Assignment 1

COMP 3801- Fall 2022

1 Instructions

Each question is worth 10 marks. The assignment is due in Brighspace before the due date. Please write clearly and answer questions precisely. As a thumb rule, the answer should be limited to ≤ 2 written pages, with ample spacing between lines and in margins, for each Question. Always start a new question on a new page, starting with Question 1, followed by Question 2, ..., Question n. Please cite all the references you have used/consulted as the source of information for each question. If a question asks you to Prove or Show, please clearly spell out the proof - the technique used - and show each step of the proof. Don't expect (partial) marks if the main idea isn't clear to us.
For a random variable X, we refer to its expectation and variance by $E[X]$ and $V[X]$, respectively.

2 Problems

1. You are given a list L of distinct elements and want to choose a random element in this list. Each element of L should have the same probability of being chosen. Unfortunately, you do not know the number of elements in L. You are allowed to make only one pass over the list. Consider the following algorithm:

   Algorithm PickRandomElement(L):
   u = first element of L;
   $i = 1$;
   while u exists
   do with probability $1/i$, set $x = u$;
   u = successor of u in L;
   $i = i + 1$
   endwhile;
   return x

   Prove that the output x of this algorithm is indeed a random element of L. In other words, prove the following: Let v be an arbitrary element of L. Then, the probability that $x = v$ after PickRandomElement(L) has terminated is equal to $1/n$, where n is the number of elements in L.

2. This question extends previous algorithm. We now want to maintain $k > 1$ elements rather than a single element. Assume the list L consists of more than k elements. Here is the modified algorithm:

Algorithm PickRandom-k-Elements(L):
Let $u_1, u_2, \ldots, u_k$ be the first k-elements of L;
Form a set $R = \{u_1, u_2, \ldots, u_k\}$;
$i = k + 1$;
Let u be the i-th element of L;
while u exists
do with probability k/i, replace a random number in R with u.
    u = successor of u in L;
    $i = i + 1$
endwhile;
return R

For this question it will be helpful to think of elements in the list L are $u_1, u_2, u_3, \ldots$, where u_1 is the first element, u_2 is second, ... Consider the i-th iteration of the While loop, i.e. the iteration when the i-th element u_i in L is considered for the first time, for any $i > k$. Let x be any of the elements among the first i elements of L (i.e., $x \in \{u_1, u_2, \ldots, u_i\}$). Prove that the probability that $x \in R$ at the end of this iteration is k/i.

3. Recall that a permutation is one-to-one and onto function $\pi : [1, 2, 3 \ldots, n] \rightarrow [1, 2, 3 \ldots, n]$ such that for every integer i, $1 \leq i \leq n$, there is exactly one integer $j \in \{1, \ldots, n\}$ such that $\pi(i) = j$. For example the permutation $[12345] \rightarrow [23154]$ represents that elements are mapped as follows: $\pi[1] = 2$, $\pi[2] = 3$, $\pi[3] = 1$, $\pi[4] = 5$, and $\Pi[5] = 4$. We can visualize a permutation as a set of directed cycles as follows. Assume that we have a vertex for each number, $1 \leq i \leq n$. If $\Pi[i] = j$, we draw a directed arc from vertex i to j. (Notice that a cycle may be a self-loop if $\pi(i) = i$.) In our example permutation we have 2 directed cycles $1 \rightarrow 2 \rightarrow 3 \rightarrow 1$ and $4 \rightarrow 5 \rightarrow 4$, and their lengths are 3 and 2, respectively. What is the expected number of cycles in a random permutation?
(Hint: Each vertex is in some directed cycle. Show that the probability that a vertex i is in cycle of length k ($1 \leq k \leq n$) is $\frac{1}{n}$, irrespective of the length of the cycle. You may also want to define a r.v. X_i for each vertex i, where $X_i = \frac{1}{k}$, if vertex i participates in the cycle of length k. Think of the quantity $X = \sum_{i=1}^{n} X_i$.)

4. Let $V[X]$ denote the variance of a random variable X. Answer the following.

   (a) In your own words, what is the meaning of the variance of a random variable?

   (b) Show that if X and Y are independent random variables than $V[X+Y] = V[X]+V[Y]$.

   (c) Using a simple example, show that if X and Y are dependent then $V[X+Y] \neq V[X]+V[Y]$. For the same example, evaluate $E[X+Y]$ and check whether that equals $E[X]+E[Y]$. (Note that the linearity of expectation says that $E[X+Y = E[X]+E[Y]$ for any pair of random variables, irrespective of whether they are dependent or independent.)

   (d) Let X be a random variable and c a constant. Show that

      i. $V[X + c] = V[X]$.
      ii. $V[cX] = c^2V[X]$.
      iii. $V[X] \geq 0$.

iv. $V[X] = 0$ if and only if the r.v. X is a constant.

5. This exercise provides an alternate proof of Chebyshev's inequality using Markov's inequality. Let X be a random variable with mean μ and variance σ^2. Let $s > 0$ be a constant.

