CSCI 5454 – Fall 2024
Design and Analysis of Algorithms

Homework 6

Instructions:

- **Due date:** This homework is due by 11:59PM Mountain Time on **Friday, October 11th**. Late assignments will not be accepted. Your lowest two homework scores for the semester will be dropped.
- You are welcome and encouraged to discuss the problems with classmates, but **you must write up and submit your own solutions and code**. You must also write the names of everyone in your group on the top of your submission.
- The primary resources for this class are the lectures, lecture slides/notes, the CLRS and Erickson algorithms textbooks, the teaching staff, your collaborators, and the class Piazza forum. We strongly encourage you only to use these resources. If you do use another resource, make sure to cite it and explain why you needed it. **Using generative AI tools (e.g., ChatGPT), Q&A forums (e.g., Stack Exchange, Quora), or cheating repositories (e.g., Chegg, CourseHero) is not allowed.** See the syllabus for a more detailed explanation.
- You must justify all of your answers unless specifically stated otherwise.
- We strongly encourage (but do not require) you to write your solutions in LaTeX, and have provided a skeleton file.[1] However, if your homework is illegible then we reserve the right not to grade it.

[1] If you have not used Latex before, it's easy to get started using Overleaf. See their 30-minute tutorial: `https://www.overleaf.com/learn/latex/Learn_LaTeX_in_30_minutes`.

Questions

Question 1. (20 points) In class, we learned about JL transforms and their properties. The purpose of this question is to run simulations of the process and compare with the theoretical predictions. You may program these in any language of your choice. **Do not include the code in your submissions**. We will ask for plots : please ensure that they are high enough resolution for us to examine them and also make sure that the axes are labeled/plots are titled. Using code that someone else has written is not allowed.

a. (5 points) Generate a fixed random 1000×1 vector $\mathbf{x}$ such that $\|\mathbf{x}\|_2 = 1$. Generate $N = 10^4$ random Johnson-Lindenstrauss (JL) transform matrices $G_1, \ldots, G_N$ of size $k \times d$ where $k = 100$ and $d = 1000$ (the size of $\mathbf{x}$). The entries of G_i must be normally distributed with mean 0 and standard deviation $\frac{1}{\sqrt{k}}$, or equivalently, $G_i = \frac{1}{\sqrt{k}} H_i$ where each entry of H_i is a random variable with mean 0 and standard deviation 1. Let $\mathbf{y}_i = G_i \mathbf{x}$.

Plot a histogram of the norms $\|\mathbf{y}_i\|_2^2$ for $i = 1, \ldots, N$.

b. (5 point) The PDF of a chi-squared random variable with k degrees of freedom is given by

$$\xi_k(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2},$$

where Γ is the Gamma function. Generate a plot the overlays the function $f(x) = k\xi_k(x \times k)$ [correction] (note that $k = 100$) on top of your histogram from part a. Please include a new plot for this part (do not share a single one with previous part since it makes grading harder on gradescope). See `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2.html` for a Python implementation of the chi-squared PDF.

Write a ≤ 5 line explanation of why the histogram from part a must follow the PDF of a chi-squared random variable with k degrees of freedom. You may look into the notes we have provided on JL transforms to supply this explanation.

c. (5 point) Repeat part a for $k \in \{100, 200, \ldots, 1000\}$ and for each k compute the fraction of points whose norms are *distorted* by more than 5% (a point $\mathbf{y}_i$ is distorted by more than 5% if $\|\mathbf{y}_i\|_2 > (1+\varepsilon)$ or $\|\mathbf{y}_i\|_2 < (1-\varepsilon)$ for $\varepsilon = 0.05$). Create a plot with values of k in the $x-$axis plotted against the fraction of points distorted in the $y-$axis. Also plot the theoretical bound of $2\exp(\frac{-k\varepsilon^2}{4})$ that we derived in class. Provide a 2-3 line explanation for why the two curves are different.

d. (5 point) Generate $n = 10$ random vectors $x_1, \ldots, x_n$ each with dimension 1000×1. Next, generate $N = 10^3$ random JL transform matrices $G_1, \ldots, G_N$, each with dimension $k \times 1000$. For each matrix G_i, check if transforming $x_1, \ldots, x_n$ using G_i preserves distances between all pairs of points i, j:

$$(1-\epsilon)\|x_i - x_j\| \leq \|Gx_i - Gx_j\| \leq (1+\epsilon)\|x_i - x_j\|.$$

Use $\epsilon = 0.1$ for this problem.

For each $k \in \{100, 200, \ldots, 1000\}$, run the procedure described above to compute the fraction of G_is that preserve distances between all pairs of points for the fixed vectors $x_1, \ldots, x_n$. Generate a plot with values of k in the x-axis and fraction of matrices that preserve all pairwise distances in the y-axis.

Solution:

Question 2. (20 points) In this problem, we will run simulations of the Morris counting scheme we covered in class.

a. (5 points) Write *pseudocode* for an algorithm that given a *fair coin* with 50% probability of heads/tails, creates a "coin" that will show up heads with probability $\frac{1}{2^k}$ and tails with $1 - \frac{1}{2^k}$ probability. Here k is a positive integer that is input to your algorithm.

b. (8 points) Implement a function that simulates the operation of the morris counter to count $n = 5000$ events. Run your counter for $h = 10,000$ simulations and plot a histogram of the estimated counts. Also include a table (see below) that reports the following information below. "Theory" here refers to the value derived in class and "simulation" refers to what was returned by your implementation.

	Theory	$h = 10^4$ simulations
Expected (Avg) Estimated Count		
Standard Deviation		

c. (7 points) Implement a function that inputs n, h, k and counts up to n by extracting the median over k separate "super-counter", where each "super-counter" is the average over h individual Morris counters. This is the median of means trick presented in class.

Run your counter for $n = 5000, k = 5, h = 10$ for $N = 1000$ times and plot a histogram of the estimated counts. Also fill out a table below:

	Theory	$h = 10(\mathsf{corrected}), k = 5$
Expected (Avg) Estimated Count		
Standard Deviation	- do not bother-	

Write 2-3 line description and explanation for the difference you see between the distribution of estimated counts in part b and this part.

Solution: