Data Analysis

Lecture at the Bad Honnef Summer School on "Efficient Algorithms in Computational Physics", September 10–14, 2012

Peter Young

(Dated: September 12, 2012)

I. INTRODUCTION

In several of the lectures at this school you will learn about efficient algorithms to generate numerical data for problems of interest. This lecture will discuss what to do with that data once you have generated it.

Typically you will generate a set of values x_i, $y_i, \cdots$, $i = 1, \cdots N$, where N is the number of measurements. The first thing you will want to do is to estimate various average values, and determine *error bars* on those estimates. This is straightforward if one wants to compute a single average, e.g. $\langle x \rangle$, but not quite so easy for more complicated averages such as fluctuations in a quantity, $\langle x^2 \rangle - \langle x \rangle^2$, or combinations of measured values such as $\langle y \rangle / \langle x \rangle^2$. Averaging of data will be discussed in Sec. II.

Having obtained good averaged data with error bars for a system of a certain size, one typically does several sizes and fits the data to some model, for example using finite-size scaling. Techniques for fitting data will be described in the second part of this lecture in Sec. III

II. AVERAGES AND ERROR BARS

A. Basic Analysis

Suppose we have a set of data from a simulation, x_i, $(i = 1, \cdots, N)$, which we shall refer to as a *sample* of data. This data will have some random noise so the x_i are not all equal. Rather they are governed by a distribution $P(x)$, *which we don't know.*

The distribution is normalized,

$$\int_{-\infty}^{\infty} P(x)\, dx = 1, \tag{1}$$

and is usefully characterized by its moments, where the n-th moment is defined by

$$\langle x^n \rangle = \int_{-\infty}^{\infty} x^n\, P(x)\, dx\,. \tag{2}$$

We will denote the average *over the exact distribution* by angular brackets. Of particular interest are the first and second moments from which one forms the mean μ and variance σ^2, by

$$\mu \equiv \langle x \rangle \tag{3a}$$

$$\sigma^2 \equiv \langle \, (x - \langle x \rangle)^2 \, \rangle = \langle x^2 \rangle - \langle x \rangle^2 \, . \tag{3b}$$

The term "standard deviation" is used for σ, the square root of the variance.

In this section we will determine the mean $\langle x \rangle$, and the uncertainty in our estimate of it, from the N data points x_i. The determination of more complicated averages and resulting error bars will be discussed in Sec. II B 2

In order to obtain error bars we need to assume that the data are uncorrelated with each other. This is a crucial assumption, without which it is very difficult to proceed. However, it is not always clear if the data points are truly independent of each other; some correlations may be present but not immediately obvious. Here, we take the usual approach of assuming that even if there are some correlations, they are sufficiently weak so as not to significantly perturb the results of the analysis. In Monte Carlo simulations, measurements which differ by a sufficiently large number of Monte Carlo sweeps will be uncorrelated. More precisely the difference in sweep numbers should be greater than a "relaxation time" (see Helmut Katzgraber's lectures at this school). This is exploited in the "binning" method in which the data used in the analysis is not the individual measurements, but rather an average over measurements during a range of sweeps, called a "bin". If the bin size is greater than the relaxation time, results from adjacent bins will be (almost) uncorrelated. A pedagogical treatment of binning has been given by Ambegaokar and Troyer [1]. Alternatively, one can do independent Monte Carlo runs, requilibrating each time, and use, as individual data in the analysis, the average from each run.

The information *from the data* is usefully encoded in two parameters, the sample mean $\overline{x}$ and the sample standard deviation s which are defined by[1]

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \, , \tag{4a}$$

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2 \, . \tag{4b}$$

[1] The factor of $N-1$ rather than N in the expression for the sample variance in Eq. (4b) needs a couple of comments. Firstly, the final answer for the error bar on the mean, Eq. (15) below, will be independent of how this intermediate quantity is defined. Secondly, the N terms in Eq. (4b) are not all independent since $\overline{x}$, which is itself given by the x_i, is subtracted. Rather, as will be discussed more in the section on fitting, Sec. III, there are really only $N-1$ independent variables (called the "number of degrees of freedom" in the fitting context) and so dividing by $N-1$ has a rational basis.

In statistics, notation is often confusing but crucial to understand. Here, an average indicated by an over-bar, $\overline{\cdots}$, is an average over the *sample of N data points*. This is to be distinguished from an exact average over the distribution $\langle \cdots \rangle$, as in Eqs. (3a) and (3b). The latter is, however, just a theoretical construct since we *don't know* the distribution $P(x)$, only the set of N data points x_i which have been sampled from it.

Next we derive two simple results which will be useful later:

1. The mean of the sum of N independent variables with the same distribution is N times the mean of a single variable, and

2. The variance of the sum of N independent variables with the same distribution is N times the variance of a single variable.

The result for the mean is obvious since, defining $X = \sum_{i=1}^{N} x_i$,

$$\langle X \rangle = \sum_{i=1}^{N} \langle x_i \rangle = N \langle x_i \rangle \boxed{= N\mu\,.} \tag{5}$$

The result for the standard deviation needs a little more work:

$$\sigma_X^2 \equiv \langle X^2 \rangle - \langle X \rangle^2 \tag{6a}$$

$$= \sum_{i,j=1}^{N} \left(\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle \right) \tag{6b}$$

$$= \sum_{i=1}^{N} \left(\langle x_i^2 \rangle - \langle x_i \rangle^2 \right) \tag{6c}$$

$$= N \left(\langle x^2 \rangle - \langle x \rangle^2 \right) \tag{6d}$$

$$\boxed{= N\sigma^2\,.} \tag{6e}$$

To get from Eq. (6b) to Eq. (6c) we note that, for $i \neq j$, $\langle x_i x_j \rangle = \langle x_i \rangle \langle x_j \rangle$ since x_i and x_j are assumed to be statistically independent. (This is where the statistical independence of the data is needed.)

Now we describe an important thought experiment. Let's *suppose* that we could repeat the set of N measurements *very many* many times, each time obtaining a value of the sample average $\overline{x}$. From these results we could construct a distribution, $\widetilde{P}(\overline{x})$, for the sample average as shown in Fig. 1.

If we do enough repetitions we are effectively averaging over the exact distribution. Hence the

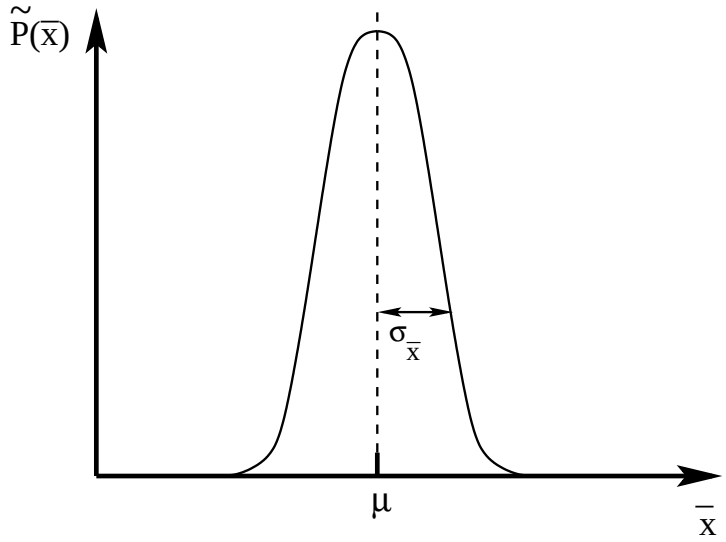

FIG. 1: The distribution of results for the sample mean $\overline{x}$ obtained by repeating the measurements of the N data points x_i many times. The average of this distribution is μ, the exact average value of x. The mean, $\overline{x}$, obtained from one sample of data typically differs from μ by an amount of order $\sigma_{\overline{x}}$, the standard deviation of the distribution $\widetilde{P}(\overline{x})$.

average of the sample mean, $\overline{x}$, over very many repetitions of the data, is given by

$$\langle \overline{x} \rangle = \frac{1}{N} \sum_{i=1}^{N} \langle x_i \rangle = \langle x \rangle \equiv \mu \,, \tag{7}$$

i.e. it is the exact average over the distribution of x, as one would intuitively expect, see Fig. 1. Eq. (7) also follows from Eq. (5) by noting that $\overline{x} = X/N$.

In fact, though, we have only the *one* set of data, so we can not determine μ exactly. However, Eq. (7) shows that

$$\boxed{\text{the best estimate of } \mu \text{ is } \overline{x},} \tag{8}$$

i.e. the sample mean, since averaging the sample mean over many repetitions of the N data points gives the true mean of the distribution, μ. An estimate like this, which gives the exact result if averaged over many repetitions of the experiment, is said to be $\boxed{\text{unbiased.}}$

We would also like an estimate of the uncertainty, or "error bar", in our estimate of $\overline{x}$ for the exact average μ. We take $\sigma_{\overline{x}}$, the standard deviation in $\overline{x}$ (obtained if one did many repetitions of the N measurements), to be the uncertainty, or error bar, in $\overline{x}$. The reason is that $\sigma_{\overline{x}}$ is the width

of the distribution $\widetilde{P}(\overline{x})$, shown in Fig. 1, and a *single* estimate $\overline{x}$ typically differs from the exact result μ by an amount of this order. The variance $\sigma^2_{\overline{x}}$ is given by

$$\sigma^2_{\overline{x}} \equiv \langle \overline{x}^2 \rangle - \langle \overline{x} \rangle^2 = \frac{\sigma^2}{N} \,, \tag{9}$$

which follows from Eq. (6e) since $\overline{x} = X/N$.

The problem with Eq. (9) is that **we don't know** σ^2 since it is a function of the exact distribution $P(x)$. We do, however, know the *sample* variance s^2, see Eq. (4b), and the average of this over many repetitions of the N data points, is equal to σ^2 since

$$\langle s^2 \rangle = \frac{1}{N-1} \sum_{i=1}^{N} \langle x_i^2 \rangle - \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \langle x_i x_j \rangle \tag{10a}$$

$$= \frac{N}{N-1} \langle x^2 \rangle - \frac{1}{N(N-1)} \left[N(N-1) \langle x \rangle^2 + N \langle x^2 \rangle \right] \tag{10b}$$

$$= \left[\langle x^2 \rangle - \langle x \rangle^2 \right] \tag{10c}$$

$$= \sigma^2 \,. \tag{10d}$$

To get from Eq. (10a) to Eq. (10b), we have separated the terms with $i = j$ in the last term of Eq. (10a) from those with $i \neq j$, and used the fact that each of the x_i is chosen from the same distribution and is statistically independent of the others. It follows from Eq. (10d) that

$$\boxed{\text{the best estimate of } \sigma^2 \text{ is } s^2 \,,} \tag{11}$$

since averaging s^2 over many repetitions of N data points gives σ^2. The estimate for σ^2 in Eq. (11) is therefore unbiased.

Combining Eqs. (9) and (11) gives

$$\boxed{\text{the best estimate of } \sigma^2_{\overline{x}} \text{ is } \frac{s^2}{N} \,.} \tag{12}$$

This estimate is also unbiased. We have now obtained, using only information from from the data, that the mean is given by

$$\boxed{\mu = \overline{x} \ \pm \ \sigma_{\overline{x}} \,.} \tag{13}$$

where we estimate

$$\boxed{\sigma_{\overline{x}} = \frac{s}{\sqrt{N}}} \tag{14}$$

which can write explicitly in terms of the data points as

$$\boxed{\sigma_{\overline{x}} = \left[\frac{1}{N(N-1)}\sum_{i=1}^{N}(x_i - \overline{x})^2\right]^{1/2}}. \tag{15}$$

Remember that $\overline{x}$ and s are the mean and standard deviation of the (one set) of data that is available to us, see Eqs. (4a) and (4b).