   (a) Show that $P(|X - \mu| \geq s) = P((X - \mu)^2 \geq s^2)$

   (b) Show that $P((X - \mu)^2 \geq s^2) \leq \frac{E[(X-\mu)^2]}{s^2}$

   (c) Show that $P(|X - \mu| \geq s) \leq \frac{\sigma^2}{s^2}$

6. Assume that our government wants to launch a new policy but wants to take a random population sample and see whether it is even worth the effort. Everybody in the population is either for or against the proposed policy. Assume there are no biases when selecting whom to survey. I.e., each individual is equally likely to be picked for the survey (and you may assume with replacement or without, whichever scenario makes it easier for the analysis). Assume that we survey a total of n people from the population. Let p be the proportion of people in the population supporting the policy, and let p' be the proportion of the people in the survey supporting the policy. Show, using Chebyshev's inequality, that for any constant $c > 0$, $Pr(|p - p'| > c) \leq \frac{1}{4nc^2}$.
   Hint: Set up an indicator random variable X_i for the i-th person in the sample. Evaluate $E[X_i]$ and $V[X_i]$. Note that $p(1 - p) \leq \frac{1}{4}$.
   For what value of n, there is at least 98% chance that $|p - p'| < 3\sigma$, where σ is the standard deviation of the random variable $X = \sum_{i=1}^{n} X_i$.

7. A popular coffee store in Amherstburg opens at 9 A.M. Many of its customers have indicated that they would prefer that the store opens at 7:30 A.M. The store has decided to estimate the fraction p of the town's population that will like an early opening by surveying town's population. They need to determine the number of people n to whom to send a survey to get within 10% of the true value of p with a probability of at least 0.95. Using Chebyshev's inequality, we will get an estimate on n. Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ be the $0-1$ indicator random variables corresponding to the set of n people picked, uniformly at random, from the town's population for the survey. Note that X_i is 1 if the i-th person prefers an early opening; otherwise it is 0. Let S_n be the number of people in $\mathcal{X}$ who will like an early opening. Define $A_n = \frac{1}{n} S_n$. Answer the following:

   (a) Show that $E[A_n] = p$, i.e., A_n is an unbiased estimator of p.

   (b) Show that variance σ of any of X_i is $V[X_i] = \sigma^2 = p(1 - p)$.

   (c) Show that $V[A_n] = \frac{1}{n}\sigma^2$.

   (d) Conclude that if n is large, $E[A_n]$ will be concentrated around p.

   Define the two parameters $\epsilon, \delta \in (0, 1)$, where ϵ accounts for the error in the estimate ($\epsilon = 0.1$ for our problem), and δ captures the desired accuracy. We are interested to evaluate $Pr(|A_n - p| \leq \epsilon p) \geq \delta$. I.e., for our problem, we want to understand what value of n will satisfy $Pr(|A_n - p| \leq 0.1 * p) \geq 0.95$? Note that, by Chebyshev's inequality, $Pr(|A_n - p| \geq \epsilon p) \leq \frac{V[A_n]}{(\epsilon p)^2}$. Answer the following.

   (a) Show that $n \geq \frac{p(1-p)}{(1-\delta)(\epsilon p)^2}$

(b) For $\epsilon = 0.1$ and $\delta = 0.95$, show that $n \geq \frac{2000(1-p)}{p}$.

(c) Show that for any $0 < p' \leq p$, $\frac{2000(1-p')}{p'} \geq \frac{2000(1-p)}{p}$

(d) If the coffee shop knows that $p' \geq 0.20$ (as gathered from the customers that come to the store daily) show that $n \approx 8000$ people are enough to survey to get the desired estimate on p.

(e) Conclude that if we know of a good lower bound p' of p, our estimate of n will be close to the optimal number of people to survey.

8. Recall the MWU method. We assumed that the algorithm gets a reward of 0 on the days it predicts correctly and loses a dollar on each day it makes the wrong prediction. Similarly, each expert gets a reward of 0 for making a correct prediction on a day and loses a dollar for making the wrong prediction. In the analysis of the MWU method, we had the expert who had at most m losses over T days. Using the potential function, we showed that the total loss M of our learning method/algorithm is at most $M \leq \frac{2}{\eta} \ln n + 2(1+\eta)m$, where n is the number of experts and $\eta \in (0, 1/2)$ is a constant.

   Suppose, instead of losses, we have gains. I.e., for each day our algorithm (or the expert) makes a correct prediction, we earn a dollar. Whereas, if the algorithm (or the expert) makes a mistake, their reward is 0 for that day. Let m be the gain of the best expert over T days, and let M be the total gain of the algorithm over T days. What will be the relationship between M, m, and n? I.e., derive an equivalent expression for the gain similar to what we did in the case of losses. Justify.

9. Here is an example of a tiny Bloom filter. It uses an array of 5 bits and two independent hash functions f and g. We want to test membership in a set S of three elements, so we hash each element using both f and g, and we set to 1 any bit that any of the three elements is hashed to by either of the hash functions. When a new element x arrives, we compute $f(x)$ and $g(x)$, and we say x is in the set S if both $f(x)$ and $g(x)$ are 1. Assume x is not in the set S. What is the probability of a false positive, i.e., the probability of saying that x is in S? Analyze using conditional probabilities. How does it compare with Bloom's analysis?

10. This problem is based on the Count-Min Sketch data structure. Suppose we want to ensure that we report all the elements that occur with a frequency of at least 3% in a data stream with probability ≥ 0.9. Try to come up with reasonable values of b and r for the Count-Min Sketch (CMS) table and justify your choice.