# PSET 0

Due on **Aug 29, 2024 at 11:59 pm ET**

**COMPSCI 2243: Algorithms for Data Science, Fall 2024** (Sitan Chen) Version: **1.0**

Submitted by **<name>**

*This assignment is meant to help gauge your level of comfort with probability and linear algebra. Your responses will be taken into consideration in the course lottery, but we hope it will also be a good way for you to assess whether you will be comfortable with the material in this course. Given that we are asking you complete this on relatively short notice and given that the semester has not yet begun, do not feel discouraged if you are unable to finish some of the questions. In cases where you feel that you see a rough path to solution but do not have the time to work out all of the technical details, we encourage you to provide a rough sketch of your intuition. As this pset is ungraded, the most important thing is for you to convince* yourself *that you know how, at least in principle, to do a decent fraction of these questions.*

**For the course application, you are only expected to submit a solution to one of the three questions. If you choose to solve more than one, we will only take into consideration the first one for which you provide solutions.**

**Note:** *In all psets for this course we will mark the trickier questions with asterisks, to help with time management.*

## 1 (18 pts.) Covariance estimation

In this problem, we will explore how many samples $n$ are needed to form an accurate estimator for the covariance matrix $\Sigma$ of an unknown distribution.

Let $q$ be an unknown probability distribution over $\mathbb{R}^d$ such that $\mathbb{E}_{X\sim q}[X] = 0$ and $\mathbb{E}_{X\sim q}[XX^\top] = \Sigma$ for some matrix $\Sigma \in \mathbb{R}^{d\times d}$. Furthermore, suppose that there is some parameter $C > 0$ such that for $X \sim q$, the bound $\|X\|_2^2 \leqslant C$ holds almost surely. Let $X_1, \dots, X_n$ be i.i.d. samples from $q$.

The main tool we will exploit is the following concentration inequality, which you may use below without proof:

**Theorem 1.1 (Matrix Bernstein inequality).** *Let $A_1, \dots, A_n$ be random $d\times d$ symmetric matrices such that for every $i$, $\mathbb{E}[A_i] = 0$ and $\|A_i\|_{\mathsf{op}} \leqslant \gamma$ almost surely. Then for all $\epsilon > 0$,*

$$\mathbf{Pr}\Big[\Big\|\sum_i A_i\Big\|_{\mathsf{op}} \geqslant \epsilon\Big] \leqslant 2d\exp\Big(-\frac{\epsilon^2}{2(\sigma^2 + \gamma\epsilon/3)}\Big), \quad \text{for} \quad \sigma^2 \triangleq \Big\|\sum_i \mathbb{E}[A_i^2]\Big\|_{\mathsf{op}}, \tag{1}$$

*where $\|\cdot\|_{\mathsf{op}}$ denotes the operator norm.*

**1.A.** (1 pts.) Given the samples $X_1, \dots, X_n$ from $q$, what matrix would you form as your estimator for $\Sigma$?

**1.B.** (5 pts.) Let $\widehat{\Sigma}$ denote your answer to the previous question. Using Theorem 1.1, show that for all $0 < \epsilon < 1$,

$$\mathbf{Pr}\big[\|\widehat{\Sigma} - \Sigma\|_{\mathsf{op}} \geqslant \epsilon \cdot \|\Sigma\|_{\mathsf{op}}\big] \leqslant 2d\exp\big(-n\epsilon^2 \cdot \Omega(\|\Sigma\|_{\mathsf{op}}/C)\big). \tag{2}$$

**1.C.** (2 pts.) Let $0 < \epsilon < 1$. Using Eq. (2), determine an upper bound (up to constant factors) on the number of samples $n$ needed to ensure that $\|\widehat{\Sigma} - \Sigma\|_{\mathsf{op}} \leqslant \epsilon \cdot \|\Sigma\|_{\mathsf{op}}$ with probability at least $99\%$.

**1.D.** (5 pts.*) One intriguing feature of Theorem 1.1 is the extra factor of $d$ on the right-hand side. Please explain why one might expect this factor to appear, and provide a specific example of a distribution over $A_1, \dots, A_n$ where this factor is unavoidable. A formal proof is not required here. (**Hint:** Consider when $A_1, \dots, A_n$ are diagonal matrices.)

