CS 6150: HW 5 – Randomized algorithms

Submission date: Thursday, November 23, 2023, 11:59 PM

This assignment has 5 questions, for a total of 50 points plus 5 bonus points. Unless otherwise specified, complete and reasoned arguments will be expected for all answers.

Question	Points	Score
Collecting coupons	15	
Brownian motion	12	
Trade-offs in sampling	6	
Satisfying ordering constraints	11	
Birthdays and applications	11	
Total:	55	

Question 1: Collecting coupons ..... **[15]**

A cereal company has decided to give out superhero stickers with boxes of its cereal. There are n superheroes in total, and suppose that each cereal box you buy has a sticker of a uniformly random superhero. What is the expected number of boxes you need to buy so that you end up with at least one copy of *all* the n stickers?

There are many ways to do this analysis; let us see one of them. We would like to write down a recurrence for the expected value. Define $f(n,k)$ to be the expected number of boxes you need to buy to end up with all the stickers, *given that you have already seen* k distinct stickers. Thus by definition, $f(n,n) = 0$, and the goal is to compute $f(n,0)$.

(a) **[6]** Use the law of conditional expectations to prove that

$$f(n,k) = \frac{n-k}{n}\left(1 + f(n,k+1)\right) + \frac{k}{n}\left(1 + f(n,k)\right).$$

Simplify this to evaluate $f(n,0)$. [*Hint:* you may use the identity $1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} = \log n + c$ for some $c \in (0,1)$.]

Describe a general solution in the right direction	3 points
Correct derivation	3 point

Answer: Note that you have already seen k distinct stickers, when you open a new box there are two different possible events. One is we see a new coupon we have not seen before (let this event be F) and the other is we do not see a new coupon (let this event be $\tilde{F}$). We can see that, given the event F,

$$Pr(F) = \frac{n-k}{n}$$

and

$$E(f(n,k)|F) = 1 + f(n,k+1)$$

since we have opened one box and now we have seen $k+1$ and need to see the rest. For the event $\tilde{F}$, we get,

$$Pr(\tilde{F}) = \frac{k}{n}$$

and

$$E(f(n,k)|\tilde{F}) = 1 + f(n,k)$$

since we have opened one box and we have only seen k.
This gives,

$$\begin{aligned} f(n,k) &= E(f(n,k)|\tilde{F})Pr(\tilde{F}) + E(f(n,k)|F)Pr(F) \\ &= \frac{n-k}{n}\left(1 + f(n,k+1)\right) + \frac{k}{n}\left(1 + f(n,k)\right) \end{aligned}$$

This gives, $\frac{n-k}{n}\left(f(n,k) - f(n,k+1)\right) = 1 \implies f(n,k) - f(n,k+1) = \frac{n}{n-k}$. Therefore, we can see that,

$$f(n,0) - f(n,n) = \sum_{k=0}^{n-1} f(n,k) - f(n,k+1) = \sum_{k=0}^{n-1} \frac{n}{n-k}$$

Note that $f(n,n) = 0$ and $\sum_{k=0}^{n-1} \frac{n}{n-k} = n\sum_{k=1}^{n} \frac{1}{k} = n(\log n + c)$ for a small constant c. Therefore, $f(n,0) = n\log n + cn$

(b) **[3]** Suppose $n > 4$. Prove that the probability that you need to buy $8n \log n$ boxes in order to see all the n stickers is $\leq 1/4$.

Describe a general solution in the right direction	2 points
Correct derivation	1 point

Answer: Note that for $n > 4$, $2n \log n \geq n \log n + cn$. Let the number of boxes we need to see be X. We know $E(X) = f(n, 0)$. Then we can see that using Markov's inequality,

$$Pr(X = 8n \log n) \leq Pr(X \geq 8n \log n) \leq \frac{f(n,0)}{8n \log n} \leq \frac{2n \log n}{8n \log n} = \frac{1}{4}$$

(c) **[6]** Use a more direct computation to bound the probability above by $\frac{1}{n^4}$. [*Hint:* what is the probability that you buy $8n \log n - 1$ boxes and you still have not seen a given sticker? Can you now use the union bound?]

[All logarithms above are natural logs. You might also find the inequality $1 - x \leq e^{-x}$ useful.]

