosome is assigned a fitness score F proportional to some goodness-of-fit parameter, e.g. $\chi^2(\boldsymbol{\theta}^{(i)})$. This fitness score F should increase as χ^2 decreases.

(3) Select the top fitness scores. These "parent" chromosomes, $\{\boldsymbol{\theta}^{best}\}$, will be used to breed the next generation ("offspring").

(4) Generate "offspring" from the parent chromosomes: $\{\boldsymbol{\theta}^{(i)}\}$, where $\boldsymbol{\theta}^{(i)} \in \mathbb{R}^p$ and $i = 1, \ldots, M$.

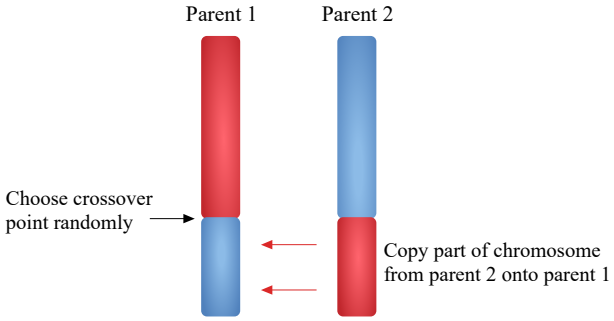


Figure 7.3. Crossover of chromosomes.

(5) Introduce random changes in the genetic code in the form of crossover and mutations. Mutations are "tweaks" to the chromosome contents (e.g. flip a 0 to a 1). Crossover involves picking a gene at random and generating a new chromosome that consists of chromosome 1 up until that gene followed by the rest of chromosome 2, as indicated in Fig. 7.3.

(6) Repeat steps 2-5 until solution is found.

The inherently parallel nature of the algorithm is embodied in steps 2 and 4, where the GAs search large numbers of candidate solutions simultaneously. A decision is made only after all the candidate solutions have been generated. The objective function used by genetic algorithms is based on actual, problem-specific information, rather than auxiliary information, such as a gradient or a Hessian.

Step 3 is analogous to natural selection. Like its natural selection counterpart in biology, the selection operation selects pairs of highly fit organisms for mating. This focus toward the highly fit individuals is what drives genetic algorithms.

The genetic algorithm analogy to mating is called crossover. The crossover operation provides a mixing of the genes from the parents, and globally it mixes the genetic material of the whole population. It is the mixing of the genes, the stirring of the pot of genetic material, that gives robustness to the genetic algorithm.

The two organisms chosen by selection are combined to form a new individual with similarities to both parents. If the mixing is done carefully, then a large amount of genetic material can be tested. Although selection focuses on the genetic algorithm, it is crossover that adds variety.

The method based on the random selection of a single crossover point described above is the simplest implementation of crossover. More complicated

implementations are possible. There is no specific guidelines as to what rule should be used. One could, for example, take the average of two chromosomes, their product, their difference or their sum. Which implementation is best suited depends on application.

Problem 124. (a) Is the genetic algorithm a "local" or "global" optimization algorithm? Explain.

(b) What element of this algorithm gives it its "local" or "global" nature?

(c) Discuss what is meant by local / global optimization, and give examples of algorithms (genetic or not) which achieve local vs. global optimization.

Problem 125. Solve analytically the following minimization problem for the unknown vector g:

$$\arg\min_{g} \|Ug - k\|^2$$

where g is a column vector with P entries whose values are unknown quantities that you are to solve for, k is a column vector with N rows. U is a $N \times P$ matrix (N rows, P columns). The minimization is to be carried out by searching over all possible vector g with finite entries.

(a) Find the exact solution to the minimization problem.

(b) Explain how this minimization problem can be used to fit experimental data to a model (linear or non-linear model). Explain what would be the roles of U, g and k in this context.

Problem 126. Solve the following equation giving the unknown x number of moles of a substance needed in a reaction:

$$4\sin(2x) + 5\log(2x^2) - 1000 + x^2\exp(5x) = 0.$$

(a) Derive an algorithm and write a working computer program or use a spreadsheet or calculator to obtain a correct value for x. Explain how to do it.

(b) Find the value x such that the left hand side of the equation is a minimum. You found a local minimum. Explain how you found a local minimum. Is it possible to find a global minimum?

(c) Find a minimum of the function $f(x, y) = x^2\cos(2x)\cos(2y)$ near the point $(x, y) = (100, 100)$. Explain all the details of how you proceeded to find the minimum.

(d) Fit data points $(95, 85)$, $(85, 95)$, $(80, 70)$, $(70, 65)$, $(60, 70)$ (data given in the form (x_i, y_i)) to a model $y(x) = A + Bx$ and find the values for A and B. Show all your work.

Solution. (a) Set $f(x) = 4\sin(2x) + 5\log(2x^2) - 1000 + x^2\exp(5x)$ and use Newton Raphson ($x_{n+1} = x_n - f(x_n)/f'(x_n)$) using some initial guess x_0.

(b) $f(x) = 4\sin(2x)+5\log(2x^2)-1000+x^2\exp(5x)$ and use Newton Raphson applied to the first derivative, $x_{n+1} = x_n - f'(x_n)/f''(x_n)$, and a suitable initial guess for x_0. You can find infinitely many local minima because of the $\sin(2x)$ term. The function does not have a global minimum due to the singularity at the origin, $\lim_{x\to 0}\log(2x^2) = -\infty$.

(c) Use any of the gradient-based search algorithms (steepest descent, Newton, Gauss-Newton or Levenberg-Marquardt. Set initial conditions to $(100, 100)$.

(d) Use the formulae, $A = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{\Delta}$, $B = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\Delta}$, $\Delta = n\sum x_i^2 - (\sum x_i)^2$, and plug in the data provided to compute A and B. I get the results: $A = 26.768$ and $B = 0.644$, i.e. the model is $y(x) = 26.768 + 0.644x$. ■

Problem 127. Write MATLAB code to find the global minimum of:

$$f(x, y) = 100(x^2 - y)^2 + (1 - x)^2 \text{ on the domain } [-3, 3]^2.$$

You may use MATLAB's `ga` command. Another option is to write your own code instead of using the built-in `ga` command. If you write you own code, and if your code is short and easy to understand, your homework will be weighed *twice* (100% bonus points). It's possible to write fairly short `ga` code (under 30 lines) to solve this problem. For the code to work efficiently, the objective function will need to be vectorized, i.e. take vectors/matrices as input, and output a vector. Element-wise operations will be required.

Problem 128. For the two functions of problem 123, compare the local (Levenberg) search to the global (`ga`) search. Modify your Levenberg code to explore the domain and find global extrema; compare results with `ga`. Comment on whether this method (using a local search algorithm to perform a global search, instead of using a true global search algorithm) is practical in the general case of arbitrary function and domain.

Problem 129. In problem 137(b), I have provided code for simulated annealing. Compare the speed of convergence for our simulated annealing code to the `ga` command, i.e. plot χ^2 vs iteration. (Usually, one uses a semi-log graph to study convergence.) In a third calculation, use the built-in MATLAB command `simulannealbnd` for this optimization:
`https://www.mathworks.com/help/gads/simulated-annealing-examples.html`
`https://www.mathworks.com/help/gads/simulannealbnd.html`
Compare performance for all 3 algorithms. Which of the 3 algorithms is fastest?

Problem 130. Use the `ga` command to fit an exponential decay function to the fake dataset of problem 139.

Chapter 8

Errors in the Fitted Parameters during Nonlinear Fitting

Given that nonlinear fitting methods are based on computer algorithms, how can we obtain estimates of the error in the fitting parameters? We no longer have the option of deriving analytical formulas. It turns out that the covariance matrix provides estimates of the errors in the fitting parameters. In this lecture we show that $cov(\boldsymbol{\theta}_k, \boldsymbol{\theta}_k) = 2(\mathbf{H}_k)^{-1}$, where $\mathbf{H}_k$ is the Hessian matrix at the k-th iteration.

8.1. Linear least squares

In linear least squares, we assume a model of the form

$$\vec{y} = \underbrace{\mathbf{A}\boldsymbol{\theta}}_{\text{model}} + \vec{\epsilon}$$

where $\vec{y} = (y_1, \ldots, y_n)^T$ is the column vector of measured data points. There are n data points collected. $\mathbf{A}$ is called the design matrix and has dimensions $n \times p$. The model parameters are stored in the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$. Finally, the errors in each measurement are stored in a vector $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$.

The errors ϵ_i are random variables assumed to be independent and identically distributed (iid rv) according to a normal law

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2), \qquad i = 1, \ldots, n$$

In general $n \gg p$, since we want a model with much fewer parameters than there are data points collected to construct it. In a sense, data fitting amounts to doing data compression. Some effort must be devoted to finding an adequate model that accurately describes the key features of the problem with as few parameters as possible. In components form, the above equation can be written as

$$y_i = \sum_{j=1}^{p} a_{ij}\theta_j + \epsilon_i$$

where a_{ij} are the elements of the design matrix $\mathbf{A}$.

Let us look at some examples of design matrix $\mathbf{A}$. This matrix is found in the expression $\vec{y} = \mathbf{A}\boldsymbol{\theta} + \vec{\epsilon}$. Consider the model

$$y_i = A + Bx_i + Cx_i^2 + \epsilon_i, \qquad i = 1, \ldots, n$$

We find, by inspection

$$\mathbf{A} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & & \\ 1 & x_n & x_n^2 \end{bmatrix}, \qquad \boldsymbol{\theta} = \begin{bmatrix} A \\ B \\ C \end{bmatrix}, \qquad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \vec{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Another example is

$$y_i = A + B\log x_i + C\cos^2(x_i) + \epsilon_i$$

We find, by inspection that

$$\mathbf{A} = \begin{bmatrix} 1 & \log x_1 & \cos^2(x_1) \\ 1 & \log x_2 & \cos^2(x_2) \\ \vdots & & \vdots \\ 1 & \log x_n & \cos^2(x_n) \end{bmatrix}$$

8.2. MLE

Recall that with least squares, the fitting parameters are obtained by applying the principle of maximum likelihood, which consists of solving for the model parameters $\boldsymbol{\theta}$ for which the probability density of the observed deviations (data minus model) is a maximum. Maximizing L is equivalent to maximizing its logarithm, l:

$$l = -\frac{\chi^2}{2} = -\frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - y(x_i|\boldsymbol{\theta}))^2}{\sigma_i^2} = -\frac{1}{2}\sum_{i=1}^{n} \frac{\epsilon_i^2}{\sigma_i^2}$$

Here, our assumption will be that ϵ_i are Gaussian iid rv. Therefore, the joint probability distribution of all $(\epsilon_1, \ldots, \epsilon_n)$ is a product of individual

distributions for each of the ϵ_i, each of which is Gaussian with mean 0 and standard deviation σ.

In this special case, the "likelihood function" is

$$L(\boldsymbol{\theta}, \sigma^2|\vec{y}) \equiv p(\vec{y}|\boldsymbol{\theta}, \sigma^2) = \frac{\prod_{i=1}^{n} \exp\left\{-\frac{1}{2}\left(y_i - \sum_{j=1}^{p} a_{ij}\theta_j\right)^2 / \sigma^2\right\}}{\sigma^n (2\pi)^{n/2}}$$
$$= \frac{e^{-\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} a_{ij}\theta_j\right)^2/\sigma^2}}{\sigma^n (2\pi)^{n/2}}$$

L is the probability density for observing the set of deviations $\{\epsilon_i\}$ or, equivalently, the set of measurements $\vec{y}$ given a model and its parameters $y(x_i|\boldsymbol{\theta})$.

Taking the natural log of this expression yields the "log likelihood"

$$l \equiv \log L = const - n \log \sigma - \frac{1}{2}\left[\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} a_{ij}\theta_j\right)^2\right] / \sigma^2.$$

where $const = -(n/2)\log(2\pi)$. This term will be ignored since it will not be needed.

The principle of "maximum likelihood" tells us we should compute the extremum of this function with respect to its parameters. The parameters are $\boldsymbol{\theta}$ and σ.

8.2.1. MLE of the Model Parameters $\vec{\beta}$. Differentiating with respect to $\boldsymbol{\theta}$, i.e. $\partial l/\partial\theta_r = 0$, gives the following equation:

$$\sum_{i=1}^{n} a_{ir}\left(y_i - \sum_{j=1}^{p} a_{ij}\theta_j\right) = 0, \qquad (r = 1, \ldots, p)$$

or,

$$\sum_{j=1}^{p}\left(\sum_{i=1}^{n} a_{ir}a_{ij}\right)\theta_j = \sum_{i=1}^{n} a_{ir}y_i.$$

This can be expressed in a more compact notation if we write $\mathbf{C} = (c_{ij})$ for the ij-th element of the $p \times p$ matrix $\mathbf{C} \equiv \mathbf{A}^T\mathbf{A}$. We note that this matrix is symmetric ($\mathbf{C}^T = \mathbf{C}$) since $\mathbf{C}^T = (\mathbf{A}^T\mathbf{A})^T = \mathbf{A}^T\mathbf{A} = \mathbf{C}$. This allows us to write the bracketed term on left hand side as

$$c_{ij} = \sum_{k=1}^{n}(A^T)_{ik}A_{kj} = \sum_{k=1}^{n} a_{ki}a_{kj}$$

so the left hand side becomes

$$\sum_{j=1}^{p} c_{rj}\theta_j = \sum_{i=1}^{n} a_{ir}y_i = \sum_{i=1}^{n}(\mathbf{A}^T)_{ri}y_i$$

or

$$(\mathbf{C}\boldsymbol{\theta})_r = (\mathbf{A}^T\vec{y})_r, \qquad r = 1, \ldots, p$$

Thus, we have obtained the so-called "normal equations"

$$\mathbf{C}\boldsymbol{\theta} = \mathbf{A}^T\vec{y} \qquad \mathbf{C} = \mathbf{A}^T\mathbf{A}$$

The solution to these normal equations is obtained by multiplying both sides of the equation on the left by the inverse of $\mathbf{C}$, which is simply $\mathbf{C}^T$ (since $\mathbf{C}$ is a symmetric matrix). Since the vector $\boldsymbol{\theta}$ has been obtaining from the extremum condition $\partial l/\partial\boldsymbol{\theta} = 0$ we denote this particular vector as $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\theta}} = \mathbf{C}^{-1}\mathbf{A}^T\vec{y} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\vec{y} \qquad \text{"normal equations"}$$

provided that the symmetric matrix $\mathbf{C} = \mathbf{A}^T\mathbf{A}$ is invertible. We note that the normal equations are linear in $\vec{y}$. They are also linear in $\boldsymbol{\theta}$, meaning that we can easily solve for $\boldsymbol{\theta}$. In the case of nonlinear models, we will see that the normal equations are generally not linear functions of the parameters.

Finally, to summarize what we have done, we note that maximizing the likelihood function L is entirely equivalent to minimizing the sum of square errors with respect to the choice of $\boldsymbol{\theta}$

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} a_{ij}\theta_j\right)^2 = (\vec{y} - \mathbf{A}\boldsymbol{\theta})^T(\vec{y} - \mathbf{A}\boldsymbol{\theta}) = \vec{\epsilon}^T\vec{\epsilon}.$$

The solution to this problem is given by the solution to the normal equations. The term $\vec{\epsilon}^T\vec{\epsilon}$ is an inner product of $\vec{\epsilon}$ with itself. This is often expressed in terms of the norm (length) of the vector $\vec{\epsilon}$ as follows: $\vec{\epsilon}^T\vec{\epsilon} = \|\vec{\epsilon}\|^2$, where $\|\vec{\epsilon}\|$ is the length of $\vec{\epsilon}$.

To illustrate the use of the normal equations in solving linear least squares problems, let us look at the example of a linear model ($y_i = A + Bx_i$). This equation must be written in the form $\vec{y}(\vec{x}|\boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\theta}$ where $\mathbf{A}$ is the design matrix. By inspection we see that

$$\vec{y}(\vec{x}|\boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\theta} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \\ x_n & 1 \end{bmatrix} \begin{bmatrix} B \\ A \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

The experimentally measured data points $\vec{y}$ are expressed as the sum of the model $\vec{y}(\vec{x}|\boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\theta}$ plus the random errors $\vec{\epsilon}$:

$$\vec{y} = \mathbf{A}\boldsymbol{\theta} + \vec{\epsilon}$$

where

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} B \\ A \end{bmatrix}, \quad \vec{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

This is nothing more than a restatement of the fact that $\vec{\epsilon}$ is simply the deviation $\vec{y} - \mathbf{A}\boldsymbol{\theta}$ of experimentally measured data $\vec{y}$ from the model $\mathbf{A}\boldsymbol{\theta}$. That is, $\vec{\epsilon} = \vec{y} - \mathbf{A}\boldsymbol{\theta}$. We recall that χ^2 involves a summation over such deviations.

We now compute the product $\mathbf{C}^{-1}\mathbf{A}^T\vec{y}$, where $\mathbf{C} = \mathbf{A}^T\mathbf{A}$. First, the products $\mathbf{C} = \mathbf{A}^T\mathbf{A}$ and $\mathbf{A}^T\vec{y}$ are, respectively

$$\mathbf{C} = \mathbf{A}^T\mathbf{A} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \\ 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \\ x_n & 1 \end{bmatrix} = \begin{bmatrix} \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i & n \end{bmatrix},$$

and

$$\mathbf{A}^T\vec{y} = \begin{bmatrix} \sum_i x_i y_i \\ \sum_i y_i \end{bmatrix}.$$

Next, the inverse $\mathbf{C}^{-1}$ of this 2×2 matrix is computed from the usual formula

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

You can easily check by direct multiplication that this matrix meets the conditions required to be the inverse: $\mathbf{C}^{-1}\mathbf{C} = \mathbf{C}\mathbf{C}^{-1} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

In our case, the inverse of $\mathbf{C} = \left[\begin{smallmatrix} \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i & n \end{smallmatrix}\right]$ is

$$\mathbf{C}^{-1} = \frac{1}{n\sum_i x_i^2 - \sum_i x_i \sum_j x_j} \begin{bmatrix} n & -\sum_i x_i \\ -\sum_i x_i & \sum_i x_i^2 \end{bmatrix}.$$

Therefore, $\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{B} \\ \hat{A} \end{bmatrix}$ is equal to

$$\mathbf{C}^{-1}\mathbf{A}^T\vec{y} = \frac{1}{n\sum_i x_i^2 - \sum_i x_i \sum_j x_j} \begin{bmatrix} n\sum_i x_i y_i - \sum_i x_i \sum_j y_j \\ -\sum_i x_i \sum_j x_j y_j + \sum_i x_i^2 \sum_j y_j \end{bmatrix}.$$

This result is nothing new: this is exactly the same result we have derived in Lecture 4 by applying the maximum likelihood principle to the minimization of χ^2. Here, we simply verified that our normal equation is in agreement with the result already derived previously.

8.3. MLE of the Parameter σ

There is yet another parameter that can be estimated via maximum likelihood, σ. While σ does not appear at first glance to be a "model" parameter in the sense of the model expressed as $\vec{y}(\vec{x}|\boldsymbol{\theta}) = \mathbf{A}(\vec{x})\boldsymbol{\theta}$, we initially stated the assumption that our deviations $\vec{\epsilon} = \vec{y} - \vec{y}(\vec{x}|\boldsymbol{\theta})$ were distributed as Gaussians with mean 0 and standard deviation σ. In this sense, σ is an unknown parameter that can be solved for in terms of the data and the model.

Differentiation of the log likelihood

$$l \equiv \log L = const - n\log\sigma - \frac{1}{2}\left[\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} a_{ij}\theta_j\right)^2\right]/\sigma^2.$$

with respect to the parameter σ gives

$$\frac{\partial l}{\partial \sigma} = 0 \qquad \Rightarrow \qquad -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} a_{ij}\theta_j\right)^2 = 0$$

At the extremum, we will denote the value of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$. Similarly, let us denote the value of σ as $\hat{\sigma}$ at the extremum. The quantities $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}$ are called *maximum likelihood estimators*. Solving for σ^2 we get, for the maximum likelihood estimator:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} a_{ij}\hat{\boldsymbol{\theta}}_j\right)^2.$$

8.4. The Covariance Matrix

An important concept that will be used later in this course is the covariance matrix. We recall near the end of Lecture 2 we had defined the covariance of two random variables X and Y:

$$cov(X,Y) \equiv \overline{(X - \overline{X})(Y - \overline{Y})} = \overline{XY} - \overline{X}\cdot\overline{Y}.$$

Covariance is important because in the case $Y = X$ it reports the variance whereas for $X \neq Y$ it tell us to what extent the random variable X is correlated to Y. The covariance matrix is used when X is a vector in which we have collected all random variables of interest, i.e. $\vec{X} = (X, Y, \ldots, Z)^T$. It is defined as the matrix whose elements are $cov(X_i, X_j)$. It can be computed most easily using the matrix form

$$cov(\vec{X}, \vec{X}) = \overline{(\vec{X} - \overline{X})(\vec{X} - \overline{X})^T}.$$

Here, we are interested in the covariance matrix of the fitted model parameters $cov(\boldsymbol{\theta}, \boldsymbol{\theta})$. The diagonal elements are the variances (errors squared) of each model parameters. The off-diagonal elements are the covariances. The matrix elements are important because we are interested in knowing from the fitting procedure what are the uncertainties in each fitted parameter and the extent to which model parameters are redundant (correlated). Using the definition $\hat{\boldsymbol{\theta}} = \mathbf{C}^{-1}\mathbf{A}^T\vec{y}$ we find:

$$\begin{aligned} cov(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}) =& \overline{\mathbf{C}^{-1}\mathbf{A}^T(\vec{y} - \overline{y})(\mathbf{C}^{-1}\mathbf{A}^T(\vec{y} - \overline{y}))^T} \\ =& \mathbf{C}^{-1}\mathbf{A}^T\overline{(\vec{y} - \overline{y})(\vec{y} - \overline{y})^T}\mathbf{A}(\mathbf{C}^{-1})^T. \end{aligned}$$

We recognize the term $\overline{(\vec{y} - \overline{y})(\vec{y} - \overline{y})^T}$ as a variance. Since[1] the elements of $\vec{y}$ are statistically independent with variance σ^2 (by assumption), the off-diagonal elements are zero and this variance evaluates to

$$\overline{(\vec{y} - \overline{y})(\vec{y} - \overline{y})^T} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{1}$$

where $\mathbf{1}$ is the unit matrix. In the second line we have used the property $(AB)^T = B^TA^T$ to write the term $(\mathbf{C}^{-1}\mathbf{A}^T(\vec{y} - \overline{y}))^T$ as

$$(\mathbf{C}^{-1}\mathbf{A}^T)^T(\vec{y} - \overline{y}) = (\vec{y} - \overline{y})^T\mathbf{A}(\mathbf{C}^{-1})^T.$$

Next, we use the fact that inverse of a matrix and transpose operations are interchangeable, i.e. $(\mathbf{C}^T)^{-1} = (\mathbf{C}^{-1})^T$. But since $\mathbf{C} = \mathbf{C}^T$ (symmetric matrix) we have:

$$cov(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}) = \mathbf{C}^{-1}\mathbf{A}^T\sigma^2\mathbf{1}\mathbf{A}\mathbf{C}^{-1} = \sigma^2\mathbf{C}^{-1}\underbrace{\mathbf{A}^T\mathbf{A}}_{\mathbf{C}}\mathbf{C}^{-1} = \sigma^2\mathbf{C}^{-1}.$$

We will see later that the covariance matrix is related to the inverse of the Hessian matrix. This is useful because we have seen in previous lectures that the Hessian matrix could be estimated from the Jacobian matrix during the course of a data fitting procedure (c.f. Gauss-Newton method).

[1] Recall that $\vec{y} = \mathbf{A}\boldsymbol{\theta} + \vec{\epsilon}$ is a sum of a deterministic (non-random) quantity $\mathbf{A}\boldsymbol{\theta}$ plus a random quantity $\vec{\epsilon}$ which we have assumed the elements of which to be iid rvs with $N(0, \sigma^2)$. Thus, $\text{var}(\vec{y}) = \text{var}(\vec{\epsilon}) = \sigma^2$.

This formula $cov(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}) = \sigma^2 \mathbf{C}^{-1}$ we have derived is simply a tool that enables us to compute the covariance matrix $cov(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})$ from the conditions of the experiment. The covariance matrix is important because its diagonal elements are the variances in the fitted parameters. Let us look at a simple example: that of the linear model $y_i = A + Bx_i$. Then, $\boldsymbol{\theta} = (B, A)^T$ and therefore

$$cov(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}) = \begin{bmatrix} \overline{(B-\overline{B})(B-\overline{B})} & \overline{(B-\overline{B})(A-\overline{A})} \\ \overline{(A-\overline{A})(B-\overline{B})} & \overline{(A-\overline{A})(A-\overline{A})} \end{bmatrix}$$
$$= \begin{bmatrix} var(B) & cov(A,B) \\ cov(B,A) & var(A) \end{bmatrix}.$$

We have also seen that

$$\mathbf{C}^{-1} = \frac{1}{n\sum_i x_i^2 - \sum_i x_i \sum_j x_j} \begin{bmatrix} n & -\sum_i x_i \\ -\sum_i x_i & \sum_i x_i^2 \end{bmatrix}.$$

Thus, we have an explicit expression for the formula $cov(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}) = \sigma^2 \mathbf{C}^{-1}$ and can use it to compute the errors in the fitted parameters A and B from the experimental conditions (design matrix). We note that the $\{y_i\}$-dependence of the errors and covariances originates from the factor σ^2.

8.5. Nonlinear Least Squares

Suppose that we have n observations $(\vec{x}_i, y_i)$, $i = 1, \ldots, n$ and a nonlinear model

$$y_i = f(\vec{x}_i|\boldsymbol{\theta}) + \epsilon_i, \qquad i = 1, \ldots, n$$

or

$$\vec{y} = \vec{f}(\vec{x}|\boldsymbol{\theta}) + \vec{\epsilon},$$

where $\vec{f}(\vec{x}|\boldsymbol{\theta})$ is a nonlinear function of $\boldsymbol{\theta}$. We assume that the elements of $\vec{\epsilon}$ are independent identically distributed random variables with mean 0 and variance σ^2. Furthermore, we assume they are normally distributed. Thus, the estimates $\hat{\boldsymbol{\theta}}$ will also be known as maximum likelihood estimators.

Given this nonlinear model, let us obtain the least squares estimates of $\boldsymbol{\theta}$ and σ, which we denote $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}$ respectively, starting from the likelihood function

$$L(\boldsymbol{\theta}, \sigma^2) \equiv p(\vec{y}|\boldsymbol{\theta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{[y_i - f(\vec{x}_i|\boldsymbol{\theta})]^2}{\sigma^2}\right)$$

The log likelihood is (ignoring constants)

$$l(\boldsymbol{\theta}, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - f(\vec{x}_i|\boldsymbol{\theta})]^2$$

8.5.1. MLE $\hat{\sigma}$. The condition $\partial l/\partial\sigma^2 = 0$ gives the maximum likelihood estimator:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(\vec{x}_i|\boldsymbol{\theta})]^2$$

which is indeed a "maximum" likelihood (for a given $\boldsymbol{\theta}$) as the second derivative is negative[2]:

$$\frac{\partial^2 l}{\partial(\sigma^2)^2} = \frac{\partial}{\partial\sigma^2}\left[\frac{-n}{2\sigma^2} + \frac{\sum_{i=1}^{n}[y_i - f(\vec{x}_i|\boldsymbol{\theta})]^2}{2(\sigma^2)^2}\right] = \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\sum_{i=1}^{n}[y_i - f(\vec{x}_i|\boldsymbol{\theta})]^2$$

in which we substitute the value of σ^2, namely $\frac{\vec{\epsilon}^T\vec{\epsilon}}{n}$ where $\vec{\epsilon}^T\vec{\epsilon} = \sum_{i=1}^{n}[y_i - f(\vec{x}_i|\boldsymbol{\theta})]^2$ to get

$$\frac{\partial^2 l}{\partial(\sigma^2)^2} = \frac{n^3}{2(\vec{\epsilon}^T\vec{\epsilon})^2} - \frac{n^3}{(\vec{\epsilon}^T\vec{\epsilon})^3}\cdot(\vec{\epsilon}^T\vec{\epsilon}) = \frac{n^3}{(\vec{\epsilon}^T\vec{\epsilon})^2}(\frac{1}{2} - 1) < 0$$

hence $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}^2$ maximize $l(\boldsymbol{\theta}|\sigma^2)$. The maximum value of $p(\vec{y}|\boldsymbol{\theta}, \sigma^2)$ is

$$p(\vec{y}|\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) = \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}}e^{-n/2}.$$

8.5.2. MLE $\hat{\boldsymbol{\theta}}$. The estimator $\hat{\boldsymbol{\theta}}$ satisfies the condition $\partial\vec{\epsilon}^T\vec{\epsilon}/\partial\theta_r|_{\hat{\boldsymbol{\theta}}} = 0$, $r = 1, 2, \ldots, p$. We now write $f_i(\boldsymbol{\theta}) = f(\vec{x}_i|\boldsymbol{\theta})$ as shorthand notation. Then,

$$\vec{f}(\boldsymbol{\theta}) = (f_1(\boldsymbol{\theta}), \ldots, f_n(\boldsymbol{\theta}))^T.$$

Since f is a vector, its gradient is a matrix:

$$\mathbf{F}(\boldsymbol{\theta}) \equiv \nabla\vec{f}(\boldsymbol{\theta}) = \frac{\partial\vec{f}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \left[\frac{\partial f_i(\boldsymbol{\theta})}{\partial\theta_j}\right] = \begin{bmatrix} \frac{\partial f_1}{\partial\theta_1} & \cdots & \frac{\partial f_1}{\partial\theta_p} \\ \frac{\partial f_2}{\partial\theta_1} & \cdots & \frac{\partial f_2}{\partial\theta_p} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial\theta_1} & \cdots & \frac{\partial f_n}{\partial\theta_p} \end{bmatrix}.$$

We now use the shorthand notation $\mathbf{F} = \mathbf{F}(\boldsymbol{\theta})$ and $\hat{\mathbf{F}} = \mathbf{F}(\hat{\boldsymbol{\theta}})$. The sum of squares that must be minimized is

$$\vec{\epsilon}^T\vec{\epsilon}(\boldsymbol{\theta}) = [\vec{y} - \vec{f}(\boldsymbol{\theta})]^T[\vec{y} - \vec{f}(\boldsymbol{\theta})] = \|\vec{y} - \vec{f}(\boldsymbol{\theta})\|^2.$$

[2]Another way to see this, since $\vec{\epsilon}^T\vec{\epsilon}(\boldsymbol{\theta}) \geq \vec{\epsilon}^T\vec{\epsilon}(\hat{\boldsymbol{\theta}})$, we have $l(\hat{\boldsymbol{\theta}}) - l(\boldsymbol{\theta}) = -(n/2)\log\hat{\sigma}^2 - (n/2) - l(\boldsymbol{\theta}) \geq -(n/2)\log\hat{\sigma}^2/\sigma^2 - (n/2) + (1/2)\vec{\epsilon}^T\vec{\epsilon}(\hat{\boldsymbol{\theta}})/\sigma^2 = -(n/2)(\log\hat{\sigma}^2/\sigma^2 + 1 - \hat{\sigma}^2/\sigma \geq 0$ as $\log x \leq x - 1$ for $x \geq 0$. We have denoted $\vec{\epsilon}^T\vec{\epsilon}(\hat{\boldsymbol{\theta}})$ for $\vec{\epsilon}^T\vec{\epsilon}$ evaluated at the point $\hat{\boldsymbol{\theta}}$.

where $\vec{\epsilon}^T\vec{\epsilon}(\boldsymbol{\theta})$ denotes $\vec{\epsilon}^T\vec{\epsilon}$ evaluated at the point $\boldsymbol{\theta}$. Setting $\partial\vec{\epsilon}^T\vec{\epsilon}(\boldsymbol{\theta})/\partial\theta_r|_{\hat{\boldsymbol{\theta}}} = 0$ leads to the result

$$\sum_i \{y_i - f_i(\boldsymbol{\theta})\} \left.\frac{\partial f_i(\boldsymbol{\theta})}{\partial\theta_r}\right|_{\hat{\boldsymbol{\theta}}} = 0 \qquad \Rightarrow \qquad 0 = \hat{\mathbf{F}}^T \underbrace{\{\vec{y} - \vec{f}(\boldsymbol{\theta})\}}_{\hat{\epsilon}}$$

We get the normal equations for this nonlinear model

$$0 = \hat{\mathbf{F}}^T \cdot \hat{\epsilon}.$$

This is the equation that must be solved for nonlinear models. We note that the model parameters are contained implicitly in the matrix $\mathbf{F}$ as partial derivatives. Thus, if the model $\vec{f}$ is nonlinear, the matrix elements of $\mathbf{F}$ could also be nonlinear functions of the parameters $\boldsymbol{\theta}$.