**1.E.** (5 pts.*) Give an example of a distribution $q$ for which the upper bound you proved in Part 1.C. is optimal for $C = d$, $\Sigma = \mathrm{Id}$, and $\epsilon = 1/2$ (up to constant factors), as well as a proof of optimality. (**Hint:** Consider $q$ supported on a finite set of points.)

## Solution:

**1.A.**

**1.B.**

**1.C.**

**1.D.**

**1.E.**

## 2 (18 pts.) Learning intervals and hyper-rectangles

In this problem we explore a toy supervised learning problem and prove some rudimentary generalization bounds.

Let $a \leqslant b$ be unknown real values, and define the function $f : \mathbb{R} \to \{0, 1\}$ by[1]

$$f(x) \triangleq \mathbb{1}[a \leqslant x \leqslant b]\,. \tag{3}$$

Suppose we are given data points $(x_1, y_1), \ldots, (x_n, y_n)$ such that $x_1, \ldots, x_n \in \mathbb{R}$ are i.i.d. over some arbitrary, unknown distribution $q$, and $y_i = f(x_i)$ for all $i$. While it is not necessary, you may assume that $q$ has a continuous density.

**2.A.** (1 pts.) Given the data points $(x_1, y_1), \ldots, (x_n, y_n)$, what would you output as your estimator for the unknown parameters $a$ and $b$?

**2.B.** (6 pts.) Let $\hat{a}$ and $\hat{b}$ denote your answer to the previous question. Given $0 < \epsilon, \delta < 1$, determine an upper bound on the number of samples $n$ needed to ensure that

$$\mathop{\mathbf{Pr}}_{x \sim q}\left[\mathbb{1}[\hat{a} \leqslant x \leqslant \hat{b}] = f(x)\right] \geqslant 1 - \epsilon \tag{4}$$

with probability at least $1 - \delta$ over the randomness of $x_1, \ldots, x_n$.

Next, we consider the natural high-dimensional generalization of the above question. Let $a_1, \ldots, a_d, b_1, \ldots, b_d \in \mathbb{R}$, and define the function $g : \mathbb{R}^d \to \{0, 1\}$ by

$$g(x) \triangleq \mathbb{1}[x \in [a_1, b_1] \times \cdots \times [a_d, b_d]]\,. \tag{5}$$

The set $[a_1, b_1] \times \cdots \times [a_d, b_d] \subset \mathbb{R}^d$ denotes the Cartesian product of the intervals and is called a *hyper-rectangle*.

Suppose we are given data points $(x_1, y_1), \ldots, (x_n, y_n)$ such that $x_1, \ldots, x_n \in \mathbb{R}^d$ are i.i.d. over some arbitrary, unknown distribution $q$, and $y_i = f(x_i)$ for all $i$.

**2.C.** (5 pts.) Propose an estimator $\hat{a}_1, \ldots, \hat{a}_d, \hat{b}_1, \ldots, \hat{b}_d$ for the unknown parameters $a_1, \ldots, a_d, b_1, \ldots, b_d$ given the data points $(x_1, y_1), \ldots, (x_n, y_n)$, and given $0 < \epsilon, \delta < 1$, determine an upper bound on the number of samples $n$ needed to ensure that

$$\mathop{\mathbf{Pr}}_{x \sim q}\left[\mathbb{1}[x \in [\hat{a}_1, \hat{b}_1] \times \cdots \times [\hat{a}_d, \hat{b}_d]] = g(x)\right] \geqslant 1 - \epsilon \tag{6}$$

with probability at least $1 - \delta$ over the randomness of $x_1, \ldots, x_n$. (**Hint:** Directly use your solution to **2.B.**)

**2.D.** (6 pts.) Suppose we additionally are given a set $\mathcal{S}$ of $M$ hyper-rectangles and are guaranteed that the parameters $a_1, \ldots, a_d$, $b_1, \ldots, b_d$ correspond to one such hyper-rectangle in $\mathcal{S}$. In this case, propose an estimator $\hat{a}_1, \ldots, \hat{a}_d, \hat{b}_1, \ldots, \hat{b}_d$ given the data points $(x_1, y_1), \ldots, (x_n, y_n)$ and $\mathcal{S}$. Determine an alternative upper bound on the number of samples $n$ needed to ensure that Eq. (6) holds with probability $1 - \delta$. This upper bound should depend on $M$ rather than $d$.

