Statistics and Numerical Method — Final Exam (due 12/30/2019)

This is a take-home exam. You are welcome to consult the lecture notes and other open resources, but you must work on the problems independently. Discussion with others is strictly prohibited.

Please submit your answers on the web learning platform by the due date as usual. However, unlike previous homeworks, we do NOT accept late submission.

1. Basic linear algebra (8 pts)

For a *general* square matrix A, list two viable methods (1 pt for each) to compute (a). its determinant; and (b). its inverse. Briefly describe the procedure (no need to show equations) and leading-order computational cost for each method. (1pt for each)

2. Eigenvectors from eigenvalues (6 pts)

Recently, a paper on the arXiv reporting that one can calculate the eigenvectors based on eigenvalues: `https://arxiv.org/abs/1908.03795`. Despite that the method has already been known, it drew substantial attention on social media. Please take a brief look at this paper and examine what has been proved (3 pts). Discuss the prospect of using this method to compute eigenvectors, and compare it with the methods we have discussed in class (3 pts).

3. Hilbert matrix (again) (6 pts)

Recall the Hilbert matrix problem in the first problem set, and consider last (bonus) question. Unfortunately, none of you gave a satisfactory answer, so here let us try again. The answer has already been provided in the solution, where two methods have been shown to work well. Here, we ask you to employ one of the two methods (SVD with psudo-inverse, and conjugate gradient).

The specific problem is, to solve $A\boldsymbol{x} = \boldsymbol{b}$ using single precision, with A being a Hilbert matrix with $n = 15$. I would like you to describe your method and show the solution $\boldsymbol{x}$ (with all significant digits). If you choose to use SVD, you are allowed to use standard SVD routines without writing your own, and discuss how many singular values you want to retain to obtain the solution. If you choose CG, please write your own CG routine, and report your initial guess and number of iterations needed for convergence. Attach your code, and provide your solution.

4. Error estimation and propagation (10 pts)

Suppose you have measured two quantities X and Y, with measured values being x and y (both are positive), and standard deviation being σ_X and σ_Y. Suppose $\sigma_X \ll x$ and $\sigma_Y \ll y$, and X and Y are independent. Estimate the standard deviation of the following quantities (give the procedures): (1). $X + Y$ (2 pts); (2). XY (2 pts); (3). $1/X$ (2 pts); (4). X/Y (2 pts); (5). $\sqrt{X^2 + Y^2}$ (2 pts).

5. Gamma-ray pulsar blind search (10 pts)

One important contribution from the Fermi gamma-ray telescope is the identification of hundreds of previously unknown pulsars. These pulsars are found by blindly searching for periodic variations from unidentified gamma-ray point sources. However, due to the intrinsically weak signal and rapid rotation, Fermi may only gather one photon over thousands of periods. In this problem, you are given a set of mock photon time-of-arrival data: toa_data.dat. The first row tells the total number of photons gathered, followed by a list of photon arrival times, in units of second. We know that the period of typical gamma-ray pulsars is between about 0.1s to 1s, and the light curve usually

contains two widely-separated pulses per period. You are asked to follow the instructions in the lecture to estimate the period of this putative pulsar, and give its phase-folded light curve.

(1). Which statistic do you want to use? (2 pts)

(2). You need to make a frequency (or period) grid for periodicity search. How would you choose grid resolution? (2 pts)

(3). Show the result of your test statistic as a function of trial period (based on your grid). (2 pts)

(4). Once you identify the period from the search, how much better can you improve the accuracy of the period? (2 pts)

(5). Once you have refined your search, show the folded phase light curve. (2 pts).

Note: the data rate in real observations is about a factor 10^2 lower, and hence the problem is even more challenging. In fact, a different method is used.

6. Performance of Monte Carlo integration in different dimensions (10 pts)

We would like to compare the performance of the Monte Carlo integration technique with the regular midpoint method. To this end, consider the integral:

$$I = \int_{\mathrm{V}} f(\vec{x}) \mathrm{d}^d \vec{x}, \tag{1}$$

where the integration domain V is a d-dimensional hypercube with $0 \leqslant x_i < 1$ for each component of the vector $\vec{x} = (x_1, x_2, ...x_d)$. The function we want to integrate is given by:

$$f(\vec{x}) = \prod_{i=1}^{d} \frac{3}{2}(1 - x_i^2). \tag{2}$$

This has an analytic solution, which is $I = 1$ independent of d, but we want to ignore this for the moment and use the problem as a test of the relative performance of Monte Carlo integration and ordinary integration techniques. To this end, calculate the integral in dimensions $d = 1, 2, ...5$, using

- The midpoint method, where you divide the volume into a set of much smaller hypercubes obtained by subdividing each axis into n intervals, and where you approximate the integral by evaluating the function at the centers of the small cubes.
- Standard Monte Carlo integration in d dimensions, using N random vectors.

For definiteness, adopt $n = 6$ and $N = 20000$. For both of the methods, report the numerical results for I and the CPU-time needed for each of the dimensions $d = 1, 2, ...5$. Please provide a table containing your results (2 pts for results from each dimensionality).

7. Rejection method (10 pts)

We would like to produce a random sample $\{x_i\}$ drawn from the probability distribution function (PDF)

$$p(x) = \frac{p_0}{(x-2)^4 + \sin^8(x-3)}, \tag{3}$$

for $0 \leqslant x < 5$, with $p(x) = 0$ outside this interval.

- Determine p_0 such that the function is normalized, i.e.,

$$\int_0^5 p(x)\mathrm{d}x = 1. \quad \text{(2pts)} \tag{4}$$

- Use the rejection method with a uniform parent distribution over the interval $0 \leqslant x < 5$ to create a sample of $N = 10^6$ numbers from this distribution. What is the rejection rate (2 pts)? Plot a histogram of the distribution of your points, using 100 bins with a width $\Delta x = 0.05$, and compare it with the target distribution function on a common plot with logarithmic y-axis (2 pts).

- Now consider a function $f(x)$ meant to provide an envelope for $p(x)$. Confirm that the piecewise linear function:

$$f(x) = \frac{y_{n+1} - y_n}{x_{n+1} - x_n}(x - x_n) + y_n, \text{ for } x_n \leqslant x \leqslant x_{n+1}, \tag{5}$$

  with $n \in \{0, 1, 2, 3, 4\}$ fulfills $p(x) \leqslant f(x)$ over the interval [0, 5) for the points $(x_0, y_0) = (0, 0.01)$, $(x_1, y_1) = (1.8, 0.15)$, $(x_2, y_2) = (2.35, 2.5)$, $(x_3, y_3) = (3.0, 0.1)$,and $(x_4, y_4) = (5.0, 0.002)$. Use $f(x)$ as an auxiliary function in the rejection method to more efficiently sample the function $p(x)$. What is the rejection rate now (2 pts)? Verify with another histogram plot that the obtained sample is correct (2 pts).