Describe a general solution in the right direction	3 points
Correct derivation	3 point

Answer: Let us consider some coupon i. The probability you do not see coupon i on when we open a box is $1 - \frac{1}{n}$. Then the probability we do not see coupon i after $8n \log n - 1$ boxes are opened is $(1 - \frac{1}{n})^{8n \log n - 1} \leq (e^{-\frac{1}{n}})^{8n \log n - 1} = e^{-8 \log n} e^{\frac{1}{n}} = \frac{1}{n^8} e^{\frac{1}{n}}$. Let this event be F_i. Then, the letting F be the event that we do not see some coupon by $8n \log n - 1$ boxes, we get,

$$Pr(F) = Pr(\cup_{i=1}^{n} F_i) \leq \sum_{i=1}^{n} Pr(F_i) \leq n \cdot \frac{1}{n^8} e^{\frac{1}{n}} \leq \frac{1}{n^4}$$

Question 2: Brownian motion ..... **[12]**

Consider a particle moving on the real line, as follows: at time $t = 0$, it is at the origin, $X_0 = 0$. If it is at position s at time t, then the position at time $t + 1$ is $(s + 1)$ with probability $1/2$ and $(s - 1)$ with probability $1/2$.

(a) **[4]** Let X_t be the random variable denoting the location of the particle at time t. For some $t \geq 1$, compute $\mathbf{E}[X_t]$.

Describe a general solution in the right direction	2 points
Correct derivation	2 point

Answer: Let $X_{t-1} = s$. We can see that, $E(X_t | X_{t-1}) = \frac{1}{2}(s+1) + \frac{1}{2}(s-1) = s = X_{t+1}$. Therefore,

$$E(X_t) = \sum_{X_{t-1}} E(X_t | X_{t-1}) Pr(X_{t-1}) = \sum_{X_{t-1}} X_{t-1} Pr(X_{t-1}) = E(X_{t-1})$$

Given this we can see that $E(X_t) = E(X_1) = X_0 = 0$.

(b) [**5**] Compute $\mathbf{E}[X_t^2]$ for some integer $t \geq 1$, and use this to prove that with probability $\geq 3/4$, we have $|X_t| \leq 2\sqrt{t}$.

Describe a general solution in the right direction	3 points
Correct derivation	2 point

Answer: Let $X_{t-1} = s$. We can see that $E(X_t^2|X_{t-1}) = \frac{1}{2}(s+1)^2 + \frac{1}{2}(s-1)^2 = s^2 + 1 = X_{t-1}^2 + 1$. Again using a similar argument to the part (a), we can see that,

$$E(X_t^2) = \sum_{x_{t-1}} E(X_t^2|X_{t-1})Pr(X_{t-1}) = \sum_{x_{t-1}} \left(X_{t-1}^2 + 1\right) Pr(X_{t-1}) = E(X_{t-1}^2) + 1$$

We can also see that $E(X_1^2) = 1$. Given this we get,

$$E(X_t^2) - E(X_1^2) = \sum_{i=1}^{t-1} E(X_{i+1}^2) - E(X_i^2) = \sum_{i=1}^{t-1} 1$$

which implies $E(X_t^2) = t$.
We can now see that $Var^2(X_t) = E(X_t^2) - E(X_t)^2 = t - 0 = t$. Now using Chebyshev inequality, $Pr(|X_t - E(X_t)| \geq a \cdot Var(X_t)) \leq \frac{1}{a^2}$, we get,

$$Pr(|X_t| \geq 2\sqrt{t}) \leq \frac{1}{4}$$

Therefore, we can see that,

$$Pr(|X_t| \leq 2\sqrt{t}) \geq Pr(|X_t| < 2\sqrt{t}) \geq \frac{3}{4}$$

(c) [**3**] Part (b) shows that the magnitude of X_t after t steps of moving around is only $O(\sqrt{t})$. This raises the question: does it "move around" pretty uniformly in the interval say $(-\sqrt{t}, +\sqrt{t})$? Run experiments on the process with $t = 4 \cdot 10^4$. On average (over say 50 runs), how many times does the particle "cross the origin"? Repeat with $t = 9 \cdot 10^4$ and $t = 16 \cdot 10^4$ and report your answers.

Describe a general solution in the right direction	2 points
Correct evidence	1 point

Answer: We can run a simple program to get the following statistics,
Interpretation 1: Going to 0

t	Mean	Std.
$4 \cdot 10^4$	161.78	130.0746
$9 \cdot 10^4$	273.04	185.4874
$16 \cdot 10^4$	312.22	248.8340

Interpretation 2: Oscilating around 0 (changing the signs)

t	Mean	Std.
$4 \cdot 10^4$	74.00	59.1273
$9 \cdot 10^4$	104.46	75.6810
$16 \cdot 10^4$	146.60	92.8097

We also get that the percentage of steps spent in $(-\sqrt{t}, \sqrt{t})$ is around 70%.
Note that what you would get could be different from the values here but the general trend would remain the same. The code is available on Colab Code

Question 3: Trade-offs in sampling ..... **[6]**

For this problem, you need to run some basic experiments and write down the results you obtained. You **do not need to submit your code**, but if you prefer, you may add a publicly accessible link to the code (e.g., on github).

