COMP 585 **Spring 2025**

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

Problem Set 2

Instruction:

- Submissions are due no later than **11:59 PM** on **Tuesday, February 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. (30 points) A group of volunteers is organizing a museum exhibit scheduled to run for n days. m visitors have signed up, and each visitor randomly chooses a day in $[n] = \{1, 2, \ldots, n\}$ (each with a probability of $1/n$) to visit the exhibit. The museum administration

aims to ensure a satisfactory number of visitors throughout the exhibit. They've set a condition: if there's no visitor for t consecutive days, they will close down the exhibit. More precisely, the exhibit ends on day i if there is no visitor on days $i-(t-1), i-(t-2), \ldots, i$.

Here, we want to calculate the number of visitors we need to avoid the exhibit closing down earlier than planned using Poisson approximation. Let X_i be the number of visitors per day. Clearly, X_i is a binomial random variable $\mathbf{Bin}(m, 1/n)$. To approximate X_i's via Poisson random variables, let X'_i for $i \in [n]$ be a Poisson random variable with mean m/n, $\mathbf{Poi}(m/n)$ that represents the number of visitors on day i in the Poissonized setting. Let Z_i denote an indicator variable that represents that the exhibit ended early on day i in the Binomial setting. The exhibit did not have any visitors in the past t days. More specifically, we define for each $i \in \{t, t+1, \ldots, n-1\}$:

$$Z_i = \begin{cases} 1 & \text{if } X_{i-(t-1)} = X_{i-(t-2)} = \cdots = X_i = 0\,; \\ 0 & \text{Otherwise}\,. \end{cases}$$

Without loss of generality, assume $Z_i = 0$ for all $i < t$ and $i = n$. Let Z'_i be the corresponding variable to Z_i's in the Poissonized setting meaning it is defined based on X'_i's.

Binomial setting:	**Poissonized setting:**
# visitors in day i:	# visitors in day i:
$X_i \sim \text{Bin}(m, 1/n)$	$X'_i \sim \text{Poi}(m/n)$
Z_i indicates early termination on day i	Z'_i indicates early termination on day i

For simplicity of your calculation, assume n is divisible by t. Moreformally, assume that there is a $k \geq 1$ such that $n = t \cdot k + t$. With these settings in mind, answer the following questions:

a. What is the expected value of Z'_i for $i \in \{t, t+1, \ldots, n-1\}$?

Hint: Note that in the Poissonized setting X'_i's are independent.

b. Argue that for every i and j in $[n]$, if $i-(t-1) > j$, then Z'_i is independent of Z'_j.

c. Consider a subset of indices $I \coloneqq \{t, 2t, \ldots, kt\}$. What is the probability that $\sum_{i \in I} Z'_i$ is larger than its expectation by a factor of $(1+\epsilon)$?

Hint: Use part b. and then focus on proving a concentration bound.

d. Consider the following partition of the indices in $\{t, t+1, \ldots, n-1\}$. For $j \in [t]$, let $I^{(j)} = \{t+j-1, 2t+j-1, \ldots, n-t+j-1\}$. Observe that:

$$\bigcup_{j=1}^{t} I^{(j)} = \{t, \ldots, n-1\}\,.$$

Prove a concentration bound for the following:

$$Z' \coloneqq \sum_{i=t}^{n-1} Z'_i \,.$$

In particular show that:

$$\mathbf{Pr}\big[\, Z' \geq (1+\epsilon)\mathbf{E}\big[\, Z'\big]\big] \leq t \cdot \exp\left(-\frac{(n-t)e^{\frac{-mt}{n}}\epsilon^2}{3t}\right) .$$

Hint: Use the union bound.

e. It is known that for any event that happens with probability p in the Poissonized setting, it happens with probability at most $2\,p$ in the standard setting under a certain condition:

Theorem 1. *Let $\mathcal{E}$ be an event whose probability is either monotonically increasing or monotonically decreasing in the number of participants. If $\mathcal{E}$ has probability p in the Poissonized setting, then $\mathcal{E}$ has probability at most $2\,p$ in the Binomial setting.*

Given this theorem, show that if

$$2\,(n-t)e^{\frac{-mt}{n}} \leq 0.01\,,$$

then we know that the exhibit does not close down early with probability at least 99% (in the standard setting).

Hint: Note that the exhibit does not close down early iff $Z' = 0$ in the Poissonized setting.

Hint: You do not need to use the previous parts here. If you need a concentration bound, Markov's inequality should be sufficient.