As an example, suppose $N = 5$ and the data points are

$$x_i = 10, 11, 12, 13, 14, \tag{16}$$

(not very random looking data it must be admitted!). Then, from Eq. (4a) we have $\overline{x} = 12$, and from Eq. (4b)

$$s^2 = \frac{1}{4}\left[(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2\right] = \frac{5}{2}. \tag{17}$$

Hence, from Eq. (14),

$$\sigma_{\overline{x}} = \frac{1}{\sqrt{5}}\sqrt{\frac{5}{2}} = \frac{1}{\sqrt{2}}. \tag{18}$$

so

$$\mu = \overline{x} \pm \sigma_{\overline{x}} = 12 \pm \frac{1}{\sqrt{2}}. \tag{19}$$

How does the error bar decrease with the number of statistically independent data points N? Equation (10d) states that the expectation value of s^2 is equal to σ^2 and hence, from Eq. (14), we see that

$$\boxed{\text{the error bar in the mean goes down like } 1/\sqrt{N}.}$$

Hence, to reduce the error bar by a factor of 10 one needs 100 times as much data. This is discouraging, but is a fact of life when dealing with random noise.

For Eq. (14) to be really useful we need to know the probability that the true answer μ lies more than $\sigma_{\overline{x}}$ away from our estimate $\overline{x}$. Fortunately, for large N, the central limit theorem, derived in Appendix A, tells us (for distributions where the first two moments are finite) that the distribution of $\overline{x}$ is a Gaussian. For this distribution we know that the probability of finding a result more than one standard deviation away from the mean is 32%, more than two standard deviations is 4.5% and more than three standard deviations is 0.3%. Hence we expect that most of the time $\overline{x}$ will be within $\sigma_{\overline{x}u}$ of the correct result μ, and only occasionally will be more than two times $\sigma_{\overline{x}}$ from

it. Even if N is not very large, so there are some deviations from the Gaussian form, the above numbers are often a reasonable guide.

However, as emphasized in appendix A, distributions which occur in nature typically have much more weight in the tails than a Gaussian. As a result, the weight in the tails of the distribution of the sum can be much larger than for a Gaussian even for quite large values of N, see Fig. 4. It follows that the probability of an "outlier" can be much higher than that predicted for a Gaussian distribution, as anyone who has invested in the stock market knows well!

B. Advanced Analysis

In Sec. II A we learned how to estimate a simple average, such as $\mu_x \equiv \langle x \rangle$, plus the error bar in that quantity, from a set of data x_i. Trivially this method also applies to a *linear* combination of different averages, $\mu_x, \mu_y, \cdots$ etc. However, we often need more complicated, *non-linear* functions of averages. One example is the fluctuations in a quantity, i.e. $\langle x^2 \rangle - \langle x \rangle^2$. Another example is dimensionless combination of moments, which gives information about the *shape* of a distribution independent of its overall scale. Such quantities are very popular in finite-size scaling (FSS) analyses since the FSS form is simpler than for quantities with dimension. An popular example, first proposed by Binder, is $\langle x^4 \rangle / \langle x^2 \rangle^2$, which is known as the "kurtosis" (frequently a factor of 3 is subtracted to make it zero for a Gaussian).

Hence, in this section we consider how to determine *non-linear functions* of averages of one or more variables, $f(\mu_x, \mu_y, \cdots)$, where

$$\mu_x \equiv \langle x \rangle \,, \tag{20}$$

etc. For example, the two quantities mentioned in the previous paragraph correspond to

$$f(\mu_x, \mu_y) = \mu_y - \mu_x^2 \,, \tag{21}$$

with $y = x^2$ and

$$f(\mu_y, \mu_z) = \frac{\mu_y}{\mu_z^2} \,, \tag{22}$$

with $y = x^4$ and $z = x^2$.

The natural estimate of $f(\mu_x, \mu_y)$ from the sample data is clearly $f(\overline{x}, \overline{y})$. However, it will take some more thought to estimate the error bar in this quantity. The traditional way of doing this is called "error propagation", described in Sec. II B 1 below. However, it is now more common to

use either "jackknife" or "bootstrap" procedures, described in Secs. II B 2 and II B 3. At the price of some additional computation, which is no difficulty when done on a modern computer (though it would have been tedious in the old days when statistics calculations were done by hand), these methods automate the calculation of the error bar.

Furthermore, the estimate of $f(\mu_x, \mu_y)$ turns out to have some *bias* if f is a non-linear function. Usually this is small effect because it is order $1/N$, see for example Eq. (29) below, whereas the statistical error is of order $1/\sqrt{N}$. Since N is usually large, the bias is generally much less than the statistical error and so can generally be neglected. In any case, the jackknife and bootstrap methods also enable one to eliminate the leading ($\sim 1/N$) contribution to the bias in a automatic fashion.

1. Traditional method

First we will discuss the traditional method, known as error propagation, to compute the error bar and bias. We expand $f(\overline{x}, \overline{y})$ about $f(\mu_x, \mu_y)$ up to second order in the deviations:

$$f(\overline{x}, \overline{y}) = f(\mu_x, \mu_y) + (\partial_{\mu_x} f)\, \delta_{\overline{x}} + (\partial_{\mu_y} f)\, \delta_{\overline{y}} + \frac{1}{2}\,(\partial^2_{\mu_x \mu_x} f)\, \delta^2_{\overline{x}} + (\partial^2_{\mu_x \mu_y} f)\, \delta_{\overline{x}} \delta_{\overline{y}} + \frac{1}{2}\,(\partial^2_{\mu_y \mu_y} f)\, \delta^2_{\overline{y}} + \cdots\,, \tag{23}$$

where

$$\delta_{\overline{x}} = \overline{x} - \mu_x, \tag{24}$$

etc.

The terms of first order in the $\delta' s$ in Eq. (23) give the leading contribution to the error, but would average to zero if the procedure were to be repeated many times. However, the terms of second order do not average to zero and so give the leading contribution to the bias. We now estimate that bias.

Averaging Eq. (23) over many repetitions, and noting that

$$\langle \delta^2_{\overline{x}} \rangle = \langle \overline{x}^2 \rangle - \langle \overline{x} \rangle^2 = \sigma^2_{\overline{x}}, \quad \langle \delta^2_{\overline{y}} \rangle = \langle \overline{y}^2 \rangle - \langle \overline{y} \rangle^2 = \sigma^2_{\overline{y}}, \quad \langle \delta_{\overline{x}} \delta_{\overline{y}} \rangle = \langle \overline{x}\, \overline{y} \rangle - \langle \overline{x} \rangle \langle \overline{y} \rangle = \sigma^2_{\overline{x}\overline{y}}, \tag{25}$$

we get

$$\langle f(\overline{x}, \overline{y}) \rangle - f(\mu_x, \mu_y) = \frac{1}{2}\,(\partial^2_{\mu_x \mu_x} f)\, \sigma^2_{\overline{x}} + (\partial^2_{\mu_x \mu_y} f)\, \sigma^2_{\overline{x}\overline{y}} + \frac{1}{2}\,(\partial^2_{\mu_y \mu_y} f)\, \sigma^2_{\overline{y}}\,. \tag{26}$$

As discussed in Sec. II A our estimate of $\sigma^2_{\overline{x}}$ is the sample variance, see Eq. (4b), which we now call s^2_{xx}, divided by N, see Eq. (12), and similarly for $\sigma^2_{\overline{y}}$. In the same way, our estimate of $\sigma^2_{\overline{x}\overline{y}}$ is the

sample *covariance* of x and y, defined by

$$s_{xy}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})\ (y_i - \overline{y})\ , \tag{27}$$

divided by N.

Hence, from Eq. (26), we have

$$f(\mu_x, \mu_y) = \langle f(\overline{x}, \overline{y}) \rangle - \frac{1}{N} \left[\frac{1}{2} \left(\partial^2_{\mu_x \mu_x} f \right) s_{xx}^2 + \left(\partial^2_{\mu_x \mu_y} f \right) s_{xy}^2 + \frac{1}{2} \left(\partial^2_{\mu_y \mu_y} f \right) s_{yy}^2 \right] , \tag{28}$$

where the leading contribution to the bias is given by the $1/N$ term. Note that the bias term is "self-averaging", i.e. the fluctuations in it are small relative to the average (by a factor of $1/\sqrt{N}$) when averaging over many repetitions of the data. It follows from Eq. (28) that if one wants to eliminate the leading contribution to the bias one should

$$\boxed{\text{estimate } f(\mu_x, \mu_y) \ \text{ from } f(\overline{x}, \overline{y}) - \frac{1}{N} \left[\frac{1}{2} \left(\partial^2_{\mu_x \mu_x} f \right) s_{xx}^2 + \left(\partial^2_{\mu_x \mu_y} f \right) s_{xy}^2 + \frac{1}{2} \left(\partial^2_{\mu_y \mu_y} f \right) s_{yy}^2 \right] .} \tag{29}$$

As claimed earlier, the bias correction is of order $1/N$ and vanishes if f is a linear function. The generalization to functions of more than two averages, $f(\mu_x, \mu_y, \mu_z, \cdots)$, is obvious.

Next we discuss the leading *error* in using $f(\overline{x}, \overline{y})$ as an estimate for $f(\mu_x, \mu_y)$. This comes from the terms linear in the δ's in Eq. (23). Since $\langle f(\overline{x}, \overline{y}) \rangle - f(\mu_x, \mu_y)$ is of order $1/N$ according to Eq. (28), we can replace $\langle f(\overline{x}, \overline{y}) \rangle$ by $f(\mu_x, \mu_y)$ to get the leading order fluctuations. Hence,

$$\begin{aligned} \sigma_f^2 &\equiv \langle\, f^2(\overline{x}, \overline{y}) \rangle - \langle f(\overline{x}, \overline{y}) \rangle^2 \\ &= \langle\, [\, f(\overline{x}, \overline{y}) - \langle f(\overline{x}, \overline{y}) \rangle\,]^2 \,\rangle \\ &= \langle\, [\, f(\overline{x}, \overline{y}) - f(\mu_x, \mu_y)\,]^2 \,\rangle \\ &= (\partial_{\mu_x} f)^2 \, \langle \delta_{\overline{x}}^2 \rangle + 2 (\partial_{\mu_x} f)\, (\partial_{\mu_y} f)\, \langle \delta_{\overline{x}} \delta_{\overline{y}} \rangle + (\partial_{\mu_y} f)^2 \, \langle \delta_{\overline{y}}^2 \rangle \,, \end{aligned} \tag{30}$$

where, to get the last line, we omitted the quadratic terms in Eq. (23) and squared it. As above, we use s_{xx}^2/N as an estimate of $\langle \delta_{\overline{x}}^2 \rangle$ and similarly for the other terms. Hence

$$\boxed{\text{the best estimate of } \sigma_f^2 \text{ is } \frac{1}{N} (\partial_{\mu_x} f)^2 \, s_{xx}^2 + 2 (\partial_{\mu_x} f)\, (\partial_{\mu_y} f)\, s_{xy}^2 + (\partial_{\mu_y} f)^2 \, s_{yy}^2 \,.} \tag{31}$$

This estimate is unbiased to leading order in N. Note that we need to keep track not only of fluctuations in x and y, characterized by their variances s_{xx}^2 and s_{yy}^2, but also cross correlations between x and y, characterized by their covariance s_{xy}^2.

Hence, still to leading order in N, we get

$$\boxed{f(\mu_x, \mu_y) = f(\overline{x}, \overline{y}) \pm \sigma_f \,,} \tag{32}$$

where we estimate the error bar σ_f from Eq. (31) which shows that it is of order $1/\sqrt{N}$. Again, the generalization to functions of more than two averages is obvious.

