COMP 585 **Spring 2025**

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

Problem Set 1

Instruction:

- Submissions are due no later than **11:59 PM** on **Tuesday, January 28, 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. (10 points) Imagine that in the United States, 10% of the population suffers from diabetes. Alex went to a laboratory in Houston for a diabetes blood test and received a positive result. Concerned about the accuracy of this result, Alex researched the lab's

reliability. An independent investigation revealed some past inaccuracies in their testing. Specifically, there's a 1% chance of a false positive (indicating diabetes when there is none) and a 5% chance of a false negative (failing to detect diabetes when it is present) in their tests. Given these factors, what is the probability that Alex actually has diabetes?

Problem 2. (10 points) Suppose we have an algorithm $\mathcal{A}$ that uses m samples from a distribution P and outputs $\hat{\mu}$ as an approximation for the true mean of the distribution, μ, with probability at least 2/3. More precisely, we have:

$$\mathbf{Pr}\left[\frac{\mu}{2} \leq \hat{\mu} \leq 2\,\mu\right] \geq 2/3\,.$$

Design an algorithm that uses $O(m \log(1/\delta))$ samples and outputs $\tilde{\mu}$, for which we have:

$$\mathbf{Pr}\left[\frac{\mu}{2} \leq \tilde{\mu} \leq 2\,\mu\right] \geq 1 - \delta\,.$$

Include the proof of performance for your algorithm.

Problem 3. (30 points) Suppose we have an array A consisting of n *distinct* elements. Our objective is to sort this array using the randomized quick-sort algorithm, as described below. In this problem, we aim to demonstrate that the randomized quick-sort algorithm operates in $O(n \log n)$ time on average. Let e_i represent the i-th smallest element in the array (note that e_i may or may not be located at position $A[i]$).

Algorithm 1 Randomized quick sort algorithm

```
1:  procedure QUICK-SORT(A)
2:      A_L and A_R ← empty arrays
3:      ℓ ← a random number in [n]
4:      for i = 1, ..., size(A) do
5:          if i = ℓ then
6:              Continue with the next i.
7:          if A[i] < A[ℓ] then
8:              Add A[i] to A_L.
9:          else
10:             Add A[i] to A_R.
11:     QUICK-SORT(A_L)
12:     QUICK-SORT(A_R)
13:     return concatenation of A_L + A[ℓ] + A_R.
```

Let e_i denote the i-th element in the sorted version of array A. Now, consider two distinct elements of the array, e_i and e_j, where $i < j$. Let X_{ij} be an indicator variable that denotes whether e_i and e_j were compared during the course of the algorithm (in Line 7). With this definition in mind, please answer the following questions:

a. In the first round of the algorithm (right before invoking the recursions), what is the probability that e_i and e_j are compared? Additionally, what is the probability that e_i

and e_j are added to two different sub-arrays, A_R and A_L, thereby implying that they will never be compared again after this round?

b. Compute the expected value of X_{ij} based on the values of i and j.

Hint: Note that in each round, there are three possible outcomes for X_{ij}:

1. e_i and e_j are compared, resulting in $X_{ij} = 1$.
2. e_i and e_j are placed into two different sub-arrays, A_R and A_L, resulting in $X_{ij} = 0$.
3. The value of X_{ij} remains undetermined and will be resolved in future rounds.

c. Can you express the expected running time of the algorithm in terms of X_{ij}? Using the expected value of X_{ij} that you computed earlier, what is the expected time complexity of this quick-sort algorithm?

Hint: You may want to use the fact that the harmonic series, $H_n \coloneqq \sum_{i=1}^{n} i^{-1}$, is in $O(\log n)$.

Problem 4. (50 points) Suppose Rice University has n colleges and has admitted m students for undergraduate studies. On the first day of classes, every student wears a hat, and the hat shouts the name of the student's college (each with probability $1/n$): "GriffinJones", "HuffleRice", "sLovettin", etc.

a. How many students, in expectation, do we need to see before we encounter at least one student from each college?

Hint: Try to break down the problem into determining the expected value of X_i, where X_i represents the number of new students you need to observe to discover the i-th college after successfully discovering $i - 1$ distinct colleges.

b. Show that, given a sufficiently large constant c, if there are $m \geq n \ln n + c \cdot n$ students, then the probability of having seen at least one student from each college is at least $1 - e^{-c}$.

c. The school administration is concerned about the unbalanced distribution of students among the colleges. Show that if $m > 3\, n \ln(n^2/2)$, then with probability at least $1 - 1/n$, the number of students per college will fall within the range:

$$\left[\frac{m}{n} - \sqrt{3 \cdot \frac{m}{n} \cdot \ln(n^2/2)},\ \frac{m}{n} + \sqrt{3 \cdot \frac{m}{n} \cdot \ln(n^2/2)}\right].$$

d. Show that if $m = n$, then with probability at least $1 - 1/n$, no college will have more than $O\left(\frac{\ln n}{\ln \ln n}\right)$ students, provided that n is sufficiently large.

Hint: You might find it helpful to use the version of the Chernoff bound that works for any $\epsilon > 0$.

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.

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 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 $\epsilon > 0$:

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

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.