Problem Set 2

Instruction:

- Submissions are due no later than **11:59 PM** on **Saturday, February 17, 2024**.
- 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.
- 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 i randomly chooses a day in $[n]$ (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 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 and Z'_i be indicator random variables that indicate if the exhibit ends on day i in the standard and Poissonized settings, respectively. With this setting in mind, answer the following questions:

a. What is the expected value of Z'_i?

b. Consider a subset of indices $I \subseteq \{t, \ldots, n-1\}$, such that for every $i, j \in I$, Z'_i and Z'_j are independent. What is the probability that $\sum_{i \in I} Z'_i$ is larger than its expectation by a factor of $(1+\epsilon)$?

c. Clearly, Z'_i is NOT independent from every Z'_j, however, it is independent from many of them (which ones?). Using this observation, try to get a similar bound to the previous part for the probability of $\sum_{i=t}^{n-1} Z'_i$ being larger than $(1+\epsilon)$ factor of its expectation.

d. It is known that for any event that happens with probability p in the Poissonized setting, it happens with probability at most $e\sqrt{m}\,p$ in the standard setting. Find n in terms of m and t such that we know that the exhibit does not close down early with probability at least 99% (in the standard setting).

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: $\hat{p}_i = \frac{Y_i}{m}$.

a. For a fixed (not randomized) vector $a \in \{-1, +1\}^n$, show that $a \cdot \hat{p}$ is close to $a \cdot p$. More precisely, for every $i \in [n]$, let

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

Show that there is a constant c such that:

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

b. 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.

c. 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 (ϵ, δ)-tester that uses $O(n \log(\delta^{-1})/\epsilon^2)$ samples and runs in $O(n \cdot t + m)$ time.

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. Find m in terms of $\|p\|_2^2$ such that Y is a $(1+\gamma)$-factor approximation of $\|p\|_2^2$. That is

$$\mathbf{Pr}\big[\,|Y - \|p\|_2^2| \geq \gamma \,\|p\|_2^2\}\big] \leq 0.9 \,.$$

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

b. As shown in part (a), estimating $\|p\|_2^2$ via Y is challenging due to the dependency of sample complexity on $\|p\|_2^2$ (which we aim to estimate). To address this issue, we would like to design an algorithm with adaptive sample complexity, which adjusts the number of samples based on the samples observed. The cost of such algorithms is often measured by their expected sample complexity. Can you develop an algorithm that provides a $(1+\gamma)$-factor approximation of $\|p\|_2^2$ with an expected sample complexity of $\tilde{O}(m)$, similar to part (a)? $\tilde{O}(m)$ here means we may have extra polylogarithmic factors hidden in the O notation.

Hint: Can you solve the problem if you were guaranteed that $\|p\|_2^2$ is in the range $[2^{-i}, 2^{-(i-1)}]$?

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:

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

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

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