In general, these normal equations cannot be solved analytically and iterative methods of the type covered previously will be necessary.

In order to better understand how to use the normal equations we illustrate this concept by looking at examples.

Consider the linear model

$$\underbrace{y_i}_{\text{data}} = \underbrace{A + Bx_i}_{\text{model}} + \epsilon_i$$

The normal equations are $0 = \hat{\mathbf{F}}^T \cdot \hat{\epsilon}$. The vector $\hat{\epsilon}$ is simply the vector of deviations between experimental data $\vec{y}$ and the model $\vec{y}(\mathbf{x}|\hat{\boldsymbol{\theta}})$:

$$\hat{\epsilon} = \vec{y} - \vec{y}(\vec{x};\hat{\boldsymbol{\theta}}) = \vec{y} - A - B\vec{x} = \begin{bmatrix} y_1 - A - Bx_1 \\ y_2 - A - Bx_2 \\ \vdots \\ y_n - A - Bx_n \end{bmatrix}$$

or, in component form, $\epsilon_j = y_i - A - Bx_i$. Next, we need the matrix $\hat{\mathbf{F}}^T$, which is computed from $\mathbf{F}(\boldsymbol{\theta})^T = \begin{bmatrix} \frac{\partial f_1}{\partial\theta_1} & \cdots & \frac{\partial f_n}{\partial\theta_1} \\ \frac{\partial f_1}{\partial\theta_2} & \cdots & \frac{\partial f_n}{\partial\theta_2} \\ \vdots & & \vdots \\ \frac{\partial f_1}{\partial\theta_p} & \cdots & \frac{\partial f_n}{\partial\theta_p} \end{bmatrix}$, where $f_i = A + Bx_i$ and $\boldsymbol{\theta} = (A, B)^T$, but with two parameters ($p = 2$)

$$\mathbf{F}(\boldsymbol{\theta})^T = \begin{bmatrix} \frac{\partial f_1}{\partial\theta_1} & \cdots & \frac{\partial f_n}{\partial\theta_1} \\ \frac{\partial f_1}{\partial\theta_2} & \cdots & \frac{\partial f_n}{\partial\theta_2} \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix}$$

The condition $\hat{\mathbf{F}}^T \cdot \hat{\epsilon} = 0$ then reads

$$\sum_i (y_i - A - Bx_i) = 0$$
$$\sum_i (y_i - A - Bx_i)x_i = 0$$

The two linear equations and two unknowns can be solved for A, B (taking care of noting that the second term in the first equation is $\sum_i A = nA$):

$$B = \frac{n\sum_i y_i x_i - \sum_i x_i \sum_j y_j}{n\sum_i x_i^2 - \sum_i x_i \sum_j x_j} \qquad A = \frac{\sum_i x_i^2 \sum_j y_j - \sum_i y_i x_i \sum_j x_j}{n\sum_i x_i^2 - \sum_i x_i \sum_j x_j}$$

This is identical to the solution derived earlier using the normal equations for the linear case.

Consider the nonlinear model

$$y_i = \alpha x_i^\beta + \epsilon_i \qquad (i = 1, \ldots, n)$$

This model can describe, for example, the dependence of mass of an object (y) on a side length (x) in β dimensions. The normal equations are:

$$\sum_i (y_i - \alpha x_i^\beta) x_i^\beta = 0$$
$$\sum_i (y_i - \alpha x_i^\beta) \alpha x_i^\beta \log x_i = 0$$

These equations do not admit analytical solutions for α and β.

8.6. Linearizing a nonlinear model

In the nonlinear case, the minimization of $\|\vec{y} - \vec{f}(\boldsymbol{\theta})\|^2$ with respect to $\boldsymbol{\theta}$ yielded the normal equation $0 = \hat{\mathbf{F}}^T \cdot \hat{\epsilon}$, where $\mathbf{F} = \partial \vec{f}/\partial \boldsymbol{\theta}$ is generally nonlinear in $\boldsymbol{\theta}$ and $\hat{\epsilon} = \vec{y} - \vec{f}(\hat{\boldsymbol{\theta}})$ is also nonlinear in $\boldsymbol{\theta}$. These "normal equations", although formally correct, generally cannot be solved algebraically. In many cases, we resort to linearizing the nonlinear model, as we have done with the Newton-Raphson method. This approach does not yield the solution in a single step, but instead allows us to get progressively closer after each iteration of an update rule.

In order to do this, we first recall the following tool that we will use later. We have previously found that the minimization of $\|\vec{\epsilon}\|^2 = \|\vec{y} - \mathbf{A}\boldsymbol{\theta}\|^2$ has the following solution

$$\hat{\boldsymbol{\theta}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\vec{y}.$$

This was obtained by minimizing $\|\vec{\epsilon}\|^2$ with respect to $\boldsymbol{\theta}$, i.e. $\partial\|\vec{\epsilon}\|^2/\partial\boldsymbol{\theta} = 0$. This solution is not specific to data fitting, but is a general result from calculus.

Taylor expanding $\vec{f}(\boldsymbol{\theta})$ near $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ to first order gives $\vec{f}(\boldsymbol{\theta}) \approx \vec{f}(\hat{\boldsymbol{\theta}}) + \mathbf{F}\Delta\boldsymbol{\theta}$ where $\Delta\boldsymbol{\theta} = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$. Thus,

$$\|\vec{y} - \vec{f}(\boldsymbol{\theta})\|^2 \approx \|\vec{y} - \vec{f}(\hat{\boldsymbol{\theta}}) - \mathbf{F}\Delta\boldsymbol{\theta}\|^2.$$

Substitution of $\hat{\epsilon} = \vec{y} - \vec{f}(\hat{\boldsymbol{\theta}})$ for $\vec{f}(\hat{\boldsymbol{\theta}})$ yields

$$\|\hat{\epsilon} - \mathbf{F}\Delta\boldsymbol{\theta}\|^2.$$

Invoking the above theorem from calculus, we find that the solution of this linearized problem is given by

$$\Delta\boldsymbol{\theta} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\hat{\epsilon}.$$

Incidentally, this condition

$$\Delta\boldsymbol{\theta} = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\hat{\epsilon} \qquad (*)$$

is exactly equivalent to the update rule we derived previously for the Gauss-Newton method

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \mathbf{H}_k^{-1}\nabla\chi^2(\boldsymbol{\theta}^{(k)})$$

where the Hessian matrix $\mathbf{H}_k$ is approximated by $2\mathbf{J}_k^T\mathbf{J}_k$ with

$$\mathbf{J}_k = \mathbf{J}(\boldsymbol{\theta}^{(k)}) = \begin{bmatrix} -\frac{1}{\sigma_1}\frac{\partial y(x_1|\boldsymbol{\theta}^{(k)})}{\partial\theta_1} & \cdots & -\frac{1}{\sigma_1}\frac{\partial y(x_1|\boldsymbol{\theta}^{(k)})}{\partial\theta_p} \\ \vdots & & \vdots \\ -\frac{1}{\sigma_n}\frac{\partial y(x_n|\boldsymbol{\theta}^{(k)})}{\partial\theta_1} & \cdots & -\frac{1}{\sigma_n}\frac{\partial y(x_n|\boldsymbol{\theta}^{(k)})}{\partial\theta_p} \end{bmatrix}$$

and $\nabla\chi^2 = 2\mathbf{J}_k^T\vec{\tilde{R}}$ with $\vec{\tilde{R}}$ is the column vector of residuals $\tilde{R}_i = (y_i - y(x_i|\boldsymbol{\theta}^{(k)}))/\sigma_i$. For the two rules to be equivalent we must take the errors to be identical $\sigma_i \equiv \sigma$. In that case,

$$\mathbf{H}_k^{-1}2\mathbf{J}_k^T\begin{bmatrix}(y_1 - y(x_1|\boldsymbol{\theta}^{(k)}))/\sigma \\ \vdots \\ (y_n - y(x_n|)\boldsymbol{\theta}^{(k)}))/\sigma\end{bmatrix} = (2\mathbf{J}_k^T\mathbf{J}_k)^{-1}2\mathbf{J}_k^T\begin{bmatrix}(y_1 - y(x_1|\boldsymbol{\theta}^{(k)}))/\sigma \\ \vdots \\ (y_n - y(x_n|\boldsymbol{\theta}^{(k)}))/\sigma\end{bmatrix}$$

and we can now write $\mathbf{J}_k = -(1/\sigma)\mathbf{F}$ to get

$$= -\sigma^2(2\mathbf{F}^T\mathbf{F})^{-1}2(1/\sigma)\mathbf{F}^T\begin{bmatrix}(y_1 - y(x_1|\boldsymbol{\theta}^{(k)}))/\sigma \\ \vdots \\ (y_n - y(x_n|\boldsymbol{\theta}^{(k)}))/\sigma\end{bmatrix} = -(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\hat{\epsilon}$$

which gives an update rule identical to $(*)$

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\hat{\epsilon}$$

8.7. Relationship between Hessian and Covariance Matrices

The covariance matrix, $cov(\boldsymbol{\theta}, \boldsymbol{\theta})$, is important because it contains information about the statistics of the fitted parameters. Let us see how it can be

computed in the case of an iterative algorithm for nonlinear data fitting. In the linear case, we found an explicit formula:

$$cov(\boldsymbol{\theta}, \boldsymbol{\theta}) = \sigma^2 \mathbf{C}^{-1} = \sigma^2 (\mathbf{A}^T \mathbf{A})^{-1}. \qquad \text{(linear case)}$$

In the nonlinear case, we can linearize such that the matrix $\mathbf{A}$ becomes $\mathbf{F} = \partial \vec{f} / \partial \boldsymbol{\theta}$, the gradient of $\vec{f}$:

$$cov(\boldsymbol{\theta}, \boldsymbol{\theta}) = \sigma^2 (\mathbf{F}^T \mathbf{F})^{-1}. \qquad \text{(nonlinear case)}$$

Let us recall the definition of χ^2

$$\chi^2 = \frac{\sum_i (y_i - f_i(\boldsymbol{\theta}))^2}{\sigma^2} = \frac{\|\vec{y} - \vec{f}(\boldsymbol{\theta})\|^2}{\sigma^2} = \frac{\hat{\epsilon}^T \hat{\epsilon}}{\sigma^2}$$

and compute the Hessian matrix,

$$\begin{aligned}
(\mathbf{H})_{ij} =& \frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_j} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \frac{\|\vec{y} - \vec{f}(\boldsymbol{\theta})\|^2}{\sigma^2} \approx \frac{\partial^2}{\partial \theta_i \partial \theta_j} \frac{\|\vec{y} - \vec{f}(\hat{\boldsymbol{\theta}}) - \mathbf{F}\Delta\boldsymbol{\theta}\|^2}{\sigma^2} \\
=& \frac{\partial^2}{\partial \theta_i \partial \theta_j} \frac{[\vec{y} - \vec{f}(\hat{\boldsymbol{\theta}}) - \mathbf{F}\Delta\boldsymbol{\theta}]^T [\vec{y} - \vec{f}(\hat{\boldsymbol{\theta}}) - \mathbf{F}\Delta\boldsymbol{\theta}]}{\sigma^2} \\
=& \frac{\partial^2}{\partial \theta_i \partial \theta_j} \frac{(\vec{y} - \vec{f}(\hat{\boldsymbol{\theta}}))^T (\vec{y} - \vec{f}(\hat{\boldsymbol{\theta}})) + 2(-\mathbf{F}\Delta\boldsymbol{\theta})^T (\vec{y} - \vec{f}(\hat{\boldsymbol{\theta}})) + (\mathbf{F}\Delta\boldsymbol{\theta})^T (\mathbf{F}\Delta\boldsymbol{\theta})}{\sigma^2}
\end{aligned}$$

The first term, $(\vec{y} - \vec{f}(\hat{\boldsymbol{\theta}}))^T (\vec{y} - \vec{f}(\hat{\boldsymbol{\theta}})$, depends on $\hat{\boldsymbol{\theta}}$ but is independent of $\boldsymbol{\theta}$ and its derivative with respect to $\boldsymbol{\theta}$ is therefore zero. The second term, $2(-\mathbf{F}\Delta\boldsymbol{\theta})^T (\vec{y} - \vec{f}(\hat{\boldsymbol{\theta}}))$ depends linearly on beta, and so its second derivative with respect to $\boldsymbol{\theta}$ vanishes. This leaves only the third term as non-zero. Using the facts that $\Delta\boldsymbol{\theta} = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$ and $\mathbf{F} = \partial \vec{f}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}}$ is independent of $\boldsymbol{\theta}$, we compute the derivative:

$$\begin{aligned}
(\mathbf{H})_{ij} =& \frac{\partial^2}{\partial \theta_i \partial \theta_j} \frac{(\mathbf{F}\Delta\boldsymbol{\theta})^T (\mathbf{F}\Delta\boldsymbol{\theta})}{\sigma^2} = \frac{1}{\sigma^2} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \boldsymbol{\theta}^T \mathbf{F}^T \mathbf{F} \boldsymbol{\theta} \\
=& \frac{1}{\sigma^2} \frac{\partial^2}{\partial \theta_i \theta_j} \sum_{m,n} \theta_m (\mathbf{F}^T \mathbf{F})_{mn} \theta_n \\
=& \frac{1}{\sigma^2} \frac{\partial}{\partial \beta_i} \Big(\sum_n (\mathbf{F}^T \mathbf{F})_{jn} \theta_n + \sum_m \theta_m (\mathbf{F}^T \mathbf{F})_{mj} \Big) \\
=& \frac{1}{\sigma^2} ((\mathbf{F}^T \mathbf{F})_{ji} + (\mathbf{F}^T \mathbf{F})_{ij}) = \frac{2}{\sigma^2} (\mathbf{F}^T \mathbf{F})_{ij},
\end{aligned}$$

where in the second equality we have replaced $\Delta\boldsymbol{\theta}$ by $\boldsymbol{\theta}$ because the derivative operation with respect to $\boldsymbol{\theta}$. This leads to:

$$\mathbf{H} = (2/\sigma^2)(\mathbf{F}^T \mathbf{F}).$$

8.7.1. Summary: relationship between Hessian and covariance matrix. Thus, we have shown that the covariance matrix is proportional to the inverse of the Hessian

$$cov(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}) = 2(\mathbf{H}_k)^{-1}$$

(The factor of 2 would go away had we defined the Hessian as the derivative of the log likelihood function; can you show this?) Recall that the covariance matrix contains variances of the fitted parameters $\boldsymbol{\theta}$ along the diagonal and covariances of these parameters as off-diagonal elements:

$$cov(\boldsymbol{\theta}, \boldsymbol{\theta}) = \begin{bmatrix} var(\theta_1) & cov(\theta_1, \theta_2) & \dots & cov(\theta_1, \theta_p) \\ cov(\theta_2, \theta_1) & var(\theta_2) & \dots & cov(\theta_1, \theta_p) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\theta_p, \theta_1) & cov(\theta_p, \theta_2) & \dots & var(\theta_p) \end{bmatrix}$$

For example, if we require the error bars on the fitted parameters, we may extract the diagonal elements of the matrix $2(\mathbf{H}_k)^{-1}$

$$\text{diag}(cov(\boldsymbol{\theta}, \boldsymbol{\theta})) = \begin{bmatrix} var(\theta_1) & 0 & \dots & 0 \\ 0 & var(\theta_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & var(\theta_p) \end{bmatrix}$$

For this to work, of course, we need the matrix $\mathbf{H}_k$ to be invertible (non-singular). This may not always be the case. We have seen in the previous demo how to compute matrix inverses and how to compute Hessians in Matlab if analytical expressions are used. In the case where the Hessian is computed numerically by finite differences we may use the approximation $2\mathbf{J}_k^T\mathbf{J}_k$.

Problem 131. The following model is to be used for data fitting

$$y(x) = Ax^2 + Bx + 2Bx^{-1}$$

where A and B are parameters to be determined. In terms of the above model: (a) Derive the Jacobian matrix.

(b) Derive the design matrix.

(c) Derive the Hessian matrix.

(d) Derive the covariance matrix.

(e) Derive the normal equations for this model.

Chapter 9

Pearson's Chi-Square Test

In this section we will describe the chi-square test. It can be used in at least two ways:

- In Regression Analysis: to determine the distance between the data and the fit. Here, the sum is over all data points.
- Test of Expected Distribution: here the test works on categorical data and we use it to determine the distance between two "histograms". Here, the sum is over all bins of the histogram.

9.1. χ^2 test

Suppose that we have ν independent random variables X_i $(i = 1, \ldots, \nu)$ each normally distributed with mean μ_i and variance σ_i^2. Chi-square is defined as:

(9.1)

$$\chi^2 \equiv \frac{(X_1 - \mu_1)^2}{\sigma_1^2} + \frac{(X_2 - \mu_2)^2}{\sigma_2^2} + \cdots + \frac{(X_\nu - \mu_\nu)^2}{\sigma_\nu^2} = \sum_{i=1}^{\nu} \frac{(X_i - \mu_i)^2}{\sigma_i^2} = \sum_{i=1}^{\nu} \tilde{R}_i^2.$$

Since $X_1, \ldots, X_\nu$ are random variables so is χ^2. Therefore, we will denote it as $\boldsymbol{\chi}^2$ (boldface type). Because the data is normally distributed, the normalized residuals are close to 1 (66% of all $\tilde{R}_i$ values are within ± 1 of 0). Hence, given a set of measurements $\{X_i\}$ $(i = 1, \ldots, \nu)$, if we have chosen the μ_i and σ_i^2 correctly, we may expect that a calculation 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 μ_i. 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^2, we may conclude that our data are not well described by our set of values μ_i.

9.2. χ^2 distribution

We note that χ^2 is a function of several independent random variables:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{y_i - y(x_i)}{\sigma_i} \right)^2 = \sum_{i=1}^{n} \tilde{R}_i^2,$$

where $\tilde{R}_i$ are each normally distributed with mean 0 and variance 1. It therefore a random variable itself. Let's derive the distribution of χ^2. In the general case of n rv's: let $Y = X_1^2 + X_2^2 + \cdots + X_n^2$. Then,

$$\mathbb{P}(Y < y) = \mathbb{P}(X_1^2 + \cdots + X_n^2 < y) = \int\limits_{\{x_1^2+\cdots+x_n^2<y\}} \cdots \int \frac{e^{-(x_1^2+\cdots+x_n^2)/2}}{(2\pi)^{n/2}} \mathrm{d}x_1 \ldots \mathrm{d}x_n.$$

The PDF is obtained by taking the limit $\mathrm{d}y \to 0$:[1]

$$\begin{aligned}
\mathbb{P}(y < Y < y + \mathrm{d}y) =& \mathbb{P}(y < X_1^2 + \cdots + X_n^2 < y + \mathrm{d}y) \\
=& \int\limits_{\{y<x_1^2+\cdots+x_n^2<y+\mathrm{d}y\}} \cdots \int \frac{1}{(2\pi)^{n/2}} e^{-(x_1^2+\cdots+x_n^2)/2} \mathrm{d}x_1 \ldots \mathrm{d}x_n \\
=& \frac{e^{-y/2}}{(2\pi)^{n/2}} \int\limits_{\{y<x_1^2+\cdots+x_n^2<y+\mathrm{d}y\}} \cdots \int \mathrm{d}x_1 \ldots \mathrm{d}x_n = \frac{e^{-y/2}}{(2\pi)^{n/2}} A \mathrm{d}R \\
=& \frac{y^{n/2-1} e^{-y/2} \mathrm{d}y}{2^{n/2}\Gamma(n/2)}.
\end{aligned}$$

where $R = \sqrt{y}$, $\mathrm{d}R = \frac{\mathrm{d}y}{2\sqrt{y}}$ and $A = \frac{2R^{n-1}\pi^{n/2}}{\Gamma(n/2)}$ is the area of the $(n-1)$-sphere. We recall that $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \mathrm{d}x$ and $\Gamma(n/2) = (n/2-1)!$ when $n \geq 1$ is even. Here, $0 \leq \chi^2 < \infty$. $\Gamma(p+1) = p\Gamma(p)$, $\Gamma(1/2) = \sqrt{\pi}$.

Let us boldface χ^2 and index it with ν, the number of degrees of freedom, i.e. $\boldsymbol{\chi}_\nu^2$. The mean of $\boldsymbol{\chi}_\nu^2$ is ν and its variance is 2ν. This can be seen as follows. For the mean:

$$\mathbb{E}(\chi^2) \;=\; \mathbb{E} \sum_{i=1}^{\nu} \frac{(X_i - \mu_i)^2}{\sigma_i^2} \;=\; \sum_{i=1}^{\nu} \frac{1}{\sigma_i^2} \mathbb{E}(X_i - \mu_i)^2 \;=\; \sum_{i=1}^{\nu} \frac{1}{\sigma_i^2} \sigma_i^2 \;=\; \nu$$

[1] Recall that $\mathbb{P}(y < Y < y + \mathrm{d}y) = \mathrm{d}\mathbb{P}(Y < y) = \mathbb{P}(Y < y + \mathrm{d}y) - \mathbb{P}(Y < y) \approx \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(Y < y) \cdot \mathrm{d}y$.

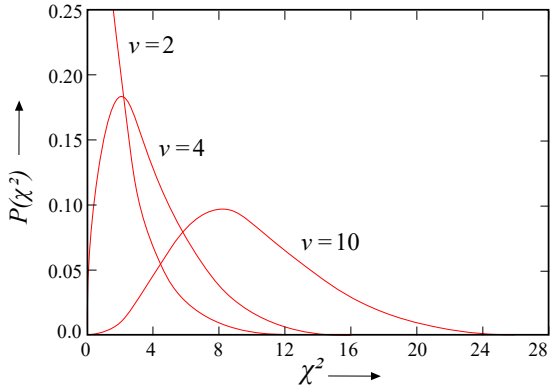


Figure 9.1. PDF of the χ^2 distribution for different values of the parameter ν.

where we have used the definition of the variance $\mathbb{E}(X_i - \mu_i)^2 = \sigma_i^2$. For the variance, $var(\chi^2)$, we have

$$var\left(\sum_{i=1}^{\nu} \frac{(X_i - \mu_i)^2}{\sigma_i^2}\right) = \sum_{i=1}^{\nu} \frac{1}{\sigma_i^4} var(X_i - \mu_i)^2$$

$$= \sum_{i=1}^{\nu} \frac{1}{\sigma_i^4} \left[\mathbb{E}(X_i - \mu_i)^4 - \sigma_i^4\right] = 2\nu$$

where we have used the 4th central moment of a Gaussian distribution, $\mathbb{E}(X_i - \mu_i)^4 = 3\sigma_i^4$.

The χ^2 distribution $p_\nu(\chi^2)$ is plotted in Fig. 9.1, for different values of ν.

9.2.1. How to use the test. Suppose that we have k experimentally measured quantities X_i and we want to test whether they are well-described by some set of hypothesized values μ_i. We form a sum using Eq. 9.1. It will contain k terms, constituting a sample value for χ^2.

The expected value of χ^2 will not be n because the $\{X_i\}$ may have been used to make estimates for the parameters of the model (e.g., μ_i and σ_i^2). Such relations (e.g., sample mean, sample variance) are said to reduce the number of degrees of freedom. If there are r such constraints then we number of degrees of freedom becomes:

$$\nu = k - r$$

and the resulting χ^2 will be one having ν (rather than k) degrees of freedom. Thus, the sum (Eq. 9.1) should be close to ν. In practice r will always be at least 1, because the data is normally used to estimate at least some parameter, such as the mean, variance, etc.

Generally speaking, $\chi^2/\nu \approx 1$ means the model represented by the μ_i and σ_i^2 is probably fine. If $\chi^2/\nu \ll 1$, we may conclude that either (i) the model is valid, but by chance, χ^2 ended up being too small. (ii) we have overestimated the values of σ_i^2. (iii) the data is fraudulent (too good to be true). If $\chi^2/\nu \gg 1$, we can conclude either that (i) the model is valid but by chance, χ^2 ended up being too high. (ii) the model is so poorly chosen that an unacceptably large value of χ^2 has resulted. (iii) data is not normally distributed. (Remember our assumption for least squares fitting that the data must be normally distributed about the model.)

9.2.2. Example 1: measuring mass of object using different types of balances. Suppose that we measure the mass of an object using four different balances (Cole-Parmer, Fischer, VWR and Nimbus). The results of the mass measurements are as follows:

Detector (scale)	mass (g)
Cole-Parmer	91.161 ± 0.013
Fischer	91.174 ± 0.011
VWR	91.186 ± 0.013
Nimbus	91.188 ± 0.013

The listed uncertainties are estimates of the σ_i, the standard deviations for each of the measurements. As can be seen, the error bars overlap. The question we would like to answer is: Can these data be well described by a single number, namely an estimate of the mass made by determining the weighted mean of the four measurements?

The weighted mean and its error are:

$$\overline{m} = \frac{\sum m_i/\sigma_i^2}{\sum 1/\sigma_i^2} \quad \text{and} \quad var(\overline{m}) = \frac{1}{(\sum 1/\sigma_i^2)^2} \sum \frac{1}{(\sigma_i^2)^2} var(m_i) = \frac{1}{\sum 1/\sigma_i^2}.$$

The value (for the 'true mean') we would report is:

$$\overline{m} \pm \sqrt{var(\overline{m})} = 91.177 \pm 0.006.$$

We then form χ^2:

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

Here we have $\nu = 4 - 1 = 3$ degrees of freedom (not 4), since we have used the weighted mean of the four measurements to estimate the value of the true mass, and this uses up one degree of freedom. We note that the errors $\{\sigma_i\}$ were not computed from the data, but instead obtained experimentally and constitute data. The value $\chi^2/\nu = 2.78/3 \approx 0.93$ is close to 1, and so we conclude that the model is probably fine. The weighted mean (and

assumption of Gaussian distribution of the measurements) is therefore valid. Note: our claim that 0.93 is close to 1 should be decided based on a statistical significance criterion (level of confidence). We will see later how to do this.

9.3. Test of Expected Distribution

The test of expected distribution (also called the χ^2 test) is used to test the nature of a statistical distribution from which some random sample is drawn. Suppose that we have k classes (e.g. designated categories or bins in a histogram), with probabilities $p_1, p_2, \ldots, p_k$ assigned to each class. If the chosen classes account for all the data, then $\sum p_i = 1$. If we take the data and plot a histogram, we classify the data: this is done by counting the number of observation falling into each of the k classes (or bins). We have n_1 counts in the first class, n_2 counts in the second class, and so on, up to n_k counts in the kth class. Suppose there is a total of N observations, so $\sum n_i = N$.

Then it can be shown that the sum

$$\frac{(n_1 - Np_1)^2}{Np_1} + \frac{(n_2 - Np_2)^2}{Np_2} + \cdots + \frac{(n_k - Np_k)^2}{Np_k} = \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i} \tag{9.2}$$

has approximately the χ^2 distribution with

$$\nu = k - r$$

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 it must be that $\sum n_i = \sum Np_i = N \sum p_i = N$. In other words, ν is given by:

- $\nu = k - 1$ if the expected frequencies can be computed without having to estimate the population parameters from sample statistics. We subtract 1 from k because of the constraint condition $\sum_j E_j = N$ (where $E_j = Np_j$), which states that if we know $k - 1$ of the expected frequencies, the remaining frequency can be determined.
- $\nu = k - 1 - m$ (i.e. $r = m + 1$) if the expected frequencies can be computed only by estimating m population parameters from sample statistics.

Since Np_i is the mean, or expected value of n_i, the form of χ^2 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. The above formula can be rewritten, using $E_i = Np_i$ and $O_i = n_i$:

$$\boxed{\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}} \tag{9.3}$$

where O_i stands for "observed value" and E_i stands for "expected value".

9.3.1. The case of Poisson counts. We note that Eq. (9.2) looks different from (9.1). However, in the case of Poisson distributed values, the mean is equal to the variance. So this link with the Poisson distribution provides the motivation for the definition (9.2). In fact, Eq. (9.3) is obtained from Eq. (9.1) by taking $\sigma_i^2 = E_i$ (the variance of the Poisson distribution is also the mean). This is justified because the vertical axis in a histogram is the number of counts whereas the number of counts is approximately Poisson distributed.

Equation (9.2) is sometimes written as the sum of observed minus expected values squared (divided by expected value). Your textbook uses O_i to denote the number of counts observed (during the ith interval or as a frequency of occurrence for some category/bin). If E_i denotes the expected number of counts (i.e. the mean value of O_i), we write:[2]

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad \text{(Poisson statistics)}$$

O_i: observed count/frequency,

E_i: expected count (model), as derived from the Poisson distribution formula. Note: E_i is not necessarily equal to $P(n;\overline{n}) = e^{-\overline{n}}\overline{n}^n/n!$; it depends on how the data is binned. However, it will be derived using $P(n;\overline{n})$.

k: number of bins/categories[3]

For good fits, the error obeys $\sqrt{E_i} \approx \sqrt{O_i}$. If

$$\chi^2 \gg \nu = k - r$$

(r: number of parameters in the distribution or the number of constraints used to compute the expected values) there is significant disagreement between observed and expected distribution. Please note that because χ^2 itself must be dimensionless, this test only makes sense if O and E are dimensionless quantities, such as counts.

Because E_i and O_i are counts (frequencies of occurrence) rather than physical quantities (with units), you must bin your data into a histogram and then compare the histogram to the distribution to which it is fitted to. In practice, the distribution needs to use experimental data to compute its parameters (e.g. obtained via maximum likelihood estimation). When such

[2]We have seen, in the case of Poisson statistics, that the error is given by $\sqrt{O_i}$. This approach is problematic for low counts, as low counts lead to large random fluctuations of the error in the denominator of the chi-square. One solution is to re-bin the data such that the counts are not too low.

[3]Or number of data points, n, if the test is applied as in the case of regression analysis, as discussed in the previous section).

parameters are computed from the data, we say that the degrees of freedom have been reduced. This is why the number of degrees of freedom is equal to the number of categories minus the number of parameters (or more precisely, the number of conditions used to estimate the parameters of the distribution from the data), $\nu = k - r$.

9.3.2. Example 2: counting trees on acres of land. Suppose that you own a large parcel of land in the desert that is divided into 48 acres (a rectangular region of 6 acres by 8 acres), and you hire someone to count the trees in each acre. The data ranges from 0 trees up to 6 trees per acre. But because there are very few acres that contain 4 or more trees, you decide to lump them all together into the same category ("4 or more trees"). You find the following data:

Observed counts					
Category	1	2	3	4	5
Number of trees in an acre	0	1	2	3	≥ 4
Frequency (number of acres in this category)	9	9	10	14	6

Suppose that for bin 5, the six counts were actually 4, 4, 5, 5, 6, 6. You want to check whether or not the counting statistics can be described by a Poisson distribution. The Poisson distribution requires the estimation of 1 parameter. The maximum likelihood estimator of the Poisson parameter is the arithmetic average (mean) of the sample data:

$$\overline{n} = \frac{(0)(9) + (1)(9) + (2)(10) + (3)(14) + (4)(2) + (5)(2) + (6)(2)}{9 + 9 + 10 + 14 + 2 + 2 + 2} = 2.10.$$

We have 48 counts and the expected number of counts in the jth category (bin) is given by $48 \times p_j$, where p_j is the probability of observing the stated counts in this category. The probabilities p_j are obtained using the Poisson distribution. Let X be the random variable that counts the number of trees in an acre. The probabilities are:

$$p_1 = \mathbb{P}(X = 0) = \frac{e^{-2.10}(2.10)^0}{0!} = e^{-2.10},$$

$$p_2 = \mathbb{P}(X = 1) = \frac{e^{-2.10}(2.10)^1}{1!} = e^{-2.10} \cdot 2.10,$$

$$p_3 = \mathbb{P}(X = 2) = \frac{e^{-2.10}(2.10)^2}{2!} = e^{-2.10} \cdot 2.205,$$

$$p_4 = \mathbb{P}(X = 3) = \frac{e^{-2.10}(2.10)^3}{3!} = e^{-2.10} \cdot 1.5435,$$

$$p_5 = \mathbb{P}(X \geq 4) = 1 - \mathbb{P}(X \leq 3) = 1 - e^{-2.10}(1 + 2.10 + 2.205 + 1.5435) = 7.74.$$

Then, the expected frequencies/counts in each category/bin are:

$$\begin{aligned} E_1 =& 48 \cdot p_1 = 5.8779, \\ E_2 =& 48 \cdot p_2 = 12.3436, \\ E_3 =& 48 \cdot p_3 = 12.9608, \\ E_4 =& 48 \cdot p_4 = 9.0726, \\ E_5 =& 48 \cdot p_5 = 7.74. \end{aligned}$$

The test of expected distribution (χ^2) can then be calculated:

$$\frac{(9-5.8779)^2}{5.8779}+\frac{(9-12.3436)^2}{12.3436}+\frac{(10-12.9608)^2}{12.9608}+\frac{(14-9.0726)^2}{9.0726}+\frac{(6-7.74)^2}{7.74},$$

which equals to 6.3097. This number should be compared to the degrees of freedom, $\nu = k - r = 5 - 1 = 4$. Is this a good fit? If the answer is yes, then we say that the data is well described by a Poisson distribution with parameter $\overline{n} = 2.10$. To determine this, we need to check whether $\chi^2/\nu \approx 1$ according to some level of confidence (see Section 9.3.4).