Problem 2. (25 points) Suppose we have an unknown distribution p and m samples from it: $X_1, X_2, \ldots, X_m$. Let Y_i denote the number of instances of element i among these samples: $Y_i \coloneqq \sum_{j=1}^{m} \mathbb{1}_{X_j=i}$. Let $\hat{p}$ be the empirical distribution obtained from these samples. More precisely, the probability of each element is defined as follows: $\hat{p}_i = \frac{Y_i}{m}$.

More precisely, for every $i \in [n]$, let

$$Z_i \coloneqq \frac{a_i \cdot Y_i}{m}\,. \tag{1}$$

Here, we introduce some auxiliary random variable that help us to show concentration for Z_i's. Consider a fixed (not randomized) vector $a \in \{-1,+1\}^n$. First, based on a, we partition the domain of the distribution ($[n]$) into two sets A and B defined as follows:

$$A \coloneqq \{i \,|\, a_i = 1\}, \qquad B \coloneqq \{i \,|\, a_i = -1\}\,.$$

A is the set of coordinates that $a_i = 1$, and B is the compliment of A. Our plan is to show that sum of Z_i's is behaving similar to a binomial random variable. For any given sample X_j, we define a random variables W_j as follows:

$$W_j = \begin{cases} 1 & \text{if } X_j \in A\,; \\ 0 & \text{if } X_j \in B\,. \end{cases}$$

Observe that by this definition, we have:

$$a_{X_j} = 2W_j - 1\,. \tag{2}$$

a. Show that

$$\sum_{i=1}^{n} Z_i = \frac{1}{m}\sum_{j=1}^{m} 2\,W_j - 1\,,$$

b. For a fixed (not randomized) vector $a \in \{-1,+1\}^n$, show that $a \cdot \hat{p}$ is close to $a \cdot p$. Show that there is a constant c such that:

$$\mathbf{Pr}\left[\sum_{i=1}^{n} Z_i - \sum_{i=1}^{n} \mathbf{E}[Z_i] \geq \epsilon\right] \leq e^{-c\,m\,\epsilon^2}\,.$$

Hint: Use Part a.

c. Show there exists a constant c' such that if we have $m \geq c' \cdot n/\epsilon^2$ samples from p, then the empirical distribution is ϵ-close to p in ℓ_1-distance with probability at least 0.9.

d. Suppose we have a property $\mathcal{P}$ that is a set of t distributions over $[n]$. Show that for any such property, there exists a $(\epsilon, \delta = 0.9)$-tester that uses $m = O(n/\epsilon^2)$ samples and runs in $O(n \cdot t + m)$ time.

Hint: Start by learning p up to error $\epsilon/2$.

Problem 3. (15 points) Consider the collision based uniformity tester, we have discussed in the lectures. Recall that we have m samples $X_1, \ldots, X_m$. For a pair of indices $i < j$, σ_{ij} denotes the indicator variable that is one if $X_i = X_j$ and zero otherwise. Our statistic was:

$$Y = \frac{1}{\binom{m}{2}} \sum_{i<j} \sigma_{ij}\,.$$

a. For a given $\gamma > 0$ and $m = c/(\gamma^2 \cdot \|p\|_2)$ for some sufficiently large constant c, show that Y is a $(1+\gamma)$-factor approximation of $\|p\|_2^2$. That is

$$\mathbf{Pr}\left[\,\left|Y - \|p\|_2^2\right| \geq \gamma\,\|p\|_2^2\right\}\right] \leq 0.1\,.$$

You may use the bounds for the expected value and the variance of Y provided in the lecture:

$$\begin{cases} \mathbf{E}[Y] = \|p\|_2^2 \\ \mathbf{Var}[Y] \leq \frac{1}{\binom{m}{2}^2}\left[\binom{m}{2}\|p\|_2^2 + 6\binom{m}{3}\|p\|_3^3\right] \end{cases}$$

b. As shown in Part a., estimating $\|p\|_2^2$ via Y is challenging because the sample complexity depends on $\|p\|_2$, the very quantity we wish to estimate. What should m be so that, for any arbitrary distribution, we are guaranteed to have a sufficient number of samples? In other words, what is the smallest possible value of $\|p\|_2$?

