Assignment 2

Corrections to Exercise 1.

- an absolute value on the inequality that is now equation (1)
- a correction to what is now equation (2) to remove the 1 on the right side.

1. Consider the function $u(x) = e^x - 1$. As background, the exponential function has the Taylor series representation

$$e^x = \sum_{n=0}^{\infty} \frac{1}{n!} \, x^n \; .$$

This implies the series of approximations

$$e^x = 1 + x + \cdots + \frac{1}{n!} \, x^n + O(x^{n+1}) \; .$$

More precisely, if $|x| \leq a$ then

$$\left| e^x - \left(1 + x + \cdots + \frac{1}{n!} \, x^n \right) \right| \leq e^a \frac{1}{(n+1)!} \; |x|^{n+1} \qquad (1)$$

We write $f\ell(y)$ for the floating point number closest to y. The difference, $f\ell(y) - y$, is *roundoff error*. We say y is "in the range of floating point arithmetic" if

$$|f\ell(y) - y| \leq \epsilon_{\text{mach}} \, |y| \; .$$

(a) We say that u is "well conditioned for evaluation" in a range of x values if rounding x then rounding u yields an accurate approximation to $u(x)$. In formulas, this is

$$|u(x) - f\ell(\, u(f\ell(x))\,)| \leq C\epsilon_{\text{mach}} \, |x| \; .$$

This presumes C is not a large number such $C = 1$ or $C = 2$. If $C = 100$ we might still say u is well conditioned, but we would know to expect the loss of at least two digits of accuracy in evaluating u. Show that u is well conditioned in this sense if $|x| \leq 2$ and x is in the range of floating point arithmetic. Be aware that what it means for x to be in the range of floating point arithmetic depends on which arithmetic is used, single precision, double precision, half precision, quad precision, whatever.

(b) Suppose $\widehat{e}(x) = fl\left(e^{fl(x)}\right)$ is the "the exact answer correctly rounded" approximation to the exponential function. Consider the natural approximation

$$\widetilde{u}(x) = \widehat{e}(x) - 1 \; .$$

Show that this has low relative accuracy for small x in the sense that

$$\frac{|\widetilde{u}(x) - u(x)|}{|u(x)|} \sim \frac{\epsilon_{\text{mach}}}{|x|} \; .$$

Here the equivalence symbol "$\sim$" means that any x in the range of normalized arithmetic has a $\widetilde{x}$ within roundoff error of x so that $\sim$ really is almost $=$.

(c) Consider a hybrid algorithm that uses $\widehat{u} = \widetilde{u}(x)$ if $|x| > \alpha$, and

$$u(x) \approx \widehat{u}(x) = x + \frac{1}{2}x^2 + \cdots + \frac{1}{n!}\, x^n \tag{2}$$

if $|x| \leq \alpha$. What α (the largest or near largest) gives relative error

$$\frac{|\widehat{u}(x) - u(x)|}{|u(x)|} \leq 10^{-4}$$

in single or double precision with $n = 4$?

(d) Is there an n and α combination that gives relative error 10^{-15} in double precision?

2. Consider the problem of estimating

$$A(r) = \mathrm{E}\left[\, e^{rX} \,\right] \quad , \qquad X \sim \mathcal{N}(0,1) \; .$$

(a) Verify the formula

$$A(r) = e^{\frac{1}{2}r^2} \; .$$

Hint. $A(r) = \frac{1}{\sqrt{2\pi}} \int e^{rx} e^{-\frac{1}{2}x^2}\, dx$. Use the "complete the square" formula $rx - \frac{1}{2}x^2 = -\frac{1}{2}\left(x^2 - 2rx + r^2\right) + \frac{1}{2}r^2$. If b is any number, then $\int e^{-\frac{1}{2}(x-b)^2}\, dx = \int e^{-\frac{1}{2}x^2}\, dx$. (Why? Justify your steps.)

(b) The random variable in the expectation is $Y = e^{rX}$. Show that $\mathrm{var}(Y) = e^{2r^2} - e^{r^2}$ (which is more or less the same as e^{2r^2} when r is large).

(c) Suppose we generate n independent samples $X_k \sim \mathcal{N}(0,1)$ and use the estimator

$$\widehat{A}_n(r) = \frac{1}{n} \sum_{k=1}^{n} e^{rX_k} \; . \tag{3}$$

Find a formula for the likely size of the relative error

$$\frac{\left|\widehat{A}_n(r) - A(r)\right|}{A(r)}$$

For the numerator, use the standard deviation of $\widehat{A}_n(r)$. Explain what this says about the probable accuracy of the estimator (3) when r is large. What sample size n does it take to get approximately 1% error when $r = 2$ and $r = 5$?

3. Suppose you have numerical data items X_k for $1 \leq k \leq n$. A *histogram* is a graph that shows how many data points are in each interval. More precisely, let the interval (a, b) be divided into m equal size bins of length $\Delta x = \frac{b-a}{m}$. A *bin* is an interval of length Δx on the $x-$axis. We denote them by

$$B_j = [\, j\Delta x \,,\, (j+1)\Delta x \,] \;.$$

The interval $[a, b]$ is the union of these bins

$$[a, b] = B_1 \cup \;\cdots\; \cup B_m \;.$$

The *bin count* for bin B_j is the number of data values in B_j:

$$N_j = \#\left\{ X_k \in B_j \mid k = 1, \cdots, n \right\} \;.$$

For this exercise, assume the X_k are independent samples from a PDF $f(x)$.