9.3.3. Example 3: height of people. Let us use the chi-square test to test whether a data sample consisting of the heights of N=66 people can be assumed to be drawn from a Gaussian distribution or not. We first arrange the data in the form of a frequency distribution, listing for each height h, the value of $n(h)$, the number of people in the sample whose height is h (h is in inches):

h	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73
$n(h)$	1	0	1	4	6	7	13	8	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 $p(h)\mathrm{d}h$ that a height falls between h and $h + \mathrm{d}h$ is given by:

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

This expression, if multiplied by N, will give, for a sample of N people, the theoretically expected number of people $n_e(h)\mathrm{d}h$ whose height should be between h and $h + \mathrm{d}h$:

$$n_e(h)\mathrm{d}h = \frac{N}{\sqrt{2\pi\sigma^2}} e^{-(h-\mu)^2/2\sigma^2}\mathrm{d}h.$$

In our example, $N = 66$. In our table we have grouped the data into bins, each of size 1 inch. A useful approximation is to take $\mathrm{d}h = 1$ inch, which gives the expected number of people having a height h_j:

$$n_e(j) = \frac{N}{\sqrt{2\pi\sigma^2}} e^{-(h-\mu)^2/2\sigma^2}. \tag{9.4}$$

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

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

Using these values we may calculate from Equation (9.4), the number expected in each bin, which the following results:

h	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73
$n_e(j)$	0.3	0.9	1.8	3.4	5.5	7.6	9.3	9.9	9.1	7.3	5.0	3.1	1.6	0.7	0.3	0.1

In applying the chi-square test to a situation of this type, it is advisable to re-group the data into new bins (classes), 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 given as follows (note the bin sizes need not be equal):

h	$\leq$61	62	63	64	65	66	67	68	$\geq$69
$n(h)$	6	6	7	13	8	11	2	7	6
$n_e(j)$	6.5	5.5	7.6	9.3	9.9	9.1	7.3	5.0	5.8

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

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

Since we grouped our data into 9 classes, and since we have used up three degrees of freedom by demanding that: (i) the sum of the n_j be equal to N, (ii) the mean of the distribution be equal to the sample mean and (iii) the variance is equal to the sample variance, the are 6 degrees of freedom left. Hence $\chi^2/\nu = 6.96/6 \approx 1.16$, which is close to 1. Therefore, we have no good reason to reject our hypothesis that our data are drawn from a Gaussian distribution function.

9.3.4. χ^2 distribution and significance tests. The χ^2 distribution is used to perform "significance tests", as explained below by way of examples. The main idea is that, as we repeat our experiment and collect values of χ^2, if our model is a valid one, these data will be clustered about the median value of χ^2_ν, with about half of them greater than the median and half less than the median. This median value, which we denote $\chi^2_{\nu,0.5}$ is determined by:

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

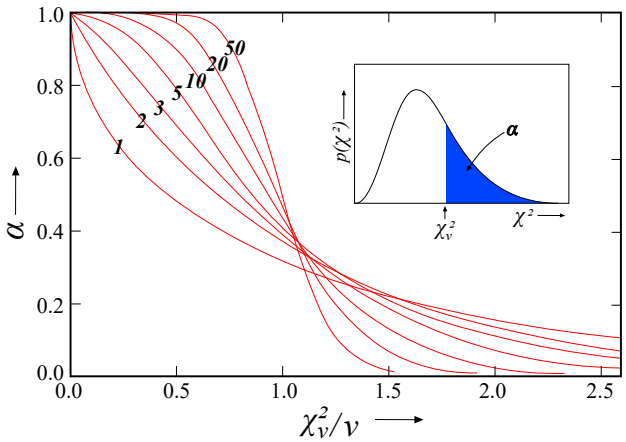


Figure 9.2. χ^2 test: plot of the upper tail probability of the chi-square distribution vs normalized chi-square (χ^2_ν/ν). The different curves correspond to values of $\nu = 1, \ldots, 50$.

In other words, we expect that a single measured value of χ^2 will have a probability of 0.5 of being greater than $\chi^2_{\nu,0.5}$. Generalizing this idea, to say that we expect a single measured value of χ^2 will have a probability α of being greater than $\chi^2_{\nu,\alpha}$, where $\chi^2_{\nu,\alpha}$ is defined by

$$\boxed{\int_{\chi^2_{\nu,\alpha}}^{\infty} p_\nu(\chi^2)\mathrm{d}\chi^2 = \alpha,} \qquad \alpha \in [0,1], \quad \chi^2_{\nu,\alpha} \in [0,\infty)$$

i.e., $\alpha \equiv \mathbb{P}(\boldsymbol{\chi}^2 > \chi^2_{\nu,\alpha})$. This definition is illustrated by the inset of Fig. 9.2, Fig. 9.2 plots α versus the normalized chi-square, 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.

In MATLAB this type of graph can be generated using the command `chi2cdf`:

```
t=0:0.1:3;
figure;
hold on;
for j=1:50
    plot(t,chi2cdf(t*j,j,'upper'));
end;
```

To use this test, we calculate χ^2 and ν. This gives the value χ^2/ν. Choose α, the significance level. Use the above figure to determine the corresponding value of $\chi^2_{\nu,\alpha}/\nu$. Compare this value with our sample value χ^2/ν. If we find that $\chi^2/\nu > \chi^2_{\nu,\alpha}/\nu$, then either: (i) a statistically improbable excursion of χ^2 has occurred; (ii) our model is poorly chosen or (iii) the data is not

normally distributed. If $\chi^2/\nu < \chi^2_{\nu,\alpha}/\nu$, we have either: (i) a valid model, but statistically improbable excursion of χ^2; (ii) the data is fraudulent ("too good to be true"). Note: a poor model can only increase the value of χ^2.

- **Example 1: measuring mass of object.** We expect this value of χ^2 to be drawn from a chi-square distribution with 3 degrees of freedom. From the above graph of α vs χ^2_ν for $\nu = 3$, the value $\chi^2/\nu = 2.78/3 \approx 0.93$ corresponds to α of about 0.42. Alternatively MATLAB can be used to obtain this value:

```
>> chi2cdf(2.78,3,'upper')

ans =

    0.4268
```

  This means that if we were to repeat the experiments we would have about a 42 percent chance of finding a χ^2 for the new measurement set larger than 2.78, assuming that the hypothesis (of a uniform distribution) is correct. We have therefore no good reasons to reject the hypothesis and conclude that the four measurements of the mass are consistent with each other. We would have had to have found χ^2 in the vicinity of 8 (leading to an α of about 0.05) to have been justified in suspecting the consistency of the measurement.[4] The fact that our value of $\chi^2/3$ is close to 1 is reassuring.

- **Example 2: counting trees on acres of land.** Let us return to the example of counting trees. There were 5 bins less 2 degrees of freedom ($\nu = 5 - 2 = 3$): (1) one degree of freedom was used to constrain the total number of acres; (2) the second degree of freedom was used to estimate the parameter of the Poisson distribution. Suppose that the confidence level is 5%, i.e. we set $\alpha = 0.05$. From the graph (at $\alpha = 0.05$ and $\nu = 3$), we find $\chi^2_\nu/\nu \sim 2.6$, which corresponds to the following critical value for the test: $\chi^2_{3,0.05} = 7.8$. Alternatively, this value can be obtained in MATLAB as follows:

```
>> chi2inv(1-0.05,3)

ans =
```

[4]From the graph, at $\alpha = 0.05$ and $\nu = 3$ the value of $\chi^2_{\nu,\alpha}/\nu = \chi^2_{3,0.05}/3$ is 2.6 (i.e., $\chi^2 \sim 7.8$). The fact that 0.9 is less than 2.6 does not mean that we have a poor model. (For that we would need a value greater than 2.6.) It could be, for example, that the data describe a statistically improbable excursion of χ^2.

```
7.8147
```

Comparing the value of $\chi^2 = 6.3$ we calculated in the example, this value is less than the critical value of the test. Thus, at the 0.05 level of significance we fail to reject the hypothesis that the data is well represented by the Poisson distribution with the estimated parameter, $\overline{n} = 2.10$.

- **Example 3: height of people.** Here $\chi^2/\nu = 6.96/6 \approx 1.16$, leading to an α of about 0.33. Therefore, we have no good reason to reject our hypothesis that our data are drawn from a Gaussian distribution function. (From the graph, $\chi^2_{\nu,0.05}/\nu$ for $\nu = 6$ is approximately equal to 2.2; and since 1.16 is less than 2.2, we cannot reject the model.)

9.4. Problem

Problem 132. The middle A note on a piano is normally tuned to the nominal value of 440 Hz. Suppose that your piano is tuned to 440 Hz on a given day. Over time, the tuning may change due to general detuning and differences in humidity and temperature. You monitor over time the exact pitch of the middle A note (normally tuned to 440 Hz). Suppose that the A note's pitch is measured on different days (day 1, 40, 80, 120) and different instruments with different error bars are used each time (i.e. instrument 1 on day 1, instrument 2 on day 40, instrument 3 on day 80, instrument 4 on day 120). Instruments 1-4 are properly calibrated (no bias), meaning that the pitch value is accurate and only its precision varies. The measurement of pitch frequency is a random variable X.

You measure the pitch at day 1 (1 day after tuning), day 40, day 80 and day 120. The results are:

Measurement Day	Instrument Used	Pitch (Hz)
Day 1	1	440.61 ± 15.11
Day 40	2	433.74 ± 0.15
Day 80	3	432.86 ± 0.10
Day 120	4	432.88 ± 0.13

(a) What is the mean of X, as calculated by the data and its error bars?

(b) Calculate the uncertainty of the mean calculated in (a).

(c) Report the pitch A as best value ± uncertainty, with the correct number of significant figures.

(d) Let's assume that the measurements are Gaussian distributed about their mean value. Is this a good assumption? (i.e. "prove" or disprove it, using the appropriate statistical test)

(e) Why would anyone assume that a measurement should be Gaussian-distributed?

Solution. (a) Use weighted average:

$$\overline{X} = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2} = 433.0581 \text{ Hz}$$

(b)

$$var\overline{X} = \frac{1}{\sum 1/\sigma_i^2} = 0.049$$

The uncertainty is the square root of this (0.0701 Hz).

(c) Keeping 1 digit for the error: 433.06 ± 0.07 Hz

(d) We use the chi square test:

$$\chi^2 = \sum_{i=1}^{4} \frac{(x_i - \overline{X})^2}{\sigma_i^2} = 26.72$$

the number of degrees of freedom is 4-1=3. (Note: there is only 1 constraint, used to obtain $\overline{X}$; the σ_i's are experimentally measured.) Dividing by 3 gives 8.9, which is much larger than 1. So Gaussian may not be a good assumption.

(e) Per central limit theorem (CLT). However, CLT does not always apply. ■

Problem 133. You arrive (by car) at the Dodgers stadium for a ball game and want to park near the entrance. However, the parking attendant redirects traffic into two threads. The first thread leads cars away from the stadium. The second thread leads to parking spots near the stadium's entrance. You were redirected far away from the stadium. You suspect that the parking attendant is being biased against economy cars and set out to test the hypothesis that the attendant's policy is to redirect more luxury cars toward the stadium and keep the economy cars away from the stadium.

(a) In order to test the hypothesis that the parking attendant was fair[5], you ask your friend to stand nearby and record the number of times that luxury cars get redirected toward the stadium. Suppose that out of 500 observations, he observed 265 luxury cars and 235 economy cars were redirected toward the stadium. Is the attendant fair or biased?

[5]By "fair" we mean no bias in car selection.

(b) After parking your car far out and walking nearly 1/2 mile back to the stadium, on your way back, you walk by the attendant and decide whether he is biased or not. Before making this decision, you want to double check the results from part (a). So you ask your friend to go and repeat this experiment. Out of 500 observations, your friend obtains an even higher count of luxury cars (270 luxury cars and 230 economy cars). Is the attendant biased?

Solution. (a)

$$\frac{(265-250)^2}{250}+\frac{(235-250)^2}{250}=1.8$$

because this number is close to 1, we cannot conclude there is bias. (Degrees of freedom: 2-1, since there are two categories, and 1 constraint for the total number of cars.)

(b)

$$\frac{(270-250)^2}{250}+\frac{(230-250)^2}{250}=3.2$$

again, this number is not orders of magnitude different than 1, so we cannot conclude there is bias. ■

Chapter 10

Machine Learning

10.1. Principal Component Analysis

Principal component analysis (PCA) is one of the oldest methods for unsupervised machine learning. Suppose we have n vectors $\vec{v}_j \in \mathbb{R}^d$, $j = 1, \ldots, n$ containing experimental data. d can be very large. The n measurements could represent, for example, the same vector measured on different days. We form the matrix:

$$C = \frac{1}{n} \sum_{j=1}^{n} \vec{v}_j \vec{v}_j^T$$

If the vectors have zero mean (easy to do by subtracting the mean), C plays the role of a covariance matrix. If we then diagonalize C and obtain the eigenvalues ω_k and corresponding eigenvectors $\vec{\omega}_k$, we can write C in terms of this eigen-decomposition:

$$C = \sum_k \omega_k \vec{\omega}_k \vec{\omega}_k^T$$

Suppose that only a small number of eigenvectors are nonzero. The matrix C is then of low rank.

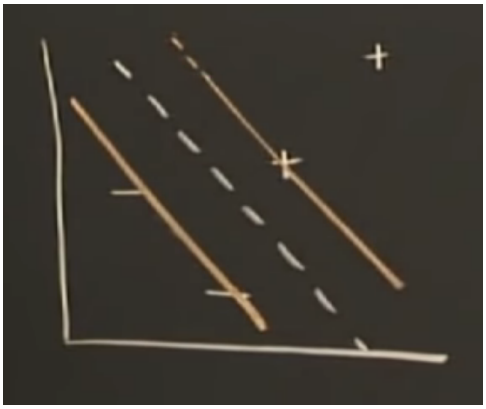
10.2. Support Vector Machines

The process of fitting data to a model is often called "learning". The output is a set of parameters that are "learned" from the data. A classification task is one where the output is a set of classes (e.g. "cat", "dog", "airplane", "boat", "positive", "negative", "yes", "no", "turn left", "turn right", etc.). Classification tasks are often used to classify images. The objects presented to the classifier are called vectors. To turn an image A, which is a 2D

matrix, into a vector, we may concatenate the columns into a long vector. This vectorization operation is often denoted $vec(A)$. The dimension of the vectors corresponds to the number of entries in the vector. A 2048×2048 image, for example, corresponds to a 4,194,304-dimensional vector.

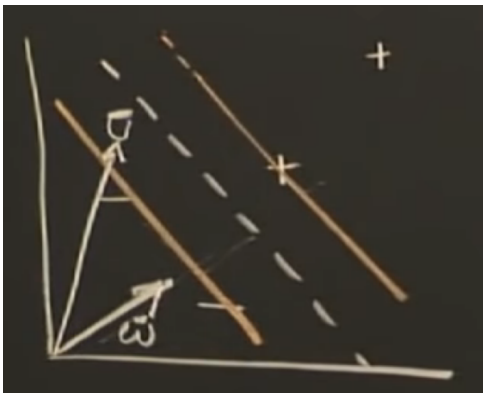
In this section we will discuss the method of support vector machines, originally developed by Vladimir Vapnik in the 1960s while he was a student in the USSR. A classifier's main purpose is to help make decisions. To make decisions we need to establish decision boundaries. Vapnik's method relies on the use of the "widest street approach" and kernel functions.

Suppose that we have a multidimensional space (illustrated below as a 2D space). Suppose also that our goal is to classify input vectors into two distinct classes: positive (+) and negative (-). Thus, we have "training data" consisting of positive and negative examples. We need a method to divide the positive examples from the negative ones. The simples approach is to draw a straight line. However, where do we draw the straight line? Vapnik suggested that the line be drawn with a view towards putting in the widest street that separates the positive samples from the negative ones:



This is called the "widest street approach". This suggests, for the decision boundary, to put in a straight line in such a way as the separation between the positive and negative examples results in the widest street between these two sets of vectors.

Consider a vector $\vec{\omega}$ that is perpendicular to the median line of the street, or equivalently, perpendicular to the gutters. We don't yet know its length. We also have some unknown vector $\vec{u}$. Our task is to decide whether this unknown is on the right or left side of the street. To do this, we want to project that vector $\vec{u}$ down on to one that's perpendicular to the street (e.g. $\vec{\omega}$), because then we'll have a distance in the direction $\vec{\omega}$. The latter is a number that's proportional to the distance in this direction ($\vec{\omega}$). The further we go, the closer we'll get to being on the right side of the street.



Let's take $\vec{\omega}$ and dot it with $\vec{u}$ and measure whether or not that number is equal to or greater than some constant c:

$$\vec{\omega} \cdot \vec{u} \geq c$$

This is the criterion for deciding whether a sample is positive or not. Or equivalently, set $c = -b$ and

$$\boxed{\vec{\omega} \cdot \vec{u} + b \geq 0 \text{ then } +} \qquad \text{(decision rule)}$$

However, we don't know what constant b to use, or which vector $\vec{\omega}$ to use. We know that $\vec{\omega}$ must be perpendicular to the median line of the street. However, there are many such possibilities of perpendicular vectors. We need some constraints to fix a particular b or $\vec{w}$.

Let us know take $\vec{\omega}$ and dot it with a positive sample x_+. We set the following constraint:

$$\vec{\omega} \cdot \vec{x}_+ + b \geq 1. \tag{10.1}$$

Likewise, a negative sample must obey:

$$\vec{\omega} \cdot \vec{x}_- + b \leq 1. \tag{10.2}$$

Let us introduce a convenient variable y_i defined such that

$$y_i = \begin{cases} +1 & \text{for } + \text{ samples} \\ -1 & \text{for } - \text{ samples} \end{cases}$$

Let use multiply Eqs. (10.1) and (eq:vap2) by y_i:

$$y_i(\vec{\omega} \cdot \vec{x}_+ + b) \geq 1,$$
$$y_i(\vec{\omega} \cdot \vec{x}_- + b) \geq 1.$$

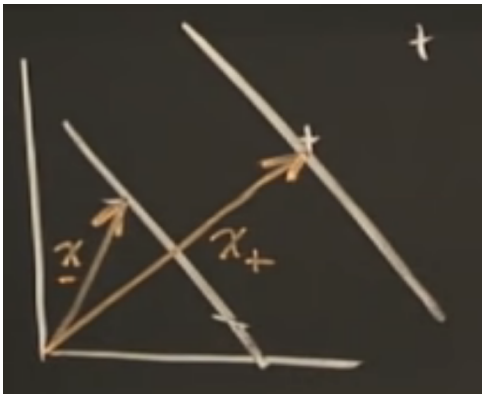
Both equations are the same, thanks to this mathematical convenience. Therefore, for any sample x_i:

$$y_i(\vec{\omega} \cdot \vec{x}_i + b) - 1 \geq 0.$$

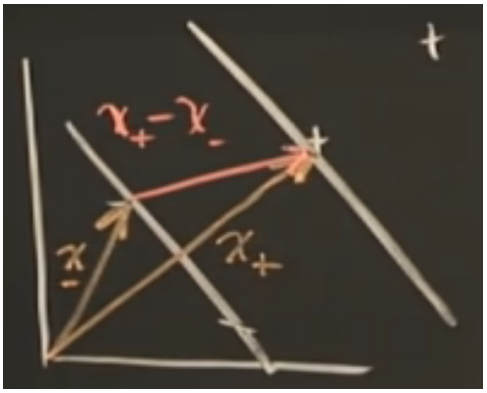
For samples in the gutter we set:

$$\boxed{y_i(\vec{\omega} \cdot \vec{x}_i + b) - 1 = 0.} \qquad \text{(gutter)} \tag{10.3}$$

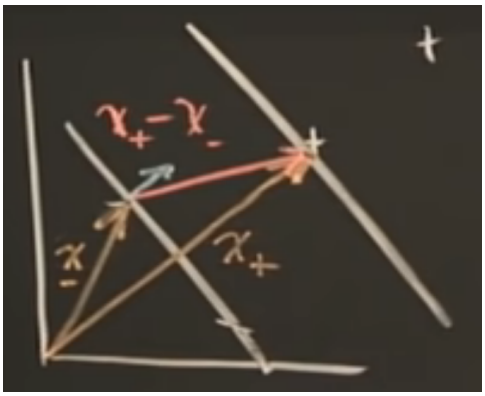
Our goal is to arrange for the line such that the street separating the two classes (+, -) is as wide as possible. For that, we need to express the distance between the two gutters. Take two samples one in each gutter:



Then take the difference between these two vectors:



If we had a unit normal that is normal to the median line of the street, then we can take the dot product of that unit normal and this difference vector, and that would be the width of the street.



$$\text{width} = (\vec{x}_+ - \vec{x}_-) \cdot \frac{\vec{\omega}}{\|\vec{\omega}\|} \tag{10.4}$$

Equation (10.3) with $\vec{x}_+$ reads $\vec{x}_+ \cdot \vec{\omega} = 1 - b$. Likewise for $\vec{x}_-$, we have $\vec{x}_- \cdot \vec{\omega} = -1 + b$. Therefore, Eq. (10.4) becomes:

$$\text{width} = (\vec{x}_+ - \vec{x}_-) \cdot \frac{\vec{\omega}}{\|\vec{\omega}\|} = (1 - b - (-1 + b)) \frac{1}{\|\vec{\omega}\|} = \frac{2}{\|\vec{\omega}\|}$$

Our goal is to maximize this width. Maximizing $2/\|\vec{\omega}\|$ is the same as maximizing $1/\|\vec{\omega}\|$, which is the same as minimizing $\|\vec{\omega}\|$. For convenience, we will instead minimize $\frac{1}{2}\|\vec{\omega}\|^2$.

Minimization of $\frac{1}{2}\|\vec{\omega}\|^2$ subject to constraints $y_i(\vec{\omega} \cdot \vec{x}_i + b) - 1 = 0$ (Eq. 10.3) can be done using the method of Lagrange multipliers:

$$L = \frac{1}{2}\|\vec{\omega}\|^2 - \sum \alpha_i \left[y_i(\vec{\omega} \cdot \vec{x}_i + b) - 1 \right].$$

Extremum is obtained by differentiating with respect to the unknowns:

$$\frac{\partial L}{\partial \vec{\omega}} = \vec{\omega} - \sum \alpha_i y_i \vec{x}_i = 0,$$

which implies that

$$\boxed{\vec{\omega} = \sum_i \alpha_i y_i \vec{x}_i.} \tag{10.5}$$

This important result tells us that $\vec{\omega}$ is a linear combination of the samples. (For some of the vectors α_i may be zero.). Next,

$$\frac{\partial L}{\partial b} = -\sum_i \alpha_i y_i = 0$$

which implies that

$$\boxed{\sum_i \alpha_i y_i = 0.}$$

Because these expression are so simple, let's see what happens if we substitute this expression for $\vec{\omega}$ into Eq. (10.5):

$$L = \frac{1}{2}\left(\sum_i \alpha_i y_i \vec{x}_i\right)\cdot\left(\sum_j \alpha_j y_j \vec{x}_j\right) - \sum_i \alpha_i y_i \vec{x}_i \cdot \left(\sum_j \alpha_j y_j \vec{x}_j\right) - \sum_i \alpha_i y_i b + \sum_i \alpha_i.$$

The term $\sum_i \alpha_i y_i b$ is zero because $\sum_i \alpha_i y_i = 0$. The first two terms are of the same form and therefore can be combined:

$$\boxed{L^* = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \vec{x}_i \cdot \vec{x}_j.} \tag{10.6}$$

This tells us that the optimal L only depends on the dot products $\vec{x}_i \cdot \vec{x}_j$ of the data vectors. Let's go back to our decision rule and substitute our newly obtained results:

$$\boxed{\sum \alpha_i y_i \vec{x}_i \cdot \vec{u} + b \geq 0 \qquad \text{then } +}$$

So far this method assumes that a line can separate + and - data. In general, however, there can be many situations where such a separation is not possible. The solution is to transform our vector space into another one: $\phi(\vec{x})$). We said earlier that maximization depends on the dot products $\vec{x}_i \cdot \vec{x}_j$. All we need is dot products in the new space $\phi(\vec{x}) \cdot \phi(\vec{y})$. This only requires a function

$$K(\vec{x}_i, \vec{x}_j) = \phi(\vec{x}_i) \cdot \phi(\vec{x}_j)$$

We don't actually need the function $\phi(\vec{x})$. All we need is the kernel function K, which provides us with the dot product of these two vectors in another space. We don't need to know the transformation. Some popular kernels are:

$$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \cdot \vec{x}_j + 1)^n \qquad \text{(polynomial kernel)}$$

$$K(\vec{x}_i, \vec{x}_j) = \exp\left(-\frac{\|\vec{x}_i - \vec{x}_j\|^2}{2\sigma}\right) \qquad \text{(radial basis function kernel)}$$

Neural tangent kernel
Neural network Gaussian process
String kernel
Kernel smoother
Graph kernel
Fisher kernel

10.3. Additional Concepts in Statistical Learning

10.3.1. KL Divergence. The Kullback-Leibler (KL) divergence can be used instead of the least squares formula for data fitting. It is used to

measure distances between probability measures. Our experimental data is drawn from a distribution.

Two distributions describing random variables with two normal distributions:

$$p(x) \sim \mathcal{N}(\mu_1, \sigma_1^2)$$

and

$$q(x) \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

The KL divergence is defined as

$$D_{KL}[p:q] = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$$

For Gaussian distributions

$$\begin{aligned}
D_{KL}[p:q] = & \int p(x) \log \frac{\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)} \mathrm{d}x \\
= & \int p(x) \log \left(\frac{\sigma_2^2}{\sigma_1^2}\right)^{1/2} \mathrm{d}x + \int p(x) \left[-\frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}\right] \mathrm{d}x \\
= & \frac{1}{2} \log \frac{\sigma_2^2}{\sigma_1^2} + \frac{1}{2\sigma_1^2} \left[-\int (x-\mu_1)^2 p(x) \mathrm{d}x + \frac{1}{2\sigma_2^2} \int (x-\mu_2)^2 p(x) \mathrm{d}x\right] \\
= & \frac{1}{2} \log \frac{\sigma_2^2}{\sigma_1^2} - \frac{\sigma_1^2}{2\sigma_2^2} + \frac{1}{2\sigma_2^2} \int (x-\mu_1+\mu_1-\mu_1-\mu_2)^2 p(x) \mathrm{d}x \\
= & \frac{1}{2} \log \frac{\sigma_2^2}{\sigma_1^2} - \frac{1}{2} + \frac{1}{2\sigma_2^2} \left[\int (x-\mu_1)^2 p(x) dx + (\mu_1-\mu_2)^2 \int p(x) dx \right. \\
& \left. + 2(\mu_1-\mu_2) \int (x-\mu_1) p(x) dx\right]
\end{aligned}$$

which simplifies to

$$= \frac{1}{2} \log \frac{\sigma_2^2}{\sigma_1^2} - \frac{1}{2} + \frac{1}{2\sigma_2^2} \left[\sigma_1^2 + (\mu_1-\mu_2)^2\right].$$

Therefore,

$$D_{KL}[p:q] = \frac{1}{2} \left[\log \frac{\sigma_2^2}{\sigma_1^2} - 1 + \frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{\sigma_2^2}\right]$$

10.3.2. Mutual Information. Given two random variables X and Y, mutual information is defined as the difference between the total entropy (of X and Y) and the joint entropy of X and Y:

$$I(X;Y) = H(X) + H(Y) - H(X,Y).$$

Another way to assess mutual information is to ask how "far away" the joint distribution of two random variables is from a product distribution. Independent random variables should share no mutual information.

Let P and Q be two probability distributions on a finite set $\mathcal{X}$. Recall that the KL-divergence of P and Q is (discrete case)

$$D_{KL}[P:Q] \equiv \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} = \mathbb{E}_p \log\frac{P}{Q}.$$

(We put $0\log(0/0)=0$ and $p\log(p/0)=\infty$ for $p>0$.). We note that this is not a metric (not symmetric, does not obey triangle inequality). Also, $D_{KL}[P:Q]\geq 0$

How do we measure mutual information using KL-divergence? Let X, Y be two $\mathcal{X}$-valued random variables with distributions $P(x)$, $P(y)$, respectively, and joint distribution $P(x,y)$:

$$I(X;Y) = D_{KL}[P(x,y):P(x)P(y)] = \sum_{x,y\in\mathcal{X}} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}$$
$$= \mathbb{E}_{P(x,y)}\log\frac{P(X,Y)}{P(X)P(Y)}.$$

Case 1: Suppose that X, Y are independent: $P(x,y)=P(x)\cdot P(y)$. Then,

$$D_{KL}[P(x)\cdot P(y):P(x)\cdot P(y)] = 0$$

Next, if $X=Y$ then

$$I(X;X) = \sum_{x\in\mathcal{X}} P(x)\cdot\log\frac{P(x)}{P(x)^2} = \sum_x P(x)\log(1/P(x)) = H(X).$$

Case 2: Let Y be a fair die. Outcomes are: $\{1,2,3,4,5,6\}$. Let $X=1$ if even, 2 if odd.

$$I(X;Y) = \sum_{x=1,2\,y=1,\ldots,6} P(x,y)\log\frac{P(x,y)}{P(x)P(y)} = 6\left(\frac{1}{6}\log\frac{1/6}{1/12}\right) = \log 2 = H(X).$$

There is a relationship between entropy and mutual information:

$$\begin{aligned} I(X;Y) =& \sum_{x,y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)} = \sum_{x,y} P(x,y)\log\frac{P(x|y)}{P(x)} \\ =& -\sum_{x,y} P(x,y)\log P(x) + \sum_{x,y} P(x,y)\log P(x|y) \\ =& -\sum_{x,y} P(x,y)\log P(x) - \left(-\sum_{x,y} P(x,y)\log P(x|y)\right) \\ =& H(X) - H(X|Y). \end{aligned}$$

This implies symmetry of information:

$I(X;Y) = H(X) + H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$

where we used the fact that:

$$H(X,Y) = H(X) + H(Y|X).$$

In particular, $I(X;X) = H(X)$. Therefore, entropy equals self-information.

Let's check that $I \geq 0$. We will need Jensen's inequality. $F : \mathbb{R} \to \mathbb{R}$ is a convex function for every $0 \leq \lambda \leq 1$ if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$

This can be rewritten as $f(\lambda_1 x + \lambda_2 y) \leq \lambda_1 f(x) + \lambda_2 f(y)$ $(\lambda_1 + \lambda_2 = 1)$ and generalized (by induction) as $f(\sum_i \lambda_i x_i) \leq \sum_i \lambda_i f(x_i)$ with $\sum_i \lambda_i = 1$. A concave function is one for this $-f$ is convex.