Note that in the simple case studied in Sec. II A where there is only one set of variables x_i and $f(\mu_x) = \mu_x$, Eq. (28) tells us that there is no bias, which is correct, and Eq. (31) gives an expression for the error bar which agrees with Eq. (14).

In Eqs. (29) and (31) we need to keep track how errors in the individual quantities like $\overline{x}$ propagate to the estimate of the function f. This requires inputting by hand the various partial derivatives into the analysis program. In the next two sections we see how *resampling* the data automatically takes account of error propagation without needing to input the partial derivatives. These approaches, known as jackknife and bootstrap, provide a *fully automatic* method of determining error bars and bias.

2. Jackknife

We define a jackknife average to be the average over all data in the sample *except one point*, i.e.

$$x_i^J \equiv \frac{1}{N-1} \sum_{j \neq i} x_j \,. \tag{33}$$

We also define corresponding jackknife averages of the function f (again for concreteness we will assume that f is a function of just 2 averages but the generalization will be obvious):

$$f_i^J \equiv f(x_i^J, y_i^J) \,. \tag{34}$$

In other words, we use the jackknife values, x_i^J, y_i^J, rather than the sample means, $\overline{x}, \overline{y}$, as the arguments of f. The same is true if we need a function of more than two averages , i.e. $f_i^J \equiv f(x_i^J, y_i^J, z_i^J, \cdots)$. The jackknife estimate of $f(\mu_x, \mu_y)$ is then the average of the f_i^J:

$$\boxed{\overline{f^J} \equiv \frac{1}{N} \sum_{i=1}^{N} f_i^J \,.} \tag{35}$$

It is straightforward to show that if f is a linear function of μ_x and μ_y then $\overline{f^J} = f(\overline{x}, \overline{y})$, i.e. the jackknife and standard averages are identical. However, when f is not a linear function, so there is bias, there *is* a difference, and we will now show the resampling carried out in the jackknife method can be used to determine bias and error bars in an automated way.

We proceed as for the derivation of Eq. (28), which we now write as

$$f(\mu_x, \mu_y) = \langle f(\overline{x}, \overline{y}) \rangle - \frac{A}{N} + \frac{B}{N^2} + \cdots \,, \tag{36}$$

where A is the term in rectangular brackets in Eq. (28), and we have added the next order correction. We find the same result for the jackknife average except that N is replaced by $N-1$, i.e.

$$f(\mu_x,\mu_y) = \langle\overline{f^J}\rangle - \frac{A}{N-1} + \frac{B}{(N-1)^2}\cdots . \tag{37}$$

We can therefore eliminate the leading contribution to the bias by forming an appropriate linear combination of $f(\overline{x},\overline{y})$ and $\overline{f^J}$, namely

$$f(\mu_x,\mu_y) = N\langle f(\overline{x},\overline{y})\rangle - (N-1)\langle\overline{f^J}\rangle + O\left(\frac{1}{N^2}\right) . \tag{38}$$

It follows that, to eliminate the leading bias without computing partial derivatives, one should

$$\boxed{\text{estimate } f(\mu_x,\mu_y) \text{ from } Nf(\overline{x},\overline{y}) - (N-1)\overline{f^J}\,.} \tag{39}$$

As mentioned earlier, bias is usually not a big problem because it is of order $1/N$ whereas the statistical error is of order $1/\sqrt{N}$ which is much bigger if N is large. In most cases, N is sufficiently large that one can use *either* the usual average $f(\overline{x},\overline{y})$, or the jackknife average $\overline{f^J}$ in Eq. (35), to estimate $f(\mu_x,\mu_y)$, since the difference between them will be much smaller than the statistical error. In other words, elimination of the leading bias using Eq. (39) is usually not necessary.

Next we show that the jackknife method gives error bars, which agree with Eq. (31) but without the need to explicitly keep track of the partial derivatives.

We define the variance of the jackknife averages by

$$\sigma^2_{f^J} \equiv \overline{(f^J)^2} - \left(\overline{f^J}\right)^2 , \tag{40}$$

where

$$\overline{(f^J)^2} = \frac{1}{N}\sum_{i=1}^{N}\left(f_i^J\right)^2 . \tag{41}$$

Using Eqs. (34) and (35), we expand $\overline{f^J}$ away from the exact result $f(\mu_x,\mu_y)$. Just including the leading contribution gives

$$\begin{aligned}
\overline{f^J} - f(\mu_x,\mu_y) &= \frac{1}{N}\sum_{i=1}^{N}\left[(\partial_{\mu_x}f)\,(x_i^J-\mu_x) + (\partial_{\mu_y}f)\,(y_i^J-\mu_y)\right] \\
&= \frac{1}{N(N-1)}\sum_{i=1}^{N}\left[(\partial_{\mu_x}f)\;\{N(\overline{x}-\mu_x)-(x_i-\mu_x)\} + (\partial_{\mu_y}f)\;\{N(\overline{y}-\mu_y)-(y_i-\mu_y)\}\right] \\
&= (\partial_{\mu_x}f)\,(\overline{x}-\mu_x) + (\partial_{\mu_y}f)\,(\overline{y}-\mu_y)\,.
\end{aligned} \tag{42}$$

Similarly we find

$$\begin{aligned}\overline{(f^J)^2} &= \frac{1}{N}\sum_{i=1}^{N}\left[f(\mu_x,\mu_y)+(\partial_{\mu_x}f)\,(x_i^J-\mu_x)+(\partial_{\mu_y}f)\,(y_i^J-\mu_y)\right]^2 \\ &= f^2(\mu_x,\mu_y)+2f(\mu_x,\mu_y)\,\left[(\partial_{\mu_x}f)\,(\overline{x}-\mu_x)+(\partial_{\mu_y}f)\,(\overline{y}-\mu_y)\right] \\ &\quad+\left\{(\partial_{\mu_x}f)^2\,\left[(\overline{x}-\mu_x)^2+\frac{s_{xx}^2}{N(N-1)}\right]+(\partial_{\mu_y}f)^2\,\left[(\overline{y}-\mu_y)^2+\frac{s_{yy}^2}{N(N-1)}\right]\right. \\ &\quad\left.+2(\partial_{\mu_x}f)(\partial_{\mu_y}f)\,\left[(\overline{x}-\mu_x)(\overline{y}-\mu_y)+\frac{s_{xy}^2}{N(N-1)}\right]\right\}. \end{aligned} \tag{43}$$

Hence, from Eq. (40), the variance in the jackknife estimates is given by

$$\sigma_{f^J}^2=\frac{1}{N(N-1)}\,\left[(\partial_{\mu_x}f)^2\,s_{xx}^2+(\partial_{\mu_y}f)^2\,s_{yy}^2+2(\partial_{\mu_x}f)(\partial_{\mu_y}f)s_{xy}\right]\,, \tag{44}$$

which is just $1/(N-1)$ times σ_f^2, the estimate of the square of the error bar in $f(\overline{x},\overline{y})$ given in Eq. (31). Hence

$$\boxed{\text{the jackknife estimate for } \sigma_f \text{ is } \sqrt{N-1}\,\sigma_{f^J}\,.} \tag{45}$$

Note that this is directly obtained from the jackknife estimates without having to put in the partial derivatives by hand. Note too that the $\sqrt{N-1}$ factor is in the *numerator* whereas the factor of $\sqrt{N}$ in Eq. (14) is in the *denominator*. Intuitively the reason for this difference is that the jackknife estimates are very close since they would all be equal except that each one omits just one data point.

If N is very large it may be better to group the N data points into $N_{\rm group}$ groups (or "bins") of data and take, as data used in the jackknife analysis, the average of data in each group. The above results clearly go through with N replaced by $N_{\rm group}$.

To summarize this subsection, to estimate $f(\mu_x,\mu_y)$ one can use either $f(\overline{x},\overline{y})$ or the jackknife average $\overline{f^J}$ in Eq. (35). The error bar in this estimate, σ_f, is related to the standard deviation in the jackknife estimates σ_{f^J} by Eq. (45).

3. Bootstrap

The bootstrap, like the jackknife, is a resampling of the N data points Whereas jackknife considers N new data sets, each of containing all the original data points minus one, bootstrap uses $N_{\rm boot}$ data sets each containing N points obtained by random (Monte Carlo) sampling of the original set of N points. During the Monte Carlo sampling, the probability that a data point is

picked is $1/N$ irrespective of whether it has been picked before. (In the statistics literature this is called picking from a set "with replacement".) Hence a given data point x_i will, *on average*, appear once in each Monte Carlo-generated data set, but may appear not at all, or twice, and so on. The probability that x_i appears n_i times is close to a Poisson distribution with mean unity. However, it is not exactly Poissonian because of the constraint in Eq. (46) below. It turns out that we shall need to include the deviation from the Poisson distribution even for large N. We shall use the term "bootstrap" data sets to denote the Monte Carlo-generated data sets.

More precisely, let us suppose that the number of times x_i appears in a bootstrap data set is n_i. Since each bootstrap dataset contains exactly N data points, we have the constraint

$$\sum_{i=1}^{N} n_i = N\,. \tag{46}$$

Consider one of N variables x_i. Each time we generate an element in a bootstrap dataset the probability that it is x_i is $1/N$, which we will denote by p. From standard probability theory, the probability that x_i occurs n_i times is given by a binomial distribution

$$P(n_i) = \frac{N!}{n_i!\,(N-n_i)!}\,p^{n_i}(1-p)^{N-n_i}\,. \tag{47}$$

The mean and standard deviation of a binomial distribution are given by

$$[n_i]_{\rm MC} = Np = 1\,, \tag{48}$$

$$[n_i^2]_{\rm MC} - [n_i]_{\rm MC}^2 = Np(1-p) = 1 - \frac{1}{N}\,, \tag{49}$$

where $[\ldots]_{\rm MC}$ denotes an exact average over bootstrap samples (for a fixed original data set x_i). For $N \to \infty$, the binomial distribution goes over to a Poisson distribution, for which the factor of $1/N$ in Eq. (49) does not appear. We assume that $N_{\rm boot}$ is sufficiently large that the bootstrap average we carry out reproduces this result with sufficient accuracy. Later, we will discuss what values for $N_{\rm boot}$ are sufficient in practice. Because of the constraint in Eq. (46), n_i and n_j (with $i \neq j$) are not independent and we find, by squaring Eq. (46) and using Eqs. (48) and (49), that

$$[n_i n_j]_{\rm MC} - [n_i]_{\rm MC}[n_j]_{\rm MC} = -\frac{1}{N} \qquad (i \neq j)\,. \tag{50}$$

First of all we just consider the simple average $\mu_x \equiv \langle x \rangle$, for which, of course, the standard methods in Sec. II A suffice, so bootstrap is not necessary. However, this will show how to get averages and error bars in a simple case, which we will then generalize to non-linear functions of averages.

We denote the average of x for a given bootstrap data set by x_α^B, where α runs from 1 to $N_{\rm boot}$, *i.e.*

$$x_\alpha^B = \frac{1}{N}\sum_{i=1}^{N} n_i^\alpha x_i \,. \tag{51}$$

We then compute the bootstrap average of the mean of x and the bootstrap variance in the mean, by averaging over all the bootstrap data sets. We assume that that $N_{\rm boot}$ is large enough for the bootstrap average to be exact, so we can use Eqs. (49) and (50). The result is

$$\overline{x^B} \equiv \frac{1}{N_{\rm boot}}\sum_{\alpha=1}^{N_{\rm boot}} x_\alpha^B \;=\; \frac{1}{N}\sum_{i=1}^{N}[n_i]_{\rm MC} x_i = \frac{1}{N}\sum_{i=1}^{N} x_i = \overline{x} \tag{52}$$

$$\sigma_{x^B}^2 \equiv \overline{(x^B)^2} - \left(\overline{x^B}\right)^2 \;=\; \frac{1}{N^2}\left(1-\frac{1}{N}\right)\sum_i x_i^2 - \frac{1}{N^3}\sum_{i\neq j} x_i x_j \,, \tag{53}$$

where

$$\overline{(x^B)^2} \equiv \frac{1}{N_{\rm boot}}\sum_{\alpha=1}^{N_{\rm boot}} \left[\left(x_\alpha^B\right)^2\right]_{\rm MC} \,. \tag{54}$$

We now average Eqs. (52) and (53) over many repetitions of the original data set x_i. Averaging Eq. (52) gives

$$\langle \overline{x^B} \rangle = \langle \overline{x} \rangle = \langle x \rangle \equiv \mu_x \,. \tag{55}$$

This shows that the bootstrap average $\overline{x^B}$ is an unbiased estimate of the exact average μ_x. Averaging Eq. (53) gives

$$\left\langle \sigma_{x^B}^2 \right\rangle = \frac{N-1}{N^2}\sigma^2 = \frac{N-1}{N}\sigma_{\overline{x}}^2 \,, \tag{56}$$

where we used Eq. (9) to get the last expression. Since $\sigma_{\overline{x}}$ is the uncertainty in the sample mean, we see that

$$\boxed{\text{the bootstrap estimate of } \sigma_{\overline{x}} \text{ is } \sqrt{\frac{N}{N-1}}\,\sigma_{x^B} \,.} \tag{57}$$

Remember that σ_{x^B} is the standard deviation of the bootstrap data sets. Usually N is sufficiently large that the square root in Eq. (57) can be replaced by unity.

As for the jackknife, these results can be generalized to finding the error bar in some possibly non-linear function, $f(\mu_x, \mu_y)$, rather than for μ_x. One computes the bootstrap estimates for $f(\mu_x, \mu_y)$, which are

$$f_\alpha^B = f(x_\alpha^B, y_\alpha^B) \,. \tag{58}$$

In other words, we use the bootstrap values, x_α^B, y_α^B, rather than the sample means, $\overline{x}, \overline{y}$, as the arguments of f. The final bootstrap estimate for $f(\mu_x, \mu_y)$ is the average of these, *i.e.*

$$\boxed{\overline{f^B} = \frac{1}{N_{\rm boot}} \sum_{\alpha=1}^{N_{\rm boot}} f_\alpha^B \,.} \tag{59}$$

Following the same methods in the jackknife section, one obtains the error bar, σ_f, in $f(\mu_x, \mu_y)$. The result is

$$\boxed{\text{the bootstrap estimate for } \sigma_f \text{ is } \sqrt{\frac{N}{N-1}}\ \sigma_{f^B}}, \tag{60}$$

where

$$\boxed{\sigma_{f^B}^2 = \overline{(f^B)^2} - \left(\overline{f^B}\right)^2 \,,} \tag{61}$$

is the variance of the bootstrap estimates. Here

$$\overline{(f^B)^2} \equiv \frac{1}{N_{\rm boot}} \sum_{\alpha=1}^{N_{\rm boot}} \left(f_\alpha^B\right)^2 \,. \tag{62}$$

Usually N is large enough that the factor of $\sqrt{N/(N-1)}$ is Eq. (60) can be replaced by unity. Equation (60) corresponds to the result Eq. (57) which we derived for the special case of $f(\mu_x, \mu_y) = \mu_x$.

Again following the same path as in the jackknife section, it is straightforward to show that Eqs. (59) and (60) are unbiased estimates for $N \to \infty$. However, if N is not too large it may be useful to eliminate the leading contribution to the bias in the mean, as we did for jackknife in Eq. (39). The result is that one should

$$\boxed{\text{estimate } f(\mu_x, \mu_y) \text{ from } 2f(\overline{x}, \overline{y}) - \overline{f^B} \,.} \tag{63}$$

The bias in Eq. (63) is of order $1/N^2$, whereas $f(\overline{x}, \overline{y})$ and $\overline{f^B}$ each have a bias of order $1/N$. However, it is not normally necessary to eliminate the bias since, if N is large, the bias is much smaller than the statistical error.

I have not systematically studied the values of $N_{\rm boot}$ that are needed in practice to get accurate estimates for the error. It seems that $N_{\rm boot}$ in the range 100 to 500 is typically chosen, and this seems to be adequate irrespective of how large N is.

To summarize this subsection, to estimate $f(\mu_x, \mu_y)$ one can either use $f(\overline{x}, \overline{y})$, or the bootstrap average in Eq. (59), and the error bar in this estimate, σ_f, is related to the standard deviation in the bootstrap estimates by Eq. (60).

III. FITTING DATA TO A MODEL

A good reference for the material in this section is Chapter 15 of Numerical Recipes [2].

Frequently we are given a set of data points $(x_i, y_i), i = 1, 2, \cdots, N$, with corresponding error bars, σ_i, through which we would like to fit to a smooth function $f(x)$. The function could be straight line (the simplest case), a higher order polynomial, or a more complicated function. The fitting function will depend on M "fitting parameters", a_α and we would like the "best" fit obtained by adjusting these parameters. We emphasize that a fitting procedure should not only

1. give the values of the fit parameters, but also

2. provide error estimates on those parameters, and

3. provide a measure of how good the fit is.

If the result of part 3 is that the fit is very poor, the results of parts 1 and 2 are probably meaningless.

The definition of "best" is not unique. However, the most useful choice, and the one nearly always taken, is "least squares", in which one minimizes the sum of the squares of the difference between the observed y-value, y_i, and the fitting function evaluated at x_i, weighted appropriately by the error bars, since, if some points have smaller error bars than others the fit should be closer to those points. The quantity to be minimized, called "chi-squared"[2], which is written mathematically as χ^2, is defined by

$$\boxed{\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2 .} \tag{64}$$

Often we assume that the distribution of the errors is Gaussian, since, according to the central limit theorem discussed in Appendix A the sum of N independent random variables has a Gaussian distribution (under fairly general conditions) if N is large. However, distributions which occur in nature usually have more weight in the "tails" than a Gaussian, and as a result, even for moderately large values of N, the probability of an "outlier" might be much bigger than expected from a Gaussian, see Fig. 4.

If the errors *are* distributed with a Gaussian distribution, χ^2 in Eq. (64) is a sum of squares of N random variables with a Gaussian distribution with mean zero and standard deviation unity.

[2] χ^2 should be thought of as a single variable rather than the square of something called χ. This notation is standard.

However, when we have minimized the value of χ^2 with respect to the M fitting parameters a_α the terms are not all independent. It turns out, see Appendix B, that the distribution of χ^2 at the minimum is that of the sum of the squares of $N - M$ (not N) Gaussian random variable with zero mean and standard deviation unity[3]. We call $N - M$ the "number of degrees of freedom" ($N_{\rm DOF}$). The χ^2 distribution is described in Appendix C, see in particular Eq. (C6).

The simplest cases are where the fitting function is a *linear function of the parameters*. We shall call this a *linear model*. Examples are a straight line ($M = 2$),

$$y = a_0 + a_1 x \,, \tag{65}$$

and an m-th order polynomial ($M = m + 1$),

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m = \sum_{\alpha=0}^{m} a_\alpha x^m \,, \tag{66}$$

where the parameters to be adjusted are the a_α. (Note that we are *not* stating here that y has to be a linear function of x, only of the fit parameters a_α.)

An example where the fitting function depends *non*-linearly on the parameters is

$$y = a_0 x^{a_1} + a_2 \,. \tag{67}$$

Linear models are fairly simply because, as we shall see, the parameters are determined by *linear* equations, which, in general, have a unique solution, and this can be found by straightforward methods. However, for fitting functions which are non-linear functions of the parameters, the resulting equations are *non-linear* which may have many solutions or none at all, and so are much less straightforward to solve. We shall discuss fitting to both linear and non-linear models in these notes.

Sometimes a non-linear model can be transformed into a linear model by a change of variables. For example, if we want to fit to

$$y = a_0 x^{a_1} \,, \tag{68}$$

which has a non-linear dependence on a_1, taking logs gives

$$\ln y = \ln a_0 + a_1 \ln x \,, \tag{69}$$

[3] Strictly speaking, this result is only if the fitting model is linear in the parameters, but is usually taken to be a reasonable approximation for non-linear models as well.

which is a *linear* function of the parameters $a_0' = \ln a_0$ and a_1. Fitting a straight line to a log-log plot is a very common procedure in science and engineering. However, it should be noted that transforming the data does not exactly take Gaussian errors into Gaussian errors, though the difference will be small if the errors are "sufficiently small". For the above log transformation this means $\sigma_i/y_i \ll 1$, i.e. the *relative* error is much less than unity.

A. Fitting to a straight line

To see how least squares fitting works, consider the simplest case of a straight line fit, Eq. (65), for which we have to minimize

$$\chi^2(a_0, a_1) = \sum_{i=1}^{N} \left(\frac{y_i - a_0 - a_1 x_i}{\sigma_i} \right)^2 , \tag{70}$$

with respect to a_0 and a_1. Differentiating χ^2 with respect to these parameters and setting the results to zero gives

$$a_0 \sum_{i=1}^{N} \frac{1}{\sigma_i^2} + a_1 \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}, \tag{71a}$$

$$a_0 \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} + a_1 \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} = \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}. \tag{71b}$$

We write this as

$$U_{00}\, a_0 + U_{01}\, a_1 = v_0, \tag{72a}$$

$$U_{10}\, a_0 + U_{11}\, a_1 = v_1, \tag{72b}$$

where

$$\boxed{U_{\alpha\beta} = \sum_{i=1}^{N} \frac{x_i^{\alpha+\beta}}{\sigma_i^2},} \quad \text{and} \tag{73}$$

$$\boxed{v_\alpha = \sum_{i=1}^{N} \frac{y_i\, x_i^\alpha}{\sigma_i^2}.} \tag{74}$$

The matrix notation, while an overkill here, will be convenient later when we do a general polynomial fit. Note that $U_{10} = U_{01}$. (More generally, later on, U will be a symmetric matrix). Equations (72) are two linear equations in two unknowns. These can be solved by eliminating one variable, which immediately gives an equation for the second one. The solution can also be determined from

$$\boxed{a_\alpha = \sum_{\beta=0}^{m} \left(U^{-1}\right)_{\alpha\beta}\, v_\beta,} \tag{75}$$

(where we have temporarily generalized to a polynomial of order m). For the straight-line fit, the inverse of the 2×2 matrix U is given, according to standard rules, by

$$U^{-1} = \frac{1}{\Delta} \begin{pmatrix} U_{11} & -U_{01} \\ -U_{01} & U_{00} \end{pmatrix} \tag{76}$$

where

$$\boxed{\Delta = U_{00}U_{11} - U_{01}^2,} \tag{77}$$

and we have noted that U is symmetric so $U_{01} = U_{10}$. The solution for a_0 and a_1 is therefore given by

$$\boxed{a_0 = \frac{U_{11}\, v_0 - U_{01}\, v_1}{\Delta},} \tag{78a}$$

$$\boxed{a_1 = \frac{-U_{01}\, v_0 + U_{00}\, v_1}{\Delta}.} \tag{78b}$$

We see that it is straightforward to determine the slope, a_1, and the intercept, a_0, of the fit from Eqs. (73), (74), (77) and (78) using the N data points (x_i, y_i), and their error bars σ_i.

