COMP 585 **Spring 2024**

Probabilistic Toolkit for Learning and Computing *Instructor: Maryam Aliakbarpour*

Problem Set 3

Instruction:

- Submissions are due no later than **11:59 PM** on **Tuesday, March 25, 2025**.
- Please upload your solution in PDF format to the course website on Canvas. You may typeset or upload a scanned version of your handwritten solution. Your solution should be legible and clear. Full credit will be given only to the correct solutions that are easy to read and understand.
- This problem set is designed to test your basic prerequisite knowledge and help you brush up on your previous knowledge. Please do not google the solutions or use Large Language Models (LLMs) to solve the problems.
- You may collaborate with other class members (group of 2-3 people), but you must mention the names of your collaborators in your solution. The idea behind collaboration is to collectively work towards finding a solution in a fair manner. Here are some guidelines for collaboration:
    - Spend a few hours thinking about the problems before engaging in discussions with others.
    - Do not collaborate with someone who has already solved the problem or is not at the same level of progress as you.
    - Exercise good judgment to prevent one person from providing the solution to another.
    - Collaboration does not permit uploading the same solution file. After discussions with team members, you must independently write your solution. Your write-up should genuinely reflect your understanding of the solution. Avoid sharing your solution with others and refrain from copying solutions, even when working together.
- Please refer to the course syllabus for information regarding the late submission policy.

Problem 1. **(35 points)** Consider a non-negative random variable that satisfies a concentration inequality of the form

$$\mathbf{Pr}[Z \geq t] \leq C \exp\left(\frac{-t^2/2}{v^2 + bt}\right) \tag{1}$$

for positive v, b and $C \geq 1$.

a. Show that
$$\mathbf{E}[Z] \leq 2v(\sqrt{\pi} + \sqrt{\log C}) + 4b(1 + \log C)\,.$$

Hint: You may find it useful to use the integral identity for expectation and Equation 2.

Hint: In the tail bound, the dominating term in the denominator of the exponent changes when $t = v^2/b$. Maybe this is good breaking point for your integral.

b. Let $X_1, \ldots, X_n$ be i.i.d. zero-mean random variables satisfying the Bernstein Condition stated in class (See Equation 3). Let $\sigma^2 = \mathbf{Var}[X_i]$. Use Part a. to show that

$$\mathbf{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right|\right] \leq \frac{2\sigma}{\sqrt{n}}(\sqrt{\pi} + \sqrt{\log 2}) + \frac{4b}{n}(1 + \log 2)$$

Problem 2. **(25 points)** Consider the random projection $M \in \mathbb{R}^{d' \times d}$ which we have used in the Johnson–Lindenstrauss (JL) lemma in our lecture. Each entry of M is a scaled Gaussian random variable:
$$M_{ij} \sim c \cdot \mathcal{N}(0,1)\,,$$
for a fixed value of $c = 1/\sqrt{d'}$. The randomized mapping in JL lemma maps every u in $\mathbb{R}^d$ to $v = Mu$ in $\mathbb{R}^{d'}$.

a. In this part, we focus on how M affects the ℓ_1-norm of a vector. Fix a vector u. Let $v = Mu$. Show that the expected ℓ_1-norm of v is bounded from above:

$$\mathbf{E}[\|v\|_1] = \mathbf{E}\left[\sum_{i=1}^{d'} |v_i|\right] \leq \sqrt{d'} \cdot \|u\|_2\,.$$

Hint: Identify the distribution of each coordinate v_i.

b. Find a similar lower bound for $\mathbf{E}[\|v\|_1]$. Then, use Part a., to come up with a set of points such that their ℓ_1-distance cannot be preserved up to a factor of two if $d' = o(d)$ (at least not in expectation).

Problem 3. **(15 points)** Suppose $\mathcal{C}$ is a finite class of binary concepts: $c : \mathcal{X} \to \{0, 1\}$. Let $\mathcal{D}$ be a target distribution over $\mathcal{X} \times \{0, 1\}$. We sample a set of m labeled instances $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X}$ from $\mathcal{D}$. For a concept $c \in \mathcal{C}$, the (true) error of c, is the probability of mislabeling a random sample via c:

$$err(c) \coloneqq \mathbf{Pr}_{(x,y)\sim\mathcal{D}}\big[c(x) \neq y\big] .$$

Show that $\mathcal{C}$ is PAC learnable in the agnostic case with $m = O\left(\frac{\log(|C|/\delta)}{\epsilon^2}\right)$ samples. That is, there exists an algorithm $\mathcal{A}$ that receives m samples, and outputs $\hat{c}$ such that with probability $1 - \delta$:

$$err(\hat{c}) \leq \min_{c \in \mathcal{C}} err(c) + \epsilon .$$

Note: The agnostic case refers to the case where there is no $c \in C$ that perfectly label all the instances ($\min_{c\in\mathcal{C}} err(c) > 0$).