Jensen's inequality states that if f is a convex function and X is a random variable then

$$\mathbb{E} f(X) \geq f(\mathbb{E} X).$$

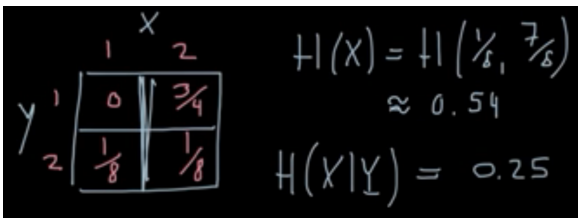
i.e., $\sum_i p_i f(x_i) \geq f(\sum_i p_i x_i)$ (discrete case) and $\int p(x) f(x) \mathrm{d}x \geq f(\int p(x) x \mathrm{d}x)$ (continuous case).

We can use the fact that log is (strictly) convex to infer that $D(P||Q) \geq 0$, with equality iff $P(a) = Q(a)$ for all $a \in A$. This is proved from Jensen's inequality. Let $S_A = \{a \in A : P(a) > 0\}$ be the support of P. Then

$$\begin{aligned} -D_{KL}[P:Q] &= -\sum_{a \in S_A} P(a) \log \frac{P(a)}{Q(a)} = \sum_{a \in S_A} P(a) \log \frac{Q(a)}{P(a)} \\ &\leq \log \sum_{a \in S_A} P(a) \frac{Q(a)}{P(a)} = \log \sum_{a \in S_A} Q(a) \leq \log \sum_{a \in A} Q(a) = 0 \end{aligned}$$

since $\sum_{a \in A} Q(a) = 1$ and log is concave (therefore, $\mathbb{E} f(X) \leq f(\mathbb{E} X)$. As a corollary, $I(X;Y) \geq 0$ with equality iff X and Y are independent.

We also note that conditioning reduces entropy. Since $I(X;Y) = H(X) - H(X|Y) \geq 0$, we have $H(X|Y) \leq H(X)$. On average, knowing another random variable Y reduces uncertainty in X.

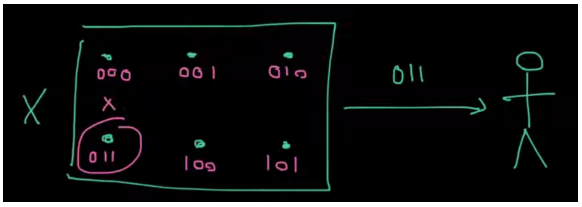


10.3.3. Information Measures. Suppose X is finite, non-empty set. We choose an element in X and want to communicate the information "which

$x \in X$ we chose" to someone else. A binary channel is one where we can transmit only strings of 0 and 1. How many bits are needed to transmit information about x, if nothing else is known about X but its cardinality? We can assign to every element in X a fixed-length binary code:

$$c : X \to \{0,1\}^n \qquad \text{(one-one)}$$

and transmit $c(x)$.



We need $n = \lceil \log_2 |X| \rceil$ bits. In this section we will write $\log_2 = \log$. The information content here are the bits needed to transmit, which equals to the log of the number of possibilities ($\log |X|$). From the probabilistic standpoint, let us assume we have probability measure on X (still finite). We draw $xx \in X$ at random according to P.

10.3.4. Algorithmic Entropy: Kolmogorov Complexity.

10.4. Natural Gradient

10.4.1. Fisher Information Matrix. Suppose we have a model parametrized by a parameter vector $\boldsymbol{\theta}$ that models a distribution $p(x|\boldsymbol{\theta})$. We normally learn $\boldsymbol{\theta}$ by maximizing the likelihood $p(x|\boldsymbol{\theta})$ with respect to the parameters $\boldsymbol{\theta}$. To assess the goodness of our estimate of $\boldsymbol{\theta}$ we define a score function:

$$s(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta}),$$

that is, score function is the gradient of log likelihood function. It is easy to check that the expected value of score with respect to our model is zero:

$$\begin{aligned}\mathbb{E}_{p(x|\boldsymbol{\theta})}[s(\boldsymbol{\theta})] =& \mathbb{E}_{p(x|\boldsymbol{\theta})}[\nabla \log p(x|\boldsymbol{\theta})] = \int \nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta}) p(x|\boldsymbol{\theta}) \mathrm{d}x \\ &= \int \frac{\nabla_{\boldsymbol{\theta}} p(x|\boldsymbol{\theta})}{p(x|\boldsymbol{\theta})} p(x|\boldsymbol{\theta}) \mathrm{d}x = \int \nabla_{\boldsymbol{\theta}} p(x|\boldsymbol{\theta}) \mathrm{d}x = \nabla \int p(x|\boldsymbol{\theta}) \mathrm{d}x = \nabla 1 = 0.\end{aligned}$$

But how certain are we to our estimate? We can define an uncertainty measure around the expected estimate. That is, we look at the covariance of score of our model. Taking the result from above:

$$\mathbb{E}_{p(x|\boldsymbol{\theta})}[(s(\boldsymbol{\theta}) - 0)(s(\boldsymbol{\theta}) - 0)^T].$$

We can then see it as an information. The covariance of score function above is the definition of Fisher Information. As we assume $\boldsymbol{\theta}$ is a vector, the Fisher

Information is in a matrix form, called Fisher Information Matrix:

$$F = \mathbb{E}_{p(x|\boldsymbol{\theta})}[\nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta}) \nabla \log p(x|\boldsymbol{\theta})^T].$$

However, usually our likelihood function is complicated and computing the expectation is intractable. We can approximate the expectation F using empirical distribution $\hat{q}(x)$, which is given by our training data $X = \{x_1, x_2, \ldots, x_N\}$. In this form, F is called empirical Fisher:

$$F = \frac{1}{N} \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \log p(x_i|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log p(x_i|\boldsymbol{\theta})^T.$$

10.4.2. Fisher and Hessian. One property of F that is not obvious is that it has the interpretation of being the negative expected Hessian of our model's log likelihood. We can check that the negative expected Hessian of log likelihood is equal to the Fisher Information Matrix F as follows:

$$\begin{aligned} H_{\log p(x|\boldsymbol{\theta})} =& J\left(\frac{\nabla p(x|\boldsymbol{\theta})}{p(x|\boldsymbol{\theta})}\right) = \frac{H_{p(x|\boldsymbol{\theta})} p(x|\boldsymbol{\theta}) - \nabla p(x|\boldsymbol{\theta}) \nabla p(x|\boldsymbol{\theta})^T}{p(x|\boldsymbol{\theta}) p(x|\boldsymbol{\theta})} \\ =& \frac{H_{p(x|\boldsymbol{\theta})} p(x|\boldsymbol{\theta})}{p(x|\boldsymbol{\theta}) p(x|\boldsymbol{\theta})} - \frac{\nabla_{\boldsymbol{\theta}} p(x|\boldsymbol{\theta}) \nabla p(x|\theta)^T}{p(x|\boldsymbol{\theta}) p(x|\boldsymbol{\theta})} \\ =& \frac{H_{p(x|\boldsymbol{\theta})}}{p(x|\boldsymbol{\theta})} - \left(\frac{\nabla_{\boldsymbol{\theta}} p(x|\boldsymbol{\theta})}{p(x|\boldsymbol{\theta})}\right) \left(\frac{\nabla_{\boldsymbol{\theta}} p(x|\boldsymbol{\theta})^T}{p(x|\boldsymbol{\theta})}\right)^T, \end{aligned}$$

where the second line is a result of applying quotient rule of derivative. Taking expectation wrt our model, we have:

$$\begin{aligned} \mathbb{E}_{p(x|\boldsymbol{\theta})}[H_{\log p(x|\boldsymbol{\theta})}] =& \mathbb{E}_{p(x|\boldsymbol{\theta})} \left[\frac{H_{p(x|\boldsymbol{\theta})}}{p(x|\boldsymbol{\theta})} - \left(\frac{\nabla_{\boldsymbol{\theta}} p(x|\boldsymbol{\theta})}{p(x|\boldsymbol{\theta})}\right) \left(\frac{\nabla_{\boldsymbol{\theta}} p(x|\boldsymbol{\theta})^T}{p(x|\boldsymbol{\theta})}\right)^T\right] \\ =& \mathbb{E}_{p(x|\boldsymbol{\theta})} \left[\frac{H_{p(x|\boldsymbol{\theta})}}{p(x|\boldsymbol{\theta})}\right] - \mathbb{E}_{p(x|\boldsymbol{\theta})} \left[\left(\frac{\nabla_{\boldsymbol{\theta}} p(x|\boldsymbol{\theta})}{p(x|\boldsymbol{\theta})}\right) \left(\frac{\nabla_{\boldsymbol{\theta}} p(x|\boldsymbol{\theta})^T}{p(x|\boldsymbol{\theta})}\right)^T\right] \\ =& \int \frac{H_{p(x|\boldsymbol{\theta})}}{p(x|\boldsymbol{\theta})} p(x|\boldsymbol{\theta}) \mathrm{d}x - \mathbb{E}_{p(x|\boldsymbol{\theta})}[\nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta})^T] \\ =& H_{\int p(x|\boldsymbol{\theta}) \mathrm{d}x} - F = H_1 - F = -F. \end{aligned}$$

Thus, we have $F = -\mathbb{E}_{p(x|\boldsymbol{\theta})}[H_{\log p(x|\boldsymbol{\theta})}]$. Knowing this result, we can see the role of F as a measure of curvature of the log likelihood function.

Fisher Information Matrix is defined as the covariance of score function. It is a curvature matrix and has interpretation as the negative expected Hessian of log likelihood function. Thus the immediate application of F is as drop-in replacement of H in second order optimization methods. One of the most exciting results of F is that it has connection to KL-divergence. This gives rise to natural gradient method, which we shall discuss now.

10.4.3. Natural Gradient Descent. Previously, we looked at the Fisher Information Matrix. We saw that it is equal to the negative expected Hessian of log likelihood. Thus, the immediate application of Fisher Information Matrix is as drop-in replacement of Hessian in second order optimization algorithm. In this article, we will look deeper at the intuition on what exactly is the Fisher Information Matrix represents and what is the interpretation of it.

10.4.4. Distribution Space. As per previous article, we have a probabilistic model represented by its likelihood $p(x|\boldsymbol{\theta})$. We want to maximize this likelihood function to find the most likely parameter $\boldsymbol{\theta}$. Equivalent formulation would be to minimize the loss function $\mathcal{L}(\boldsymbol{\theta})$, which is the negative log likelihood.

Usual way to solve this optimization is to use gradient descent. In this case, we are taking step which direction is given by $-\mathcal{L}(\boldsymbol{\theta})$. This is the steepest descent direction around the local neighborhood of the current value of $\boldsymbol{\theta}$ in the parameter space. Formally, we have

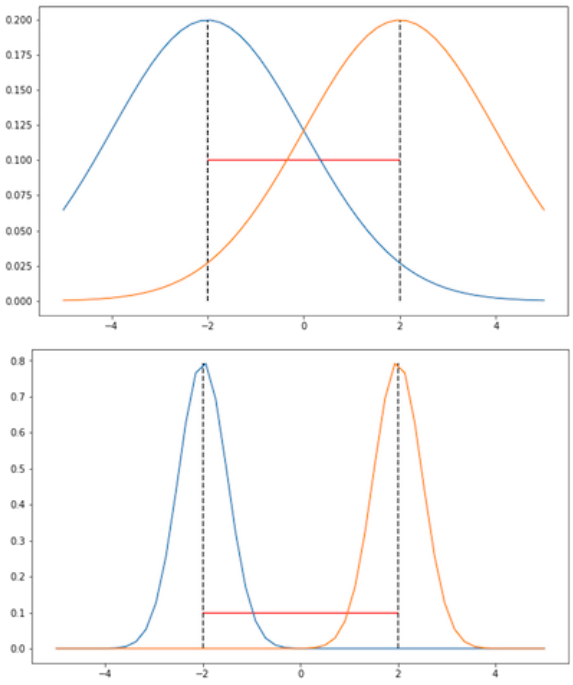
$$\frac{-\mathcal{L}(\boldsymbol{\theta})}{\|\mathcal{L}(\boldsymbol{\theta})\|} = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \underset{d \text{ s.t. } \|d\| \leq \epsilon}{\arg\min} \mathcal{L}(\boldsymbol{\theta} + d).$$

The above expression is saying that the steepest descent direction in parameter space is to pick a vector d, such that the new parameter $\boldsymbol{\theta} + d$ is within the ϵ-neighborhood of the current parameter $\boldsymbol{\theta}$, and we pick d that minimize the loss. Notice the way we express this neighborhood is by the means of Euclidean norm. Thus, the optimization in gradient descent is dependent to the Euclidean geometry of the parameter space.

Meanwhile, if our objective is to minimize the loss function (maximizing the likelihood), then it is natural that we taking step in the space of all possible likelihood, realizable by parameter $\boldsymbol{\theta}$. As the likelihood function itself is a probability distribution, we call this space distribution space. Thus it makes sense to take the steepest descent direction in this distribution space instead of parameter space.

Which metric/distance then do we need to use in this space? A popular choice would be KL-divergence. KL-divergence measure the "closeness" of two distributions. Although as KL-divergence is non-symmetric and thus not a true metric, we can use it anyway. This is because as d goes to zero, KL-divergence is asymptotically symmetric. So, within a local neighborhood, KL-divergence is approximately symmetric.

We can see the problem when using only Euclidean metric in parameter space from the illustrations below. Consider a Gaussian parameterized by only its mean and keep the variance fixed to 2 and 0.5 for the first and second image respectively:



In both images, the distance of those Gaussians are the same, i.e. 4, according to Euclidean metric (red line). However, clearly in distribution space, i.e. when we are taking into account the shape of the Gaussians, the distance is different in the first and second image. In the first image, the KL-divergence should be lower as there is more overlap between those Gaussians. Therefore, if we only work in parameter space, we cannot take into account this information about the distribution realized by the parameter.

10.4.5. Fisher Information and KL-divergence. One question still needs to be answered is what exactly is the connection between Fisher Information Matrix and KL-divergence? It turns out, Fisher Information Matrix defines the local curvature in distribution space for which KL-divergence is the metric.

It is easy to show that Fisher Information Matrix F is the Hessian of KL-divergence between two distributions $p(x|\boldsymbol{\theta})$ and $p(x|\boldsymbol{\theta}')$ with respect to $\boldsymbol{\theta}'$, evaluated at $\boldsymbol{\theta}' = \boldsymbol{\theta}$. The KL-divergence can be decomposed into entropy and cross-entropy terms, i.e.

$$D_{KL}[p(x|\boldsymbol{\theta}) : p(x|\boldsymbol{\theta}')] = \mathbb{E}_{p(x|\boldsymbol{\theta})}[\log p(x|\boldsymbol{\theta})] - \mathbb{E}_{p(x|\boldsymbol{\theta})}[\log p(x|\boldsymbol{\theta}')].$$

The first derivative wrt $\boldsymbol{\theta}'$ is:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}'} D_{KL}[p(x|\boldsymbol{\theta}) : p(x|\boldsymbol{\theta}')] =& \nabla_{\boldsymbol{\theta}'} \mathbb{E}_{p(x|\boldsymbol{\theta})}[\log p(x|\boldsymbol{\theta})] - \nabla_{\boldsymbol{\theta}'} \mathbb{E}_{p(x|\boldsymbol{\theta})}[\log p(x|\boldsymbol{\theta}')] \\ =& - \nabla_{\boldsymbol{\theta}'} \mathbb{E}_{p(x|\boldsymbol{\theta})}[\log p(x|\boldsymbol{\theta}')] = - \int p(x|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}'} \log p(x|\boldsymbol{\theta}') \mathrm{d}x.\end{aligned}$$

The second derivative is:

$$\nabla^2_{\boldsymbol{\theta}'} D_{KL}[p(x|\boldsymbol{\theta}) : p(x|\boldsymbol{\theta}')] = - \int p(x|\boldsymbol{\theta}) \nabla^2_{\boldsymbol{\theta}'} \log p(x|\boldsymbol{\theta}') \mathrm{d}x.$$

Thus, the Hessian wrt $\boldsymbol{\theta}'$ evaluated at $\boldsymbol{\theta}' = \boldsymbol{\theta}$ is:

$$\begin{aligned}H_{D_{KL}[p(x|\boldsymbol{\theta}):p(x|\boldsymbol{\theta}')]} =& - \int p(x|\boldsymbol{\theta}) \nabla^2_{\boldsymbol{\theta}'} \log p(x|\boldsymbol{\theta}')|_{\boldsymbol{\theta}'=\boldsymbol{\theta}} \mathrm{d}x = - \int p(x|\boldsymbol{\theta}) H_{\log p(x|\boldsymbol{\theta})} \mathrm{d}x \\ =& - \mathbb{E}_{p(x|\boldsymbol{\theta})}[H_{\log p(x|\boldsymbol{\theta})}] = F.\end{aligned}$$

The last line follows from the previous result about Fisher Information Matrix, in which we showed that the negative expected Hessian of log likelihood is the Fisher Information Matrix.

10.4.6. Steepest Descent in Distribution Space. Now we are ready to use the Fisher Information Matrix to enhance the gradient descent. But first, we need to derive the Taylor series expansion for KL-divergence around $\boldsymbol{\theta}$.

Letting $d \to 0$, the second order Taylor expansion of KL-divergence can be shown to be $D_{KL}[p(x|\boldsymbol{\theta}) : p(x|\boldsymbol{\theta} + d)] \approx \frac{1}{2} d^T F d$. Using $p_{\boldsymbol{\theta}}$ as a shortcut for $p(x|\boldsymbol{\theta})$ and expanding

$$\begin{aligned}D_{KL}[p(x|\boldsymbol{\theta}) : p(x|\boldsymbol{\theta} + d)] \approx& D_{KL}[p_{\boldsymbol{\theta}} : p_{\boldsymbol{\theta}}] + (\nabla_{\theta'} D_{KL}[p_{\boldsymbol{\theta}} : p_{\boldsymbol{\theta}'}|_{\boldsymbol{\theta}'=\boldsymbol{\theta}})^T d + \frac{1}{2} d^T F d \\ =& D_{KL}[p_{\boldsymbol{\theta}} : p_{\boldsymbol{\theta}}] - \mathbb{E}_{p(x|\boldsymbol{\theta})}[\nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta})]^T d + \frac{1}{2} d^T F d.\end{aligned}$$

Notice that the first term is zero as it is the same distribution. Furthermore, from the previous article, we saw that the expected value of the gradient of log likelihood, which is exactly the gradient of KL-divergence as shown in the previous proof, is also zero. Thus the only thing left is:

$$D_{KL}[p(x|\boldsymbol{\theta}) : p(x|\boldsymbol{\theta} + d)] \approx \frac{1}{2} d^T F d.$$

Now, we would like to know what is update vector d that minimizes the loss function $\mathcal{L}(\boldsymbol{\theta})$ in distribution space, so that we know in which direction decreases the KL-divergence the most. This is analogous to the method of steepest descent, but in distribution space with KL-divergence as metric, instead of the usual parameter space with Euclidean metric. For that, we do this minimization:

$$d^* = \underset{d \text{ s.t. } D_{KL}[p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}+d}] = c}{\arg\min} \mathcal{L}(\boldsymbol{\theta} + d),$$

where c is some constant. The purpose of fixing the KL-divergence to some constant is to make sure that we move along the space with constant speed, regardless the curvature. Further benefit is that this makes the algorithm more robust to the re-parametrization of the model, i.e. the algorithm does not care how the model is parametrized, it only cares about the distribution induced by the parameter.

If we write the above minimization in Lagrangian form, with constraint KL-divergence approximated by its second order Taylor series expansion and approximate $\mathcal{L}(\boldsymbol{\theta} + d)$ with its first order Taylor series expansion, we get:

$$\begin{aligned} d^* &= \arg\min_d \mathcal{L}(\boldsymbol{\theta} + d) + \lambda(D_{KL}[p_{\boldsymbol{\theta}} : p_{\boldsymbol{\theta}+d}] - c) \\ &\approx \arg\min_d \mathcal{L}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})^T d + \frac{1}{2}\lambda d^T F d - \lambda c. \end{aligned}$$

To solve this minimization, we set its derivative wrt d to zero:

$$0 = \frac{\partial}{\partial d}\mathcal{L}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})^T d + \frac{1}{2}\lambda d^T F d - \lambda c = \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) + \lambda F d$$

so that $\lambda F d = -\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$ and $d = -\frac{1}{\lambda}F^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$. Up to constant factor of $\frac{1}{\lambda}$, we get the optimal descent direction, i.e. the opposite direction of gradient while taking into account the local curvature in distribution space defined by F_{-1}. We can absorb this constant factor into the learning rate.

Definition: Natural gradient is defined as

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = F^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}).$$

As corollary, we have the following algorithm:

Algorithm: Natural Gradient Descent

(1) Repeat:
   (a) Do forward pass on our model and compute loss $\mathcal{L}(\boldsymbol{\theta})$
   (b) Compute the gradient $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$
   (c) Compute the Fisher Information Matrix F, or its empirical version (wrt training data).
   (d) Compute the natural gradient $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = F^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$
   (e) Update the parameter: $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha\nabla\mathcal{L}(\boldsymbol{\theta})$ where α is the learning rate.

(2) Until convergence.

In the above very simple model with low amount of data, we saw that we can implement natural gradient descent easily. But how easy is it to do this in the real world? As we know, the number of parameters in deep learning models is very large, within millions of parameters. The Fisher Information Matrix for these kind of models is then infeasible to compute, store, or

invert. This is the same problem as why second order optimization methods are not popular in deep learning.

One way to get around this problem is to approximate the Fisher/Hessian instead. Method like ADAM computes the running average of first and second moment of the gradient. First moment can be seen as momentum which is not our interest in this article. The second moment is approximating the Fisher Information Matrix, but constraining it to be diagonal matrix. Thus in ADAM, we only need $O(n)$ space to store (the approximation of) F instead of $O(n^2)$ and the inversion can be done in $O(n)$ instead of $O(n^3)$. In practice ADAM works really well and is currently the de facto standard for optimizing deep neural networks.

Chapter 11

Writing Custom Code for Data Fitting and Optimization in MATLAB

11.1. Basics of the MATLAB command environment

MATLAB is a scientific computing software platform that boasts millions of users worldwide. Its broad user base includes chemists, physicists, computer scientists, engineers (including finance), life scientists and mathematicians. It allows you to run tasks by executing scripts (lists of commands) or to do computations one at a time by typing commands in its console. In the MATLAB environment you can store variables (scalars, vectors, matrices, strings, structures, etc.) whereas MATLAB itself features thousands of different commands that you can use. Variables can be created by setting their value. For example:

```
a=1
```

```
a =

     1

>>
```

This command creates a new variable `a` (if it does not exist already), and sets its value equal to 1. To check that you have created this variable, type:

```
whos
```

Arrays of numbers can be stored as matrices or vectors. The following command creates a row vector:

```
a=[ 4 7 10 12 18 8 7 6 4.5 2.7 13.8 ]
```

```
a =

  Columns 1 through 9

    4.0000    7.0000   10.0000   12.0000   18.0000    8.0000
       7.0000    6.0000    4.5000

  Columns 10 through 11

    2.7000   13.8000
```

This could be, for example, a set of experimental values. MATLAB has many internal commands that are available to us for purposes of doing computations. For example, the standard deviation of the entries stored in `a` can be computed using the std command:

```
std(a)
```

```
ans =

    4.6228
```

Internal MATLAB commands are scripts (lists of commands) that can be viewed. For example, entering:

```
type std
```

we get:

```
function y = std(varargin)
%STD Standard deviation.
%   For vectors, Y = STD(X) returns the standard deviation.  For ...
    matrices,
```

```
%   Y is a row vector containing the standard deviation of each ...
    column.  For
%   N-D arrays, STD operates along the first non-singleton ...
    dimension of X.
%
%   STD normalizes Y by (N-1), where N is the sample size.  This is the
%   sqrt of an unbiased estimator of the variance of the population ...
    from
%   which X is drawn, as long as X consists of independent, identically
%   distributed samples.
%
%   Y = STD(X,1) normalizes by N and produces the square root of ...
    the second
%   moment of the sample about its mean.  STD(X,0) is the same as ...
    STD(X).
%
%   Y = STD(X,FLAG,DIM) takes the standard deviation along the ...
    dimension
%   DIM of X.  Pass in FLAG==0 to use the default normalization by ...
    N-1, or
%   1 to use N.
%
%   std(..., MISSING) specifies how NaN (Not-A-Number) values are ...
    treated.
%   The default is 'includenan':
%
%   'includenan' - the standard deviation of a vector containing NaN
%                  values is also NaN.
%   'omitnan'    - elements of X or W containing NaN values are ...
    ignored.
%                  If all elements are NaN, the result is NaN.
%
%   Example: If X = [4 -2 1; 9  5 7]
%     then std(X,0,1) is [3.5355 4.9497 4.2426] and std(X,0,2) is ...
    [3.0; 2.0]
%
%   Class support for input X:
%      float: double, single
%
%   See also COV, MEAN, VAR, MEDIAN, CORRCOEF.

%   Copyright 1984-2014 The MathWorks, Inc.

% Call var(x,flag,dim) with as many of those args as are present.
y = sqrt(var(varargin{:}));
```

As you can see, there is only 1 command that does anything: `y = ... sqrt(var(varargin:));`. It calls another command, `var`, to calculate the sample variance. Then it takes the square root of this variance and returns it as the standard deviation. You can view the code for `var` by typing

```
type var
```

The mean is calculate as follows:

```
mean(a)
```

```
ans =

    8.4545
```

Likewise, there are commands for calculating the median:

```
median(a)
```

```
ans =

     7
```

There is also a command to calculate percentiles. Type `help prctile` for more information. For example:

```
prctile(a,50)
```

```
ans =

     7
```

which is the median. Also,

```
prctile(a,10)
```

```
ans =

    3.4800
```

Vector addition:

```
a=[1 2 3]
```

```
a =

     1     2     3
```

```
b=[4 5 6]
```

```
b =

     4     5     6
```

```
a+b
```

```
ans =

     5     7     9
```

Multiplication of vector (`a`) by a scalar (`2`):

```
a
```

```
a =

     1     2     3
```

```
2*a
```

```
ans =

     2     4     6
```

Matrices can be created by separating the different rows using semicolons:

```
>> a=[1 2 ; 3 4 ]
```

```
a =

     1     2
     3     4
```

```
b=[5 6 ; 7 8]
```

```
b =

     5     6
     7     8
```

```
whos
```

```
  Name      Size            Bytes  Class     Attributes

  a         2x2                32  double
  b         2x2                32  double
```

Matrix addition (element by element addition):

```
a+b
```

```
ans =

     6     8
    10    12
```

Matrix multiplication:

```
a*b
```

```
ans =

    19    22
    43    50
```

For pointwise (element by element) multiplication, we use `.*` instead of `*`:

```
a.*b
```

```
ans =

     5    12
    21    32
```

We can even take the elements of `a` and elevate them to the powers of the elements of `b`:

```
a.^b
```

```
ans =

           1          64
        2187       65536
```

Let's now multiply a 2×2 matrix by a 2×1 (column) vector:

```
c=[6 ; 8]
```

```
c =

     6
     8
```

The multiplication of `a` and `c` is:

```
a*c
```

```
ans =

    22
    50
```

We can take the natural log of the elements of `a`:

```
log(a)
```

```
ans =

         0    0.6931
    1.0986    1.3863
```

log means natural log. To get log base 10, there is a command `log10`. Or use the property $\log_a b = \log(b)/\log(a)$.

Suppose you created a row vector,

```
c=[6 8]
```

```
c =

     6     8
```

when it reality you needed a column vector. You can use the `transpose` command to do that:

```
transpose(c)
```

```
ans =

     6
```

```
     8
```

A shorthand is to add a prime after `c`:

```
c'
```

```
ans =

     6
     8
```

If you try to multiply `a` and `c` you get an error message

```
a*c
```

```
Error using  *
Inner matrix dimensions must agree.
```

because `c` must be a column vector, not a row vector. Instead, what works is:

```
a*transpose(c)
```

```
ans =

    22
    50
```

or

```
a*c'
```

```
ans =

    22
    50
```

Ranges can be created by using the colon:

```
1:10
```

```
ans =

     1     2     3     4     5     6     7     8     9    10
```

If a different step size than 1 is desired, sandwich the step size between two colons:

```
1:0.5:10
```

```
ans =

  Columns 1 through 9

    1.0000    1.5000    2.0000    2.5000    3.0000    3.5000
        4.0000    4.5000    5.0000

  Columns 10 through 18

    5.5000    6.0000    6.5000    7.0000    7.5000    8.0000
        8.5000    9.0000    9.5000

  Column 19

   10.0000
```

If you don't know the step size but know how many points are needed, use `linspace`:

```
linspace(1,10,5)
```

```
ans =

    1.0000    3.2500    5.5000    7.7500   10.0000
```

11.1.1. Taylor series approximation to the matrix exponential. The exponential of a matrix M is defined by its Taylor series expansion:

$$\exp(M) = \sum_{j=0}^{\infty} \frac{M^j}{j!}.$$

We can use MATLAB to see how this Taylor series converges to $\exp(M)$. Let's compute partial sums (truncated Taylor series) and compare the results as the number of terms is increased. We will compare the results to the `EXPM` command in MATLAB, which uses the scaling and squaring algorithm (with Padé approximant) to compute the matrix exponential:

```
>> help expm
```

```
expm  Matrix exponential.
   expm(A) is the matrix exponential of A and is computed using
   a scaling and squaring algorithm with a Pade approximation.

   Although it is not computed this way, if A has a full set
   of eigenvectors V with corresponding eigenvalues D then
   [V,D] = EIG(A) and expm(A) = V*diag(exp(diag(D)))/V.

   EXP(A) computes the exponential of A element-by-element.

   See also logm, sqrtm, funm.
```

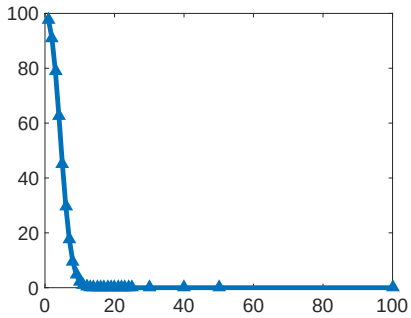
The following code below will compute the partial sum up to j=NT (a value which you can set manually), for a 2×2 matrix m:

```
m=[ 1 2 ; 3 4 ]; % matrix to be exponentiated
NT=100;
w=eye(2);
for j=1:NT,
    w = w + (m^j)/factorial(j);
end;
disp('Result:');
w
disp('True value:');
expm(m)
disp('Distance:');
d=w-expm(m);
dis=sum(abs(d(:)))/4
```

In order to assess the convergence of the series, we repeat the calculation for several different values of NT (from 1 to 100) and measure the distance of the results (w) from the true value (expm(m)), using the distance metric:

$$d = \sum_{ij} |w_{ij} - \mathrm{expm}(m)_{ij}|.$$

where the sum runs over all matrix elements. Other metrics are possible. A plot of d versus NT is shown below:



We conclude that only about 10 terms are needed for convergence and calculation of the infinite series is not required. The exact number of terms needed depends on the desired precision.

11.1.2. Detection of Outliers: Mean vs Median, STD vs MAD. Experimental data may contain outliers. There are many different methods for detecting outliers. In order to be able to remove outliers, we need to know how statistical estimators depend on outliers. In this section we will look at the simplest examples based on using the median.

11.1.2.1. *Mean vs Median.* Given a random sample $x_1, x_2, \ldots, x_N$ of a random variable X, the sample mean is defined as:

$$\mu_X = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

For example, if the random sample is 1, 2, 3, the mean is $\frac{1}{3}(1+2+3) = 6/3 = 2$. The median, on the other hand, the median is defined as the value x_m such that:

$$\int_{x_m}^{\infty} p(x)\mathrm{d}x = 0.5$$

Substituting the empirical distribution,

$$p(x) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - x_i),$$

the median is the value x_m for which

$$\int_{x_m}^{\infty} \frac{1}{N}\sum_{i=1}^{N} \delta(x - x_i)\mathrm{d}x = 0.5,$$

or

$$\frac{\#\text{ samples above } x_m}{N} = 0.5.$$

The median is the value of x_m such that the fraction of measurements (samples) is half. For example, if the random sample is 5, 1, 7, 3, 10^6, 4, 2, the

median is obtained by ordering the numbers $(1, 2, 3, 4, 5, 7, 10^6)$ and picking the middle value ($x_m = 4$). If the number of measurements is even, we take the average of the two numbers in the middle of the sorted list. Clearly, the selection of the middle number completely disregards all other values in the list. Most importantly, it disregards outliers such as 10^6.

