COMPSCI390.01: Algorithmic Foundations of Data Science
Midterm II

NAME:

Prob #	Score	Max Score
1		25
2		25
3		25
4		25
Total		100

Instructions:

1. If you write pseudocode, make sure to describe your idea in words.
2. Analyze the time complexity of your algorithms if asked.
3. You may use any algorithm covered in the class without detailed description, but you should be explicit about the input and output.
4. For any change that you make to an algorithm covered in class, you should describe the changes precisely.

Problem 1 [25pts]: Suppose we have collected a matrix X of data, and we wish to use principal component analysis (PCA) to perform dimensionality reduction.

(i) [7pts] Suppose we use the singular value decomposition to decompose X as follows:

$$X = U\Sigma V^T = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \end{bmatrix}.$$

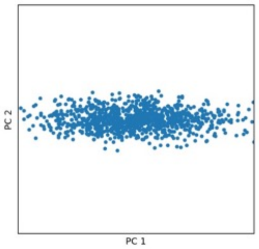
Find the *second* principal component of X. Show your work.

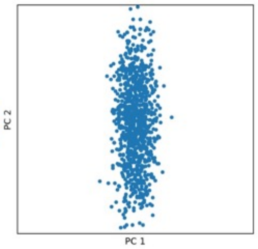
(ii) Suppose we want to perform PCA on *any* matrix X of data with SVD $X = U\Sigma V^T$. (The matrices in this part are unrelated to those in (i).) For each of the following statements or questions, provide a *brief* response — one or two sentences suffice.

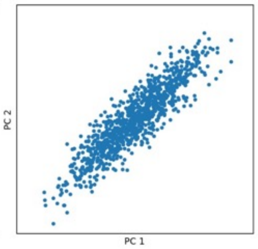
(a) [5pts] Suppose the data points lie exactly on a 2-D plane embedded in $\mathbb{R}^3$; what will PCA "discover," i.e., do you predict anything special about the PCA result?

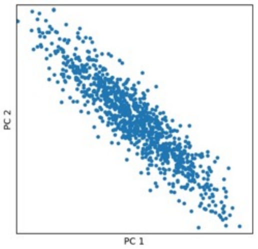
(b) [6pts] Are the matrices AV and $U\Sigma$ always equal?

(iii) [7pts] Suppose we project a matrix X of data onto the directions of its *first two* principal components. Which of the following could *possibly* display the projected data with the first PC plotted along the *horizontal* axis and the second PC along the *vertical* axis? Briefly justify your answer.









Problem 2 [25pts]: Let $X = \{x_1, \ldots, x_n\}$ be a set of n real values, and let $\varepsilon \in (0, 1)$ be a parameter. Our goal is to compute a median of X.

(i) [16pts] We choose a random sample $R \subseteq X$ of size $r = c/\epsilon^2$, where $c > 1$ is a constant, compute the median x_R of R, and return x_R as an ε-approximate median of X. Argue that the rank of x_R in X lies in the range

$$\left[(1-\varepsilon)\frac{n}{2},\ (1+\varepsilon)\frac{n}{2}\right]$$

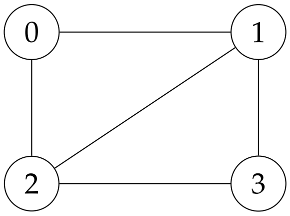
with probability at least $1/2$, provided that c is chosen sufficiently large. (Hint: use ε-approximation.)

(ii) [9pts] Suppose we receive a stream $x_1, x_2, \ldots$ of real values, and let ε be a parameter. For any $t > 0$, let $X_t = \langle x_1, \ldots, x_t \rangle$. Describe an algorithm that uses $O(1/\varepsilon^2)$ space and that at any given time $t > 0$, maintains an element x^* of X_t that is an ε-approximate median of X_t, i.e., the rank of x^* in X_t lies in the range $[(1-\varepsilon)t/2,\ (1+\varepsilon)t/2]$, with probability at least $1/2$. The algorithm may use $O(1/\varepsilon^2)$ time per update, but it cannot use more than $O(1/\varepsilon^2)$ space.

Problem 3 [25pts]: Consider the undirected graph $G = (V, E)$ shown in the figure.

(i) [10pts] Write the 4×4 transition matrix P corresponding to the simple random walk on the graph. What is the stationary distribution here?

(ii) [15pts] Suppose we want a stationary distribution where $\pi(0) = 1/2, \pi(1) = 1/4$, and $\pi(2) = \pi(3) = 1/8$. Find a transition matrix P that achieves this goal. (**Hint:** *Metropolis-Hastings.*)



Problem 4 [25pts]: Suppose there is a sensitive data set $X \subset \mathbb{R}^1$ consisting of real values, which we can only access by asking interval queries, namely, given a query interval $I = [a, b]$, return $|X \cap [a, b]|$, the number of points inside I. The system whishes to guarantee $(0.1, 0)$-differentially privacy, which is implemented using a Laplace mechanism $\text{Lap}(b)$ for some parameter $b > 0$, so for a query I, it returns a value $y = |X \cap I| + z$, where z is a random value drawn from the Laplace distribution $\text{Lap}(b)$.

The goal is to build a histogram of X consisting of $k = 200$ buckets, by asking k queries of the form $I_j = [a_j, a_{j+1}]$, for $1 \leq j < k$, where a_j's are the histogram boundaries. Let $y_j = |X \cap I_j|$ and $Y = [y_1, \ldots, y_k]$. Let $\tilde{y}_i$ be the answer returned by the system for the query I_j, so the histogram we construct is $\tilde{Y} = [\tilde{y}_1, \ldots, \tilde{y}_k]$.

(a) [10pts] What value of b should the system use to ensure $(0.1, 0)$-differential privacy?

(b) [15pts] What error $\|\tilde{Y} - Y\|_\infty$ can you bound with probability at least 99%, under a $(0.1, 0)$-DP Laplace mechanism? Recall that for $y \sim \text{Lap}(b)$, $\Pr[y > t \cdot b] < e^{-t}$.