Assignment 1, due September 9, before class. Correction to Exercise (3d)

Please follow instructions on the assignments page

About this assignment: This assignment is review of background material that is prerequisite for this course. This includes familiarity with relevant parts of Python as well as background mathematics and elementary statistics.

1. (*An exercise in calculus*) Suppose a dataset consists of N measurements X_k for $k = 1, \ldots, N$. The sample mean is

$$\overline{X} = \frac{1}{N} \sum_{k=1}^{N} X_k \ .$$

The sample variance is (pay attention to the factor N rather then $N-1$ in the denominator)

$$s^2 = \frac{1}{N} \sum_{k=1}^{N} \left(X_k - \overline{X} \right)^2 \ . \tag{1}$$

The likelihood function is

$$L(\mu, \sigma^2, \text{data}) = \left(2\pi\sigma^2 \right)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{k=1}^{N} \left(X_k - \mu \right)^2 \right) \tag{2}$$

The first class covers the derivation of (2), but that is not important for this exercise. Show that

$$\overline{X} = \arg \max_{\mu} L(\mu, \sigma^2, \text{data}) \ ,$$
$$s^2 = \arg \max_{\sigma^2} L(\mu, \sigma^2, \text{data}) \ .$$

Hint. I find it easier to express L in terms of the variance, $v = \sigma^2$. Differentiate with respect to v and set the derivative equal to zero. I find it more complicated to differentiate with respect to σ.

2. (*Practice with approximations using derivatives*) [**Instructions and explanation for this exercise:** Please do all the calculations by hand. The AI can do them faster, but you won't learning anything doing that. Please hand in your solution *in your own handwriting* rather than using LaTeX.] Consider the (gaussian) function

$$f(\sigma, x) = \frac{1}{\sigma} e^{-\frac{x^2}{2\sigma^2}} \ .$$

The $(2\pi)^{-\frac{1}{2}}$ has been left off because it's a constant that just gets in the way of the calculations and can easily be included at the end if needed. The exercise involves Taylor approximations of f to varying degrees. Define $\sigma_0 = 1$ and $x_0 = 2$. Nearby parameter values are $\sigma = \sigma_0 + \Delta\sigma$ and $x = x_0 + \Delta x$. The change in f is

$$\Delta f = f(\sigma, x) - f(\sigma_0, x_0) \ .$$

The Taylor approximations of Δf to various orders involve

$$\begin{aligned}
\Delta f_1 &= f_{0\sigma}\Delta\sigma + f_{0x}\Delta x \\
\Delta f_2 &= \frac{1}{2}f_{0\sigma\sigma}\Delta\sigma^2 + f_{0\sigma x}\Delta\sigma\Delta x + \frac{1}{2}f_{0xx}\Delta x^2 \\
\Delta f_3 &= \frac{1}{6}f_{0\sigma\sigma\sigma}\Delta\sigma^3 + \cdots \text{ (all cubic terms)} .
\end{aligned}$$

In the notation $f_{0\sigma}$, the subscript 0 means that the expression is evaluated at σ_0 and x_0. The subscript σ means partial derivative with respect to σ. Other expressions are similar, for example,

$$f_{0\sigma x} = \frac{\partial^2 f(\sigma_0, x_0)}{\partial\sigma\partial x} .$$

(a) Take $\Delta\sigma = .3$ and $\Delta x = .2$. Calculate the exact Δf and compare it to the Δf estimated as $\Delta f \approx \Delta f_1$, and $\Delta f \approx \Delta f_1 + \Delta f_2$, and $\Delta f \approx \Delta f_1 + \Delta f_2 + \Delta f_3$. You will need to use Python scripts to do the arithmetic. Comment on the accuracy of the various approximations.

(b) Now try to choose Δx as a function of $\Delta\sigma$ so that $\Delta f = 0$.

3. (*Practice with probability and analysis*). This exercise illustrates probabilistic modeling of random variables. A model only a little more complicated has been used for daily changes in stock prices. The steps in the analysis are reminders of some ideas and tricks of calculus. In this model, a random variable X is represented as a standard normal random variable Z rescaled by a weight multiplier W. We assume the weight and the standard normal are independent, and that the weight is uniformly distributed in the interval $[0, 1]$. In formulas, $X = WZ$, with $Z \sim \mathcal{N}(0, 1)$ and $W \sim \text{unif}[0, 1]$.

(a) Let $g(w, z)$ be the joint density (PDF) of the two component random variable (W, Z). In formulas: $(W, Z) \sim g(w, z)$. Write a formula for g,

(b) Let $F(x)$ be the cumulative distribution function (CDF) of X. In formulas:

$$F(x) = \Pr(X \le x) .$$

Express $F(x)$ as a double integral involving g over region in (w, z) space. Show that this may be written in the form

$$F(x) = \frac{1}{\sqrt{2\pi}} \int \left(\int_{-\infty}^{b(x,w)} e^{-\frac{1}{2}z^2}\, dz \right) dw .$$

Make this concrete by finding an explicit formula for b.

(c) The PDF of X will be called $f(x)$. This is related to the CDF by $f(x) = \frac{d}{dx}F(x)$. Differentiate the expression from part (b) with respect to x to show that if $x > 0$, then

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \frac{1}{y} e^{-\frac{1}{2}y^2}\, dy . \tag{3}$$

Hint. One way to do this is to first derive $f(x) = \int_0^1 \cdots dw$, and then to do the change of variables $w = \frac{x}{y}$.