Problem 4. **(25 points)** Consider a variant of the PAC model in which there are two example oracles: one that generates positive examples and one that generates negative examples, both according to the underlying distribution $\mathcal{D}$ on $\mathcal{X}$. Formally, given a target function[1] $c^* : \mathcal{X} \to \{0, 1\}$, let $\mathcal{D}^+$ be the distribution over $\mathcal{X}^+ = \{x \in \mathcal{X} : c^*(x) = 1\}$ defined by $\mathcal{D}^+(A) = \frac{\mathcal{D}(A)}{\mathcal{D}(\mathcal{X}^+)}$ for every $A \subseteq \mathcal{X}^+$. Similarly, $\mathcal{D}^-$ is the distribution over $\mathcal{X}^-$ induced by $\mathcal{D}$.

The definition of PAC learnability in the two-oracle model is the same as the standard definition of PAC learnability except that here the learner has access to $m^+(\epsilon, \delta)$ i.i.d. examples from $\mathcal{D}^+$ and $m^-(\epsilon, \delta)$ i.i.d. examples from $\mathcal{D}^-$. The learner's goal is to output $\hat{c}$ s.t. with probability at least $1 - \delta$ (over the choice of the two training sets, and possibly over the nondeterministic decisions made by the learning algorithm), the error is low according to both distributions:

$$err^+(\hat{c}) = \mathbf{Pr}_{x\sim\mathcal{D}^+}\big[\hat{c}(x) \neq 1\big] \leq \epsilon \,, \text{and} \qquad err^-(\hat{c}) = \mathbf{Pr}_{x\sim\mathcal{D}^-}\big[\hat{c}(x) \neq 0\big] \leq \epsilon \,.$$

a. Show that if $\mathcal{C}$ is PAC learnable (in the standard one-oracle model), then $\mathcal{C}$ is PAC learnable in the two-oracle model.

   Hint: Assume there exists a function $m(\epsilon, \delta)$, and an algorithm $\mathcal{A}$ such that $\mathcal{A}$ receives $m = m(\epsilon, \delta)$ labeled samples from a distribution $\mathcal{D}$ and outputs $\hat{c}$. With probability at least $1 - \delta$, the error of mislabeling by $\hat{c}$ is at most ϵ. Then use $\mathcal{A}$ to show that there exist two functions $m^+(\epsilon, \delta)$ and $m^-(\epsilon, \delta)$ and an algorithm $\mathcal{A}'$ such that for every ϵ and δ in $(0, 1)$, $\mathcal{A}'$ receives $m^+ = m^+(\epsilon, \delta)$ positive examples and $m^- = m^-(\epsilon, \delta)$ negative examples, and it outputs $\hat{c}$ such that the error of $\hat{c}$ is at most ϵ according to both $\mathcal{D}^+$ and $\mathcal{D}^-$ with probability at least $1 - \delta$.

b. Define c^+ to be the always-plus hypothesis and c^- to be the always-minus hypothesis. Assume that $c^+, c^- \in \mathcal{C}$. Show that if $\mathcal{C}$ is PAC learnable in the two-oracle model, then $\mathcal{C}$ is PAC learnable in the standard one-oracle model.

[1] For this problem, assume we are in the realizable case. Thus, such c^* always exists.

Hint: Assume there exist two functions $m^+(\epsilon, \delta)$ and $m^-(\epsilon, \delta)$, and an algorithm $\mathcal{A}'$ such that $\mathcal{A}'$ receives $m^+ = m^+(\epsilon, \delta)$ positive samples from $\mathcal{D}^+$ and $m^- = m^-(\epsilon, \delta)$ negative samples from $\mathcal{D}^-$ and outputs $\hat{c}$. With probability at least $1 - \delta$, the error of mislabeling by $\hat{c}$ is at most ϵ according to both $\mathcal{D}^+$ and $\mathcal{D}^-$. Then design an algorithm $\mathcal{A}$ that uses $m(\epsilon, \delta)$ samples and $\mathcal{A}'$ to PAC learn $\mathcal{C}$. In particular, you may set:

$$m(\epsilon, \delta) \coloneqq \left\lceil \frac{8 \ \max\left(m^+(\epsilon, \delta/2), m^-(\epsilon, \delta/2)\right) \cdot \log(4/\delta)}{\epsilon} \right\rceil .$$

You may have other values of m as long as they are equal to m up to constant factors.