Problem 4. (30 points)

Suppose X_1 and X_2 are zero-mean sub-Gaussian radnom variables with parameters K_1 and K_2 respectively.

a. Show that if X_1 and X_2 are independent, then $X_1 + X_2$ is $\sqrt{K_1^2 + K_2^2}$-sub-Gaussian random variable.

b. Prove that, without the need to assume independence between the random variables X_1 and X_2, the sum $X_1 + X_2$ is sub-Gaussian random variable, with its sub-Gaussian parameter bounded above by $\sqrt{2(K_1^2 + K_2^2)}$.

c. Suppose we have a series of potentially infinitely many sub-Gaussian random variables $X_1, X_2, \ldots$, and for each X_i, we have:

$$\mathbf{Pr}[|X| \geq t] \leq 2\exp\left(-\frac{t^2}{K_i^2}\right).$$

Let $K \coloneqq \max_i K_i$. Show that there exists a constant c such that:

$$\mathbf{E}\left[\max_i \frac{|X_i|}{\sqrt{1+\ln(i^2)}}\right] \leq c \cdot K$$

Hint: You may find it helpful to use the integral identity for the expectation of non-negative random variables

Hint: In Your calculation, at some point you may need the following relationship for when $t \geq K_i$:

$$\begin{aligned} \mathbf{Pr}\left[|X_i| > t \cdot \sqrt{1+\ln(i^2)}\right] &\leq \exp\left(-\frac{t^2(1+\ln(i^2))}{K_i^2}\right) \quad \text{(Via sub-Gaussianity of } X_i\text{'s)} \\ &\leq \exp\left(-\frac{t^2}{K_i^2} - \ln(i^2)\right) \end{aligned}$$

d. Using part (c), show that for every integer $n \geq 2$, there exists a constant c' such that:

$$\mathbf{Pr}\left[\max_i^n |X_i| \leq c' \cdot K\sqrt{\ln n}\right] \geq 0.9$$

Cheat sheet

Here are some tools that you might find useful in this problem set. We have discussed most of them in class, but I wanted to make sure you have access to them. You can use them in your solution without proving them.

Bayes' Theorem For two probabilistic events A and B, the theorem is stated as:

$$\mathbf{Pr}[A|B] = \frac{\mathbf{Pr}[B|A]\mathbf{Pr}[A]}{\mathbf{Pr}[B]}$$

where $\mathbf{Pr}[A|B]$ is the probability of A given B, $\mathbf{Pr}[B|A]$ is the probability of B given A, $\mathbf{Pr}[A]$ is the probability of A, and $\mathbf{Pr}[B]$ is the probability of B.

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.

Total Law of Expectation The Total Law of Expectation can be stated as:

$$\mathbb{E}[X] = \sum_i \Pr(B_i) \cdot \mathbb{E}[X|B_i]$$

where $\mathbb{E}[X]$ is the expected value of X, B_i represents the partitions of the sample space, and $\mathbb{E}[X|B_i]$ is the conditional expectation of X given B_i. Another variant of this law, is the following for two random variables X and Y:

$$\mathbb{E}[X] = \mathbf{E}_Y[\mathbf{E}_X[X|Y]]\,.$$

This states that the expectation of X is equal to the expectation of the conditional expectation of X given Y.

Expectation of non-negative random variables If Y is a non-negative random variable:

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

Union Bound The Union Bound is a fundamental concept in probability theory used to provide an upper bound on the probability of the union of events. It is stated as:

$$\mathbf{Pr}\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n} \mathbf{Pr}[A_i]$$

where A_i are events in a probability space. It is important to note that this bound does NOT rely on any assumptions regarding the independence of the events A_i's.

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

Let $X_1, X_2, \ldots, X_m$ be independent non-identically distributed Bernoulli trials such that $\mathbf{Pr}[X_i = 1] = p_i$ and $\mathbf{Pr}[X_i = 0] = 1 - p_i$. 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 $\epsilon > 0$:

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

Chebyshev's inequality Let X be a random variable with non-zero variance σ^2 and mean μ. Then for any real number $k > 0$:

$$\mathbf{Pr}[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

Markov's inequality If X is a non-negative random variable and $\alpha > 0$, then the probability that X is at least α is at most the expectation of X divided by α:

$$\mathbf{Pr}[X \geq \alpha] \leq \frac{\mathbb{E}[X]}{\alpha}$$

Poisson distribution The *Poisson distribution* is a discrete probability distribution that models the number of events occurring in a fixed interval of time or space, under the assumption that these events occur with a constant mean rate λ and independently of the time

since the last event. Its probability mass function is given by

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

A key property of the Poisson distribution is that both its mean and variance are equal to λ.

Inequalities The following inequality holds for every real number x:

$$1 - x \leq e^{-x}$$

For more inequalities like this, you may refer to this link.