COMPSCI 514: Problem Set 1

Due: 9/22 by 11:59pm in Gradescope.

Instructions:

- You are allowed to work on this problem set in a group of up to three members.
- 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 should separately submit the core competency problems from any challenge problems you choose to complete. These do not necessarily need to be submitted with the same groups.
- You may talk to members of other groups at a high level about the problems but **not work through the solutions in detail together**.
- You must show your work/derive any answers as part of the solutions to receive full credit.

Core Competency Problems

1. Concentration Bound Practice (10 points)

1. (2 points) On a given day, 1000 users visit your website. User i makes a purchase with probability p_i. You are given that $\sum_{i=1}^{1000} p_i = 100$. Let $\mathbf{S}$ be the total number of purchases that are made. What is $\mathbb{E}[\mathbf{S}]$?
2. (2 points) Use Markov's inequality to give an upper bound on the probability that at least 500 purchases are made.
3. (2 points) Let $\mathbf{S}_i$ be an indicator random variable which is 1 if user i makes a purchase and 0 otherwise. Show that $\text{Var}[\mathbf{S}_i] \leq p_i$.
4. (2 points) What additional information do you need to use part (3) to give an upper bound on $\text{Var}[\mathbf{S}]$? Assuming this additional information, compute such an upper bound.
5. (2 point) Use part (4) and Chebyshev's inequality to give a tighter upper bound (as compared to part (2)) on the probability that at least 500 purchases are made.

2. Probability and Expectation Practice (10 points)

1. (2 points) Prove that if $\mathbf{X}$ and $\mathbf{Y}$ are independent random variables, $\mathbb{E}[\mathbf{X} \cdot \mathbf{Y}] = \mathbb{E}[\mathbf{X}] \cdot \mathbb{E}[\mathbf{Y}]$.
2. (2 points) Design two random variables $\mathbf{X}$ and $\mathbf{Y}$ that satisfy $\text{Var}(\mathbf{X}) + \text{Var}(\mathbf{Y}) > \text{Var}(\mathbf{X} + \mathbf{Y})$. Design two different random variables that satisfy $\text{Var}(\mathbf{X}) + \text{Var}(\mathbf{Y}) < \text{Var}(\mathbf{X} + \mathbf{Y})$
3. (2 points) For a random variable $\mathbf{X}$, let $D[\mathbf{X}] = \mathbb{E}[|\mathbf{X} - \mathbb{E}[\mathbf{X}]|]$ denote the mean absolute deviation of $\mathbf{X}$ from its mean (this is like the variance but without squaring the deviation). True or False: if $\mathbf{X}, \mathbf{Y}$ are independent then $D[\mathbf{X} + \mathbf{Y}] = D[\mathbf{X}] + D[\mathbf{Y}]$. If true, prove it. If false, give a counterexample.

4. (2 points) For any $t > 0$ exhibit a non-negative random variable $\mathbf{X}$ for which Markov's inequality is tight. I.e., for which $\Pr[\mathbf{X} \geq t] = \frac{\mathbb{E}[\mathbf{X}]}{t}$.

5. (2 points) Consider storing n items in a hash table with $m = 4n$ buckets, using a fully random hash function $\mathbf{h} : [n] \to [4n]$ (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. Population Size Estimation via Mark-and-Recapture (11 points)

You want to estimate the number of individuals in a large population (e.g., a population of animals in some area), 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 w individuals, which you assume are drawn independently and uniformly at random with replacement from a population of size n. Let $\mathbf{C}$ denote the number of pairs of captured individuals that are the same (which you can count by observing if you have already tagged any captured individuals). Prove that $\mathbb{E}[\mathbf{C}] = \frac{\binom{w}{2}}{n}$.

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

   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. **Hint:** Use that $\text{Var}[\mathbf{X}] = \mathbb{E}[\mathbf{X}^2] - \mathbb{E}[\mathbf{X}]^2$ and that when $\mathbf{X}, \mathbf{Y}$ are independent, $\mathbb{E}[\mathbf{X} \cdot \mathbf{Y}] = \mathbb{E}[\mathbf{X}] \cdot \mathbb{E}[\mathbf{Y}]$.

3. (2 points) Use the above fact to show that $\text{Var}[\mathbf{C}] \leq \frac{\binom{w}{2}}{n}$. **Hint:** First compute $\text{Var}[\mathbf{C}_{i,j}]$.

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

5. (2 points) Consider estimating the network size as $\tilde{n} = \frac{\binom{w}{2}}{\mathbf{C}}$. Prove that if $|\mathbf{C} - \mathbb{E}[\mathbf{C}]| \leq \frac{\epsilon}{2} \cdot \mathbb{E}[\mathbf{C}]$ 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. (1 point) Conclude that for any $\epsilon, \delta \in (0, 1)$, setting $w \geq \frac{4\sqrt{n}}{\epsilon\sqrt{\delta}}$ suffices to estimate the population size to error ϵn with probability at least $1 - \delta$.

Challenge Problems

C1. Network Size Estimation via Colliding Crawls

Hint: It will be helpful to solve Problem 3 before tackling this problem.

You want to estimate the number of nodes n in a large network (the Facebook social network, the web, etc.), by randomly crawling the network. Assume that after enough steps of randomly hopping from node to node in the network, your random crawl lands on node i with probability $p_i = \frac{d_i}{m}$ where d_i is the degree of node i and $m = \sum_{i=1}^{n} d_i$ is the sum of degrees. This is known as the 'steady state' distribution of a random walk.