## Solution:

**2.A.**

**2.B.**

**2.C.**

**2.D.**

[1] Here the notation $\mathbb{1}[\cdot]$ denotes the indicator function, i.e. if $x \in \mathbb{R}$ satisfies $a \leqslant x \leqslant b$, then $\mathbb{1}[a \leqslant x \leqslant b] = 1$, and otherwise $\mathbb{1}[a \leqslant x \leqslant b] = 0$.

## 3 (18 pts.) Boolean and Gaussian vectors

Let $d \in \mathbb{N}$ be even. Let $x$ be sampled uniformly at random from $\{\pm 1\}^d$, and let $g$ be a standard Gaussian vector, i.e. $g \sim \mathcal{N}(0, \mathrm{Id}_d)$. In this problem we will explore the extent to which projections of these vectors are similar. Let $v^* \in \mathbb{R}^d$ be the unit vector given by $(1/\sqrt{d}, \ldots, 1/\sqrt{d})$.

Given two probability distributions $p, q$ over $\mathbb{R}$, define the *total variation distance*

$$d_{\mathsf{TV}}(p, q) \triangleq \sup_{S \subset \mathbb{R}} \left| p(S) - q(S) \right|, \tag{7}$$

where $p(S)$ denotes the probability that $x \sim p$ satisfies $x \in S$, and $q(S)$ is defined similarly. Define the *Kolmogorov distance*

$$d_{\mathsf{K}}(p, q) \triangleq \sup_{a \in \mathbb{R}} \left| \mathop{\mathbf{Pr}}_{x \sim p}[x \leqslant a] - \mathop{\mathbf{Pr}}_{x \sim q}[x \leqslant a] \right|. \tag{8}$$

**3.A.** (2 pts.) True/False: The total variation distance between $\langle v^*, x \rangle$ and $\langle v^*, g \rangle$ converges to zero as $d \to \infty$. If true, what is this theorem called? Otherwise, prove that it is false.

**3.B.** (4 pts.) Prove the following two bounds:

$$\mathbf{Pr}[|\langle v^*, x \rangle| > t] \leqslant 2 \exp(-\Omega(t^2)) \tag{9}$$

$$\mathbf{Pr}[|\langle v^*, g \rangle| > t] \leqslant 2 \exp(-\Omega(t^2)). \tag{10}$$

For the next few questions, we will consider the following result:

**Theorem 3.1 (Berry-Esseen).** *Let $X_1, \ldots, X_n$ be i.i.d. random variables satisfying $\mathbb{E}[X_i] = 0$, $\mathbb{E}[X_i^2] = 1$, and $\mathbb{E}[|X_i|^3] = \rho < \infty$ for all $i$. If we define $Y \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$, then the Kolmogorov distance between the standard normal distribution $\mathcal{N}(0,1)$ and the distribution over $Y$ and is at most $\frac{\rho}{2\sqrt{n}}$.*

**3.C.** (3 pts.) Using Theorem 3.1, prove a bound on the Kolmogorov distance between the distributions over $\langle v^*, x \rangle$ and $\langle v^*, g \rangle$.

**3.D.** (6 pts.*) Using the result of **3.C.**, give a proof that for all sufficiently large even $d$,

$$\binom{d}{d/2} = \Theta\left(\frac{2^d}{\sqrt{d}}\right). \tag{11}$$

**3.E.** (3 pts.) Among the results established in **3.A.**, **3.B.**, **3.C.**, which ones generalize when $v^*$ is replaced with an arbitrary unit vector, and why?

## Solution:

**3.A.**

**3.B.**

**3.C.**

**3.D.**

**3.E.**