11.1.2.2. *STD vs MAD.* The sample standard deviation is defined in terms of sums of square differences:

$$\sigma_X = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_X)^2}.$$

We may use the p-norm to generalize the deviation as follows:

$$\sigma_X(p) = \left(\frac{1}{N-1}\sum_{i=1}^{N}|x_i - \mu_X|^p\right)^{1/p}, \qquad p = 1, 2, \ldots$$

The special case $p = 1$ is called mean absolute deviation, $\frac{1}{N-1}\sum_{i=1}^{N}|x_i - \mu_X|$. It is easy to see that outliers will dominate the sum when p is large, making the deviation more sensitive to outliers. On the other hand, $p = 1$ is the least sensitive to outliers.

To further reduce the dependence on outliers, we can replace the mean by the median. The median absolute deviation (MAD) is defined as:

```
MAD = MEDIAN(ABS(X-MEDIAN(X)))
```

where `X` is the vector whose entries are the random sample.

11.1.2.3. *MATLAB example.* The following MATLAB example below simulates the measurement of a nominal 20 V voltage. The random sample is stored in the vector `nois`. In the vector `noiso` we have corrupted `nois` by adding an outlier. As can be seen below, the mean is sensitive to outliers but the median is not. Likewise, the standard deviation is sensitive to outliers but the MAD is not. To detect the outlier, we could take the median plus and minus some multiple of the MAD (e.g. `median(X)` $\pm$ `5*mad(X,1)`) and consider any points outside this range to be an outlier.

```
>> nois=randn([1 100])+20;
>> figure;plot(nois,'o-');
>> axis([0 100],[0 25]);
>> noiso=nois; noiso(50)=1e6;
>> figure;plot(noiso);
>> std(noiso)

ans =
```

```
   9.9998e+04

>> median(noiso)

ans =

   20.1582

>> mean(noiso)

ans =

   1.0020e+04

>> mad(noiso,1)

ans =

    0.6377

>> mad(noiso)

ans =

   1.9800e+04
```

In the above code, we created a "constant" 20 V signal and added Gaussian noise with zero mean and standard deviation of 1. The only estimators that come close to these value are the median and `MAD(X,1)`.

11.1.3. Trendline Removal. Suppose you want to estimate the noise in your experiment. However, the noise is additive and sits on top of a measured signal which drifts over time (non-stationary). For example, you may be measuring the fluorescence decay of a fluorophore, or a free-induction decay (FID) in a nuclear magnetic resonance experiment. Such signals typically decay exponentially. One way is to use a long acquisition window and use the tail of the signal (when it is flat), compute the standard deviation of the tail and use this as our estimate of the noise. This is problematic if the acquisition window is too short to capture the tail of the signal. It is also problematic if the noise of interest only exists while the signal is present. In these cases, we want to extract the noise during the part of the signal that decays exponentially. But taking the standard deviation of an exponentially decaying signal will reflect the signal decay envelope, not just the noise.

11.1.3.1. *Subtracting the trendline obtained from a model.* One approach would be to fit an exponential decay to the noisy signal, then subtract the signal from the model to get the residuals and analyze the noise from the residuals. This requires us to have a model properly describing the signal behavior. It also would only work if the noise is much weaker than the signal, otherwise the fit will be poor. If we are unable to fit the model accurately, our noise estimate will be worthless.

11.1.3.2. *High-pass filtering.* A second approach would be to filter the signal and remove its "low frequency components" (slowly varying part). This is also known by engineers as applying a "high-pass filter". In MATLAB there are many built-in filters that can be applied to a signal. For example, see the page:

`https://www.mathworks.com/help/signal/ug/filtering-data-with-signal-processing-toolbox.html`

For our purposes the simplest high-pass filter is obtained by taking the derivative of the signal. Taking the derivative removes any constant baseline (since the derivative of a constant is 0). It also removes a number of low-frequency components (slowly varying parts). This is most easily understood when the signal is represented in terms of its frequency components.

The Fourier transform of a square-integrable signal[1] $f(x)$ is defined as:

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi} \mathrm{d}x$$

In MATLAB the `fft` command can be used to obtain discrete Fourier transformation of a digital signal. The Fourier transform extracts the frequency components of a signal. In the above expression, $\hat{f}(\xi)$ is the amplitude of the component at frequency ξ. This component is obtained by integrating $f(x)$ times $e^{-2\pi i x \xi}$ in order to measure the overlap between those two functions. $e^{-2\pi i x \xi}$ is a complex trigonometric function that oscillates at frequency ξ. For more information about the Fourier transform, see:

`https://en.wikipedia.org/wiki/Fourier_transform`

Compare this to the derivative of $f(x)$, which we denote $f'(x)$. The Fourier transform of $f'(x)$ is:

$$\hat{f}'(\xi) = \int_{-\infty}^{\infty} \frac{\mathrm{d}f(x)}{\mathrm{d}x} e^{-2\pi i x \xi} \mathrm{d}x = 2\pi i \xi \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi} \mathrm{d}x = (2\pi i)\xi \hat{f}(\xi),$$

where the second equality follows by integration by parts. This result tells us that taking the derivative of $f(x)$ amounts to multiplication of $\hat{f}(\xi)$ by ξ in the frequency domain. Therefore, low frequencies are scaled down, whereas

[1] Square integrable: $\int_{-\infty}^{\infty} |f(x)|^2 \mathrm{d}x < \infty$. This implies that $|f(x)|$ decays to 0 at infinity.

high frequency components are scaled up, hence the terminology "high-pass filter".

The following MATLAB example creates an exponentially decaying signal with additive Gaussian noise, takes its derivative (`diff` command) and plots the result:

```
>> t=0:0.01:10;
>> u=10*exp(-t/2)+50+randn(size(t));
>> figure;plot(t,u);
>> v=diff(u);
>> td=t(1:end-1);
>> figure;plot(td,v);
```

We can then plot the power spectral density (`pwelch` command), which is the square of the Fourier transform plotted on a log scale (in dB), to visualize the signal in the frequency domain:

```
>> figure;plot(pwelch(u));
>> figure;plot(pwelch(v));
```

As can be seen, the low frequencies are scaled down. For a signal in units of Volts (V) the one-sided power spectral density (PSD) in V^2/Hz is defined as:

$$\mathrm{PSD}(f) = \frac{2|X(f)|^2}{(t_2 - t_1)}$$

where $X(f)$ is the Fourier transform of the signal $x(t)$ defined[2] over the time range (t_1, t_2):

$$X(f) \equiv \int_{t_1}^{t_2} x(t) e^{-2\pi i f t} \mathrm{d}t,$$

for any frequency f in the two-sided frequency domain $(-F, F)$. If $x(t)$ is expressed in units of V, $X(t)$ is expressed in units of V/Hz. The PSD divides up the total power of the signal. To see this, we integrate the PSD over its entire one-sided frequency domain $(0, F)$:

$$\begin{aligned}\int_0^F \mathrm{PSD}(f)\mathrm{d}f = \int_0^F \frac{2|X(f)|^2}{t_2 - t_1}\mathrm{d}f = \frac{1}{t_2 - t_1}\int_{-F}^{F} |X(f)|^2 \mathrm{d}f \\ = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} |x(t)|^2 \mathrm{d}t,\end{aligned}$$

where the last step follows from the Parseval's theorem. The result is precisely the average power of the signal in the time range $(t_2 - t_1)$.

[2]The factor 2 is due to adding the contributions from positive and negative frequencies.

We often express the PSD in dB relative to some (given) reference signal value x_{ref} (units: V):

$$\mathrm{PSD}_{\mathrm{dB}}(f) = 10\log_{10}\left[\frac{\mathrm{PSD}(f)}{x_{ref}^2}\right].$$

Note: since the argument of the logarithm is in units of Hz^{-1}, this spectral measure can loosely be said to be in units of "dB/Hz".

Those who are not familiar with the Fourier transform may gain some appreciation by applying the short-term Fourier transform (STFT)

https://en.wikipedia.org/wiki/Short-time_Fourier_transform

$$X(\tau,\omega) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-i\omega t}\mathrm{d}t \qquad w(t):\ \text{window function (continuous case)}$$

$$X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-i\omega n} \qquad \text{(discrete case)}$$

to an audio signal:

```
handel = load('handel');
figure;
[sp,fp,tp] = pspectrum(handel.y,handel.Fs,'spectrogram');
mesh(tp,fp,sp)
view(-15,60)
xlabel('Time (s)')
ylabel('Frequency (Hz)')
soundsc(handel.y, handel.Fs) % hear the audio clip
```

The output is a plot of power spectral density as function of time and frequency. This is obtained by computing the Fourier transform on short time windows, as a function of time. At each point in time, a spectrum displays which frequencies (pitches) are present in the sound heard. This is similar to a musical score, where time flows along the horizontal direction and frequency is shown by the notes along the vertical direction (lines of a staff).

https://www.mathworks.com/help/signal/ref/bandpass.html

https://www.mathworks.com/help/signal/ref/bandstop.html

https://www.mathworks.com/help/signal/ref/highpass.html

https://www.mathworks.com/help/signal/ref/lowpass.html

```
fs = 1e3;
t = 0:1/fs:1;
x = [2 1 2]*sin(2*pi*[50 150 250]'.*t) + randn(size(t))/10;
bandpass(x,[100 200],fs)
fs = 2e3;
t = 0:1/fs:0.3-1/fs;
```

```
l = [0 130.81 146.83 164.81 174.61 196.00 220 246.94];
m = [0 261.63 293.66 329.63 349.23 392.00 440 493.88];
h = [0 523.25 587.33 659.25 698.46 783.99 880 987.77];
note = @(f,g) [1 1 1]*sin(2*pi*[l(g) m(g) h(f)]'.*t);
mel = [3 2 1 2 3 3 3 0 2 2 2 0 3 5 5 0 3 2 1 2 3 3 3 3 2 2 3 2 1]+1;
acc = [3 0 5 0 3 0 3 3 2 0 2 2 3 0 5 5 3 0 5 0 3 3 3 0 2 2 3 0 1]+1;
song = [];
for kj = 1:length(mel)
    song = [song note(mel(kj),acc(kj)) zeros(1,0.01*fs)];
end
song = song/(max(abs(song))+0.1);
% To hear, type sound(song,fs)
pspectrum(song,fs,'spectrogram','TimeResolution',0.31, ...
    'OverlapPercent',0,'MinThreshold',-60)
pong = bandpass(song,[230 450],fs);
% To hear, type sound(pong,fs)
bandpass(song,[230 450],fs)
figure
pspectrum(pong,fs,'spectrogram','TimeResolution',0.31, ...
    'OverlapPercent',0,'MinThreshold',-60)
```

11.1.4. Noise Removal.

11.1.4.1. *1D signal: Low-pass filtering.* Noisy data leads to errors in the fit. Suppose that you want to keep the trendline and remove the noise. This "denoising" operation is often carried out using a low-pass filter. An example of low-pass filter that removes noise is the moving average:

$$\overline{x}_i = \frac{1}{2M+1} \sum_{j=-M}^{M} x[i+j],$$

where $2M+1$ is the number of time points used to compute the moving average. Let's write MATLAB code to implement the moving average. I will take a one-sided moving average, i.e. $\overline{x}_i = \frac{1}{M+1} \sum_{j=-M}^{0} x[i+j]$, which would have applications when the data is acquired (and analyzed) in real-time and data points in the future are not yet available.

```
% moving average

data = thingSpeakRead(276806,'DateRange',[datetime('January 3, 2019 ...
    0:0:0') datetime('January 4, 2019 ...
    0:0:0')],'Fields',1,'outputFormat','timetable');
   %load stock market data (MATLAB dataset)

lag = 6; % 6-pt moving average
simple = movavg(data.Last,'simple',lag); % let MATLAB compute mov avg
plot(data.Timestamps,data.Last, data.Timestamps,simple);
legend('Last Price','6 Pt. Average');
ylabel('Last Stock Price');
title('Last Price & Moving Average');

```

```
% calculate moving average ourselves, compare w/MATLAB

M=6; % 6-pt moving average
clear mav
for j=M:length(data.Last),
    prev=data.Last(j-M+1:j);
    mp=mean(prev);
    mav(j-M+1)=mp;
end;

figure;plot(data.Timestamps,data.Last,'b');
hold on; plot(data.Timestamps(M:length(data.Last)),mav,'r');
```

A moving average is a form of a convolution often used in time series analysis to smooth out noise in data by replacing a data point with the average of neighboring values in a moving window. A moving average is essentially a low-pass filter because it removes short-term fluctuations to highlight a deeper underlying trend.

For those not familiar with convolution: the convolution of two functions f and g is denoted $f * g$ and calculated as

$$(f * g)(x) \equiv \int_{-\infty}^{\infty} f(y)g(x-y)\mathrm{d}y.$$

In discrete form we have:

$$(f * g)[n] = \sum_{m=-\infty}^{\infty} f[m]g[n-m].$$

For more information about convolutions, including visual explanations, see

`https://en.wikipedia.org/wiki/Convolution`

In MATLAB, a number of variants exist: `movmean`, `movmad`, `movmedian`. The moving median is less susceptible to outliers (e.g. rare events such as spikes). The moving average is also implemented in the command `smoothdata`:

`https://www.mathworks.com/help/matlab/ref/smoothdata.html`

The MATLAB link provides useful examples of signal filtering. The first example provided is the moving average:

```
x = 1:100;
A = cos(2*pi*0.05*x+2*pi*rand) + 0.5*randn(1,100);
B = smoothdata(A);
plot(x,A,'-o',x,B,'-x')
legend('Original Data','Smoothed Data')
```

Gaussian filtering is also a popular method for removing noise. The example provided illustrates filtering using two different window lengths (4 and 20):

```
x = 1:100;
A = cos(2*pi*0.05*x+2*pi*rand) + 0.5*randn(1,100);
[B, window] = smoothdata(A,'gaussian');
window
C = smoothdata(A,'gaussian',20);
plot(x,B,'-o',x,C,'-x')
legend('Small Window','Large Window')
```

Try these examples. It is generally useful to visualize the effects of the filter in Fourier space, i.e. use the `pwelch` command.

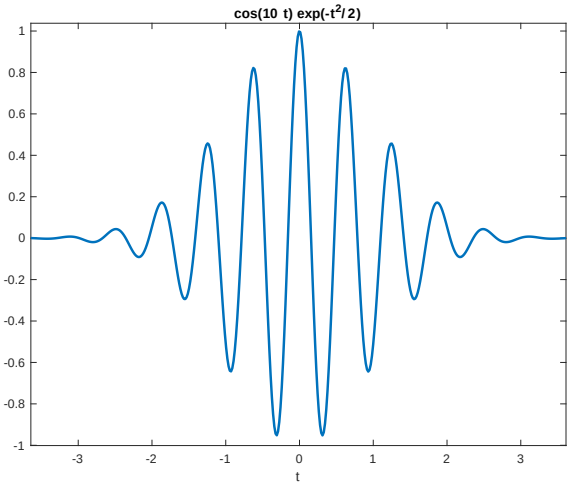
11.1.4.2. *1D signal: Wavelet denoising.* Another powerful tool for signal processing is the wavelet transform. A wavelet is a function of the form

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right), \qquad a, b \in \mathbb{R}$$

where ψ is the "mother wavelet", a is called the dilation (scale) parameter and b is the translation (position) parameter. An example mother wavelet is the Morlet wavelet:

$$\psi(t) = e^{i\omega_0 t}e^{-t^2/2},$$

whose real part is shown below:



Wavelets functions are localized in time. They are translated along the signal (parameter b). They are also scaled (parameter a). The scaling parameter stretches or compresses the wavelet (in time) and is akin to varying the frequency.

The continuous wavelet transform (CWT) of a signal $x(t)$ is given by

$$\hat{x}(a,b) = \langle x, \psi_{a,b}\rangle = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} x(t)\cdot\psi^*\left(\frac{t-b}{a}\right)\mathrm{d}t.$$

Here, $\langle x, \psi_{a,b} \rangle$ is called the L^2 inner product. Here, a and b vary continuously and hence the CWT generates a lot of data. When a and b are chosen to be discrete, the analysis is easier. The discrete wavelet transform (DWT) is obtained by choosing a to be integer powers of a fixed dilation parameter $a_0 > 1$, i.e., $a = a_0^m$ (varying m alters the width of the wavelets), and translating the narrow wavelets by small steps and wider wavelets by larger steps, $b = nb_0a_0^m$, $b_0 > 0$ fixed and $n \in \mathbb{Z}$. The wavelets are discretely labeled:

$$\psi_{m,n}(k) = a_0^{-m/2}\psi(a_0^{-m}(k - nb_0a_0^m)), \qquad m, n \in \mathbb{Z}.$$

The DWT of $x(k)$ is defined as

$$\hat{x}[m, n] = \langle f, \psi_{m,n} \rangle = a_0^{-m/2} \sum_{k=-\infty}^{\infty} x(k)\psi^*(a_0^{-m}k - nb_0).$$

If the scales and positions are chosen based on powers of two (Dyadic), the analysis becomes more efficient. In 1988 Stéphane Mallat remarked that for special choice of $\psi(k)$ and a_0, b_0 the $\psi_{m,n}(k)$ constitute an orthonormal basis for $L^2(\mathbb{R})$. In particular, if $a_0 = 2$, $b_0 = 1$ there exist $\psi(k)$ with good time-frequency localization properties such that the

$$\psi_{m,n}(k) = 2^{-m/2}\psi(2^{-m}k - n), \qquad m, n \in \mathbb{Z},$$

constitutes an orthonormal basis for $L^2(\mathbb{R})$. The DWT is then

$$\hat{x}[m, n] = \langle f, \psi_{m,n} \rangle = 2^{-m/2} \sum_{k=-\infty}^{\infty} x(k)\psi^*(2^{-m}k - n).$$

The parameters a, b result in a time-frequency analysis similar to the STFT except that the size of the time window being analyzed by the wavelet varies depending on the scale parameter a. A downside of STFT is that it has a fixed resolution. The width of the windowing function relates to how the signal is represented. It determines whether there is good frequency or time resolution. A wide window[3] gives better frequency resolution but poor time resolution. A narrower window gives good time resolution but poor frequency resolution. Time and frequency are related by an uncertainty principle.[4] This is one of the reasons for the creation of the wavelet transform

[3]Frequency resolution Δf is related to the length of the time window T according to an inverse relationship $\Delta f = 1/T$.

[4]Let $x(t)$ be a signal and $X(f)$ its Fourier transform. Define the following PDF: $p_x(t) = \frac{|x(t)|^2}{\|x(t)\|^2}$, where $\|x(t)\|^2 = \int_{-\infty}^{\infty} |x(t)|^2 \mathrm{d}t$. Then, $p_x(t) > 0$ and $\int_{-\infty}^{\infty} p_x(t)\mathrm{d}t = 1$. The spread of the signal $x(t)$ over time (i.e. locality of the dispersion of $x(t)$) can be measured as the variance

$$\sigma_t^2 = \int_{-\infty}^{\infty} (t - \mu_t)^2 p_x(t)\mathrm{d}t = \frac{1}{\|x(t)\|^2} \int_{-\infty}^{\infty} (t - \mu_t)^2 |x(t)|^2 \mathrm{d}t,$$

(multiresolution analysis). The latter offers good time resolution for high-frequency events and good frequency resolution for low-frequency events, a combination suited for many signals.

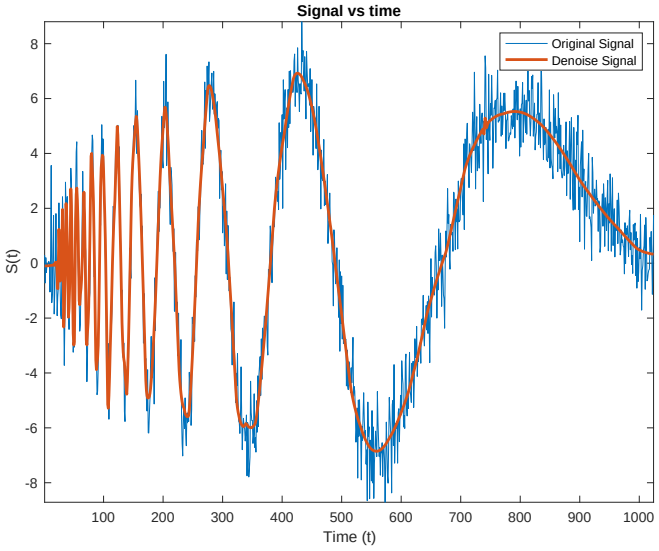
For more information see:

`https://en.wikipedia.org/wiki/Short-time_Fourier_transform`
`https://en.wikipedia.org/wiki/Discrete_wavelet_transform`
`https://en.wikipedia.org/wiki/Complex_wavelet_transform`
`https://en.wikipedia.org/wiki/Continuous_wavelet_transform`
`https://en.wikipedia.org/wiki/Wavelet`

The following MATLAB code takes a noisy Doppler signal (`noisdopp`) and cleans it up using the `wdenoise` command

```
load noisdopp
xden=wdenoise(noisdopp);
h1=plot([noisdopp' xden']);title('Signal vs time');
h1(2).LineWidth=2;xlabel('Time (t)');ylabel('S(t)');
legend('Original Signal','Denoised Signal');
```

as shown in the graph below:

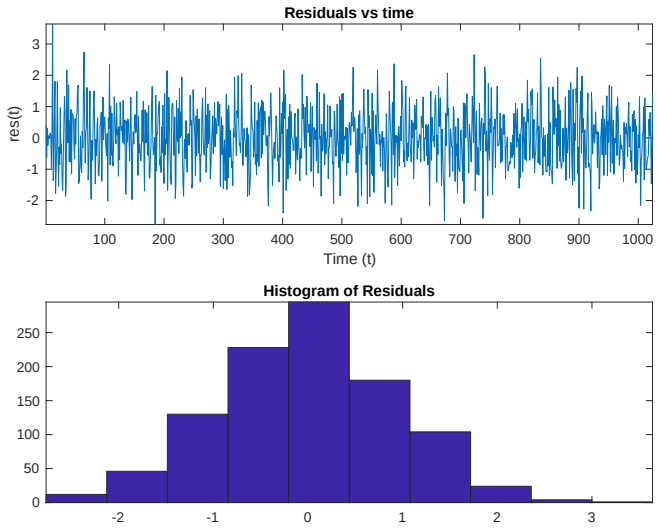


where μ_t is the mean $\mu_t = \int_{-\infty}^{\infty} tp_x(t)\mathrm{d}t = \frac{1}{\|x(t)\|^2}\int_{-\infty}^{\infty} t|x(t)|^2\mathrm{d}t$. Similarly we define $\sigma_f^2 = \frac{1}{\|X(f)\|^2}\int_{-\infty}^{\infty}(f-\mu_f)^2|X(f)|^2\mathrm{d}f = \frac{1}{\|x(t)\|^2}\int_{-\infty}^{\infty}(f-\mu_f)^2|X(f)|^2\mathrm{d}f$, where we used Parseval's identity $\|X(f)\|^2 = \|x(t)\|^2$ and $\mu_f = \frac{1}{\|X(f)\|^2}\int_{-\infty}^{\infty} f|X(f)|^2\mathrm{d}f$. The uncertainty principle is $\sigma_t^2\sigma_f^2 \geq (16\pi^2)^{-1}$.

If we subtract the noisy signal from the cleaned-up signal we are left with the noise. The noise statistics are visualized by plotting the histogram of the residuals:

```
res=noisdopp-xden; % residuals
figure;subplot(2,1,1);plot(res);title('Residuals vs time');
xlabel('Time (t)');ylabel('res(t)');
subplot(2,1,2);hist(res);title('Histogram of Residuals');
```

as shown below:



This method can be used to isolate and analyze the signal and noise separately.

In some versions of MATLAB the above dataset (`noisdopp`) does not appear to be available. If that is the case, try this code below, which creates a fake Doppler signal and adds noise to it:

```
% chirp signal denoising

t = 0:0.001:2; % 2 s at 1 kHz sample rate
y = chirp(t,0,1,150); % create chirp
spectrogram(y,256,250,256,1E3); % plot spectrogram

ts=t(1:200); % apply to first 200 points only
ys=y(1:200); % so we can visualize a few oscillations
figure;plot(ts,ys);

ysn=ys+0.3*randn(size(ys)); % add noise
```

```
12
13 figure;plot(ts,ysn); % plot noisy signal
14
15 xden=wdenoise(ysn); % denoise
16
17 hold on;
18 plot(ts,xden,'g'); % plot denoised signal
```

11.1.4.3. *2D signal: Image denoising.* There is a wide range of methods that are available for image denoising. The reader is referred to the MATLAB documentation for examples:

- 2D Gaussian filtering of images:
  `https://www.mathworks.com/help/images/ref/imgaussfilt.html`
- 2D adaptive noise-removal filtering:
  `https://www.mathworks.com/help/images/ref/wiener2.html`
- 2D median filtering:
  `https://www.mathworks.com/help/images/ref/medfilt2.html`
- Additional algorithms implemented in MATLAB:
  `https://www.mathworks.com/help/images/linear-filtering.html`

In this section we present some basic examples of filtering in the Fourier domain. We start by zero-padding the outer region of k-space, i.e. nulling out the high-frequency components. Then we perform Gaussian filtering, i.e. multiplication of the k-space representation of the image by a Gaussian filter. In both cases, the idea is to reduce the amplitude of the high-frequency components, as the latter mostly contain noise. Unfortunately, the high-frequency components contain information about sharp edges and by reducing their amplitude, such filters smooth out the edges. The denoised image ends up blurry.

```
1  % image of vegs
2  originalRGB = imread('peppers.png');
3  imr=squeeze(originalRGB(:,:,1));
4  figure;imagesc(imr);
5  colormap(gray);colorbar;
6
7  % add noise
8  imrn=double(imr)+10*randn(size(imr));
9  figure;imagesc(imrn);
10 colormap(gray);colorbar;
11
12 % go to Fourier space & filter
13 fi=fftshift(fft2(imrn));
14 figure;imagesc(abs(fi));
15 colormap(gray);colorbar;
16 caxis([0 3e4]);
17
```

```
% box filter
bf=zeros(size(fi));
%bf(100:300,150:400)=fi(100:300,150:400);
bf(150:250,200:325)=fi(150:250,200:325);
figure;imagesc(abs(bf));
colormap(gray);colorbar;
caxis([0 3e4]);

% 2D gaussian filter
[XX,YY]=meshgrid(1:512,1:384);
XX=XX-512/2;
YY=YY-384/2;
bf=fi.*exp(-( (XX.^2)/10000 + (YY.^2)/10000 ));
figure;imagesc(abs(bf));
colormap(gray);colorbar;
caxis([0 3e4]);

% go back to real-space, show image
imf=ifft2(fftshift(bf));
figure;imagesc(abs(imf));
colormap(gray);colorbar;
%caxis([0 3e4]);
```

There are many algorithms capable of noise removal without blurring edges. One such example is the method of *non-local means*:

`https://www.mathworks.com/help/images/ref/imnlmfilt.html`

The method is described here:

- Buades, A., B. Coll, and J.-M. Morel. "A Non-Local Algorithm for Image Denoising." 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 2, June 2005, pp. 60–65.
- `https://en.wikipedia.org/wiki/Non-local_means`

```
% non-local means filtering (greyscale)
I = imread('cameraman.tif');
noisyImage = imnoise(I,'gaussian',0,0.0015);
[filteredImage,estDoS] = imnlmfilt(noisyImage);
montage({noisyImage,filteredImage});
title(['Estimated Degree of Smoothing, ',...
    'estDoS = ',num2str(estDoS)]);
```

The method of Wavelets described in the previous section can be used to denoise not only 1D signals but also (2D) images. Because wavelets localize features in your data to different scales, we can preserve important signal or image features while removing noise. The basic idea behind wavelet denoising, or wavelet thresholding, is that the wavelet transform leads to a sparse representation for many real-world signals and images. What this

means is that the wavelet transform concentrates signal and image features in a few large-magnitude wavelet coefficients. Wavelet coefficients which are small in value are typically noise and we can "shrink" those coefficients or remove them without affecting the signal or image quality. After you threshold the coefficients, we reconstruct the data using the inverse wavelet transform. This method of wavelet denoising has many advantages over Fourier-based filtering methods.

To denoise images you would use the command `wdenoise2`. For example:

```
1 load('jump.mat')
2 wdenoise2(jump)
```

More signal processing examples can be found in the MATLAB documentation for the Wavelet Toolbox

`https://www.mathworks.com/help/wavelet/`

11.1.5. Empirical Distribution and Histograms. Let X be a random variable. We measure X by collecting a random sample $x_1, x_2, \ldots, x_N$. Recall that the empirical distribution of this random sample is the PDF:

$$p(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x_i - x),$$

where $x \in \mathcal{X}$. $\mathcal{X}$ here is the set where the random variable X takes its values from (i.e. the "range" of X). The histogram is obtained by partitioning the set $\mathcal{X}$ into n_b intervals $I_1, I_2, \ldots, I_{n_b} \subset \mathcal{X}$ such that $\cup_{j=1}^{n_b} I_j = \mathcal{X}$ and $I_i \cap I_j = \emptyset$ $(i \neq j)$ and counting the number of samples falling into each interval (bin). Mathematically, this is equivalent to integrating the empirical distribution over each interval (bin) to get the bin count (frequency):

$$bin(j) \equiv \int_{I_j} p(x)dx = \frac{1}{N}\int_{I_j}\sum_{i=1}^{N}\delta(x_i - x)\mathrm{d}x = \frac{\#\text{ samples in bin } j}{N}$$

The histogram itself is a bar graph plotting the bin count $bin(j)$ vs the bin index j.

11.1.6. Plotting histograms: Illustration of the Central Limit Theorem. The command in MATLAB to plot histograms is `hist` on earlier versions of MATLAB and `histogram` in the later versions. Another useful function is `histfit`, which will not only plot a histogram, but also will fit a Gaussian PDF to it. Let's look at the central limit theorem (CLT).

We know the CLT predicts that the sum of random variables (regardless of their distribution, as long as they have finite mean and variance) eventually

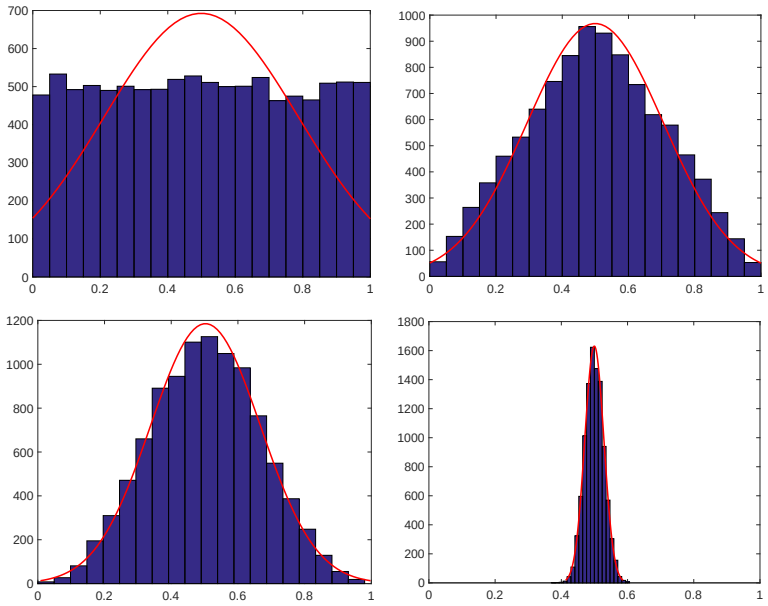


Figure 11.1. Illustration of the central limit theorem. Uniformly distributed random variables X_i (on $[0,1]$) are added together to form the average $(X_1 + \cdots + X_n)/n$. The cases $n = 1, 2$ (top row) and $n = 3, 100$ (bottom row) are shown. These histograms are plotted by generating 10,000 realizations of each random variable X_i.

converges to a normal distribution in the limit of large n:

$$\frac{X_1 + \cdots + X_n}{n} \to \mathcal{N}(\mu, \sigma^2/n) \quad \text{as } n \to \infty.$$

In the code below, `NA` uniformly distributed random variables are added together. The results for `NA=1,2,3,100` are shown in Fig 11.1. As can be seen, the $n = 1$ case is uniform, the $n = 2$ case yields a tent function, whereas $n = 3$ looks increasingly more like a Gaussian. The $n = 100$ case looks Gaussian. Notice also the width of the distribution narrows. This is due to the variance of the arithmetic average, σ^2/n, which decreases as $1/n$.

```
v=zeros([1 10000]);
NA=100; % of rv's to be added in sum
for j=1:NA,
    v=v+rand([1 10000]); % randomly generate 10,000 pts
end; % and add them together (sum NA such rv's)
v=v/NA; % normalize to the # of rv's added
figure; histfit(v,20); % plot histogram, fit a gaussian to it
set(gca,'xlim',[0 1]);
set(gca,'fontsize',16);
```

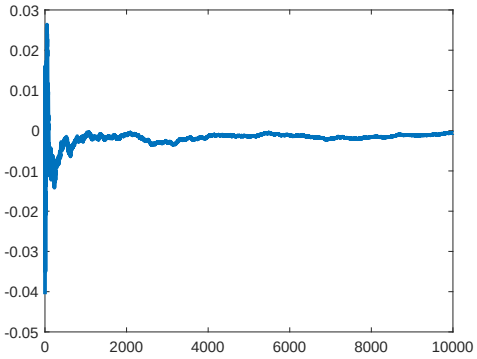


Figure 11.2. Illustration of the law of large numbers. Plot of $\frac{1}{n}\sum_{i=1}^{n}\sin(X_i)$ vs n, from $n = 1, \ldots, 10000$ (horizontal axis). As n increases, the value of the sum approaches 0, the value of the integral.

