COMPSCI 514: Problem Set 1

Due: 9/20 by 11:59pm in Gradescope.

Instructions:

- You are allowed to work on this problem set in a group of up to three members.
- You should choose your group from within your own class (either online or in-person).
- You may talk to members of other groups at a high level about the problems but **not work through the solutions in detail together**.
- Each group should **submit a single solution set**: one member should upload a pdf to Gradescope, marking the other members as part of their group in Gradescope.
- You must show your work/derive any answers as part of the solutions to receive full credit.

Core Competency Problems

1. Probability Practice (12 points)

1. (2 points) Let $\mathbf{X}$ and $\mathbf{Y}$ be independent random variables taking values in the sets S and T respectively. For any subsets $S' \subseteq S$ and $T' \subseteq T$ let A be the event that $\mathbf{X} \in S'$ and B be the event that $\mathbf{Y} \in T'$. Prove that A and B are independent events. I.e., that $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$.

   Hint: As a concrete example to help parse the notation: $\mathbf{X}$ and $\mathbf{Y}$ could be independent six-sided dice rolls. We would have $S = T = \{1, 2, 3, 4, 5, 6\}$. We could let $S' = \{1, 3, 5\}$, in which case, A is the event that the first die is odd. We could let $T' = \{1, 2, 3\}$, in which case, B is the event that the second die has value at most 3.

2. (2 points) Consider storing n items in a hash table with $m = 2n$ buckets, using a fully random hash function $\mathbf{h} : [n] \to [2n]$ (i.e., each item is assigned independently to a uniform random bucket). What is the probability that a given item lands in its own bucket (i.e., that it does not collide with any other items)? What is the limit of this probability as $n \to \infty$?

3. (2 points) Consider a list of n distinct numbers: $x_1, \ldots, x_n$. We say that x_i and x_j are 'inverted' if $i < j$ but $x_i > x_j$: i.e., if x_i and x_j are out of order. The Bubble Sort algorithm sorts a list by moving from left to right and swapping adjacent inverted numbers until the list is sorted. If $x_1, \ldots, x_n$ is a uniformly random permutation, what is the expected number of swaps that Bubble Sort makes? **Hint:** Use linearity of expectation.

4. (2 points) Consider an oversimplified model of a stock price over time: a stock starts at price $\mathbf{q}_0 = 1$. Each day, with probability 1/2 its value increases by a multiplicative factor of $r > 1$ and with probability 1/2 it decreases by the same multiplicative factor. I.e., $\mathbf{q}_{i+1} = \mathbf{q}_i \cdot r$ with probability 1/2 and $\mathbf{q}_{i+1} = \mathbf{q}_i / r$ with probability 1/2. Assume that the fluctuations are

independent across days. What is $\mathbb{E}[\mathbf{q}_n]$? What is $\text{Var}[\mathbf{q}_n]$? **Hint:** You may want to use that $\text{Var}[\mathbf{X}] = \mathbb{E}[\mathbf{X}^2] - \mathbb{E}[\mathbf{X}]^2$.

5. (1 point) Over a long period of time, is the above stock a good investment, a bad one, or a neutral one? Explain why in a sentence or two.

6. (2 points) Consider a similar oversimplified model of a stock price over time: a stock starts at price $\mathbf{q}_0 = 1$. Each day, with probability 1/2 its value increases by an additive factor of $r > 0$ and with probability 1/2 it decreases by the same additive factor. I.e., $\mathbf{q}_{i+1} = \mathbf{q}_i + r$ with probability 1/2 and $\mathbf{q}_{i+1} = \mathbf{q}_i - r$ with probability 1/2. Assume that the fluctuations are independent across days. What is $\mathbb{E}[\mathbf{q}_n]$? What is $\text{Var}[\mathbf{q}_n]$? **Note:** Under this model the stock price may become negative.

7. (1 point) Over a long period of time, is the above stock a good investment, a bad one, or a neutral one? Explain why in a sentence or two.

2. Concentration Bound Practice (8 points)

1. (2 points) Suppose you have a randomized algorithm $\mathcal{A}$ for solving some problem that has expected running time T on any input and always outputs the correct answer. Design an algorithm that has *worse-case* running time $5T$ and outputs the correct answer on any input with probability at least 4/5.

2. (2 points) Suppose you have a random variable $\mathbf{X}$ with $\mathbb{E}[\mathbf{X}] = 100$. Suppose also that $\mathbf{X}$ always lies in the range $[50, 500]$. Give the best (i.e., the smallest) upper bound you can on $\Pr[\mathbf{X} \geq 200]$. Argue that it is not possible to give a tighter bound without more information about $\mathbf{X}$.