1. Consider sending out w independent random crawls that end at w random nodes. Let $\mathbf{C}$ denote the number of pairwise collisions between these nodes. What is $\mathbb{E}[\mathbf{C}]$? **Hint:** The expression will depend on w, m, and the node degrees.

2. Consider sending out w independent random crawls that end at w random nodes, $i_1, i_2, \ldots, i_w$. Let $\mathbf{D} = \sum_{j=1}^{w} d_{i_j}$ be the sum of degrees of the sampled nodes. What is $\mathbb{E}[\mathbf{D}]$.?

3. Consider sending out w independent random crawls that end at w random nodes, $i_1, i_2, \ldots, i_w$. Let $\mathbf{I} = \sum_{j=1}^{w} 1/d_{i_j}$ be the sum of *inverse node degrees* of the sampled nodes. What is $\mathbb{E}[\mathbf{I}]$.?

4. Describe an estimator $\tilde{n}$ for n based on the statistics $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{I}$.

5. Show that for any $\epsilon \in (0, 1)$, if $|\mathbf{C} - \mathbb{E}[\mathbf{C}]| \leq \frac{\epsilon}{6} \cdot \mathbb{E}[\mathbf{C}]$, $|\mathbf{D} - \mathbb{E}[\mathbf{D}]| \leq \frac{\epsilon}{6} \cdot \mathbb{E}[\mathbf{D}]$, and $|\mathbf{I} - \mathbb{E}[\mathbf{I}]| \leq \frac{\epsilon}{6} \cdot \mathbb{E}[\mathbf{I}]$, then $|n - \tilde{n}| \leq \epsilon n$. **Hint:** Don't worry too much about getting the absolute tightest result here. If you can show the bound with ϵn replaced by $c\epsilon n$ for any constant c, that is enough for full credit.

6. Pick one of $\mathbf{C}, \mathbf{D}$, or $\mathbf{I}$. Use Chebyshev's inequality to show how large must you set w such that the bound in part (5) holds with probability at least 99/100. Discuss (informally) how you might expect this sample size to compare to n for a large social network. Will it be significantly smaller than n (and thus our algorithm will be much faster than simply scanning the full network)?

   Hint: Do not worry about optimizing constants here. We are looking for the right scaling of w rather than the precise constant factor.

C2. Implementing Mark-and-Recapture[1]

The link https://en.wikipedia.org/wiki/Special:Random will bring you to a random article on English language Wikipedia.

1. Use this link to implement the mark-and-recapture algorithm from Problem 3 and evaluate the claim that English language Wikipedia has 6.7 million unique articles (see e.g., https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia.)

2. Include your code, all relevant results and calculations, and a discussion of how accurate you think your estimate is. We should be able to understand your approach and results without looking at the code. The code is just as a sanity check that you completed the problem.

[1] This problem is taken from Chris Musco's NYU Course: *CS-GY 6763: Algorithmic Machine Learning and Data Science.*

3. Describe your methodology for choosing the number of random articles sampled to make your estimate. Did you choose as single value up front? Did you have to adjust it at all?

4. How do you think the results are impacted by the fact that the random article feature doesn't return a truly uniformly random articles (see discussion at https://en.wikipedia.org/wiki/Wikipedia:FAQ/Technical#random)? Do you think this biases your estimate to be too low or too high and why? You might want to look at Problem C1 when thinking about this.

Hint: In Python, you can obtain a random url by running:

```
 import requests
response = requests.get("https://en.wikipedia.org/wiki/Special:Random")
random_url = response.url
```

In my experiments, downloading 5000 articles took around 30 minutes, so your code might take a bit to run. But compare this to scanning all possible articles to check the claim, which would take roughly 25 days (if your IP isn 't blocked for scrapping).

C3. Job Search 🌶🌶🌶

You have posted a job and received n applications. You plan to interview the candidates and use these interviews to hire the best candidate for the job. When you interview a candidate, you give them a score, with the highest score being the best. Assume there are no tied scores.

Unfortunately, due to your company's rules, you interview candidates one by one and after you interview the i^{th} candidate, you have to either offer them the job or reject them on the spot. You can never hire a candidate after you reject them.

Under this constraint, you still want to hire the best candidate with as high a probability as possible. Consider the following strategy: you interview the candidates in a random order (chosen uniformly at random from all $n!$ possible orderings). For some value $m > 0$, you reject all of the first m candidates. After the m^{th} candidate, you hire the first candidate who is better than all previously interviewed candidates.

1. For any $i \geq m+1$, what is the probability of the event that the best candidate is assigned to the i^{th} interview and further, that they are hired?

2. Use part (1) to show that the probability of hiring the best candidate with this strategy is equal to $\frac{m}{n} \cdot \sum_{i=m+1}^{n} \frac{1}{i-1}$.

3. Show that if you set $m = n/e$, then you succeed in hiring the best candidate with probability at least $1/e$. **Hint:** First prove that $\sum_{i=m+1}^{n} \frac{1}{i-1} \geq \ln n - \ln m$ by comparing this sum to an integral. You may assume for simplicity that m is an integer.

4. Consider the problem where you are allowed to make two job offers. You still need to either reject a candidate or make them an offer immediately after interviewing them. Describe a strategy that improves on the above bound of $1/e$ by as much as you can. Aim for success probability at least .45. My answer succeeds with probability $\geq 1/2$.