B. Fitting to a polynomial

Frequently we need to fit to a higher order polynomial than a straight line, in which case we minimize

$$\chi^2(a_0, a_1, \cdots, a_m) = \sum_{i=1}^{N} \left(\frac{y_i - \sum_{\alpha=0}^{m} a_\alpha x_i^\alpha}{\sigma_i} \right)^2 \tag{79}$$

with respect to the $(m+1)$ parameters a_α. Setting to zero the derivatives of χ^2 with respect to the a_α gives

$$\boxed{\sum_{\beta=0}^{m} U_{\alpha\beta}\, a_\beta = v_\alpha,} \tag{80}$$

where $U_{\alpha\beta}$ and v_α have been defined in Eqs. (73) and (74). Eq. (80) represents $m+1$ *linear* equations, one for each value of α. Their solution is given formally by Eq. (75), which is expressed in terms of the inverse matrix U^{-1}.

In addition to the best fit values of the parameters we also need to determine the error bars in those values. Interestingly, this information is *also* contained in the matrix U^{-1}.

To derive the error bars in the fit parameters we start by noting that, if the y_i change by an amount δy_i, then a_α changes by an amount

$$\delta a_\alpha = \sum_{i=1}^{N} \frac{\partial a_\alpha}{\partial y_i}\, \delta y_i \,. \tag{81}$$

Averaging over fluctuations in the y_i we get the variance of the value of a_α to be

$$\sigma_\alpha^2 \equiv \langle \delta a_\alpha^2 \rangle = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial a_\alpha}{\partial y_i} \right)^2 , \tag{82}$$

where we used that the data points y_i are statistically independent. Writing Eq. (75) explicitly in terms of the data values,

$$a_\alpha = \sum_\beta \left(U^{-1}\right)_{\alpha\beta} \sum_{i=1}^{N} \frac{y_i\, x_i^\beta}{\sigma_i^2} \,, \tag{83}$$

and noting that U is independent of the y_i, we get

$$\frac{\partial a_\alpha}{\partial y_i} = \sum_\beta \left(U^{-1}\right)_{\alpha\beta} \frac{x_i^\beta}{\sigma_i^2} \,. \tag{84}$$

Substituting into Eq. (82) gives

$$\sigma_\alpha^2 = \sum_{\beta,\gamma} \left(U^{-1}\right)_{\alpha\beta} \left(U^{-1}\right)_{\alpha\gamma} \left[\sum_{i=1}^{N} \frac{x_i^{\beta+\gamma}}{\sigma_i^2} \right] . \tag{85}$$

The term in rectangular brackets is just $U_{\beta\gamma}$ and so the last equation reduces to

$$\boxed{\sigma_\alpha^2 = \left(U^{-1}\right)_{\alpha\alpha} \,,} \tag{86}$$

where we remind the reader that the elements of the matrix U are given by Eq. (73). We emphasize that the uncertainty in a_α is σ_α, not the square of this.

We also need a parameter to describe the quality of the fit. A useful quantity is the probability that, given the fit, the data could have occurred with a χ^2 greater than or equal to the value found. This is generally denoted by Q and is given by Eq. (C9) assuming the data have Gaussian noise. Note that the effects of *non-Gaussian* statistics is to increase the probability of outliers, so fits with a fairly small value of Q, say around 0.01, may be considered acceptable. However, fits with a *very* small value of Q should not be trusted and the values of the fit parameters are probably meaningless in these cases.

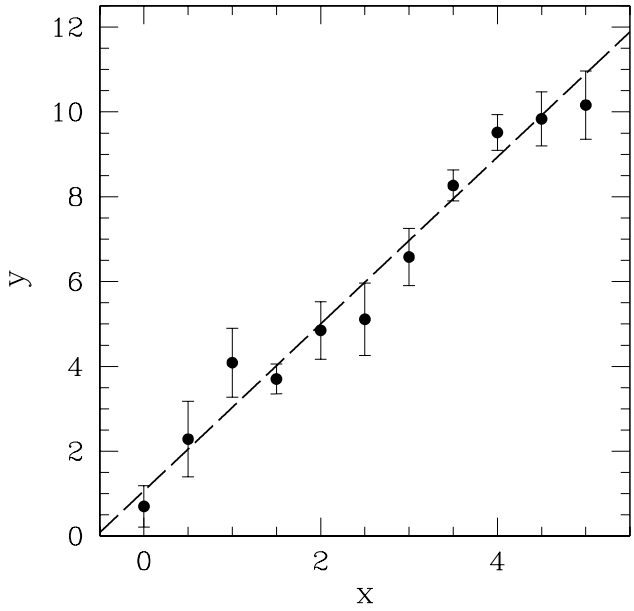

FIG. 2: An example of a straight-line fit to a set of data with error bars.

For the case of a straight line fit, the inverse of U is given explicitly in Eq. (76). Using this information, and the values of (x_i, y_i, σ_i) for the data in Fig. 2, I find that the fit parameters (assuming a straight line fit) are

$$a_0 = 0.84 \pm 0.32, \tag{87}$$

$$a_1 = 2.05 \pm 0.11, \tag{88}$$

in which the error bars on the fit parameters on a_0 and a_1, which are denoted by σ_0 and σ_1, are determined from Eq. (86). I had generated the data by starting with $y = 1 + 2x$ and adding some noise with zero mean. Hence the fit should be consistent with $y = 1 + 2x$ within the error bars, and it is. The value of χ^2 is 7.44 so $\chi^2/N_{\rm DOF} = 7.44/9 = 0.866$ and the quality of fit parameter, given by Eq. (C9), is $Q = 0.592$ which is good.

We call U^{-1} the "*covariance matrix*". Its off-diagonal elements are also useful since they contain information about correlations between the fitted parameters. More precisely, one can show, following the lines of the above derivation of σ_α^2, that the correlation of fit parameters α and β, known mathematically as their "covariance", is given by the appropriate off-diagonal element

of the covariance matrix,

$$\mathrm{Cov}(\alpha,\beta) \equiv \langle \delta a_\alpha \, \delta a_\beta \rangle = \left(U^{-1}\right)_{\alpha\beta} . \tag{89}$$

The correlation coefficient, $r_{\alpha\beta}$, which is a dimensionless measure of the correlation between δa_α and δa_β lying between -1 and 1, is given by

$$r_{\alpha\beta} = \frac{\mathrm{Cov}(\alpha,\beta)}{\sigma_\alpha \sigma_\beta} . \tag{90}$$

A good fitting program should output the correlation coefficients as well as the fit parameters, their error bars, the value of χ^2/N_{DOF}, and the goodness of fit parameter Q.

For a linear model, χ^2 is a quadratic function of the fit parameters and so the elements of the "*curvature matrix*"[4], $(1/2)\,\partial^2\chi^2/\partial a_\alpha \partial a_\beta$ are constants, independent of the values of the fit parameters. In fact, we see from Eqs. (73) and (79) that

$$\frac{1}{2}\,\frac{\partial^2\chi^2}{\partial a_\alpha \partial a_\beta} = U_{\alpha\beta} , \tag{91}$$

so *the curvature matrix is equal to* U, given by Eq. (73) for a polynomial fit.

If we fit to a *general* linear model, writing

$$f(x) = \sum_{\alpha=1}^{M} a_\alpha \, X_\alpha(x), \tag{92}$$

where $X_1(x), X_2(x), \cdots, X_M(x)$ a fixed functions of x called basis functions, the curvature matrix is given by

$$\boxed{U_{\alpha\beta} = \sum_{i=1}^{N} \frac{X_\alpha(x_i)\, X_\beta(x_i)}{\sigma_i^2} .} \tag{93}$$

Similarly, the quantities v_α in Eq. (74) become

$$v_\alpha = \sum_{i=1}^{N} \frac{y_i \, X_\alpha(x_i)}{\sigma_i^2} , \tag{94}$$

for a general set of basis functions.

C. Fitting to a non-linear model

As for linear models, one minimizes χ^2 in Eq. (64). The difference is that the resulting equations are non-linear so there might be many solutions or non at all. Techniques for solving the coupled

[4] It is conventional to include the factor of 1/2.

non-linear equations invariably require specifying an initial value for the variables a_α. The most common method for fitting to non-linear models is the Levenberg-Marquardt (LM) method, see e.g. Numerical Recipes [2]. Implementing the Numerical Recipes code for LM is a little complicated because it requires the user to provide a routine for the derivatives of χ^2 with respect to the fit parameters as well as for χ^2 itself, and to check for convergence. Alternatively, one can use the fitting routines in the scipy package of python or use gnuplot. (For the latter, see the comment in Appendix D about getting the error bars in the parameters correct. This applies when fitting to linear as well as non-linear models.)

One difference from fitting to a linear model is that the curvature matrix, defined by the LHS of Eq. (91) is not constant but is a function of the fit parameters. Nonetheless, the covariance matrix, which is the inverse of the curvature matrix *at the minimum of* χ^2, still gives information about error bars on the fit parameters. This is discussed more in the next subsection, in which we point out, however, that,

> in general, error bars for non-linear models should be determined by investigating the *global* behavior of χ^2 as each of the parameters is varied in turn, rather than getting them from the *local* curvature at the minimum.

As a reminder:

- The *curvature matrix* is defined in general by the LHS of Eq. (91), which, for a linear model is equivalent to Eq. (93) (Eq. (73) for a polynomial fit.)

- The *covariance matrix* is the inverse of the curvature matrix at the minimum of χ^2 (the last remark being only needed for a non-linear model). Its diagonal elements give error bars on the fit parameters according to Eq. (86) (but see the indented caveat in the previous paragraph for non-linear models) and its off-diagonal elements give correlations between fit parameters according to Eqs. (89) and (90).

D. Confidence limits

In the last two subsections we showed that the diagonal elements of the covariance matrix give an error bar on the fit parameters. In this section we discuss more fully what this error bar means and introduce the notion of "confidence limits".

We assume that the data is described by a model with a particular set of parameters $a_\alpha^{\rm true}$ which, unfortunately, we don't know. If we were, somehow, to have many data sets each one would

give a different set of fit parameters $a_\alpha^{(i)}, i = 0, 1, 2, \cdots$, clustered about the true set $a_\alpha^{\rm true}$, because of noise in the data. The scatter in these parameters is a measure of the uncertainty (error bar) in determining these parameters from the *one* data set available to us, which is what we want to know.

We denote here our one data set by $y_i^{(0)}$. This gives one set of fit parameters which we call here $a_\alpha^{(0)}$. Suppose, however, we were to generate *many* simulated data sets *assuming that* $a_\alpha^{(0)}$ *is the true set of parameters*. Fitting each simulated dataset would give different values for the a_α, clustered now about the $a_\alpha^{(0)}$. We now come to an important, but rarely discussed, point:

> We assume that the scatter of the fit parameters of these simulated data sets about $a_\alpha^{(0)}$, which, as we will see, we can calculate from the single set of data available to us, is very similar to the scatter of the fit parameters of real data sets $a_\alpha^{(i)}$ about $a_\alpha^{\rm true}$, which is what we *really* want to know (since it is our estimate of the error bar on a_α) but can't calculate directly.

According to Numerical Recipes [2] there is a theorem which states that if we take simulated data sets assuming the actual parameters are the $a_\alpha^{(0)}$ and the noise is Gaussian, and for each data set compute the fit parameters $a_\alpha^{S(i)}, i = 1, 2, \cdots$, then the probability distribution of a^S is given by the multi-variable Gaussian distribution

$$\boxed{P(\vec{a}^S) \propto \exp\left(-\frac{1}{2}\sum_{\alpha,\beta} \delta a_\alpha^S \, U_{\alpha\beta} \, \delta a_\beta^S\right) ,} \tag{95}$$

where $\delta a_\alpha^S \equiv a_\alpha^{S(i)} - a_\alpha^{(0)}$ and U is the curvature matrix defined in terms of the second derivative of χ^2 at the minimum according to Eq. (91). A proof of this result is given in Appendix E. It applies for a linear model, and also for a non-linear model if the uncertainties in the parameters do not extend outside a region where an effective linear model could be used. (This does not preclude using a non-linear routine to *find* the best parameters.)