(a) Write a method that returns a one index numpy array with n independent exponential random variables X_k with rate parameter λ using the sampler $X_k = -\frac{1}{\lambda}\log(U_k)$, $U_k \sim \text{unif}(0, 1)$. *Warning.* You have to *instantiate* (create) the random number generator object `rng` (called a *bit generator* in the numpy documentation) outside this method and pass `rng` as an argument to this method. Otherwise you will get the same samples every time you call the method. The main program should instantiate the random number generator object and pass it to any methods that need it.

(b) Write a method that takes a one index numpy array together with parameters a, b, m, and returns a numpy integer one index array of bin counts.

(c) Let $x_j = (j + \frac{1}{2})\Delta x$ be the midpoint of B_j. Show that if Δx is small, then

$$N_j \approx n\Delta x f(x_j) \;.$$

Use this to show that bin counts may be used to estimate the PDF:

$$f(x_j) \approx \widehat{f}_j = \frac{N_j}{n\Delta x} \;.$$

(d) Write a method that creates a plot with the true values $f(x_j)$ and the estimated values $\widehat{f}_j$ for n independent samples of the exponential distribution with rate parameter λ. In the graph, use a solid curve for the true f and dots for the estimates $\widehat{f}_j$. If the estimates are very accurate then the dots lie on the solid curve. [Recall coding standards, which include numerical labels in plot titles.] Do some experiments to show how the accuracy depends on m and n. You should see the effects of having bins that are too small (m too large, Δx too small), but improving accuracy for a given Δx if n is very large. If $Deltax$ is too large, then $f(x_j)\Delta x$ is not an accurate approximation of $\Pr(X \in B_j)$.

4. This exercise is a computational exploration of the central limit theorem, or CLT. There are various versions of "the" CLT that differ in their technical hypotheses and conclusions. One version involves i.id. random variables $Y_1, \cdots, Y_n$ and their sum $X = Y_1 + \cdots + Y_n$. We write Y for a generic random variable with the distribution of the Y_k and write $Y_k \sim Y$ to express the fact that Y and Y_k have the same probability distribution. For this exercise, "the same probability distribution" just means that they have the same probability density.[1] "The" CLT says that if n is large and if $\sigma_Y^2 = \mathrm{var}(Y) < \infty$, then X is approximately Gaussian in the sense that the PDF of X is approximately the Gaussian PDF. The PDF of a Gaussian random variable with mean μ_X and variance σ_X^2 is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-\frac{(x-\mu_X)^2}{2\sigma_x^2}} . \tag{4}$$

This PDF is also written $\mathcal{N}(\mu_X, \sigma_X^2)$. We write $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ if X has the PDF (4).

(a) Suppose $Z \sim \mathcal{N}(0,1)$ and $X = \mu_X + \sigma_X Z$. Show that $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$. *Hint.* Suppose z and x are related by $x = \mu_X + \sigma_X x$. Then $\Pr(x \leq X \leq x + dx) = f(x)dx$, and $x \leq X \leq x + dx$ is equivalent to $z \leq Z \leq z + dz$, with a proper relation between dz and dx. This gives a formula for the PDF of X in terms of the PDF of Z. *Explanation.* This Z is called *standard* normal. The exercise shows that any normal can be obtained from a standard normal by shifting (adding μ_X) and scaling (multiplying by σ_X). Another important point: the PDF (4) is completely determined by the parameters μ_X and σ_X. Thus, the distribution of a normal random variable is completely determined by its mean and variance. If you know X is Gaussian, its distribution is completely determined by μ_x and σ_X.

(b) Show that if T is exponential with rate parameter λ, then

$$\mu_T = \mathrm{E}[T] = \frac{1}{\lambda} \ , \quad \sigma_T^2 = \mathrm{var}(T) = \frac{1}{\lambda^2} \ . \tag{5}$$

[1] It is possible that a random variable does not have a probability density. An example of this is the random variable with $Y = 0$ or $Y = 1$ each with probability $\frac{1}{2}$. If there were a PDF for this Y, it would be infinite at $y = 0$ and $y = 0$ and zero for all other y. There is an informal *delta function*, written $\delta(x)$ that represents this idea, but it is not a proper function.

For the variance formula, it is convenient to use the formula (verify this)

$$\mathrm{var}(T) = \mathrm{E}\Big[\,(T-\mu_T)^2\,\Big] = \mathrm{E}\big[\,T^2\,\big] - \mu_T^2\ .$$

The formulas (5) have the right units because λ has units of $1/T$.

(c) Write a method to make a sample of the *scaled* and *centered* sum involving n independent exponential random variables with the same rate parameter:

$$X = \frac{1}{\sqrt{n}\,\sigma_T}\left(T_1 + \cdots T_n - n\mu_T\right)\ .$$

Do this using one call to the exponential generator from Exercise 3. The CLT says the PDF of X should be approximately standard normal. Estimate the PDF of X using the histogram method of Exercise 3. For that you need a large number of independent samples of X, which you get by calling the X sampler many times. The warning of Exercise 3a applies here too. Experiment with different values of n to see when the approximation is poor, when it is getting better, and when it is pretty accurate. Note that the PDF of X is not exactly symmetric, but is more nearly symmetric for large n. You will need to make many X samples and have reasonably small bins to produce an accurate estimate of the PDF of X. The slowness of scalar Python should be noticeable. *Warning.* Be careful to distinguish the n in the CLT from the n as used in Exercise 3. You might want to use subscripts or different letters?