3. (2 points) There are 100 small towns in Western Massachusetts. Each has a 1/2 chance of getting snow this November. Let $\mathbf{X}$ be the number of these towns that get snow this November. So $\mathbb{E}[\mathbf{X}] = 50$. By Markov's inequality, the probability that all of the towns get snow is at most $50/100 = 1/2$. Explain why this bound cannot be improved if no further information is given.

4. (2 points) Suppose you take a random walk on a number line – you start at position 0 and at each step, with probability 1/2 you increment your position by 1. With probability 1/2 you decrement it by 1. Your movements in each step are independent of each other. Prove that after n steps, with probability at least 15/16, your position has magnitude at most $4\sqrt{n}$. **Hint:** Apply Chebyshev's inequality.

3. Mark-and-Recapture Analysis (12 points)

You want to estimate the number of individuals in a large population (e.g., a population of animals), by randomly capturing individuals from the population, tagging them, and observing if you re-capture them in the future. The less re-captures you see, the higher your estimate for the population size will be. This idea is widely employed in ecology for population size estimation, and is similar to the CAPTCHA database example discussed in class.

1. (2 points) Consider capturing m individuals, which you assume are drawn independently and uniformly at random with replacement from a population of size n. Let $\mathbf{D}$ denote the number of pairs of captured individuals that are the same. What is $\mathbb{E}[\mathbf{D}]$?.

2. (2 points) For any $i < j$, let $\mathbf{D}_{i,j}$ be a random variable, which is 1 if the i^{th} and j^{th} captured individuals are the same, and 0 otherwise. Prove that the $\mathbf{D}_{i,j}$ random variables are *pairwise independent*. I.e., for any two pairs (i,j) and (k,ℓ) that differ in at least one element, $\mathbf{D}_{i,j}$ and $\mathbf{D}_{k,\ell}$ are independent.

3. (2 points) Prove that for any set of pairwise independent random variables $\mathbf{X}_1, \ldots, \mathbf{X}_z$,

$$\text{Var}\left[\sum_{i=1}^{z} \mathbf{X}_i\right] = \sum_{i=1}^{z} \text{Var}[\mathbf{X}_i].$$

   This is, pairwise independence suffices for linearity of variance to hold.

   Use this fact to show that $\text{Var}[\mathbf{D}] \leq \frac{\binom{m}{2}}{n}$.

4. (2 points) Prove that for any $\epsilon, \delta \in (0,1)$ if we set $m \geq \frac{2\sqrt{n}}{\epsilon\sqrt{\delta}}$ then with probability at least $1-\delta$, $|\mathbf{D} - \mathbb{E}[\mathbf{D}]| \leq \epsilon\mathbb{E}[\mathbf{D}]$. **Hint:** I used that fact that for $m \geq 2$, $\binom{m}{2} \geq \frac{m^2}{4}$ to simplify my calculations. Do not worry about getting the absolute tightest bound here. If you can show the above bound for $m \geq \frac{c\sqrt{n}}{\epsilon\sqrt{\delta}}$ for any fixed constant c, then you will receive full credit.

5. (2 points) Consider estimating the population size as $\tilde{n} = \frac{\binom{m}{2}}{\mathbf{D}}$. Prove that if $|\mathbf{D} - \mathbb{E}[\mathbf{D}]| \leq \frac{\epsilon}{2} \cdot \mathbb{E}[\mathbf{D}]$ for some $\epsilon \in (0,1)$, then $|\tilde{n} - n| \leq \epsilon n$. **Hint:** Use that for any $x \in (0, 1/2)$, $\frac{1}{1-x} \leq 1+2x$ and that for any $x \in (0,1)$, $\frac{1}{1+x} \geq 1-x$.

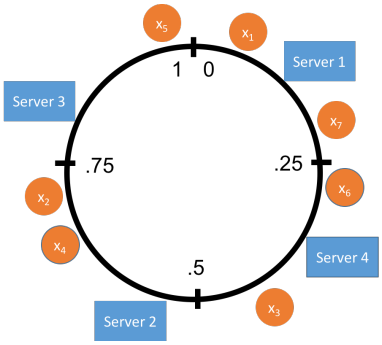
6. (2 points) Conclude that for any $\epsilon, \delta \in (0,1)$, setting $m \geq \frac{4\sqrt{n}}{\epsilon\sqrt{\delta}}$ suffices to estimate the population size to error ϵn with probability at least $1-\delta$.