From Eq. (91) the change in χ^2 as the parameters are varied away from the minimum is given by

$$\Delta\chi^2 \equiv \chi^2(\{a_\alpha^{S(i)}\}) - \chi^2(\{a_\alpha^{(0)}\}) = \sum_{\alpha,\beta} \delta a_\alpha^S \, U_{\alpha\beta} \, \delta a_\beta^S \, , \tag{96}$$

in which the χ^2 are all evaluated from the single (actual) data set $y_i^{(0)}$. Equation (95) can therefore be written as

$$\boxed{P(\vec{a}^S) \propto \exp\left(-\frac{1}{2}\Delta\chi^2\right) .} \tag{97}$$

Hence the probability of a particular deviation, δa_α^S, of the fit parameters in a simulated data set away from the parameters in the actual data set, depends on how much this change increases χ^2 (evaluated from the actual data set) away from the minimum. In general a "confidence limit" is the range of fit parameter values such that $\Delta\chi^2$ is less than some specified value. The simplest case, and the only one we discuss here, is the variation of *one* variable at a time, though multi-variate confidence limits can also be defined, see Numerical Recipes [2].

We therefore consider the change in χ^2 when one variable, a_1^S say, is held at a specified value, and all the others ($\beta = 2, 3, \cdots, M$) are varied in order to minimize χ^2. Minimizing $\Delta\chi^2$ in Eq. (96) with respect to a_β^S gives

$$\sum_{\gamma=1}^{M} U_{\beta\gamma}\, \delta a_\gamma^S = 0, \qquad (\beta = 2, 3, \cdots, M)\,. \tag{98}$$

The corresponding sum for $\beta = 1$, namely $\sum_{\gamma=1}^{M} U_{1\gamma}\, \delta a_\gamma^S$, is not zero because δa_1 is fixed. It will be some number, c say. Hence we can write

$$\sum_{\gamma=1}^{M} U_{\alpha\gamma}\, \delta a_\gamma^S = c_\alpha, \qquad (\alpha = 1, 2, \cdots, M)\,, \tag{99}$$

where $c_1 = c$ and $c_\beta = 0\,(\beta \neq 1)$. The solution is

$$\delta a_\alpha^S = \sum_{\beta=1}^{M} \left(U^{-1}\right)_{\alpha\beta} c_\beta\,. \tag{100}$$

For $\alpha = 1$ this gives

$$c = \delta a_1^S / \left(U^{-1}\right)_{11}\,. \tag{101}$$

Substituting Eq. (100) into Eq. (96), and using Eq. (101) we find that $\Delta\chi^2$ is related to $\left(\delta a_1^S\right)^2$ by

$$\Delta\chi^2 = \frac{(\delta a_1^S)^2}{(U^{-1})_{11}}. \tag{102}$$

(Curiously, the coefficient of $(\delta a_1)^2$ is one over the 11 element of the inverse of U, rather than U_{11} which is how it appears in Eq. (96) in which the $\beta \neq 1$ parameters are free rather than adjusted to minimize χ^2.)

From Eq. (97) we finally get

$$P(a_1^S) \propto \exp\left(-\frac{1}{2}\,\frac{(\delta a_1^S)^2}{\sigma_1^2}\right)\,, \tag{103}$$

where

$$\sigma_1^2 = \left(U^{-1}\right)_{11}\,. \tag{104}$$

According to the assumption made earlier in this section, we take Eq. (103) to also apply to the probability for $\delta a_1 \equiv a_1^{\rm true} - a_1^{(0)}$, where we remind the reader that $a_1^{(0)}$ is the fit parameter obtained from the actual data. Hence the probability that the true value is equal to $a_1^{\rm true}$ is given by

$$\boxed{P(a_1^{\rm true}) \propto \exp\left(-\frac{1}{2}\,\frac{(\delta a_1)^2}{\sigma_1^2}\right).} \tag{105}$$

This gives $\langle(\delta a_1)^2\rangle = \sigma_1^2 = \left(U^{-1}\right)_{11}$, in agreement with what we found earlier in Eq. (86). We emphasize that Eq. (105) assumes Gaussian noise on the data points, and either the model is linear or, if non-linear, that the range of uncertainty in the parameters is small enough that a description in terms of an effective linear model is satisfactory.

However we have not *just* recovered our earlier result, Eq. (86), by more complicated means since we have gained *additional* information. From the properties of a Gaussian distribution we now know that, from Eq. (105), the probability that a_α lies within one standard deviation σ_α of the value which minimizes χ^2 is 68%, the probability of its being within two standard deviations is 95.5%, and so on. Furthermore, from Eq. (102), which applies to δa_1 as well as δa_1^S according to our assumption, we see that

> *if a single fit parameter is one standard deviation away from its value at the minimum of χ^2 (the other fit parameters being varied to minimize χ^2), then $\Delta\chi^2 = 1$.*

Unfortunately this last sentence, and Eqs. (95) and (105) which are used to derive it, are not valid for a non-linear model if the uncertainties of the parameters extends outside the range where an effective linear model can be used. Nonetheless, we argue that it is still better to use the range where $\delta\chi^2 \leq 1$ as the confidence range for each parameter rather than the error bar determined from the curvature of χ^2 at the minimum.

This is illustrated in Fig. 3. The left hand figure is for a linear model, for which the curve of $\Delta\chi^2$ against δa_1 is a parabola, and the right hand figure is for a non-linear model, for which it is not a parabola though it has a quadratic variation about the minimum shown by the dashed curve.

For the linear case, the values of δa_1 where $\Delta\chi^2 = 1$ are the *same* as the values $\pm\sigma_1$, where σ_1 is the standard error bar obtained from the *local* curvature in the vicinity of the minimum. However, for the non-linear case, the values of δa_1 where $\Delta\chi^2 = 1$ are *different* from $\pm\sigma_1$, and indeed the values on the positive and negative sides, σ_1^+ and σ_1^-, are not equal. For the data in the figure, it is clear that the value of a_1 is more tightly constrained on the positive side than the negative

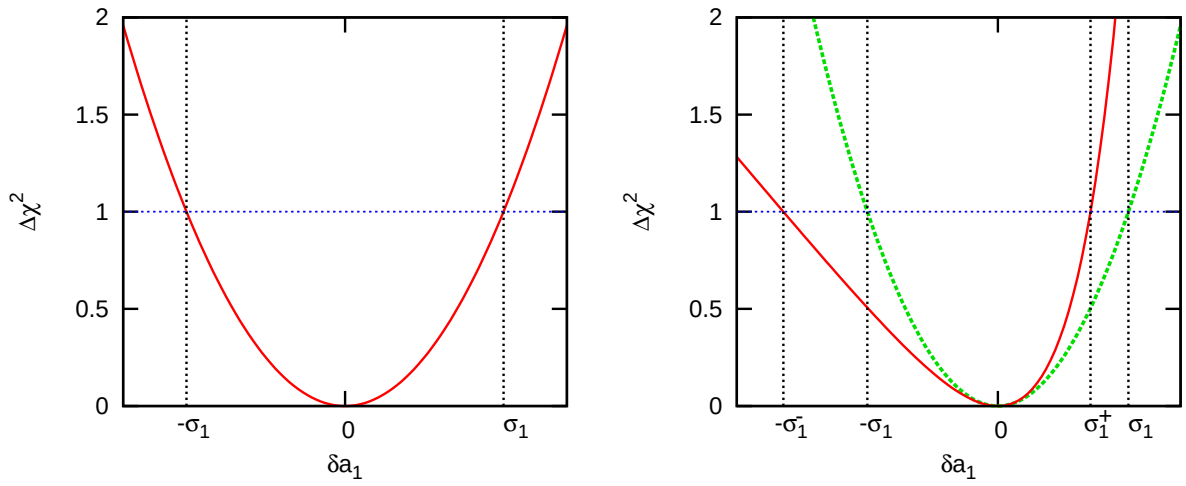

FIG. 3: **Left:** The change in χ^2 as a fit parameter a_1 is varied away from the value that minimizes χ^2 for a *linear* model. The shape is a parabola for which $\Delta\chi^2 = 1$ when $\delta a = \pm\sigma_1$, where σ_1 is the error bar. **Right:** The solid curve is the change in χ^2 for a *non-linear* model. The curve is no longer a parabola and is even non-symmetric. The dashed curve is a parabola which fits the solid curve at the minimum. The fitting program only has information about the *local* behavior at the minimum and so gives an error range $\pm\sigma_1$ where the value of the parabola is 1. However, the parameter a_1 is clearly more tightly constrained on the plus side than on the minus side, and a better way to determine the error range is to look *globally* and locate the values of δa_1 where $\Delta\chi^2 = 1$. This gives an error bar σ_1^+ on the plus side, and a different error bar, σ_1^-, on the minus side, both of which are different from σ_1.

side, and so it is better to give the error bars as $+\sigma_1^+$ and $-\sigma_1^-$, obtained from the range where $\Delta\chi^2 \leq 1$, rather the symmetric range $\pm\sigma_1$.

Based on this example, we conclude that, even if the noise on the data points is not Gaussian, and/or the model is non-linear and the range of parameter uncertainty extends outside the range where an effective linear model is satisfactory, it is still useful to define a "confidence interval" for each parameter as the region where $\Delta\chi^2$ is less than a specified value. However, Eq. (105) is no longer valid so one can not convert the value of $\Delta\chi^2$ to a probability that the true value lies within this range.

If one wants to get a confidence interval *with a specified probability* in the case where the model is non-linear and the range of parameter uncertainty extends outside the range where an effective linear model is adequate (but still assuming that the noise is Gaussian to a good enough approximation) one could generate many additional data sets, assuming the the initial fitted parameters

are the true ones, but with noise on the data. In other words, the y_i are generated with the probability distribution given in the curly brackets in Eq. (E4). One then fits each data set and looks at the distribution of results for the fitted parameters. The confidence interval for a particular fitted parameter could then be the range which includes 68% of the datasets (i.e. 16% missing on either side of the minimum). Unfortunately, this does not seem to be a standard procedure in the statistical physics community.

Appendix A: Central Limit Theorem

In this appendix we give a proof of the central limit theorem.

We assume a distribution that falls off sufficiently fast at $\pm\infty$ that the mean and variance are finite. This *excludes*, for example, the Lorentzian distribution:

$$P_{\rm Lor} = \frac{1}{\pi}\frac{1}{1+x^2}\,. \tag{A1}$$

A common distribution which *does* have a finite mean and variance is the Gaussian distribution,

$$P_{\rm Gauss} = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]\,. \tag{A2}$$

Using standard results for Gaussian integrals you should be able to show that the distribution is normalized and that the mean and standard deviation are μ and σ respectively.

Consider a distribution, *not necessarily Gaussian*, with a finite mean and distribution. We pick N random numbers x_i from such a distribution and form the sum

$$X = \sum_{i=1}^{N} x_i.$$

The determination of the distribution of X, which we call $P_N(X)$, will involve the Fourier transform of $P(x)$, called the "characteristic function" in the context of probability theory. This is defined by

$$Q(k) = \int_{-\infty}^{\infty} P(x)e^{ikx}\,dx\,.$$

Expanding out the exponential we can write $Q(k)$ in terms of the moments of $P(x)$

$$Q(k) = 1 + ik\langle x\rangle + \frac{(ik)^2}{2!}\langle x^2\rangle + \frac{(ik)^3}{3!}\langle x^3\rangle + \cdots\,.$$

It will be convenient in what follows to write $Q(k)$ as an exponential, i.e.

$$\begin{aligned} Q(k) &= \exp\left[\ln\left(1 + ik\langle x\rangle + \frac{(ik)^2}{2!}\langle x^2\rangle + \frac{(ik)^3}{3!}\langle x^3\rangle + \cdots\right)\right] \\ &= \boxed{\exp\left[ik\mu - \frac{k^2\sigma^2}{2!} + \frac{c_3(ik)^3}{3!} + \frac{c_4(ik)^4}{4!} + \cdots\right]\,,} \end{aligned} \tag{A3}$$

where c_3 involves third and lower moments, c_4 involves fourth and lower moments, and so on. In general, the c_n are *cumulant* averages.

