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:

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

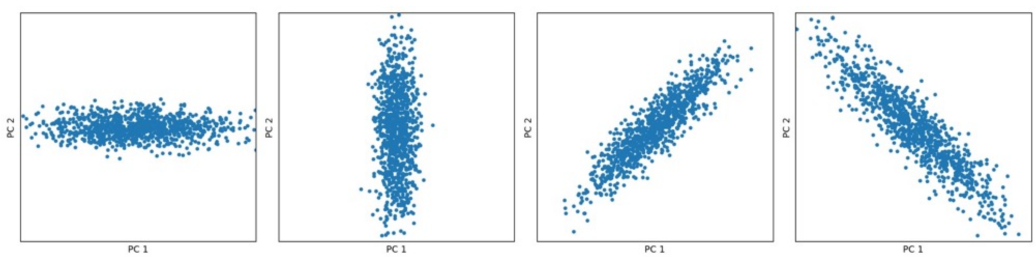
Find the *second* principal component of X. Show your work.

(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 XV 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.



Solution.

(i) The second loading (principal direction) is the second column of V, namely $v_2 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T$. Correspondingly the second principal component score is $Xv_2 = \begin{bmatrix} 2 & 0 & 0 \end{bmatrix}^T$. Since we did not specify the one we are looking for, either will receive credit.

(ii) (a) In this case, two singular values are nonzero and the third one is zero. PCA will therefore recover exactly two meaningful components, and the third component will have zero variance. In linear-algebraic terms, $\text{rank}(X) = 2$.

(b) Yes. Since V is orthogonal, we know $XV = U\Sigma V^T V = U\Sigma$.

(iii) Only the first diagram *could* be right. By definition, the variance captured by PC1 is at least that of PC2, so we can immediately discard plot 2 which is more spread out vertically than horizontally. Further, by construction PC1 and PC2 should be uncorrelated, so the projection cannot exhibit a linear relationship (or a "tilt") as in the last two plots.

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 $r = c/\varepsilon^2$, 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.

Solution.

(i) Let $\mathcal{I}$ be the set of all intervals in $\mathbb{R}^1$. Consider the range space $\Sigma = (X, \{X \cap I \mid I \in \mathcal{I}\})$. As discussed in the class, VC-dim$(\Sigma) = 2$. Therefore a random subset $R \subseteq X$ of size $r = O(1/\varepsilon^2)$ is a $(\varepsilon/2)$-approximation of Σ with probability at least $1/2$. Consider the interval $I = (-\infty, x_R]$, and let $A = X \cap I$ be the range corresponding to I. The rank of x_R in X is $|A|$. By construction, $\frac{|A \cap R|}{|R|} = \frac{1}{2}$.

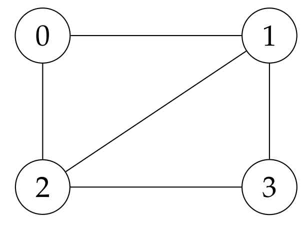
If A is an $(\varepsilon/2)$-approximation of Σ, then by the definition of ε-approximation,

$$\left|\frac{|A|}{n} - \frac{1}{2}\right| \le \frac{\varepsilon}{2} \Rightarrow \left||A| - \frac{n}{2}\right| \le \frac{\varepsilon}{2}n \Rightarrow |A| \in \left[(1-\varepsilon)\frac{n}{2},\ (1+\varepsilon)\frac{n}{2}\right],$$

as desired.

(ii) At time t, the algorithm maintains a random sample R of X_t of size $r^* = \min\{t, r\}$, where r is the same as in part (i), using reservoir sampling as discussed in the class. The algorithm also maintains the median x^* of R. Whenever R changes, it recomputes the median of R in $O(1/\varepsilon^2)$ time. By (i), x^* is an ε-approximate median of X_t with probability at least $1/2$ at any given time.

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.*)

Solution.

(i) Recall that in this scenario, at each vertex u, we randmly move to one of its neighbors with probability $\deg(u))$. Hence, we obtain the following transition matrix:

$$P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix}.$$

The stationary distribution is proportional to degree, i.e., $\pi_v \propto \deg(v)$:

$$\pi = \frac{1}{2|E|}(\deg(0), \deg(1), \deg(2), \deg(3)) = (0.2, 0.3, 0.3, 0.2).$$

Lazy random walk or reasonable variants will also receive credit.

(ii) We apply Metropolis-Hastings to the target distribution π. The max degree is $r = 3$. Our target distribution is already normalized, and the transition probabilities are now governed by $P_{i,j} = 1/r \cdot \min\{1, \pi(j)/\pi(j)\}$ for $j \neq i$. Once all of these values are computed, we set $P_{i,i} = 1 - \sum_{j \neq i} P_{i,j}$. This gives

$$P = \frac{1}{3}\begin{bmatrix} * & 1/2 & 1/4 & 0 \\ 1 & * & 1/2 & 1/2 \\ 1 & 1 & * & 1 \\ 0 & 1 & 1 & * \end{bmatrix} = \begin{bmatrix} * & 1/6 & 1/12 & 0 \\ 1/3 & * & 1/6 & 1/6 \\ 1/3 & 1/3 & * & 1/3 \\ 0 & 1/3 & 1/3 & * \end{bmatrix} = \begin{bmatrix} 3/4 & 1/6 & 1/12 & 0 \\ 1/3 & 1/3 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 1/3 \\ 0 & 1/3 & 1/3 & 1/3 \end{bmatrix}.$$

A quick verification indeed yields $\pi P = \pi$. *Note the answer is not unique. Any P with $\pi P = \pi$ will receive credit.*

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 wishes to guarantee $(0.1, 0)$-differential privacy, which is implemented using a Laplace mechanism $\mathrm{Lap}(b)$ 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 $\mathrm{Lap}(b)$.

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 $1 \le j < k$, 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] Give an upper bound on the error $\|\tilde{Y} - Y\|_\infty = \max_{1 \le j \le k} |\tilde{y}_i - y_i|$ that holds with probability at least 99% under a $(0.1, 0)$-DP Laplace mechanism. Recall that for $z \sim \mathrm{Lap}(b)$, $\Pr[z > t \cdot b] < e^{-t}$.

Solution.

(i) For a dataset X, let $f(X) = [y_1, \ldots, y_k]$. Since the query intervals I_j's are pairwise disjoint, ℓ_1-sensitivity of f, Δf, is 1, as the change of one item in X will change at most one y_i by 1. Recall that the Laplace mechanism choose $b = \Delta f / \varepsilon$ to ensure $(\varepsilon, 0)$-differential privacy. Therefore, $b = 1/0.1 = 10$.

(ii) As discussed in the class,

$$\Pr\left[\|\tilde{Y} - Y\|_\infty > \ln\left(\frac{k}{\delta}\right) \cdot \frac{\Delta f}{\varepsilon}\right] < \delta.$$

In our setting, $\delta = 0.01$, $k = 200$, $\varepsilon = 0.1$, and $\Delta f = 1$. Therefore, the error with probability 0.99 is at most

$$\ln\left(\frac{200}{0.01}\right) \frac{1}{0.1} = 10 \times \ln(2 \times 10^4) \approx 99.$$