11.1.7. Cumulative Sums: Illustrating the Law of Large Numbers. When dealing with summations of n terms, it is often useful to look at the partial results of the sum as a way to study convergence of the sum. To this end, the command `cumsum` can be used. The code below illustrates the example of computing the following integral numerically, using a random number generator. Let's compute the average of $\sin(x)$ from 0 to 2π:

$$I = \frac{1}{2\pi}\int_0^{2\pi} \sin(x)\mathrm{d}x.$$

This integral is of the form $\int f(x)p(x)\mathrm{d}x$, where $p(x) = (2\pi)^{-1}\mathbf{1}_{[0,2\pi]}(x)$ is the PDF for the uniform distribution over the interval $[0, 2\pi]$. Thus, it is the expectation value $\mathbb{E}_p(\sin(X))$, which can be approximated by the law of large numbers (LLN) as a sum:

$$I = \mathbb{E}_p(\sin(X)) \approx \frac{1}{n}\sum_{i=1}^{n}\sin(X_i), \qquad X_i \sim p(x).$$

Thus, I can be approximated by generating uniformly distributed random numbers $\{X_i\}$ and taking the arithmetic average of the $\{\sin(X_i)\}$. The code below generates the summation for different n.

```
NP=10000; % max # of points in LLN sum (n=1,...,NP)
x=(2*pi)*rand([1 NP]); % sample integrand uniformly from 0 to 2*pi
sx=(1/2/pi)*sin(x); % function to be integrated
NL=linspace(1,NP,NP); % normalize to # of pts. in sum
ysx=cumsum(sx)./NL;
figure; plot(ysx); % convergence of sum vs n
set(gca,'fontsize',16);
```

11.1.8. Fitting a trendline to data plotted in a figure window. Data can be analyzed using the basic curve fitting tools available from the pull-down menus of a figure window. To show how this works, let us start by creating a fake data set. The horizontal axis is:

```
t=1:100;
```

Create a slope:

```
slope=10;
```

This command will create a straight line graph, with gaussian noise added:

```
f=t*slope + 60*randn([1 100]);
```

To get more information about random numbers type `help randn`. Plot the results as follows:

```
figure; plot(t,f,'o');
```

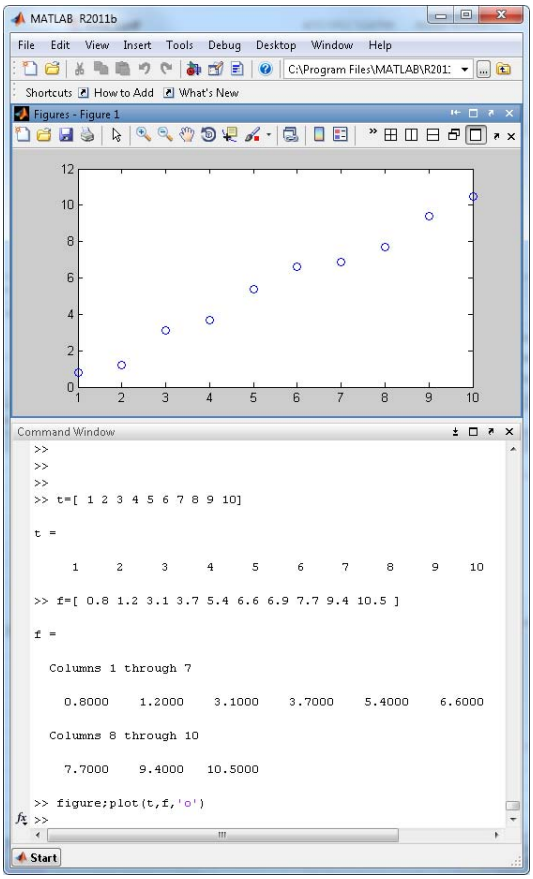
In the next section we shall analyze this data, as well as data from the following function (straight line plus quadratic component):

```
g=f+ 0.1*t.^2;
figure; plot(t,g,'o');
```

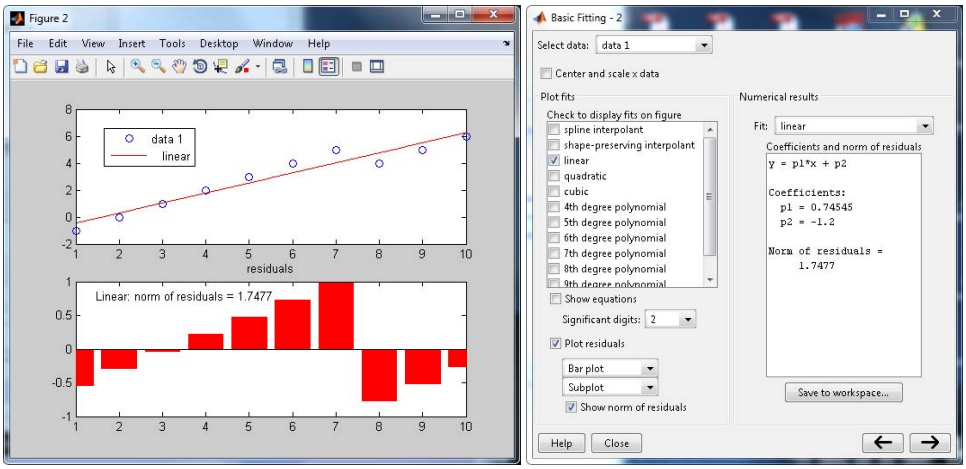
and this one (straight line plus sinusoidal component):

```
h=f + 50*sin(0.2*t);
figure; plot(t,h,'o');
```

The straight line data looks like this:



Basic curve fitting can be done by choosing `Tools` in the `Figure` menu, followed by `Basic Fitting`. If we expect the data to be described by a straight line, select `linear` from the menu. MATLAB will plot the trendline and list the results from the fit in a box.



If you check the box `Plot residuals`, MATLAB will divide the window into two sections and add a plot of the differences between the trendline and the fitted curve. Residuals are a good way to see if the model chosen (in this case, a straight line) is suitable for describing the experimental data. The residuals are shown in the above example. If the data is a good fit to the model, the residuals should be normally distributed (Gaussian) with mean 0.

More formally, the i-th residual r_i is defined as:

$$r_i = y(x_i) - y_i$$

where $y(x_i)$ is the model and y_i is the experimentally-measured data point. MATLAB also returns a quantity called `Norm of residuals`. MATLAB defines this quantity as

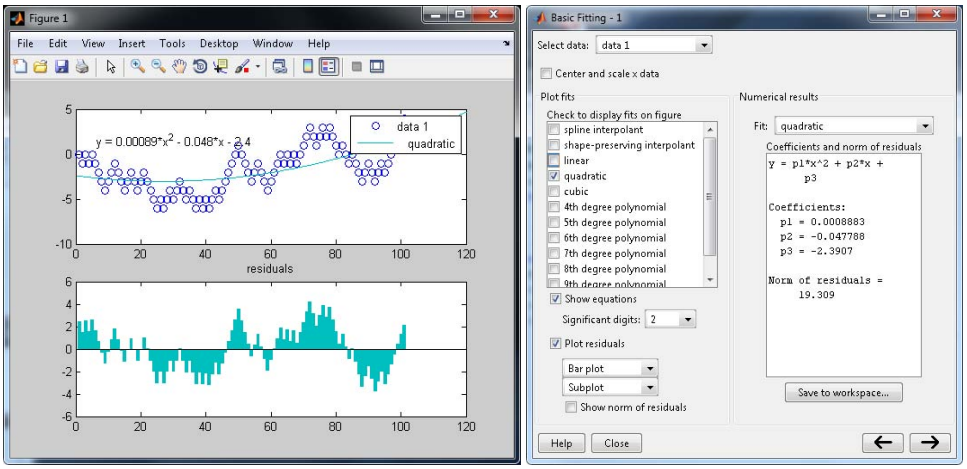
$$\texttt{Norm of residuals} = \sqrt{\sum_{i=1}^{N} r_i^2},$$

where N is the number of data points. This provides a measure of the goodness-of-fit. A perfect fit has zero residuals and the `Norm of residuals` will also be zero. A model for the fit should be chosen as to minimize the `Norm of residuals`. Choosing any model for the sake of minimizing `Norm of residuals` is generally not a good strategy.

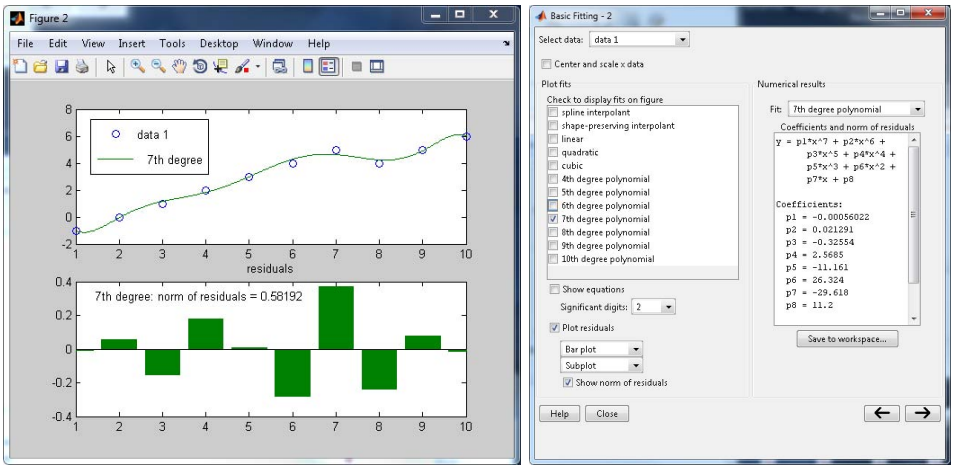
The following link contains useful information on interactive data fitting

`http://www.mathworks.com/help/techdoc/data_analysis/f1-15377.html`

An example of poor data fitting is shown below. The best judge for this is visual inspection: the quadratic model clearly isn't a good match to the data. The poor fit is also evidenced by the large value of `Norm of residuals`.



A bad choice of fitting function is shown below

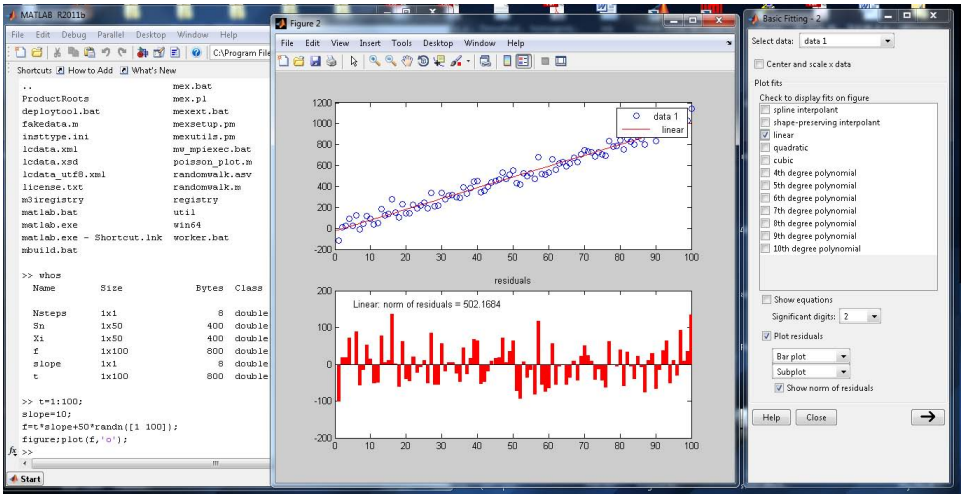


where a 7th order polynomial was used to fit the above linear data set. The fit to the 7th order polynomial gives a better value for the `Norm of residuals`. But it is also completely unphysical and yields no useful information about the experiment.

In general, you want to pick a model based on the physics of the system you are studying. You also should pick the simplest model possible, i.e. one with fewest parameters. As you increase the number of fitting parameters (i.e. by choosing more complicated functions), the fit look better (visually or according to `Norm of residuals`), but the model is meaningless from a physical standpoint.

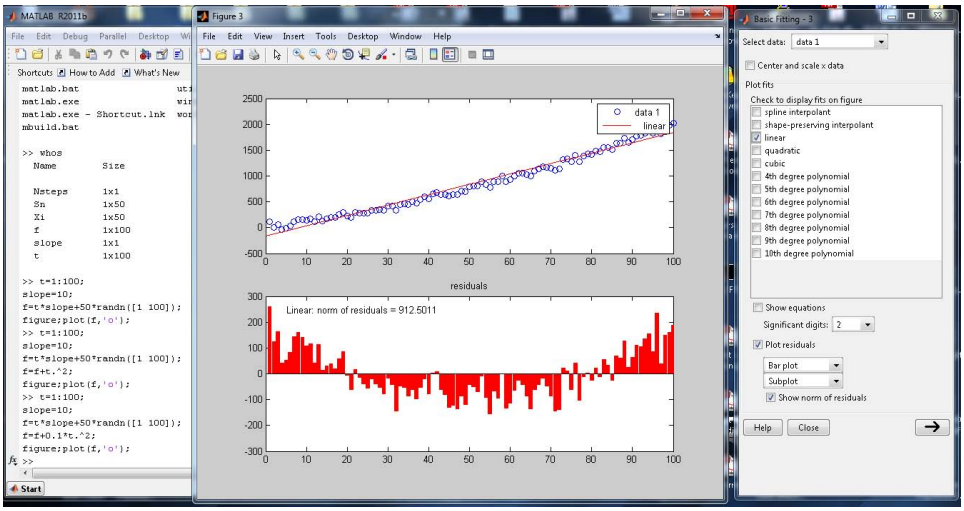
To re-iterate this point: when fitting data, we want to minimize the residuals, but this should not be done at the expense of choosing an inadequate model. Start from the simplest model and build up gradually in complexity as needed. Examples are shown below.

In the following example, a noisy data set is fit well by a straight line. Inspection of the residuals show that the fit does not appear to be missing any underlying trends, i.e. the residuals are just noise and don't exhibit any particular structure.

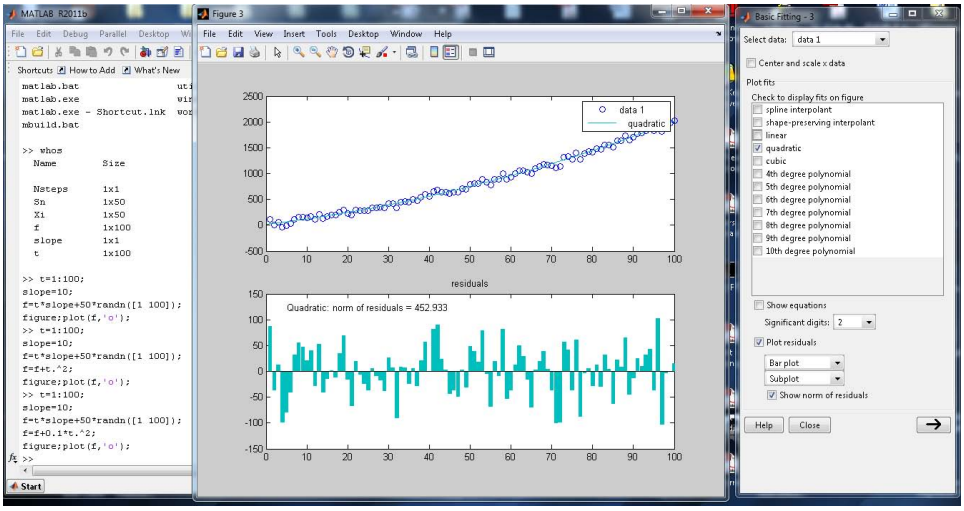


The second example below shows a similar data set, but one in which the linear fit leaves out residuals. Inspection of the structure of these residuals suggests that we are missing a quadratic component. You would then be better off to fit the data using a second order polynomial.

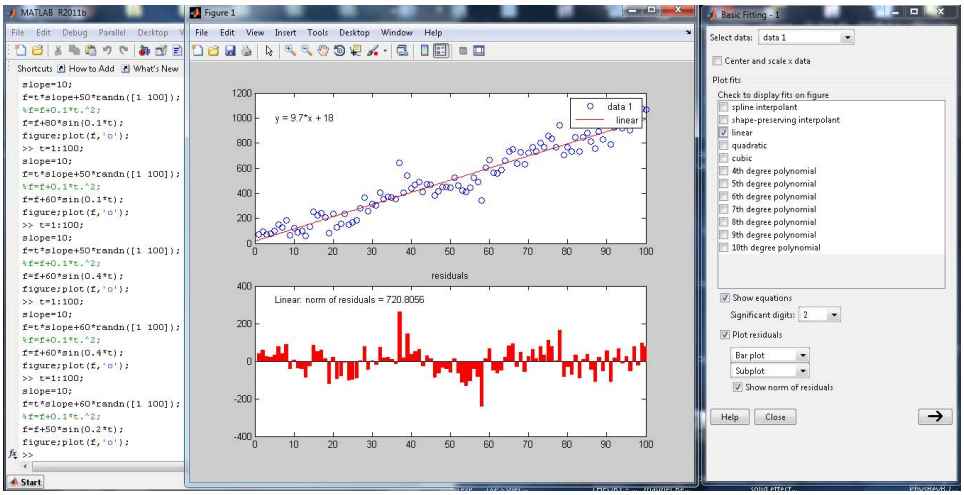
In such cases, you are left with the task of explaining the physical origin of this quadratic dependence. In the case of Hooke's law, one possible explanation would be the over-streching of the spring which is known to lead to non-linear behavior. Each experimental situation is different and must be examined in its own light.



The following interactive fit shows that a quadratic model leaves residuals that are just noise (no specific underlying structure). As an exercise, we could plot the histogram of these residuals and check that they are distributed normally with mean 0.



The final example is an experimental data set which is mostly a straight line, but where the residuals reveal an underlying sinusoidal behavior.



11.1.9. Can residuals plots help identify systematic errors? We remark that plotting the residuals can be a useful method for identifying systematic errors in your experiment. When the laws of physics state that the behavior should be modeled by a particular set of equations, but that the residuals of the fit to this model point to additional trends in the data, this is a sign that there may be systematic errors.

For example, a sinusoidal trend in the residuals that does not follow from the expected model may be suggestive of the presence of vibrations in your experimental apparatus. In this case, systematic errors could be mitigated by fitting the data to the original model, plus a sinusoidal term. The parameters of the model are extracted and those of the sinusoidal fit are discarded. A brute-force, but perhaps costly, way to reduce systematic errors is to improve the physical apparatus.

A careful analysis of the origins of systematic errors may enable you to identify and subtract these errors without having to re-design the experiment.

11.2. cftool – a GUI-based curve fitting tool

The `cftool` command in MATLAB is part of the Curve Fitting toolbox and provides a graphical user interface (GUI) for data fitting. In this section, we will demonstrate its use.

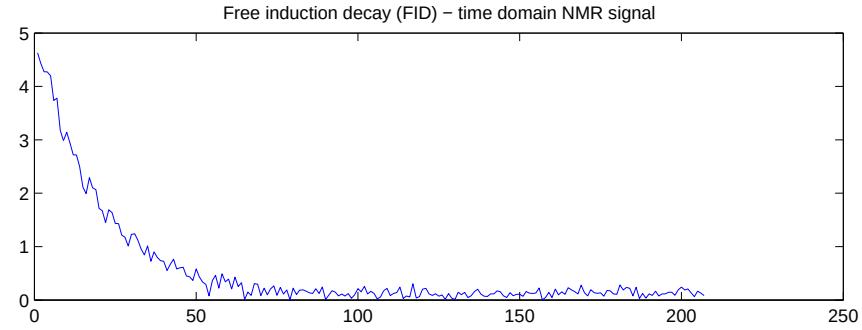
Let us first load a data set from the file `noise.mat`. The variable `signoise` contains a complex-valued vector which is the free induction decay (FID) from a nuclear magnetic resonance (NMR) experiment done on a test tube of water (a single resonance). The last two commands below will plot the signal in the time domain (the FID) and its Fourier transform (the NMR spectrum).

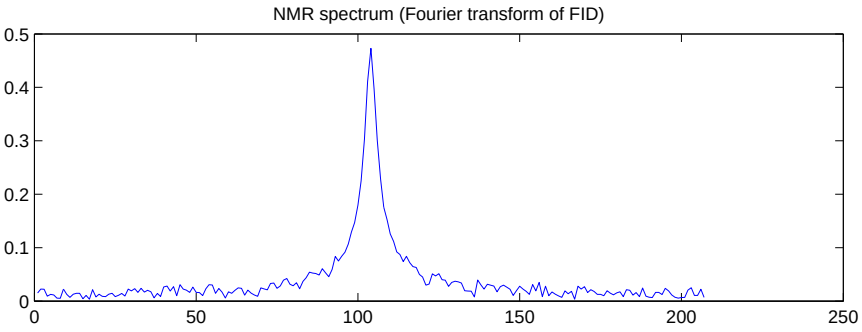
```
load noise
whos
```

```
Name              Size              Bytes  Class     Attributes

addnoise        207x1                3312  double    complex
multnoise         1x951             15216  double    complex
signoise          1x207              3312  double    complex
```

```
figure;plot(abs(signoise));
figure;plot(abs(fftshift(ifft(signoise))));
```

The result is:





This follows from the Fourier transform of an exponential decay

$$s(t) = e^{-2\pi|t|/\tau},$$

which is Lorentzian[5]

$$\hat{s}(f) = \frac{1}{\pi}\frac{\tau^{-1}}{f^2 + \tau^{-2}}.$$

[5]For a proof of this statement, see:
`http://mathworld.wolfram.com/FourierTransformExponentialFunction.html`

τ is inversely related to the width of the Lorentzian.

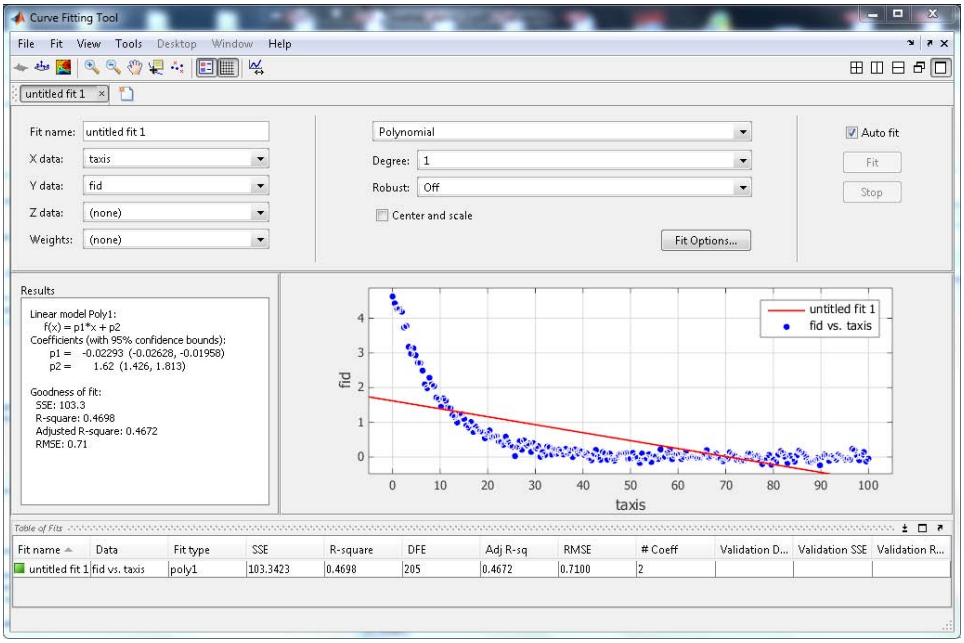
Let us fit only the real part. We shall determine the decay constant.

```
figure;plot(real(signoise));
fid=real(signoise);
```

Let us also create a time axis. For example, suppose this FID was sampled over a 100ms time window. We create a vector containing 207 time values from 0 to 100 ms (207 is the number of points in the FID):

```
taxis=linspace(0,100,207);
```

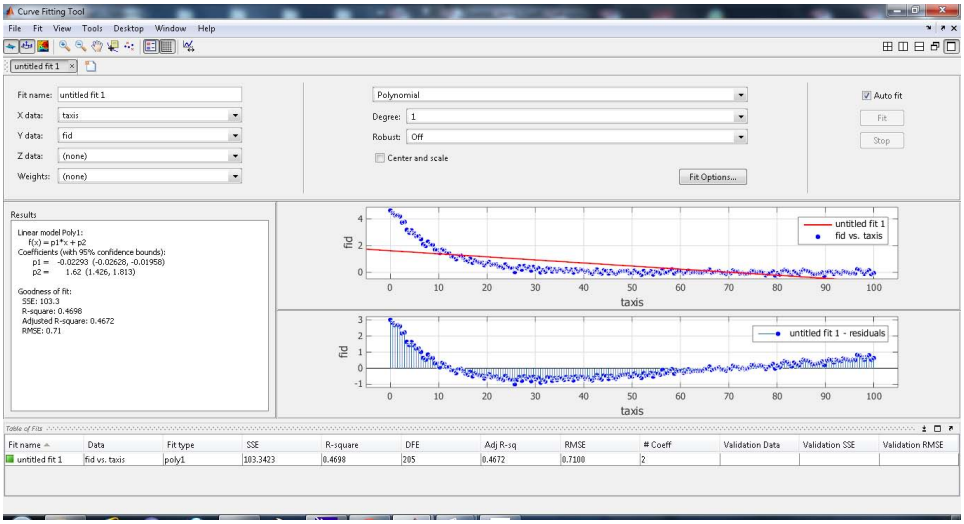
To launch the curve fitting GUI type `cftool`. In the drop-down box "X data" select the dataset "taxis" and in the box "Y data" select "fid". `cftool` immediately attempts to fit a first degree polynomial (straight line):



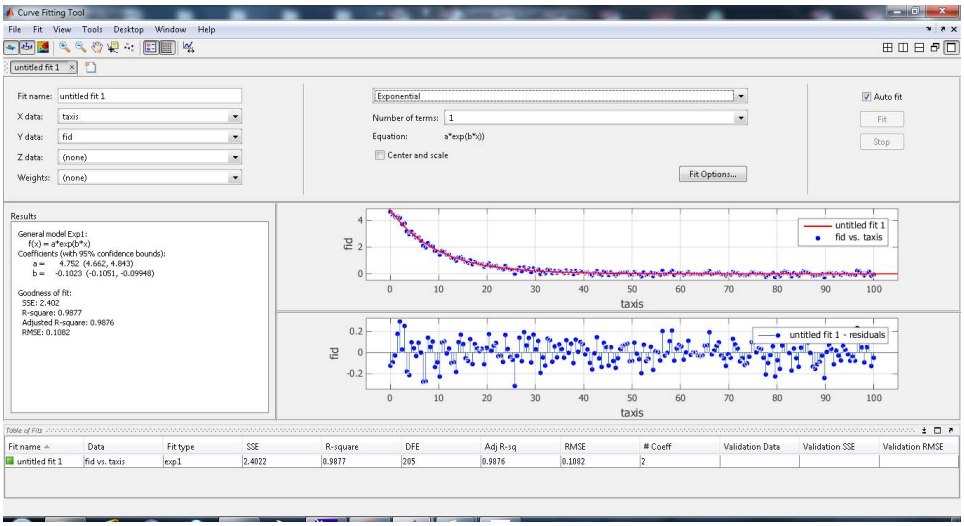
Obviously, the fit is not a very good one, as evidenced by the low R^2 value of 0.4698. The other sign of the bad fit is the very large error bars on the fitted parameters `p1` and `p2`.

Now let's plot the residuals by hitting the "Residuals plot" button of the `cftool` window (second icon in the top left corner of the window). As you can see the residuals indicate that our model inadequately describes this

data.



Let's now select an appropriate model to fit this data. You can select from the drop-down menu "Exponential", which will fit the data to a function of the form `a*exp(b*x)`. The result is much better (R^2 value is 0.9876):

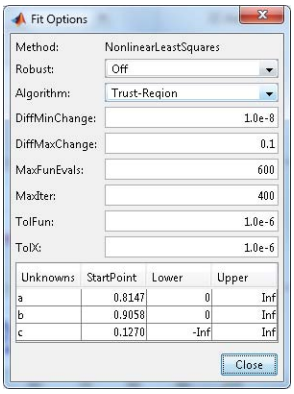


The residuals look fine, i.e. do not show any obvious structure. Let us now check whether or not adding a "baseline" to the fitting function leads to improvement. From the drop-down box, choose "Custom Equation" and in

the box below enter the function `a*exp(-b*x)+c`. On my computer, I get the following error message:[6]

```
Inf computed by model function, fitting cannot continue.
Try using or tightening upper and lower bounds on coefficients.
```

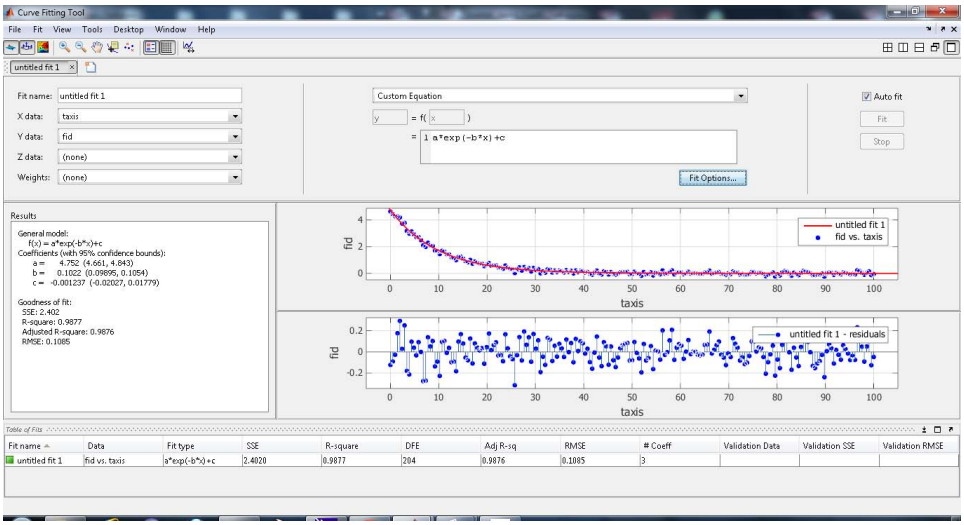
MATLAB is telling us that it cannot fit the data properly - some parameters go to infinity. We need to provide upper and lower bounds on the parameters. Click on the box "Fit Options". You will see a new dialog box which provides several options for the fit. Since we know that the parameters "b" and "a" should be positive numbers, we enter 0 as the lower bound.



You can also check that in the drop-down box for "Algorithm" there are two options: `Trust-Region` and `Levenberg-Marquardt`.

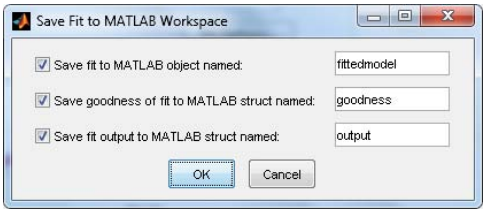
These lower bounds are enough to cause the fit to converge:

[6] It's possible that you don't get this error. It depends on the particular version of MATLAB you are using; the default parameters may differ.