For the important case of a Gaussian, the Fourier transform is obtained by "completing the square". The result is that the Fourier transform of a Gaussian is also a Gaussian, namely,

$$\boxed{Q_{\rm Gauss}(k) = \exp\left[ik\mu - \frac{k^2\sigma^2}{2}\right],} \tag{A4}$$

showing that the higher order cumulants, c_3, c_4, etc. in Eq. (A3) *all vanish* for a Gaussian.

The distribution $P_N(x)$ can be expressed as

$$P_N(x) = \int_{-\infty}^{\infty} P(x_1)dx_1 \int_{-\infty}^{\infty} P(x_2)dx_2 \cdots \int_{-\infty}^{\infty} P(x_N)dx_N\, \delta(X - \sum_{i=1}^{N} x_i)\,. \tag{A5}$$

We evaluate this by using the integral representation of the delta function

$$\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ikx}\, dk\,, \tag{A6}$$

so

$$P_N(X) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} \int_{-\infty}^{\infty} P(x_1)dx_1 \int_{-\infty}^{\infty} P(x_2)dx_2 \cdots \int_{-\infty}^{\infty} P(x_N)dx_N\, \exp[ik(x_1 + x_2 + \cdots x_N - X)] \tag{A7}$$

$$= \int_{-\infty}^{\infty} \frac{dk}{2\pi} Q(k)^N e^{-ikX}\,, \tag{A8}$$

showing that the Fourier transform of $P_N(x)$, which we call $Q_N(k)$, is given by

$$\boxed{Q_N(k) = Q(k)^N\,.} \tag{A9}$$

Consequently

$$Q_N(k) = \exp\left[ikN\mu - \frac{Nk^2\sigma^2}{2} + \frac{Nc_3(ik)^3}{4!} + \frac{Nc_4(ik)^4}{4!} + \cdots\right]. \tag{A10}$$

Comparing with Eq. (A3) we see that

> the mean of the distribution of the sum of N variables (the coefficient of $-ik$ in the exponential) is N times the mean of the distribution of one variable, and the variance of the distribution of the sum of N variables (the coefficient of $-k^2/2!$) is N times the variance of the distribution of one variable.

These are general statements applicable for *any* N and have already been derived in Sec. II A.

However, if N is *large* we can now go further. The distribution $P_N(X)$ is the inverse transform of $Q_N(k)$, see Eq. (A8), so

$$P_N(X) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left[-ikX' - \frac{Nk^2\sigma^2}{2!} + N\frac{c_3(ik)^3}{3!} + \frac{Nc_4(ik)^4}{4!} + \cdots\right] dk\,, \tag{A11}$$

where

$$X' = X - N\mu\,. \tag{A12}$$

Looking at the $-Nk^2/2$ term in the exponential in Eq. (A11), we see that the integrand is significant for $k < k^\star$, where $N\sigma^2(k^\star)^2 = 1$, and negligibly small for $k \gg k^\star$. However, for $0 < k < k^\star$ the higher order terms in Eq. (A11), (i.e. those of order k^3, k^4 etc.) are very small since $N(k^\star)^3 \sim N^{-1/2}, N(k^\star)^4 \sim N^{-1}$ and so on. Hence the terms of higher order than k^2 in Eq. (A11), do not contribute for large N and so

$$\lim_{N\to\infty} P_N(X) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left[-ikX' - \frac{Nk^2\sigma^2}{2}\right] dk\,. \tag{A13}$$

In other words, for large N the distribution is the Fourier transform of a Gaussian, which, as we know, is also a Gaussian. Completing the square in Eq. (A13) gives

$$\begin{aligned}\lim_{N\to\infty} P_N(X) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left[-\frac{N\sigma^2}{2}\left(k - \frac{iX'}{N\sigma^2}\right)^2\right] dk\ \exp\left[-\frac{(X')^2}{2N\sigma^2}\right] \\ &= \boxed{\frac{1}{\sqrt{2\pi N}\,\sigma}\exp\left[-\frac{(X-N\mu)^2}{2N\sigma^2}\right]}\,,\end{aligned} \tag{A14}$$

where, in the last line, we used Eq. (A12). This is a Gaussian with mean $N\mu$ and variance $N\sigma^2$. Eq. (A14) is the central limit theorem in statistics. It tells us that,

> for $N \to \infty$, the distribution of the sum of N variables is a *Gaussian* of mean N times the mean, μ, of the distribution of one variable, and variance N times the variance of the distribution of one variable, σ^2, independent of the form of the distribution of one variable, $P(x)$, provided only that μ and σ are finite.

The central limit theorem is of such generality that it is extremely important. It is the reason why the Gaussian distribution has such a preeminent place in the theory of statistics.

Note that if the distribution of the individual x_i is Gaussian, then the distribution of the sum of N variables is *always* Gaussian, even for N small. This follows from Eq. (A9) and the fact that the Fourier transform of a Gaussian is a Gaussian.

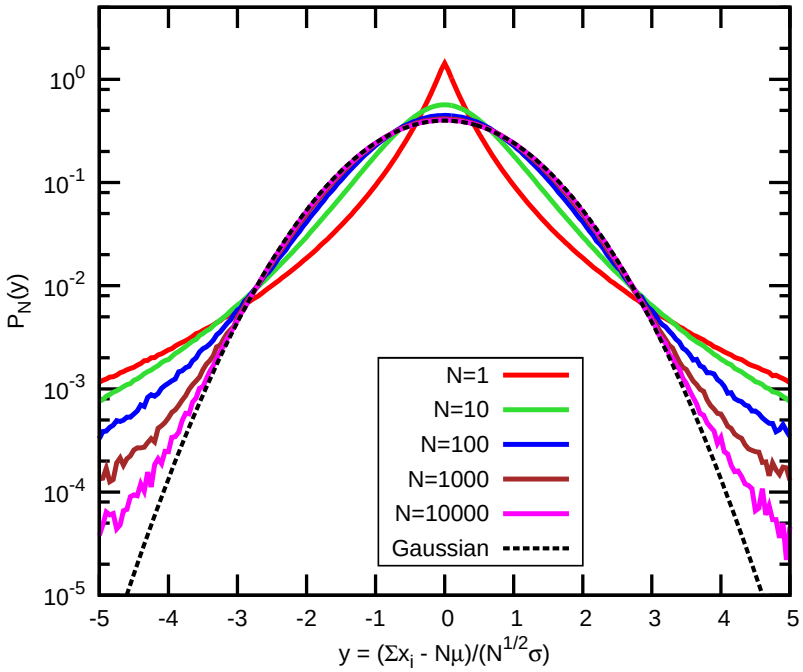

FIG. 4: Figure showing the approach to the central limit theorem for the distribution in Eq. (A15), which has mean, μ, equal to 0, and standard deviation, σ, equal to 1. The horizontal axis is the sum of N random variables divided by $\sqrt{N}$ which, for all N, has zero mean and standard deviation unity. For large N the distribution approaches a Gaussian. However, convergence is non-uniform, and is extremely slow in the tails.

In practice, distributions that we meet in nature, have a much broader tail than that of the Gaussian distribution which falls off very fast at large $|x-\mu|/\sigma$. As a result, even if the distribution of the sum approximates well a Gaussian in the central region for only modest values of N, it might take a much larger value of N to beat down the weight in the tail to the value of the Gaussian. Hence, even for moderate values of N, the probability of a deviation greater than σ can be significantly larger than that of the Gaussian distribution which is 32%. This caution will be important in Sec. III when we discuss the quality of fits.

We will illustrate the slow convergence of the distribution of the sum to a Gaussian in Fig. (4), in which the distribution of the individual variables x_i is

$$P(x) = \frac{3}{2}\frac{1}{(1+|x|)^4}\,. \tag{A15}$$

This has mean 0 and standard deviation 1, but moments higher than the second do not exist because the integrals diverge. For large N the distribution approaches a Gaussian, as expected, but convergence is very slow in the tails.

Appendix B: The number of degrees of freedom

Consider, for simplicity, a straight line fit, so we have to determine the values of a_0 and a_1 which minimize Eq. (70). The N terms in Eq. (70) are not statistically independent at the minimum because the values of a_0 and a_1, given by Eq. (71), depend on the data points (x_i, y_i, σ_i).

Consider the "residuals" defined by

$$\epsilon_i = \frac{y_i - a_0 - a_1 x_i}{\sigma} \,. \tag{B1}$$

If the model were exact and we use the exact values of the parameters a_0 and a_1 the ϵ_i would be independent and each have a Gaussian distribution with zero mean and standard deviation unity.

However, choosing the *best-fit* values of a_0 and a_1 *from the data* according to Eq. (71) implies that

$$\sum_{i=1}^{N} \frac{1}{\sigma_i} \epsilon_i = 0 \,, \tag{B2a}$$

$$\sum_{i=1}^{N} \frac{x_i}{\sigma_i} \epsilon_i = 0 \,, \tag{B2b}$$

which are are two *linear constraints* on the ϵ_i. This means that we only need to specify $N-2$ of them and we know them all. In the N dimensional space of the ϵ_i we have eliminated two directions, so there can be no Gaussian fluctuations along them. However the other $N-2$ dimensions are unchanged, and will have the same Gaussian fluctuations as before. Thus χ^2 has the distribution of a sum of squares of $N-2$ Gaussian random variables. This makes sense since, if there were only two points, the fit would go perfectly through them, so $\chi^2 = 0$ in this case.

Clearly this argument can be generalized to any fitting function which depends *linearly* on M fitting parameters, with the result that χ^2 has the distribution of a sum of squares of $N_{\rm DOF} = N - M$ Gaussian random variables, in which the quantity $N_{\rm DOF}$ is called the "number of degrees of freedom".

Even if the fitting function depends non-linearly on the parameters, this last result is often taken as a reasonable approximation.