(d) There does not seem to be a simple formula for $f(x)$ that doesn't involve integration. It is possible to figure out the behavior of $f(x)$ for small $x > 0$. It might seem that f becomes large for small x because when w is small, multiplying by even a reasonably large Z may give a small result. Also, the integral (3) is infinite if $x = 0$. So, how large is it when x is

only slightly larger than zero? A trick (one of several) for estimating this is integration by parts:

$$\frac{1}{y} = \frac{d}{dy}\log(y) \; .$$

Use this (or another trick you come up with) to show that

Corrected version below $\quad f(x) = \dfrac{1}{\sqrt{2\pi}}\log(x) + \text{ (something bounded) as } x \to 0 \; .$

$$f(x) = -\frac{1}{\sqrt{2\pi}}\log(x) + \text{ (something bounded) as } x \to 0 \; .$$

Any formula for the remainder term (something bounded) will involve an integral that involves x as a parameter. But this integral does not blow up as $x \to 0$. The original formula without the $-$ cannot be right because $\log(x) \to -\infty$ as $x \to 0$. But f is a probability density, which cannot be negative.

4. (*A computational experiment*) This exercise is a computational experiment to determine the distribution of s^2 in (1). Try to understand as much as you can about how the distribution of the random variable s^2 depends on the number of samples N. Be careful, each of the n experiments needs $N \neq n$ independent samples X_k. These are independent for each k and each experiment i. For small N, the distribution is not symmetric about its maximum point. It is less likely to have s^2 smaller than $s^2_{\max}$ than larger (or maybe the other way? The asymmetry becomes less when N is larger. It may be easier to see the asymmetry in the potential than the PDF itself. Experiment with the number of experiments n (the experiment sample size) and the bin size to see how small the bins can be and still get reliable density and potential estimates.

   The general description of histogram density estimation involves n independent samples of a random variable Y with PDF $f(y)$. There are bins of size Δy and bin centers b_j. The bin itself is an interval of length Δy with b_j in the center: $[b_j - \frac{1}{2}\Delta y, \, b_j + \frac{1}{2}\Delta y]$. The samples are $Y_1, \cdots, Y_n$. The *bin count*, N_j, is the number of samples in the j^{th} bin interval

$$N_j = \# \left\{ Y_k \in [b_j - \frac{1}{2}\Delta y, \, b_j + \frac{1}{2}\Delta y] \right\} \; .$$

   The probability that a sample lands in bin j is

$$p_j = \Pr\left(b_j - \frac{1}{2}\Delta y < Y < b_j + \frac{1}{2}\Delta y \right) = \int_{b_j - \frac{1}{2}\Delta y}^{b_j + \frac{1}{2}\Delta y} f(y)\, dy \approx \Delta y f(b_j) \; .$$

   The last approximation is accurate if Δy is small. Evaluating the PDF at the bin center makes the approximation more accurate (draw a picture to see this). If there are n trials, each with probability p_j, then the expected number of hits is

$$N_j \approx n p_j$$

   This gives the density estimate

$$f(b_j) \approx \frac{p_j}{\Delta y} \approx \frac{N_j}{n \Delta y} \; . \tag{4}$$

   Physical scientists often think of probabilities in terms of potential functions. If $u(x)$ is a potential function, the corresponding PDF is

$$f(x) = \frac{1}{Z} e^{-u(x)} \; .$$

Thus, a point x is more likely if its probability potential is lower. The most likely point are the ones with the lowest potential. The *normalization factor*, Z, is a constant that makes $\int f(x)dx = 1$ without having to adjust the potential function. As we will see, it is common to know the potential function but not the normalization constant. The potential function for a Gaussian is quadratic. $f(x) \propto e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ comes from the potential function $u(x) = \frac{1}{2\sigma^2}(x-\mu)^2$. One reason to look at potential functions is to judge "how gaussian" a random variable is. The central limit theorem says certain random variables (sample means, for example) are approximately Gaussian. You can assess the accuracy of this by plotting the potential function to see how closely it resembles a parabola. Plots of the PDF itself may look "bell shaped" while the potential is far from being parabolic.

The code `Assignment1.py` has many of the ingredients needed for this assignment. It estimates the density of a random variable that happens to be Gaussian with a quadratic potential function. Please download this and use it as a starting point for your code. You will need to replace the simple experiment with a more complicated one that returns the sample variance of N independent standard normals. As will be explained in class in a few weeks, it is important that the random number generator object be instantiated in the main program and passed as an argument to the experiment function. You will also have to change the wording in the titles, axis labels, and legend. It is important that the title contain values of all important parameters used to make the plot.

Note about grading. Once you have a working code, please "play" with it to explore the qualitative questions in the assignments. Once you think you see something (asymmetry, noisy or smooth histogram, etc.), play with parameters to find values that illustrate your observation clearly. Write your observations in a few sentences, not as comments in the code. You will be graded in the clarity and interest of your observations and on how well they are illustrated in the plots you include. Please select plots with care so there are only as many as needed to support your observations. Imagine that you are preparing a professional presentation (a conference paper or talk) for people who with some background but who have not done the experiments you're explaining.