We note that the R^2 value (0.9877) only shows marginal improvement. Thus, the "baseline" parameter was superfluous in this case.

Finally, you can check that the residuals follow a Gaussian distribution. To do this, we will follow several steps. First, choose "Save to Workspace" from the "Fit" menu and click OK:



You can check that MATLAB has created new variables in the workspace. Clicking OK results in:

```
Variables have been created in the current workspace.
```

Then check by typing:

```
whos
```

and see:

```
  Name                Size                Bytes  Class     Attributes

  addnoise          207x1                  3312  double    complex
  fid                 1x207                1656  double
  fittedmodel         1x1                   961  cfit
```

```
  goodness          1x1                920  struct
  multnoise         1x951            15216  double    complex
  output            1x1                8834  struct
  signoise          1x207              3312  double    complex
  taxis             1x207              1656  double
```

One of these variables is the summary of results from the fit:

```
fittedmodel
```

```
fittedmodel =

     General model:
     fittedmodel(x) = a*exp(-b*x)+c
     Coefficients (with 95% confidence bounds):
       a =       4.752  (4.661, 4.843)
       b =      0.1022  (0.09895, 0.1054)
       c =   -0.001237  (-0.02027, 0.01779)
```

The second variable contains values related to the goodness of fit such as R^2:

```
goodness
```

```
goodness =

           sse: 2.4020
       rsquare: 0.9877
           dfe: 204
    adjrsquare: 0.9876
          rmse: 0.1085
```

The third variable is a structure which contains the full output from the fit:

```
output
```

```
output =

           numobs: 207
         numparam: 3
        residuals: [207x1 double]
         Jacobian: [207x3 double]
         exitflag: 3
```

```
firstorderopt: 2.0225e-004
   iterations: 6
    funcCount: 28
 cgiterations: 0
    algorithm: 'trust-region-reflective'
      message: [1x86 char]
```

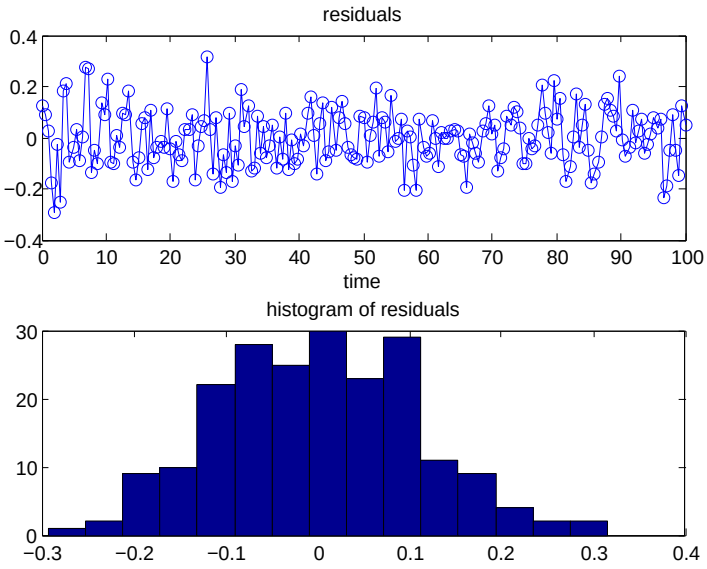
To get the residuals, we use the notation `output.residuals`

```
figure;plot(taxis,output.residuals,'o-');xaxis('time');title('residuals');
```

We can also plot a histogram of the residuals by typing:

```
figure;hist(output.residuals,15);
```

The result is:



11.3. Steepest Descent Algorithm: implementation from scratch

We will now implement the steepest descent algorithm without the help of any built-in MATLAB fitting routines. This implementation could be ported to a low-level programming language such as FORTRAN or C with minimal effort. The update rule for this iterative algorithm is:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \lambda \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)})$$

where λ is the learning rate and $\boldsymbol{\theta}^{(k)}$ is the vector of parameters at the k-th iteration. If λ is too large, the solutions will oscillate, and if it is too small, the algorithm will take too long to converge. χ^2 is defined as the sum of square errors normalized to the square of errors in each of the n data points $\{y_i\}$:

$$\chi^2(\{(x_i, y_i)\}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{(y_i - y(x_i|\boldsymbol{\theta}))^2}{\sigma_i^2}$$

where $y(x_i|\boldsymbol{\theta})$ is our fitting model.

11.3.1. Fit to a straight line. For a straight line model,

$$y(x|A, B) = Ax + B.$$

there are two parameters $\boldsymbol{\theta} = (A, B)^T$. The gradient of χ^2 has only two components:

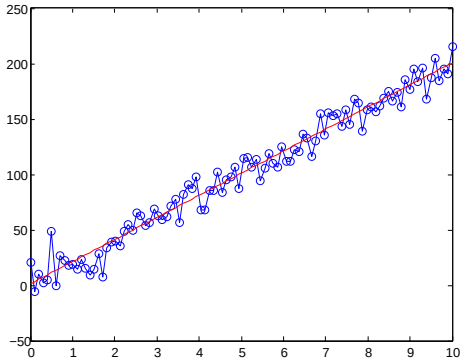
$$(\nabla\chi^2)_A = \frac{\partial\chi^2}{\partial A} = \sum_{i=1}^{n} \frac{-2}{\sigma_i^2}(y_i - Ax_i - B)x_i$$

$$(\nabla\chi^2)_B = \frac{\partial\chi^2}{\partial B} = \sum_{i=1}^{n} \frac{-2}{\sigma_i^2}(y_i - Ax_i - B)$$

Let's create a MATLAB script by typing the command `edit fit3` (`fit3` is an arbitrary filename). Paste the following commands into the script:

```
tdata=linspace(0,10,100);
a=20; b=2;
ydata=a*tdata+b; % generate a fake data set
sigma=10; % with noise added
ydata=ydata+sigma*randn(size(ydata));
figure;plot(tdata,ydata,'o-'); % visualize
theta_k=[30 3]; % starting point
sigmai=sigma*ones(size(ydata)); % assume error is constant for all ...
    measurements
lambda=0.0001; % scaling parameter beta (if too large, solutions ...
    oscillate)
for j=1:777,
    A=theta_k(1); % extract current value of A from the vector as
    B=theta_k(2); % extract value of B
    f=A*tdata+B; % fitting model function
    figure(1); hold off; plot(tdata,ydata,'bo-'); hold on;
    plot(tdata,f,'r'); drawnow;
    dx2da=sum(-(2./(sigmai.^2)).*(ydata-A*tdata-B).*tdata);
    dx2db=sum(-(2./(sigmai.^2)).*(ydata-A*tdata-B));
    grad_chi=[ dx2da dx2db ]; % construct gradient vector
    theta_k = theta_k - lambda*grad_chi; % bs+1=bs+h
    nchi2=sum((1./(sigmai.^2)).*(ydata-A*tdata-B).^2)/length(sigmai);
    fprintf(' iter=%d A=%f B=%f  chi2=%f \r', j, theta_k(1), ...
        theta_k(2), nchi2);
end
```

We obtain the following output



and fitting results:

```
fit3
```

```
 iter=1 A=29.930332 B=2.989497  chi2=37.344001
 iter=2 A=29.861141 B=2.979067  chi2=36.849307
...
 iter=665 A=19.937750 B=1.469166  chi2=1.118404
 iter=666 A=19.937034 B=1.469032  chi2=1.118350
```

In this case, the steepest descent method performs reasonably well.

11.3.2. Nonlinear fit to an exponential decay. Let's consider the following model:

$$y(x|A,B,C) = A\exp(-x/B) + C.$$

The χ^2 function is

$$\chi^2(\{(x_i,y_i)\}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{(y_i - A\exp(-x_i/B) - C)^2}{\sigma_i^2}$$

where $\boldsymbol{\theta} = (A,B,C)^T$.

We can easily compute the gradient of χ^2. It has three components, which are the partial derivatives of χ^2 with respect to each fitting parameter: A,

B and C:

$$(\nabla\chi^2)_A = \frac{\partial\chi^2}{\partial A} = \sum_{i=1}^{n} \frac{-2}{\sigma_i^2}(y_i - A\exp(-x_i/B) - C)\exp(-x_i/B)$$

$$(\nabla\chi^2)_B = \frac{\partial\chi^2}{\partial B} = \sum_{i=1}^{n} \frac{-2}{\sigma_i^2}(y_i - A\exp(-x_i/B) - C)\frac{A}{B^2}\exp(-x_i/B)x_i$$

$$(\nabla\chi^2)_C = \frac{\partial\chi^2}{\partial C} = \sum_{i=1}^{n} \frac{-2}{\sigma_i^2}(y_i - A\exp(-x_i/B) - C)$$

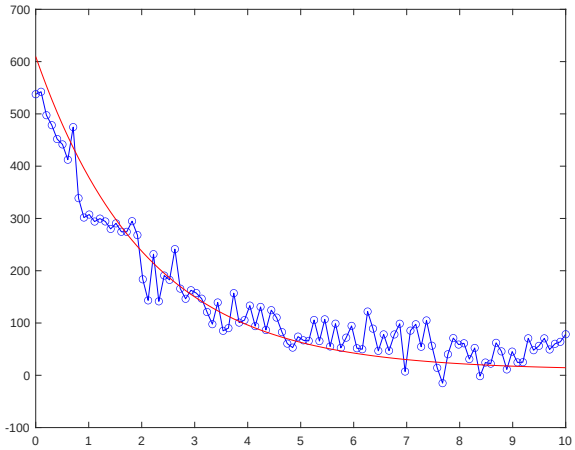
Let's implement this algorithm. First we create a fake data set:

```
tdata=linspace(0,10,100);
a=500; b=2; c=40;
ydata=a*exp(-tdata/b)+c;
sigma=30; % standard dev. of noise
ydata=ydata+sigma*randn(size(ydata));
figure;plot(tdata,ydata,'o-');
```

Create a new MATLAB script file by typing `edit fit2` and copying the above code for generating our 'fake' data set. Then continue with the following commands:

```
theta_k=[600 10 10]; % starting point
sigmai=sigma*ones(size(ydata)); % assume error is constant for all ...
    measurements
lambda=0.0001; % scaling parameter beta (if too large, solutions ...
    oscillate)
for j=1:777,
    A=theta_k(1); % extract current value of A from the vector as
    B=theta_k(2); % extract value of B
    C=theta_k(3); % extract C
    f=A*exp(-tdata/B)+C; % fitting model function
    figure(1); hold off; plot(tdata,ydata,'bo-'); hold on;
    plot(tdata,f,'r');
    dx2da=sum(-(2./(sigmai.^2)).*(ydata-A*exp(-tdata/B)-C) ...
            .*exp(-tdata/B));
    dx2db=sum(-(2./(sigmai.^2)).*(ydata-A*exp(-tdata/B)-C)*(A/B^2) ...
            .*exp(-tdata/B).*tdata);
    dx2dc=sum(-(2./(sigmai.^2)).*(ydata-A*exp(-tdata/B)-C));
    grad_chi=[ dx2da dx2db dx2dc ]; % construct gradient vector
    theta_k = theta_k - lambda*grad_chi; % bs+1=bs+h
    nchi2=sum((1./(sigmai.^2)).*(ydata-A*exp(-tdata/B)-C).^2) ...
            /length(sigmai);
    fprintf(' iter=%d A=%f B=%f C=%f  chi2=%f \r', j, theta_k(1), ...
        theta_k(2), theta_k(3), nchi2);
end
```

After almost 600 iterations we get the following output



and fitting results are:

```
fit2
```

```
iter=1 A=599.996539 B=9.910959 C=9.994502   chi2=72.083371
iter=2 A=599.993108 B=9.821373 C=9.989035   chi2=71.283909
...
iter=554 A=599.794302 B=2.063676 C=9.800685   chi2=1.750119
iter=555 A=599.794191 B=2.063674 C=9.800858   chi2=1.750114
```

The curve is fit, however, the fit results are not as good as with the `FIT` command. This is because steepest descent is not a very good algorithm compared to the one used by `FIT`. To find out which algorithm is used by `FIT` type:

```
help fitoptions
```

```
FITOPTIONS  Create/modify a fit options object.
   F = FITOPTIONS(LIBNAME) creates the fitoptions object F
   with the option parameters set to the default values for the
   library model LIBNAME. See CFLIBHELP for more
   information on LIBNAME.
```

and scroll down to the 'Algorithm' section:

```
Algorithm  - Algorithm to be used in FIT
  [{'Levenberg-Marquardt'} | 'Gauss-Newton' | 'Trust-Region']
```

Thus, the default algorithm is Levenberg-Marquardt which is inherently better than the steepest descent method. Other options are the Gauss-Newton and Trust-Region algorithms. If you'd like to try and compare any of these algorithms, you can read the manual pages for `FIT` and `FITOPTIONS` to find out how to change the default method.

Something you may want to experiment with is trying different initial guesses. You will see that if the initial guess is too far from the actual values, this algorithm will fail miserably: in some cases it will get stuck in a local minimum and in other cases the solution may diverge. The second parameter you may want to experiment with is the value of η. Try a value 100 times larger or 100 times smaller.

11.4. Marquardt-Levenberg Algorithm

Marquardt-Levenberg update rule:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - (\mathbf{H}_k + \lambda \text{diag}[\mathbf{H}_k])^{-1} \nabla \chi^2(\boldsymbol{\theta}^{(k)})$$

$\mathbf{H}_k$ is proportional to the curvature of χ^2, i.e. large steps are made in the direction of low curvature (flat terrain) and small steps in the direction with high curvature (steep incline). $\mathbf{H}_k$ is called the curvature matrix.

The parameter λ is adjusted at each iteration. We stop iterating when χ^2 does not change appreciably. Here is a possible implementation of the Levenberg-Marquardt method:

- Pick initial ($k = 1$) guess for set of fitted parameters $\boldsymbol{\theta}^{(k)}$.
- Compute $\chi^2(\boldsymbol{\theta}^{(k)})$.
- Pick a modest value for λ, say $\lambda = 0.001$.
- (*) Solve for $\delta\boldsymbol{\theta}^{(k)} = (\mathbf{H}_k + \lambda \text{diag}[\mathbf{H}_k])^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)})$ and evaluate $\chi^2(\boldsymbol{\theta}^{(k)} + \delta\boldsymbol{\theta}^{(k)})$.
- If $\chi^2(\boldsymbol{\theta}^{(k)} + \delta\boldsymbol{\theta}^{(k)}) \geq \chi^2(\boldsymbol{\theta}^{(k)})$, we increase λ by a factor of 10, set $k = k+1$ and go back to (*).
- If $\chi^2(\boldsymbol{\theta}^{(k)} + \delta\boldsymbol{\theta}^{(k)}) < \chi^2(\boldsymbol{\theta}^{(k)})$, we decrease λ by a factor of 10, update the trial solution $\boldsymbol{\theta}^{(k)} \rightarrow \boldsymbol{\theta}^{(k)} + \delta\boldsymbol{\theta}^{(k)}$, set $k = k + 1$ and go back to (*).
- Stopping criterion: changes in parameters that yield changes in χ^2 by an amount $\ll 1$ are not statistically meaningful.
- When finished, use the Hessian to compute the estimated covariance matrix ($cov(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}) = 2\mathbf{H}_k^{-1}$). The latter allows you to obtain errors in the fitted parameters or correlations among them.

11.4.1. Computer Implementation: Straight Line Fit. The update rule is:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - (\mathbf{H}_k + \lambda \text{diag}[\mathbf{H}_k])^{-1} \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)})$$

where

$$\mathbf{H}_k \equiv \mathbf{H}(\boldsymbol{\theta}^{(k)}) = \begin{bmatrix} \frac{\partial^2 \chi^2}{\partial \theta_1^2} & \cdots & \frac{\partial^2 \chi^2}{\partial \theta_1 \partial \theta_p} \\ \vdots & & \vdots \\ \frac{\partial^2 \chi^2}{\partial \theta_p \partial \theta_1} & \cdots & \frac{\partial^2 \chi^2}{\partial \theta_p^2} \end{bmatrix}$$

We have already seen how to compute the gradient vector $\nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}^{(k)}) = (\frac{\partial \chi^2}{\partial \theta_1}, \ldots, \frac{\partial \chi^2}{\partial \theta_p})^T$.

In the case of the linear model,

$$y(x|\boldsymbol{\theta}) = Ax + B.$$

The gradient of χ^2

$$\chi^2(\{(x_i, y_i)\}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{(y_i - Ax_i - B)^2}{\sigma_i^2}$$

has the two components we previously found:

$$(\nabla \chi^2)_A = \frac{\partial \chi^2}{\partial A} = \sum_{i=1}^{n} \frac{-2}{\sigma_i^2}(y_i - Ax_i - B)x_i$$

$$(\nabla \chi^2)_B = \frac{\partial \chi^2}{\partial B} = \sum_{i=1}^{n} \frac{-2}{\sigma_i^2}(y_i - Ax_i - B)$$

whereas the Hessian matrix has 4 components:

$$\frac{\partial^2 \chi^2}{\partial A^2} = \sum_{i=1}^{n} \frac{2x_i^2}{\sigma_i^2} \qquad \frac{\partial^2 \chi^2}{\partial A \partial B} = \sum_{i=1}^{n} \frac{2x_i}{\sigma_i^2}$$

$$\frac{\partial^2 \chi^2}{\partial B \partial A} = \frac{2x_i}{\sigma_i^2} \qquad \frac{\partial^2 \chi^2}{\partial B^2} = \sum_{i=1}^{n} \frac{2}{\sigma_i^2}$$

The matrix $\mathbf{H}_k$ and its diagonal are:

$$\mathbf{H}_k = \begin{bmatrix} \sum_{i=1}^{n} \frac{2x_i^2}{\sigma_i^2} & \sum_{i=1}^{n} \frac{2x_i}{\sigma_i^2} \\ \frac{2x_i}{\sigma_i^2} & \sum_{i=1}^{n} \frac{2}{\sigma_i^2} \end{bmatrix}, \qquad \text{diag}[\mathbf{H}_k] = \begin{bmatrix} \sum_{i=1}^{n} \frac{2x_i^2}{\sigma_i^2} & 0 \\ 0 & \sum_{i=1}^{n} \frac{2}{\sigma_i^2} \end{bmatrix}.$$

The inverse of a 2×2 matrix is easily obtained from the formula

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

For larger matrices, there are formulas to compute them. In the implementation of the algorithm on a computer, however, you will often resort to numerical techniques to obtain matrix inverses. In MATLAB you can

use the INV command. There is also the pseudo-inverse PINV which works better when matrices are nearly singular.

For example, try

```
>> A=[1 2 ; 3 4]
```

```
A =

     1     2
     3     4
```

```
inv(A)
```

```
ans =

   -2.0000    1.0000
    1.5000   -0.5000
```

and compare with the above formula. You should also check that this matrix inverse satisfies the definition of the inverse of a matrix: $A^{-1}A = \mathbf{1} = AA^{-1}$:

```
inv(A)*A
```

```
ans =

    1.0000         0
    0.0000    1.0000
```

```
A*inv(A)
```

```
ans =

    1.0000         0
    0.0000    1.0000
```

See Problem 137 for a complete implementation.

11.5. Curve Fitting by Simulated Annealing

Here is a possible MATLAB implementation of the Metropolis Monte-Carlo algorithm for curve fitting in the case of our previous example of an exponential decay function with noise.

```
% FIT_EXP_MC - Metropolis Monte-Carlo implementation of the curve
% fitting method

% create a 'fake' data set
tdata=linspace(0,10,100);
a=500; b=2; c=40;
% let's do an exponential decay
ydata=a*exp(-tdata/b)+c;
% and then add some gaussian random noise
sigma=30; % standard dev. of noise
ydata=ydata+sigma*randn(size(ydata));
figure;plot(tdata,ydata,'o-'); % visualize

theta_k=[1 1 1]; % starting point, can pick randomly or not
% theta_k=50*rand([1 3]); % randomly pick starting point
sigmai=sigma*ones(size(ydata)); % assume error is constant for all
                                % measurements
T=1; % start with high temperature
for j=1:9999,
    A=theta_k(1); % extract current value of A from the vector as
    B=theta_k(2); % extract value of B
    C=theta_k(3); % extract C
    f=A*exp(-tdata/B)+C; % fitting model function
    figure(1); hold off; plot(tdata,ydata,'bo-'); hold on;
    plot(tdata,f,'r'); drawnow;
    chi2=sum((1./(sigmai.^2)).*(ydata-A*exp(-tdata/B)-C).^2) ...
             /length(sigmai);
    Δtheta=0.05*randn([1 3]); % random displacement vector
    theta_kp1 = theta_k + theta_k.*Δtheta; % make random change in ...
        parameters
    An=theta_kp1(1); Bn=theta_kp1(2); Cn=theta_kp1(3);
    chi2n=sum((1./(sigmai.^2)).*(ydata-An*exp(-tdata/Bn)-Cn).^2)/ ...
          length(sigmai);
    ΔE=chi2n-chi2;
    if (ΔE < 0), % accept the move
        theta_k=theta_kp1;
    else, % else reject the move with probability P
        R=rand; P=exp(-ΔE/T);
        if (P > R),
            theta_k=theta_kp1;
        end;
    end;
    T=T-0.01*T; % cooling schedule
    fprintf(' iter=%d A=%f B=%f C=%f  chi2=%f ΔE=%f  T=%f \r', ...
             j, theta_k(1), theta_k(2), theta_k(3), chi2, ΔE, T);
end
```

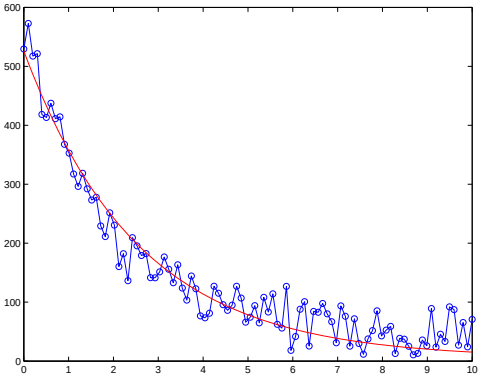
A sample run proceeds as follows:

```
fit_exp_mc
```

```
iter=1 A=332.450122 B=9.823360 C=9.608963  chi2=11.692912
 deltaE=1.817395  T=0.990000
```

```
iter=2 A=332.450122 B=9.823360 C=9.608963  chi2=13.510307
 deltaE=2.439683  T=0.980100
iter=3 A=317.456609 B=9.534128 C=9.311598  chi2=13.510307
 deltaE=-1.658223  T=0.970299
iter=4 A=303.398269 B=9.540513 C=9.180057  chi2=11.852084
 deltaE=-0.779770  T=0.960596

.
.
.

iter=665 A=521.159989 B=2.552933 C=4.891518  chi2=1.126674
 deltaE=0.202851  T=0.001251
iter=666 A=521.159989 B=2.552933 C=4.891518  chi2=1.126674
 deltaE=0.158424  T=0.001239
```

We have limited the number of iterations. However, in real implementations, a better criterion can be used, monitoring the value of χ^2 to see if it changes significantly or not. I get something like this:



You should also try different initial guesses and check that the method is less sensitive to the choice of initial conditions as compared to the deterministic algorithms. It is easy to check, for example, that the steepest descent method does not work for the following choice of initial parameters, `[1 ... 1 1]`, because these parameters are too far from their true values.

11.6. Curve Fitting by Genetic Algorithm

The code below uses the `ga` command in MATLAB to fit experimental data to a model. Notice how the particular example below uses the l_1 norm instead of the l_2 norm (χ^2). Recall the definition (Eq. 3.3) of l_p norm of a

vector $\vec{x}$:

$$\|\vec{x}\|_p = \left(|x_1|^p + \cdots + |x_n|^p\right)^{1/p}.$$

In data fitting problems the length of the vector $\vec{x}$ equals the number of data points. When we have a norm, we can define a distance function as:

$$d(\vec{x}, \vec{y}) = \|\vec{x} - \vec{y}\|_p.$$

With `ga` or simulated annealing there is no need to limit our discussion to χ^2, because we do not need to compute derivatives of the cost function, as was the case for least squares methods. Global optimization algorithms only need to compute the value of the cost function, not its derivatives. This means that any suitable distance function (distance between model and data) may be used.

```
xx=linspace(0,10,100);
yy=50*(1-2*exp(-0.5*xx));
yy=yy+10*randn(size(yy));
figure;
axes;
plot(xx, yy, 'b+');

fun = @(p) p(1)*(1-p(2)*exp(-p(3)*xx));
objFun = @(p) norm(fun(p)-yy,1);

options = optimoptions('ga','ConstraintTolerance', ...
    1e-6,'PlotFcn', @gaplotbestf,'MaxStallGenerations', ...
    500,'MaxGenerations',1000);

sol = ga(objFun, 3,[],[],[],[],[],[],[],options);

figure;
axes;
plot(xx, yy, 'b+');
hold on
plot(xx, fun(sol), 'r-');
legend({'Data points', 'Fitted Curve'});
```

11.7. Extrema Search by Newton Raphson Method

See Problem 137.

11.8. Extrema Search by Simulated Annealing

Can you explain what the code below does and how it works? What search (what type of extrema) is it performing? See also Problem 137.

```
clear
syms x
f=@(x) sin(x*0.01*pi)+0.25*sin(x*0.01*pi*5)+0.15*sin(x*0.01*pi*12);
```

```
4
5  figure(1); ezplot(f,[60 160]);
6
7  x=70; % start point
8  stepsize=5; % std dev
9  % measurements
10 T=1; % start with high temperature
11 for j=1:600,
12     xn=x+stepsize*randn(size(x));
13
14     % keep x within bounds!
15     if (xn > 160), xn=160; end;
16     if (xn < 60), xn=60; end;
17
18     figure(2); clf; ezplot(f,[60 160]); hold on;
19     plot(x,f(x),'ro'); drawnow; pause(0.1); % slow down!
20
21     DeltaE=f(xn)-f(x); % DeltaE = final - initial
22     if (DeltaE < 0), x=xn; % accept move
23     else, % else reject the move with probability P
24       R=rand; P=exp(-DeltaE/T);
25       if (P > R), x=xn; end;
26     end;
27     T=T-0.01*T; % cooling schedule
28     fprintf([' iter=%d f=%f DeltaE=%f T=%f \r'],j,f(x),DeltaE,T);
29 end;
```

11.9. Problems

Problem 134. Create a fake data set to simulate a linear relationship $y(x) = A + Bx$ between two random variables X and Y (plus noise). Use the analytical formulae (Eqs. 5.1 and 5.2) we obtained in Chapter 5 for the coefficients A and B to compute their values.

Solution. Here is MATLAB code that will do the job:

```
1 A=10; B=5; n=10000;
2 x=linspace(0,1,n);
3 y=A+B*x + 1.0*randn(size(x));
4 figure; plot(x,y,'o');
5 % compute slope and intercept (least squares, linear regression)
6 D=n*sum(x.^2)-(sum(x))^2;
7 A=(1/D)*(sum(x.^2)*sum(y)-sum(x)*sum(x.*y))
8 B=(1/D)*(n*sum(x.*y)-sum(x)*sum(y))
```

The output is as follows:

```
A =
```

```
    9.9941

B =

    5.0030
```

The intercept (9.9941) is fairly close to the true value (10). The slope (5.0030) is close to the actual one (5). ■

Problem 135. Create a fake data set for an exponential decay plus baseline, plus additive Gaussian noise:

```
tdata=linspace(0,10,100);
a=500; b=2; c=40;
ydata=a*exp(-tdata/b)+c;
sigma=30; % standard dev. of noise
ydata=ydata+sigma*randn(size(ydata));
figure;plot(tdata,ydata,'o-');
```

(Addition of the noise with the `randn` command will introduce errors in the fitted parameters which you must determine.) Choose the model to be:

$$y(x) = A\exp(-x/B) + C$$

Write a MATLAB program to implement the Gauss-Newton method (where the Hessian matrix is approximated by $2\mathbf{J}_k^T\mathbf{J}_k$, where $\mathbf{J}_k$ is the Jacobian matrix at the k-th iteration). Compare its performance to the Newton method. Does it converge faster or slower, how accurate are the results, and what do the errors in the fitting parameters compare? Errors in the fitting parameters are obtained as the square root of the diagonal elements of the covariance matrix, $cov(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}) = 2\mathbf{H}_k^{-1}$, where $\mathbf{H}_k$ is the Hessian matrix of $\chi^2(\boldsymbol{\theta}^{(k)})$ and $\mathbf{H}_k^{-1}$ is its inverse (see Sections 8.7.1 and 6.18).

Repeat the same exercise, except write MATLAB code to implement the Levenberg method. Compare method performance to Newton and Gauss-Newton methods.

Repeat the same exercise, except write MATLAB code to implement the Marquardt-Levenberg method. Compare method performance to Newton, Gauss-Newton and Levenberg methods.

Solution. These problems do not have unique solutions (everyone's code can be different). However, the calculations for Jacobian matrix and other

quantities are straightforward and follow procedures similar to what we did in class at the science learning center. ■

Problem 136. Write your own MATLAB code (as simple as possible) to implement the Newton-Raphson method as a way:

(a) to find the zeros of the function

$$f(x) = x^3 - 9x^2 + 2x + 8 \qquad \text{on the interval } [-2, 9].$$

Does the ability to find zeros depend on choice of initial condition? What algorithm do you think the MATLAB command `fzero` uses by default? Compare the performance of your code against the output of the `fzero` command.

(b) use Newton Raphson (and comparing with `fzero`) to find the point x where $\exp(-\exp(-x)) = x$.

Solution. (a) Here is my code for the `fzero` command:

```
1 t=linspace(-2,9,100);
2 for j=1:length(t),
3     y(j) = fzero( @(x) (x^3-9*x^2+2*x+8), t(j) );
4 end;
```

The idea here is that `fzero` does a local search (see documentation: `help fzero`) and requires a starting point for x. It will then find a nearby zero. If there are more than 1 zeros, we must loop this starting point over the domain of x. In the above code the vector `y` contains a list of zeros that were found. In the case of the above function, there are 3 zeros over the domain $[-2, 9]$: -0.8070, 1.1444 and 8.6625. (The algorithm finds the zeros within some finite precision; hence the obtained numerical values in `y` corresponding to the same zero may differ beyond some significant figure.). The algorithm used is discussed in the MATLAB documentation

`https://www.mathworks.com/help/matlab/ref/fzero.html`

```
Algorithms
The fzero command is a function file. The algorithm,
created by T. Dekker, uses a combination of bisection,
secant, and inverse quadratic interpolation methods.
An Algol 60 version, with some improvements,
is given in [1]. A Fortran version, upon which fzero
is based, is in [2].
```

Here is my code for the Newton-Raphson method:

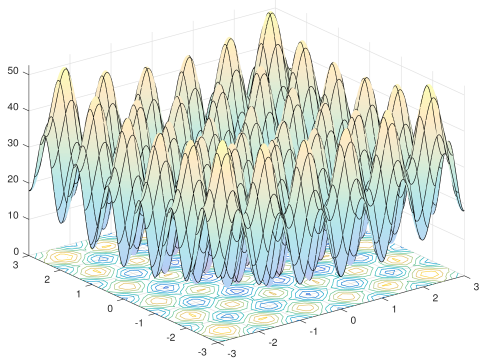
```
clear
syms x

f=@(x) x.^3-9*x.^2+2*x+8;

fp=matlabFunction(diff(f,x)); % f'(x)
fpp=matlabFunction(diff(fp,x)); % f''(x) only needed for newton method

figure(1);
subplot(3,1,1),ezplot(f,[-2 9]);
subplot(3,1,2),ezplot(fp,[-2 9]);
subplot(3,1,3),ezplot(fpp,[-2 9]);

xn=108; % start point
for j=1:50,
  figure(2); clf; e=ezplot(f,[-2 9]); hold on; plot(xn,f(xn),'ro');
  drawnow; pause(0.1);
% xn=xn-0.1*(fp(xn)/fpp(xn)); % newton method (find extrema)
  xn=xn-0.1*(f(xn)/fp(xn)); % newton-raphson method  (find zeros)
  if (xn>9), xn=9; end;
  if (xn<-2), xn=-2; end;
end;
xn
```

(b) Left as an exercise. (The above code can be modified accordingly.) ■

Problem 137. The following function of two independent variables (x_1, x_2),

$$f(x_1, x_2) = 20 + x_1^2 + x_2^2 - 10(\cos 2\pi x_1 + \cos 2\pi x_2),$$

is plotted below:



As the plot shows, this function has many local minima. However, the function has just one global minimum, which occurs at the point [0 0] in the x_1-x_2 plane, as indicated by the vertical line in the plot, where the value of the function is 0. At any local minimum other than [0 0], the value of the function is greater than 0. The farther the local minimum is from the origin, the larger the value of the function is at that point. As you can imagine, because this function has many local minima, it is difficult for standard, gradient-based methods to find the global minimum.