Appendix C: The χ^2 distribution and the goodness of fit parameter Q

The χ^2 distribution for m degrees of freedom is the distribution of the sum of m independent random variables with a Gaussian distribution with zero mean and standard deviation unity. To determine this we write

$$P(x_1, x_2, \cdots, x_m)\, dx_1 dx_2 \cdots dx_m = \frac{1}{(2\pi)^{m/2}}\, e^{-x_1^2/2}\, e^{-x_2^2/2} \cdots e^{-x_m^2/2}\, dx_1 dx_2 \cdots dx_m\,, \tag{C1}$$

convert to polar coordinates, and integrate over directions, which gives the distribution of the radial variable to be

$$\widetilde{P}(r)\, dr = \frac{S_m}{(2\pi)^{m/2}}\, r^{m-1}\, e^{-r^2/2}\, dr\,, \tag{C2}$$

where S_m is the surface area of a unit m-dimensional sphere. To determine S_m we integrate Eq. (C2) over r, noting that $\widetilde{P}(r)$ is normalized, which gives

$$S_m = \frac{2\pi^{m/2}}{\Gamma(m/2)}\,, \tag{C3}$$

where $\Gamma(x)$ is the Euler gamma function defined by

$$\Gamma(x) = \int_0^\infty t^{x-1}\, e^{-t}\, dt\,. \tag{C4}$$

From Eqs. (C2) and (C3) we have

$$\widetilde{P}(r) = \frac{1}{2^{m/2-1}\Gamma(m/2)}\, r^{m-1} e^{-r^2/2}\,. \tag{C5}$$

This is the distribution of r but we want the distribution of $\chi^2 \equiv \sum_i x_i^2 = r^2$. To avoid confusion of notation we write X for χ^2, and define the χ^2 distribution for m variables as $P^{(m)}(X)$. We have $P^{(m)}(X)\, dX = \widetilde{P}(r)\, dr$ so the χ^2 distribution for m degrees of freedom is

$$\begin{aligned} P^{(m)}(X) &= \frac{\widetilde{P}(r)}{dX/dr} \\ &\boxed{= \frac{1}{2^{m/2}\Gamma(m/2)}\, X^{(m/2)-1}\, e^{-X/2} \qquad (X > 0)\,.} \end{aligned} \tag{C6}$$

The χ^2 distribution is zero for $X < 0$. Using Eq. (C4) and the property of the gamma function that $\Gamma(n+1) = n\Gamma(n)$ one can show that

$$\int_0^\infty P^{(m)}(X)\, dX = 1\,, \tag{C7a}$$

$$\langle X \rangle \equiv \int_0^\infty X\, P^{(m)}(X)\, dX = m\,, \tag{C7b}$$

$$\langle X^2 \rangle \equiv \int_0^\infty X^2\, P^{(m)}(X)\, dX = m^2 + 2m\,, \quad \text{so} \tag{C7c}$$

$$\langle X^2 \rangle - \langle X \rangle^2 = 2m\,. \tag{C7d}$$

From Eqs. (C7b) and (C7d) we see that typically χ^2 lies in the range $m - \sqrt{2m}$ to $m + \sqrt{2m}$. For large m the distribution approaches a Gaussian according to the central limit theory discussed in Appendix A. Typically one focuses on the value of χ^2 per degree freedom since this should be around unity independent of m.

The goodness of fit parameter is the probability that the specified value of χ^2, or greater, could occur by random chance. From Eq. (C6) it is given by

$$Q = \frac{1}{2^{m/2}\Gamma(m/2)} \int_{\chi^2}^{\infty} X^{(m/2)-1}\, e^{-X/2}\, dX\,, \tag{C8}$$

$$\boxed{= \frac{1}{\Gamma(m/2)} \int_{\chi^2/2}^{\infty} y^{(m/2)-1}\, e^{-y}\, dy\,,} \tag{C9}$$

which is known as an incomplete gamma function. Code to generate the incomplete gamma function is given in Numerical Recipes [2]. There is also a built-in function to generate the goodness of fit parameter in the scipy package of python and in the graphics program gnuplot. Note that $Q = 1$ for $\chi^2 = 0$ and $Q \to 0$ for $\chi^2 \to \infty$. Remember that m is the number of degrees of freedom, written as $N_{\rm DOF}$ elsewhere in these notes.

Appendix D: Asymptotic standard error

Sometimes one does not have error bars on the data. Nonetheless, one can still use χ^2 fitting to get an *estimate* of those errors (assuming that they are all equal) and thereby also get an error bar on the fit parameters. The latter is called the "asymptotic standard error". Assuming the same error bar $\sigma_{\rm ass}$ for all points, we determine $\sigma_{\rm ass}$ from the requirement that χ^2 per degree of freedom is precisely one, i.e. the mean value according to Eq. (C7b). This gives

$$1 = \frac{\chi^2}{N_{\rm DOF}} = \frac{1}{N_{\rm DOF}} \sum_{i=1}^{N} \left(\frac{y_i - f(x_i)}{\sigma_{\rm ass}} \right)^2\,, \tag{D1}$$

or, equivalently,

$$\boxed{\sigma_{\rm ass}^2 = \frac{1}{N_{\rm DOF}} \sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2\,.} \tag{D2}$$

The error bars on the fit parameters are then obtained from Eq. (86), with the elements of U given by Eq. (73) in which σ_i is replaced by $\sigma_{\rm ass}$. Equivalently, one can set the σ_i to unity in determining U from Eq. (73), and estimate the error from

$$\boxed{\sigma_\alpha^2 = (U)^{-1}_{\alpha\alpha}\, \sigma_{\rm ass}^2\,,} \quad \text{(asymptotic standard error).} \tag{D3}$$

A simple example of the use of the asymptotic standard error in a situation where we don't know the error on the data points, is fitting to a constant, i.e. *determining the average of a set of data*, which we discussed in detail in Sec. II. In this case we have

$$U_{00} = N, \qquad v_0 = \sum_{i=1}^{N} y_i, \tag{D4}$$

so the only fit parameter is

$$a_0 = \frac{v_0}{U_{00}} = \frac{1}{N} \sum_{i=1}^{N} y_i = \overline{y}, \tag{D5}$$

which gives, naturally enough, the average of the data points, $\overline{y}$. The number of degrees of freedom is $N - 1$, since there is one fit parameter, so

$$\sigma^2_{\rm ass} = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \overline{y})^2 \ , \tag{D6}$$

and hence the square of the error on a_0 is given, from Eq. (D3), by

$$\sigma_0^2 = \frac{1}{U_{00}}\, \sigma^2_{\rm ass} = \frac{1}{N(N-1)} \sum_{i=1}^{N} (y_i - \overline{y})^2 \ , \tag{D7}$$

which is precisely the expression for the error in the mean of a set of data given in Eq. (15).

As an aside, I mention that a popular plotting program, gnuplot, which also does fits, presents error bars on the fit parameters in a confusing way. Even if there are error bars on the points, gnuplot presents the "asymptotic standard error" on the fit parameters. Gnuplot calculates the elements of U correctly from Eq. (73) including the error bars, but then apparently also determines an "assumed error" from an expression like Eq. (D2) but including the error bars, i.e.

$$\sigma^2_{\rm ass} = \frac{1}{N_{\rm DOF}} \sum_{i=1}^{N} \left(\ \frac{y_i - f(x_i)}{\sigma_i} \ \right)^2 \ = \ \frac{\chi^2}{N_{\rm DOF}}, \qquad \text{(gnuplot)}\,. \tag{D8}$$

Hence gnuplot's $\sigma^2_{\rm ass}$ is just the chi-squared per degree of freedom. The error bar (squared) quoted by gnuplot is $(U)^{-1}_{\alpha\alpha}\ \sigma^2_{\rm ass}$, as in Eq. (D3). However, this is wrong since the error bars on the data points have already been included in calculating the elements of U, so the error on the fit parameter α should be $(U)^{-1}_{\alpha\alpha}$. Hence,

> to get the correct error bars on fit parameters from gnuplot, you have to divide gnuplot's asymptotic standard errors by the square root of the chi-squared per degree of freedom (which, fortunately, gnuplot gives correctly).

Appendix E: Derivation of Eq. (95)

The fitting function to an arbitrary linear model is

$$f(x) = \sum_{\alpha=1}^{M} a_\alpha X_\alpha(x), \tag{E1}$$

where the $X_\alpha(x)$ are the M basis functions, and the a_α are the fitted parameters. These parameters are given by the solution of the linear equations

$$\sum_{\beta=1}^{M} U_{\alpha\beta}\, a_\beta = v_\alpha\, a_\alpha\,, \tag{E2}$$

where

$$U_{\alpha\beta} = \sum_{i=1}^{N} \frac{X_\alpha(x_i)\, X_\beta(x_i)}{\sigma_i^2}\,. \tag{E3a}$$

$$v_\alpha = \sum_{i=1}^{N} \frac{y_i\, X_\alpha(x_i)}{\sigma_i^2}\,. \tag{E3b}$$

We do a fit for *one* set of y-values, $y_i^{(0)}$, and obtain a set of fit parameters $a_\alpha^{(0)}$. We then assume that the $a_\alpha^{(0)}$ are the correct parameters and generate an *ensemble* of simulated data sets based on this assumption, using the error bars on the data and assuming that the noise is Gaussian. We ask what is the probability that the fit to one of the new data sets has parameters $\vec{a}$.

This probability distribution is given by

$$P(\vec{a}) = \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi}\sigma_i} \int_{-\infty}^{\infty} dy_i\, \exp\left[-\frac{\left(y_i - \vec{a}^{(0)} \cdot \vec{X}(x_i) \right)^2}{2\sigma_i^2} \right] \right\} \prod_{\alpha=1}^{M} \delta\left(\sum_\beta U_{\alpha\beta} a_\beta - v_\alpha \right) \det U\,, \tag{E4}$$

where the factor in curly brackets is (an integral over) the probability distribution of the data points y_i, and the delta functions project out those sets of data points which have a particular set of fitted parameters. The factor of $\det U$ is a Jacobian to normalize the distribution. Using the integral representation of the delta function, and writing explicitly the expression for v_α from Eq. (E3b), one has

$$P(\vec{a}) = \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi}\sigma_i} \int_{-\infty}^{\infty} dy_i\, \exp\left[-\frac{\left(y_i - \vec{a}^{(0)} \cdot \vec{X}(x_i) \right)^2}{2\sigma_i^2} \right] \right\} \times \tag{E5}$$

$$\prod_{\alpha=1}^{M} \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} dk_\alpha \exp\left[ik_\alpha \left(\sum_\beta U_{\alpha\beta} a_\beta - \sum_{i=1}^{N} \frac{y_i\, X_\alpha(x_i)}{\sigma_i^2} \right) \right] \right) \det U\,. \tag{E6}$$

We carry out the y integrals by "completing the square"

$$P(\vec{a}) = \prod_{\alpha=1}^{M} \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} dk_\alpha \right) \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi}\sigma_i} \int_{-\infty}^{\infty} dy_i \, \exp\left[-\frac{\left(y_i - \vec{a}^{(0)} \cdot \vec{X}(x_i) + i\vec{k} \cdot \vec{X}(x_i) \right)^2}{2\sigma_i^2} \right] \right\} \times \tag{E7}$$

$$\exp\left[-\frac{1}{2\sigma_i^2} \left(\left(\vec{k} \cdot \vec{X}(i) \right)^2 + 2i \left(\vec{k} \cdot \vec{X}(x_i) \right) \left(\vec{a}^{(0)} \cdot \vec{X}(x_i) \right) \right) \right] \times \exp\left[i \sum_{\alpha,\beta} k_\alpha \, U_{\alpha\beta} \, a_\beta \right] \det U \, . \tag{E8}$$

Doing the y-integrals, the factors in curly brackets are equal to unity. Using Eq. (E3a) we then get

$$P(\vec{a}) = \prod_{\alpha=1}^{M} \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} dk_\alpha \right) \exp\left[-\frac{1}{2} \sum_{\alpha,\beta} k_\alpha \, U_{\alpha\beta} \, k_\beta + i \sum_{\alpha,\beta} k_\alpha \, U_{\alpha\beta} \, \delta a_\beta \right] \det U \, , \tag{E9}$$

where

$$\delta a_\beta \equiv a_\beta - a_\beta^{(0)} \, . \tag{E10}$$

The k-integrals are done by working in the basis in which U is diagonal. The result is

$$\boxed{P(\vec{a}) = \frac{(\det U)^{1/2}}{(2\pi)^{m/2}} \exp\left[-\frac{1}{2} \sum_{\alpha,\beta} \delta a_\alpha U_{\alpha\beta} \delta a_\beta \right]} \, , \tag{E11}$$

which is Eq. (95), including the normalization constant in front of the exponential. In Eq. (95) we emphasized that the a_α are fits to simulated data sets by adding a superscript S. Here we have omitted this superscript to avoid cluttering up the notation too much.

---

[1] V. Ambegaokar and M. Troyer, *Estimating errors reliably in Monte Carlo simulations of the Ehrenfest model*, Am. J. Phys. **78**, 150 (2009).

[2] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C, 2nd Ed.* (Cambridge University Press, Cambridge, 1992).