Challenge Problems (Complete 1 of 2)

C1. Dynamic Load Balancing (10 points)

Consider a large scale distributed database, storing items coming from some universe U. A random hash function $\mathbf{h} : U \to [m]$ is used to assign each item x to one of m servers. When that item is queried in the future, the hash function is used to identify which server it is stored on. In many applications, the number of servers scales dynamically, depending on the storage load, availability, etc. If a new server is added to the current set of m, since $\mathbf{h}$ maps only to $[m]$, we will have to pick a new hash function, rehash and move all the stored items.

Consider the following solution: pick a random hash function $\mathbf{h}$ which maps both items and servers to values chosen independently and uniformly at random in the range $[0,1]$. Each item is assigned to the first server to its right, with wrap-around. I.e., item x is assigned to the server with the smallest hash value larger than $\mathbf{h}(x)$. If there are no servers with a larger hash value, the item is assigned to the server with the smallest hash value. When a new server is added to the system it is hashed to $[0,1]$ and any items that should be assigned to it are moved.



1. (2 points) Assume there are n items and m servers, what is the expected number of items that is assigned to each server? If a new server is added to the system, what is the expected number of items that must be moved?

2. (3 points) Show that with probability $\geq 9/10$, with m servers in total, no server is assigned a hash range of width greater than $\frac{10 \ln m}{m}$. **Hint:** Fix a single server and then think about how the remaining $m-1$ servers must be hashed so that its hash range has size $\geq \frac{10 \ln m}{m}$.

3. (3 points) Show using a concentration bound that with probability $\geq 9/10$ the maximum load on a server is $\leq \frac{20n \ln m}{m}$.

4. (2 points) Let $\mathbf{X}_i$ be the width of the hash range assigned to server i when there are m servers in total. Give an exact formula for $\text{Var}[\mathbf{X}_i]$. **Hint:** You may want to use that for a non-negative random variable $\mathbf{Y}$, $\mathbb{E}[\mathbf{Y}] = \int_0^\infty \Pr[\mathbf{Y} \geq x]\, dx$.

You may assume that m and $\frac{n}{m}$ are both large, say > 30 to get your bounds to hold. You may want to use the helpful inequality: for any $x > 0$ and $c > 0$ with $\frac{c}{x} \leq 1$, $\left(1 - \frac{c}{x}\right)^x \leq e^{-c}$.

C2. Testing Uniform Samples with Duplicates (10 points)

Note: You may want to complete Core Problem 3 before attempting this problem.

Let P be a distribution over $[n]$ that places probability p_i on outcome i. We would like to take a small number of samples from P and determine if P is close to uniform – i.e., if it is close to placing probability $1/n$ on each outcome. We let $\Delta(P) \stackrel{\text{def}}{=} \sum_{i=1}^n |p_i - 1/n|$ denote P's distance to uniformity.

1. (2 points) Consider taking m independent samples from P. Let $\mathbf{D}$ be the number of pairwise duplicate samples we observe. Prove that $\mathbb{E}[\mathbf{D}] = \binom{m}{2} \cdot \sum_{i=1}^n p_i^2$.

2. (2 points) Prove that $\text{Var}[\mathbf{D}] \leq \binom{m}{2} \cdot \sum_{i=1}^n p_i^2 + 6\binom{m}{3} \cdot \sum_{i=1}^n p_i^3$.

3. (2 points) Argue that for any $\gamma \in (0,1)$, if we take $m = \frac{c\sqrt{n}}{\gamma^2}$ samples for a large enough constant c, then, with probability at least $9/10$, $|\mathbf{D} - \mathbb{E}[\mathbf{D}]| \leq \gamma \mathbb{E}[\mathbf{D}]$.

4. (2 points) Let $\Delta_2(P) = \sum_{i=1}^n (p_i - 1/n)^2$. Prove that $\sum_{i=1}^n p_i^2 = 1/n + \Delta_2(P)$. In turn, prove that $\sum_{i=1}^n p_i^2 \geq \frac{1+\Delta(P)^2}{n}$.

5. (2 points) Describe and analyze an algorithm that takes $O(\sqrt{n}/\epsilon^4)$ samples from P and satisfies: a) if P is in fact the uniform distribution, the algorithm outputs YES with probability at least $9/10$; b) if $\Delta(P) \geq \epsilon$, the algorithm outputs NO with probability at least $9/10$.