(a) Write MATLAB code to execute a gradient-based search (steepest descent) and show that the algorithm can at best find a local minimum near the starting point. Choose a starting point far from the global minimum, i.e. $|x_1| > 3$ and $|x_2| > 3$. On the other hand, show that if you pick a starting point near the origin (but not *at* the origin), the gradient search can find the global minimum. How far away from the origin can you pick the starting point, such that the search will still find the global minimum (i.e. such that beyond this radius, only a local minimum can be found)? In your algorithm, plot the function as a surface or contour plot, and indicate the position of the gradient-based search as function of time, so we can visualize the progress of the gradient search.

(b) Write simulated annealing code to search for the global minimum and show that the global minimum can always be located, regardless of the starting point.

Solution. (a) Here you are free to implement steepest descent, Newton, Gauss-Newton or Levenberg-Marquardt. We provide 2 examples; as you can see the code is fairly simple. For the Newton method:

```
1  % Newton method in 2D
2  syms x1 x2
3  f = @(x1,x2) 20+x1.^2+x2.^2-10*(cos(2*pi*x1)+cos(2*pi*x2)); % f
4  G=matlabFunction(gradient(f,[x1,x2])); % gradient of f
5  H=matlabFunction(hessian(f,[x1,x2])); % hessian of f
6
7  x=1*randn([2 1]); % initial position (x1,x2)
8  for j=1:50,
9      x=x-pinv(H(x(1),x(2)))*G(x(1),x(2)); % Newton update rule
10     figure(1); clf; ezcontour(f,[-4,4]); hold on;
11     plot(x(1),x(2),'ro'); drawnow; pause(0.5);
12     fprintf(' iter=%d  f=%f x1=%f x2=%f \n',...
13         j,f(x(1),x(2)),x(1),x(2));
14 end;
```

Note: in line 7, initial conditions are chosen randomly (`randn`). This can cause problems if the initial conditions are close to the edge of the figure (the domain of display is `[-4 4]` in line 10). It's possible that the 'red dot' may disappear from the screen. Feel free to experiment with manual settings for the initial conditions, i.e. replace line 7 with deterministic conditions such as `x=[0.2 0.3];`

For the Levenberg-Marquardt method:

```
1  % Marquardt-Levenberg method in 2D
2  syms x1 x2
```

```
3  f = @(x1,x2) 20+x1.^2+x2.^2-10*(cos(2*pi*x1)+cos(2*pi*x2)); % f
4  G=matlabFunction(gradient(f,[x1,x2])); % gradient of f
5  H=matlabFunction(hessian(f,[x1,x2])); % hessian of f
6
7  x=1*randn([2 1]); % initial position (x1,x2)
8  lambda=0.001;
9  for j=1:50,
10     hess=H(x(1),x(2));
11     f_before=f(x(1),x(2));
12     xn=x-pinv(hess+lambda*diag(diag(hess)))*G(x(1),x(2)); % update rule
13     f_after=f(xn(1),xn(2));
14     figure(1); clf; ezcontour(f,[-4,4]); hold on;
15     plot(x(1),x(2),'ro'); drawnow; pause(0.5);
16     fprintf(' iter=%d  f=%f x1=%f x2=%f lambda=%f \n',...
17         j,f(x(1),x(2)),x(1),x(2),lambda);
18     if (f_after ≥ f_before), lambda=lambda*10;
19     else lambda=lambda/10; x=xn;
20     end;
21 end;
```

Note: the MATLAB command `diag` is invoked twice in a row (see line 12), i.e. `diag(diag(hess))`. The first use of `diag` extracts the diagonal elements of the matrix `hess` and returns a vector. The second instance of `diag` takes this vector and creates a matrix with all zeros everywhere except along the diagonal. The net result is the creation of a matrix whose diagonal is equal to the diagonal of `hess` and zero values off-diagonal.

(b) The following code converges in less than 4,000 steps. Initial temperature was `T=12`. Step size was `0.4`. Cooling schedule was set to `-0.1%` per step.

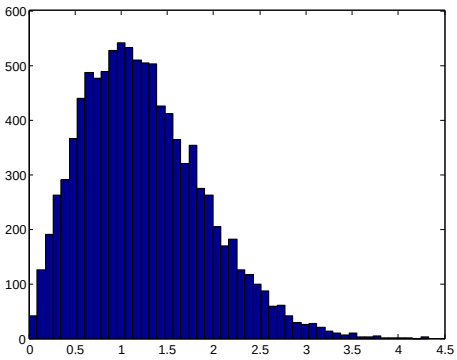
```
1  syms x1 x2
2  f = @(x1,x2) 20+x1.^2+x2.^2-10*(cos(2*pi*x1)+cos(2*pi*x2)); % f
3  x=1*randn([2 1]); % initial position (x1,x2)
4  T=12; % starting temperature
5  for j=1:20000,
6      figure(1); clf; ezcontour(f,[-4,4]); hold on;
7      plot(x(1),x(2),'ro'); drawnow; %pause(0.1);
8      dx=0.4*randn(size(x)); % make random move
9      f_before=f(x(1),x(2));
10     f_after=f(x(1)+dx(1),x(2)+dx(2));
11     dE=f_after-f_before; % change in energy
12     if (dE < 0), x=x+dx; % accept the move
13     else, % else reject move with probability P
14      R=rand; P=exp(-dE/T);
15      if (P > R), x=x+dx; end;
16     end;
17     T=T-0.001*T; % cooling schedule
18     fprintf(' iter=%d x1=%f x2=%f E=%f dE=%f T=%f \n', ...
19         j,x(1),x(2),f(x(1),x(2)),dE,T);
20 end
```

■

Problem 138. Let X and Y be independent, identically distributed normal (Gaussian) random variables with mean 0 and variance 1. Write a computer program to generate random values for X and Y. (Generate a very large number of pairs X, Y.) For each pair (X, Y), compute the radius $R = \sqrt{X^2 + Y^2}$. Plot the probability density (i.e., histogram) for R. Obtain the resulting distribution.

Solution. R follows a Rayleigh distribution. Also known as "Rice" or "Rician" distribution. Only 4 lines of code are needed to plot this distribution:

```
1  X=randn([1 10000]);
2  Y=randn([1 10000]);
3  R=sqrt(X.^2+Y.^2);
4  figure;hist(R,50);
```



For those interested in a formal proof: consider the two-dimensional vector (x, y) which has components that are Gaussian distributed, that is, $p(x) = \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$, and the same for $p(y)$. Denote with r the length $r = \sqrt{x^2 + y^2}$. It is distributed as

$$p(r|\sigma) = \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y \, e^{-x^2/2\sigma^2} e^{-y^2/2\sigma^2} \delta(r - \sqrt{x^2 + y^2}).$$

(The Dirac delta function is used to enforce the constraint that r must equal the length of the vector (x, y).) By transforming to polar coordinates one has

$$p(r|\sigma) = \frac{1}{2\pi\sigma^2} \int_0^{2\pi} \mathrm{d}\phi \int_0^{\infty} \mathrm{d}\rho \; \delta(r - \rho) \rho e^{-\rho^2/2\sigma^2} = \frac{r}{\sigma^2} e^{-r^2/2\sigma^2}$$

which is the Rayleigh distribution. ■

Problem 139. Write a MATLAB program to fit an exponentially decaying signal. The signal (`ydata`) can be simulated using the following sequence of commands (the code below generates the signal and adds noise):

```
tdata=linspace(0,10,100);
a=500; b=2; c=40;
ydata=a*exp(-tdata/b)+c;
sigma=30; % standard dev. of noise
ydata=ydata+sigma*randn(size(ydata));
figure;plot(tdata,ydata,'o-');
```

You are allowed to use the following built-in MATLAB commands to fit the data: `fit`, `fminsearch`, `fminbnd`, `fminunc`, `fmincon`, `lsqnonlin`, ... `lsqcurvefit`. You are not allowed to use graphical user interfaces such as `cftool` - although it might be a useful exercise to experiment with it, as a way to double-check the performance of your method.

Problem 140. A number x is called a zero of a function f if and only if $f(x) = 0$. One way to find zeros is to sketch the graph of the function. For example, consider the function $f(x) = x^3 - 4x^2 - 11x + 30$. The following code could be used for this purpose:

```
f=@(x) x.^3-4*x.^2-11*x+30;
x=linspace(-10,10,200);
y=f(x);
plot(x,y);
axis([-10,10,-100,100]);
grid on;
```

As you can see, the zeros are evident from the plot. You can also double-check that the zeros are indeed zeros. Suppose that the zeros are $-3, 2, 5$. Then type:

```
x=[-3,2,5];
f(x)
```

(a) In MATLAB check that $x = -1$ and $x = 3$ are zeros of the function $f(x) = x^2 - 2x - 3$ by plotting the function and identifying the zero crossings. (Plotting can be done as explained above, or an even easier way is to use the command `ezplot`.)

(b) Find the zeros of $f(x) = x^4 - 29x^2 - 132$.

(c) Use the MATLAB command `fzero` to find the zeros for the functions in (a) and (b). Does `fzero` give you a more accurate result than the graphical method?

(d) Use the command `fzero` to solve the following equation $5-2x = e^{-0.25x}$ for x.

(e) Use the `fzero` command to find the zeros of the functions below on the given domain:

$$f(x) = 9 - 4x - x^2 \text{ on } [-7, 3]$$
$$f(x) = 2x^2 - x - 8 \text{ on } [-3, 4]$$
$$f(x) = x^3 - 9x^2 + 2x + 8 \text{ on } [-2, 9]$$

(You will need to read the instructions on how to use the `fzero` command on an interval.)

Problem 141. We previously found that for an unrestricted random walk, the statistics of the random walker after n steps (histogram of S_n) follow a Gaussian distribution whose variance scales with n. Write two MATLAB codes to simulate 1) a random walk with and 2) without boundary. For boundary: Place the boundary at some distance from the origin, but close enough, so that the particle hits the boundary. The boundary is reflective: when the particle hits the boundary, instead of crossing it, it is reflected back to the direction of origin. Plot the statistics of this random walk. The resulting distribution is clearly non-Gaussian. What distribution do you find?

Problem 142. Explain what this MATLAB code does. (`pinv` is nearly the same as `inv`.) What type of problem is it solving?

```
syms x1 x2
f = @(x1,x2) 20+x1.^2+x2.^2-10*(cos(2*pi*x1)+cos(2*pi*x2)); % f
G=matlabFunction(gradient(f,[x1,x2])); % gradient of f
H=matlabFunction(hessian(f,[x1,x2])); % hessian of f
x=1*randn([2 1]); % initial position (x1,x2)
for j=1:50, % # of iterations
    x=x-pinv(H(x(1),x(2)))*G(x(1),x(2));
    figure(1); clf; ezcontour(f,[-4,4]); hold on; %plot
    plot(x(1),x(2),'ro'); drawnow;
    fprintf(' iter=%d  f=%f x1=%f x2=%f \n',...
        j,f(x(1),x(2)),x(1),x(2));
end;
```

Solution. This code implements the Newton method to iteratively find a local extremum of the function f. ■

In the three problems below you will need a data set to work with. Choose a data set to analyze, either: 1) by finding data on the internet (a search for 'free public data sets' will turn up many results), 2) pick a data set from a lab experiment if you have such data sets or 3) create a fake data set

$\{(x_i, y_i)\}$ as we did previously, but make sure there is enough noise contamination so that the fitting task is somewhat challenging. It may be easier to use the same data set and models for both local and global optimization problems.

Problem 143. Write MATLAB code to implement the Gauss-Newton method. Pick a suitable model $\{y(x_i|\boldsymbol{\theta})\}$ with at least 3 parameters (e.g. no straight lines fit please). The model must be nonlinear in the fitting parameters to justify the use of this nonlinear least squares method. Derive the necessary equations for the Jacobian matrix entries. Program the resulting expressions into your MATLAB code. Every line of code must be justified (insert comments with % as needed). Do not use a `for` loop to loop over iterations as we have done in class. Instead, use a `while` loop (see MATLAB documentation for the `while` statement). Loop over iterations until the fractional change in χ^2 relative to its weighted average for the last 10 iterations changes by an amount less than 10^{-3}, i.e.

```
while abs((chi2-chi2last5)/chi2last5) > 1e-3,
      ...
end
```

where `chi2` is the last value of `chi2` and `chi2last5` is the average value of `chi2` for the previous 5 iterations. If the value 10^{-3} is too restrictive and convergence is too slow, relax this requirement and explain why you had to do so.

Problem 144. Fit a data set to an appropriate model using a global search algorithm. The model should have at least 3 fitting parameters. You may impose reasonable constraints on the search, such as positivity of certain coefficients (if applicable). However, if you choose to enforce constraints, make sure they are grounded in sound physical arguments (as opposed to designing constraints that force the algorithm to find a pre-determined solution). You may use the MATLAB global optimization toolbox and any of the algorithms provided in it. One such algorithm is the Genetic Algorithm (command `ga`); there are others too.

Problem 145. Local optimization algorithms stop at the nearest extremum, whereas global optimization algorithms can find the global optimum. Illustrate this in the context of Problems 143 and 144 (or in a separate demonstration) by showing that global optimization finds the global extremum regardless of initial conditions whereas the solutions found by local searches strongly depend on the choice of initial conditions. Try at least 5 wildly

different choice of initial conditions for the fitting parameters and record how many iterations are needed to reach a solution. Compare the case of local vs global optimization. If you use MATLAB's `ga` command the initial parameters are called `Initial Population`. See:

`https://mathworks.com/help/gads/genetic-algorithm-options.html`

In earlier versions of MATLAB one sets the `InitialPopulationMatrix`

`https://mathworks.com/help/gads/gaoptimset.html`

`https://mathworks.com/help/optim/ug/`

`optim.problemdef.optimizationproblem.optimoptions.html`

Chapter 12

Review of Math Concepts

12.1. Solving Systems of 2 Equations and 2 Unknowns

To solve for N unknowns in N equations we can use the matrix inverse. The formula for the inverse of a 2×2 matrix is:

$$\boxed{\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.}$$

The cross product $ad - bc$ is the determinant of the 2×2 matrix. Students unfamiliar with matrix inverses should check that the inverse of this 2×2 matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, denoted A^{-1}, satisfies the following conditions: $AA^{-1} = I$ and $A^{-1}A = I$. Here, I is the 2×2 identity matrix, $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

In the Section 5.1.2 we needed to solve:

$$An + B \sum x_i = \sum y_i$$
$$A \sum x_i + B \sum x_i^2 = \sum x_i y_i.$$

This is done by rewriting it in matrix form:

$$\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}.$$

Multiplying both sides on the left by the inverse of $\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$, we solve for A, B:

$$\begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}. \tag{12.1}$$

Here, the inverse of that matrix is:

$$\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} = \frac{1}{\Delta} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}. \tag{12.2}$$

Substitution of (12.2) into (12.1) gives the final result:

$$\boxed{\begin{pmatrix} A \\ B \end{pmatrix} = \frac{1}{\Delta} \begin{pmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ n \sum x_i y_i - \sum x_i \sum y_i \end{pmatrix}, \quad \Delta = n \sum x_i^2 - (\sum x_i)^2.}$$

12.2. Changing Variables Under the Integral Sign

We often need to change variables under the integral sign to express the probabilistic integral in terms of *known* probability densities. At the same time, we should use a convenient coordinate system where the integrals can be computed.

Suppose that we want to integrate $f(u, v)$ over a region R. Under the inverse of the transformation $u = u(x, y)$, $v = v(x, y)$ the region R becomes S and the double integral becomes

$$\boxed{\iint_R f(u, v)\, \mathrm{d}u\, \mathrm{d}v = \iint_S f(u(x, y), v(x, y)) \left| \frac{\partial(u, v)}{\partial(x, y)} \right| \mathrm{d}x\, \mathrm{d}y,}$$

where $\frac{\partial(u,v)}{\partial(x,y)}$ is the Jacobian determinant:

$$\frac{\partial(u, v)}{\partial(x, y)} \equiv \begin{Vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{Vmatrix} = \frac{\partial u}{\partial x} \frac{\partial v}{\partial y} - \frac{\partial u}{\partial y} \frac{\partial v}{\partial x}.$$

If we look at the area elements, we see that the Jacobian plays the role of a scaling factor indicating by how much the area element is scaled in the new coordinates:

$$\mathrm{d}u\, \mathrm{d}v = \left| \frac{\partial(u, v)}{\partial(x, y)} \right| \mathrm{d}x\, \mathrm{d}y.$$

This idea extends to multi-dimensional integrals, where the Jacobian represents the scaling of the volume element, etc. To avoid cluttering, I will often use the shorthand notation $\partial_x u = \frac{\partial u}{\partial x}$, etc. for partial derivatives.

For those who don't remember the change-of-variables formula, there is another method which does not require you to remember anything except for

the alternating algebra of differential forms. $\mathrm{d}u$ and $\mathrm{d}v$ are differential 1-forms (covector fields). The product $\mathrm{d}u\mathrm{d}v$ should instead be viewed as a "wedge product" $\mathrm{d}u\wedge\mathrm{d}v$, which is a 2-form. A 2-form $\mathrm{d}u\wedge\mathrm{d}v$ is an oriented area element spanned by the covectors $\mathrm{d}u$ and $\mathrm{d}v$. In the left hand side we replace $dudv$ by the wedge product $\mathrm{d}u\wedge\mathrm{d}v$

$$\iint_R f(u,v)\,\mathrm{d}u\,\mathrm{d}v = \iint_R f(u,v)\,\mathrm{d}u\wedge\,\mathrm{d}v$$

Then, viewing u and v as functions of x and y, we expand $\mathrm{d}u$ and $\mathrm{d}v$ as total differentials:

$$\mathrm{d}u(x,y) = (\partial_x u)\mathrm{d}x + (\partial_y u)\mathrm{d}y$$

and

$$\mathrm{d}v(x,y) = (\partial_x v)\mathrm{d}x + (\partial_y v)\mathrm{d}y.$$

Then we form the wedge product of $\mathrm{d}u$ and $\mathrm{d}v$:

$$\mathrm{d}u\wedge\mathrm{d}v = [(\partial_x u)\mathrm{d}x + (\partial_y u)\mathrm{d}y]\wedge[(\partial_x v)\mathrm{d}x + (\partial_y v)\mathrm{d}y]\,,$$

When distributing the product, we apply the rules of the alternating algebra: since $\mathrm{d}x\wedge\mathrm{d}y$ is an oriented area element spanned by the covectors $\mathrm{d}x$ and $\mathrm{d}y$, we have that $\mathrm{d}x\wedge\mathrm{d}y = -\mathrm{d}y\wedge\mathrm{d}x$ (sign flipped because we have an "oriented area element" and this amounts to changing from the left hand rule to the right hand rule in a cross product), $\mathrm{d}x\wedge\mathrm{d}x = 0$ and $\mathrm{d}y\wedge\mathrm{d}y = 0$ (zero because the area spanned by two collinear vectors is zero). We are left with:

$$\mathrm{d}u\wedge\mathrm{d}v = (\partial_x u)(\partial_y v)\mathrm{d}x\wedge\mathrm{d}y + (\partial_y u)(\partial_x v)\mathrm{d}y\wedge\mathrm{d}x = (\partial_x u\,\partial_y v - \partial_y u\,\partial_x v)\mathrm{d}x\wedge\mathrm{d}y.$$

You will recognize the coefficient of $\mathrm{d}x\wedge\mathrm{d}y$ on the right hand side as the Jacobian determinant $\frac{\partial(u,v)}{\partial(x,y)}$. Thus, the alternating algebra of differential forms took care of calculating the determinant for us. This works in any number of dimensions.

Let us work out an example. Suppose that we have an integral

$$\iint_R f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

and want to change from Cartesian to polar coordinates, i.e.

$$x = r\cos\theta, \qquad y = r\sin\theta.$$

The total differentials are:

$$\mathrm{d}x(r,\theta) = \partial_r x\,\mathrm{d}r + \partial_\theta x\,\mathrm{d}\theta = \cos\theta\,\mathrm{d}r - r\sin\theta\,\mathrm{d}\theta$$

$$\mathrm{d}y(r,\theta) = \partial_r y\,\mathrm{d}r + \partial_\theta y\,\mathrm{d}\theta = \sin\theta\,\mathrm{d}r + r\cos\theta\,\mathrm{d}\theta.$$

Forming the wedge product $\mathrm{d}x\wedge\mathrm{d}y$:

$$\mathrm{d}x\wedge\mathrm{d}y = [\cos\theta\,\mathrm{d}r - r\sin\theta\,\mathrm{d}\theta]\wedge[\sin\theta\,\mathrm{d}r + r\cos\theta\,\mathrm{d}\theta]\,.$$

Applying the rules $\mathrm{d}\theta \wedge \mathrm{d}\theta = 0$, $\mathrm{d}r \wedge \mathrm{d}r = 0$ and $\mathrm{d}r \wedge \mathrm{d}\theta = -\mathrm{d}\theta \wedge \mathrm{d}r$, we are left with:

$$\mathrm{d}x \wedge \mathrm{d}y = r\cos^2\theta\,\mathrm{d}r \wedge \mathrm{d}\theta - r\sin^2\theta\,\mathrm{d}\theta \wedge \mathrm{d}r = r\,\mathrm{d}r \wedge \mathrm{d}\theta,$$

which is the familiar area element in polar coordinates.

Now let us return to the example of the previous section where we had the integral $\iint p_U(u)p_V(v)\mathrm{d}u\mathrm{d}v$, where $v = z$ and $u = zy$. Writing $\mathrm{d}u\mathrm{d}v$ as a wedge product $\mathrm{d}u \wedge \mathrm{d}v$, expanding the total differentials: $\mathrm{d}u(y,z) = z\mathrm{d}y + y\mathrm{d}z$ and $\mathrm{d}v(y,z) = \mathrm{d}z$ yields $\mathrm{d}u \wedge \mathrm{d}v = z\mathrm{d}y \wedge \mathrm{d}z$, where z is the Jacobian determinant that was sought and $\mathrm{d}y \wedge \mathrm{d}z$ are the new integration variables. The alternating algebra of differential forms automatically computes the determinant for us.

12.3. Leibniz Formula

The Leibniz formula for differentiation of integrals (the Leibniz integral rule) is:

$$\frac{\mathrm{d}}{\mathrm{d}y}\left(\int_{a(y)}^{b(y)} f(x,y)\,\mathrm{d}x\right) = \underbrace{\int_{a(y)}^{b(y)} \frac{\partial}{\partial y} f(x,y)\,\mathrm{d}x}_{1} + \underbrace{f\big(b(y),y\big)\cdot b'(y)}_{2} - \underbrace{f\big(a(y),y\big)\cdot a'(y)}_{3}$$

which consists of the sum of three terms: in the first one the differentiation is carried out inside the integral; the remaining two terms are surface (boundary) terms. This formula will help you compute PDFs from CDFs.

Let's look at an example of obtaining the PDF from the CDF $\mathbb{P}(Y < y)$ when $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = e^X$, by differentiating with respect to y. Since the y dependence occurs only in the upper limit of the integral, only the second term in the Leibniz formula is non-zero:

$$p_Y(y) \equiv \frac{\mathrm{d}\mathbb{P}(Y<y)}{\mathrm{d}y} = \frac{\mathrm{d}}{\mathrm{d}y}\int_{-\infty}^{\log y} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}\mathrm{d}x$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\log y-\mu)^2/2\sigma^2}\cdot\frac{1}{y}.$$

This particular PDF is also known as the log-normal distribution.

12.4. Infinitesimals

It is important to know how to work with infinitesimals. Let us take the link between CDF and PDF as an example. Denoting the CDF as $\mathbb{P}(X < x) = F(x)$, let us Taylor expand $F(x+\epsilon)$ about the point x (here $\epsilon > 0$ is

a small quantity):

$$F(x+\epsilon) = F(x) + \epsilon \cdot F'(x) + o(|\epsilon|) \tag{12.3}$$

where $F'(x) = \mathrm{d}F(x)/\mathrm{d}x$ and $o(|\epsilon|)$ denotes higher order terms (in this case ϵ^2 and higher powers of ϵ) which decay to 0 faster than ϵ in the limit $\epsilon \to 0$:

$$\lim_{\epsilon \to 0} \frac{o(|\epsilon|)}{|\epsilon|} = 0,$$

so that taking the limit $\epsilon \to 0$ in Eq. (12.3) leads to $F(x+\epsilon) = F(x) + \epsilon \mathrm{d}F(x)/\mathrm{d}x$. Taking $\epsilon = \mathrm{d}x$ (infinitesimal) this can be rewritten as:

$$F(x+\mathrm{d}x) - F(x) = \cancel{F(x)} + \mathrm{d}x \cdot F'(x) + o(|\mathrm{d}x|) - \cancel{F(x)} = \mathrm{d}F(x) + o(|\mathrm{d}x|),$$

since $F'(x) = \mathrm{d}F(x)/\mathrm{d}x$. Thus, as $\mathrm{d}x \to 0$ (without being equal to 0) the term $o(|\mathrm{d}x|)$ vanishes and we have that

$$\boxed{F(x+\mathrm{d}x) - F(x) = \mathrm{d}F(x).} \tag{12.4}$$

(This is only true if $\mathrm{d}x$ is an infinitesimal. In that case, $\mathrm{d}F$ is the total differential of F.)

Denote the CDF as $F(x) \equiv \mathbb{P}(X < x)$ and recall the interpretation of the PDF. Given a random variable X its PDF $p_X(x)$ times $\mathrm{d}x$ gives the probability that X will lie in the interval $(x, x+\mathrm{d}x)$:

$$\boxed{p_X(x)\mathrm{d}x = \mathbb{P}(x \le X \le x+\mathrm{d}x) = \mathbb{P}(X \le x+\mathrm{d}x) - \mathbb{P}(X \le x) = \mathrm{d}\mathbb{P}(X \le x),}$$

where $\mathrm{d}F(x) = F(x+\mathrm{d}x) - F(x)$ was used in the last step. In the second equality we have made use of the definition of the "probability function" $\mathbb{P}(\cdot)$ as an integral of the PDF, i.e.

$$\begin{aligned}\mathbb{P}(a \le X \le b) = \int_a^b p_X(x)\mathrm{d}x = \int_{-\infty}^b p_X(x)\mathrm{d}x - \int_{-\infty}^a p_X(x)\mathrm{d}x \\ = \mathbb{P}(X \le b) - \mathbb{P}(X \le a)\end{aligned}$$

hence $\mathbb{P}(x \le X \le x+\mathrm{d}x) = \mathbb{P}(X \le x+\mathrm{d}x) - \mathbb{P}(X \le x)$. Another way to look at the quantity $\mathbb{P}(x \le X \le x+\mathrm{d}x)$ is the integral of the PDF from x to $x+\mathrm{d}x$

$$\mathbb{P}(x \le X \le x+\mathrm{d}x) = \int_x^{x+\mathrm{d}x} p_X(x')\mathrm{d}x' = p_X(x)\mathrm{d}x.$$

The last equality follows because the integral is a Riemann sum containing only 1 term. It contains only 1 term because the interval $[x, x+\mathrm{d}x]$ where the integral is carried out is infinitesimally small.

So integrating from a to b we get the probability that X takes values between a and b:

$$\begin{aligned}\int_a^b p_X(x')\mathrm{d}x' = \int_a^b \mathbb{P}(x \le X \le x+\mathrm{d}x) &= \int_a^b [\mathbb{P}(X \le x+\mathrm{d}x) - \mathbb{P}(X \le x)] \\ &= \int_a^b \mathrm{d}\mathbb{P}(X \le x) = \mathbb{P}(X \le b) - \mathbb{P}(X \le a) \\ &= \mathbb{P}(a \le X \le b).\end{aligned}$$

If the interval (a, b) is chosen to be $(-\infty, x)$ we get the CDF:[1]

$$\int_{-\infty}^{x} p_X(x')\mathrm{d}x' = \mathbb{P}(-\infty \le X \le x) = \mathbb{P}(X \le x).$$

Differentiating with respect to x yields a method for obtaining the PDF from the CDF:

$$\boxed{\frac{\mathrm{d}\mathbb{P}(X \le x)}{\mathrm{d}x} = p_X(x).} \tag{12.5}$$

So now you know how to go from PDF to CDF or from CDF to PDF. The two concepts are related to each other by an integral or a derivative. If you are asked to obtain the probability distribution of a rv you can derive either the PDF or the CDF. In general, obtaining the CDF is easier because fewer steps are needed and the interpretation of the CDF in terms of probability is also simpler.

12.5. Taylor's Theorem in Several Variables

Because nonlinear optimization methods make extensive use of partial derivatives, here we review partial derivatives and the Taylor's theorem in multiple variables. We will show how to compute the partial derivatives of $1/r$, where $r = |\mathbf{r}|$ and $\mathbf{r}$ has components $\mathbf{r} = (x, y, z)$. r is its Euclidean length:

$$|\mathbf{r}| \equiv r \equiv \sqrt{x^2 + y^2 + z^2}.$$

12.5.1. Einstein summation convention. To keep the notation neat (uncluttered), we will use the Einstein summation convention. Whenever two indices are repeated in the same term, a summation is implied. For example, in the dot product of $\mathbf{u} = (u_x, u_y, u_z)$ and $\mathbf{v} = (v_x, v_y, v_z)$ we have:

$$\mathbf{u} \cdot \mathbf{v} = u_\alpha v_\alpha \equiv \sum_{i=1}^{3} u_i v_i = u_x v_x + u_y v_y + u_z v_z.$$

It is simpler to write $u_\alpha v_\alpha$ than the entire summation.

[1] Notice that we wrote $\mathbb{P}(X \le x)$ instead of $\mathbb{P}(-\infty \le X \le x)$ because the statement that $X \ge -\infty$ is always true and therefore, unnecessary or redundant.

12.5.2. Multivariate Taylor expansion. In 1D the Taylor expansion of $f(x+h)$ at x is:

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + \frac{1}{3!}f^{(3)}(x)h^3 + \dots$$

In n-D, a scalar-valued function $f(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^n$, is expanded as:

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \sum_{i=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_i} h_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} h_i h_j + \frac{1}{3!}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial^3 f(\mathbf{x})}{\partial x_i \partial x_j \partial x_k} h_i h_j h_k + \dots$$

12.5.3. Abbreviation for Partial Derivatives. We will often use the shorthand notation to abbreviate the notation for partial differentiation:

$$\partial_\alpha \equiv \frac{\partial}{\partial x_\alpha}$$

In this notation, and using the summation convention, the multivariate Taylor expansion looks particularly neat:

$$\boxed{f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + h_i \partial_i f(\mathbf{x}) + \frac{1}{2} h_i h_j \partial_i \partial_j f(\mathbf{x}) + \frac{1}{3!} h_i h_j h_k \partial_i \partial_j \partial_k f(\mathbf{x}) \dots}$$

(summation convention). The vectors $\mathbf{h}$ have been moved to the left-hand-side of all derivative operators to avoid any possible confusion about which quantity is differentiated.

12.5.4. Example: Derivative of $1/r$. The first order partial derivative of

$$\frac{1}{r} \equiv \frac{1}{|\mathbf{r}|} = \frac{1}{\sqrt{x^2+y^2+z^2}}$$

with respect to x is:

$$\frac{\partial}{\partial x}\left(\frac{1}{r}\right) \equiv \partial_x \left(\frac{1}{r}\right) = -\frac{1}{2}\frac{(2x)}{(x^2+y^2+z^2)^{3/2}} = -\frac{x}{r^3}.$$

Similar expressions are found for differentiation with respect to y or z. Thus, for any component $\alpha = x, y, z$ we have:

$$\boxed{\frac{\partial}{\partial r_\alpha}\left(\frac{1}{r}\right) \equiv \partial_\alpha \left(\frac{1}{r}\right) = -\frac{r_\alpha}{r^3}.}$$

Bibliography