Cheat Sheet

You may find the following facts useful for this problem set.

Integral identity for expectation: For a non-negative random variable X, we have:

$$\mathbf{E}[X] = \int_0^\infty \mathbf{Pr}[X > t]\, dt\,.$$

A useful integral: While generally integrals of the form $\int e^{-x^2} dx$ do not have a simple closed form, we have the following for every $a > 0$:

$$\int_0^\infty e^{-x^2/a}\, dx = \frac{\sqrt{\pi \cdot a}}{2} \tag{2}$$

You may also verify this by looking at the PDF of the Gaussian distribution.

Bernstein condition: Suppose X is a random variable with mean $\mathbf{E}[X] = \mu$ and variance $\mathbf{Var}[X] = \sigma^2$. We say X satisfy the Bernstein condition with parameter b if for every integer $i \geq 2$ the following holds:

$$\mathbf{E}\left[(X-\mu)^i\right] \leq \frac{1}{2} i!\, \sigma^2\, b^{i-2} \tag{3}$$

A useful formula for variance: For a random variable X, we have:

$$\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2\,.$$

Jensen's inequality: For every convex function f, and a random variable X, we have:

$$f\left(\mathbf{E}[X]\right) \leq \mathbf{E}[f(X)]\,.$$

Chernoff Bound The Chernoff Bound is a probabilistic bound that provides an exponentially decreasing bound on tail distributions of the sum of random variables.

Let $X_1, X_2, \ldots, X_m$ be independent Poisson trials such that $\mathbf{Pr}[X_i = 1] = p_i$ and $\mathbf{Pr}[X_i = 0] = 1 - p_i$. (Poisson trials are 0-1 random variables like the Bernoulli trials. However, the Poisson trials do not have to have the same success probability.) Let $X = \sum_{i=1}^m X_i$ be the sum of these m random variables, and let p denote the mean of p_i's: $p = \sum_{i=1}^m p_i/m$. The bound is expressed as follows for every $\epsilon \in [0,1]$:

$$\mathbf{Pr}\left[\frac{X}{m} \geq (1+\epsilon)\cdot p\right] \leq e^{-mp\epsilon^2/3}, \text{ and}$$
$$\mathbf{Pr}\left[\frac{X}{m} \leq (1-\epsilon)\cdot p\right] \leq e^{-mp\epsilon^2/2}$$

Here, is another version of this inequality which holds for any $\epsilon > 0$:

$$\mathbf{Pr}\left[\frac{X}{m} \geq (1+\epsilon)\cdot p\right] \leq \left(\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right)^{mp}.$$

Hoeffding bound: The Hoeffding bound bounds the probability that the sum of independent bounded random variables deviates from its expected value. Mathematically, for independent variables $X_1, X_2, \ldots, X_n$ with bounds $a_i \leq X_i \leq b_i$, the bound is given by:

$$\mathbf{Pr}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}[X]\right| \geq t\right] \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right),$$

where $\mathbf{E}[X]$ is the expected value of the mean of X_i's, and t is the margin of deviation.

If X_i's are Bernoulli random variable with success probability p then we have:

$$\mathbf{Pr}\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geq p+\epsilon\right] \leq \exp\left(-2n\epsilon^2\right)$$

$$\mathbf{Pr}\left[\frac{1}{n}\sum_{i=1}^{n} X_i \leq p-\epsilon\right] \leq \exp\left(-2n\epsilon^2\right)$$

Total Law of Probability The Total Law of Probability is a fundamental rule that relates marginal probabilities to conditional probabilities. It is expressed as:

$$\mathbf{Pr}[A] = \sum_{i} \mathbf{Pr}[A|B_i]\mathbf{Pr}[B_i]$$

where $\mathbf{Pr}[A]$ is the total probability of event A, and $\mathbf{Pr}[A|B_i]$ is the probability of A given B_i, with B_i being a partition of the sample space.