Suppose we have a population of size 1 million, and suppose 52% of them vote $+1$ and 48% of them vote -1. Now, randomly pick samples of size (a) 20, (b) 100, (c) 400, and evaluate the probability that $+1$ is majority even in your sample (by running the experiment say 100 times and taking the average). Write down the values you observe for these probabilities in the cases (a-c).

Next, what is the size of the sample you need for this probability to become 0.9?

Describe a general solution in the right direction	3 points
Correct, convincing evidence	3 point

Answer: We can see the following empirical probabilities from the experiments (after 100 trials).

Sample size	Empirical Probability
20	0.41
100	0.64
400	0.83

Note that the values you get might be different but the pattern should be the same. The code is available on Colab Code

Running the experiment with a sample size of approximately 1000, you should be able to get 0.9.

Question 4: Satisfying ordering constraints ..... **[11]**

Suppose we have n elements, labelled $1, 2, \ldots, n$, and our goal is to place them in some order on the line (thus the goal is to find a permutation π). We are also given m constraints. Each constraint has a triple (a, b, c), and the constraint is said to be *satisfied* if in the ordering we find, a does **not** lie "between" b and c (it need not be that b is to the left of c or vice versa). For example, if $n = 4$ and we consider the ordering 2431, then the constraint $(1, 4, 3)$ is satisfied, but $(3, 1, 2)$ is not.

Given the constraints, the goal is to find an ordering that satisfies as many constraints as possible (for simplicity, assume in what follows that m **is a multiple of** 3). For large m, n, this problem becomes very difficult.

(a) **[6]** As a baseline, let us consider a *uniformly random* ordering. What is the expected number of constraints that are satisfied by this ordering? [*Hint:* define appropriate random variables whose sum is the quantity of interest, and apply the linearity of expectation.]

Describe a general solution in the right structure	3 points
Correct solution	3 point

Answer: Let the Y_i be an indicator variable where $Y_i = 1$ if constraint i is satisfied (and 0 otherwise). Note that given a constraint i of the form (a, b, c), the possible arrangements of the constraints in the ordering are $(a, b, c), (a, c, b), (b, a, c), (b, c, a), (c, a, b), (c, b, a)$. Note that since we consider a uniformly random sampling any of a, b, c can be at any position with equal probability. Therefore, all six outcomes are equally likely and only 4 of them satisfy the constraint. Therefore, we can see that, $Pr(Y_i = 1) = 4/6 = 2/3$. Given this, letting X be the number of constraints satisfied, we can see that $X = \sum_{i=1}^{m} Y_i$ and therefore, $E(X) = \sum_{i=1}^{m} E(Y_i) = \sum_{i=1}^{m} \frac{2}{3} = \frac{2m}{3}$.

(b) **[5]** Let X be the random variable which is the number of constraints satisfied by a random ordering, and let E denote its expectation (which we computed in part (a)). Now, Markov's inequality tells us, for example, that $\Pr[X \geq 2E] \leq 1/2$. But it does not say anything that lets us argue that $\Pr[X \geq E]$ is "large enough" (which we need if we want to say that generating a few random orderings and picking the best one leads to many constraints being satisfied with high probability).

Describe a general solution with the right structure	3 points
Correct solution	2 point

Use the definition of X above to conclude that $\Pr[X \geq E] \geq 1/m$.

Answer: We know from part (a), $E = E(X) = \frac{2m}{3}$. Note that since we need the the $Pr(X \geq E)$ to be high, we should try to argue about the $Pr(Y = m - X \leq m - E)$ being high which is equivalent to $Pr(Y = m - x > m - E)$ being low. Since Y is an integer, we can see that, $Pr(Y > m - E) = Pr(Y \geq m - E + 1) \leq \frac{m-E}{m-E+1} = \frac{m}{m+3} = 1 - \frac{3}{m+3} \leq 1 - \frac{1}{m}$ (as long as $m \geq 2$). Given this, we can see that $Pr(X \geq E) = 1 - Pr(Y \leq m - E) \geq \frac{1}{m}$.

Alternative Answer: Assume $Pr(X \geq E) < 1/m$. Let this be ϵ. Then, we can see that $E \leq (E-1)Pr(X < E) + mPr(X \geq E) = (E-1)(1-\epsilon) + \epsilon m = E - 1 + \epsilon + \epsilon(m - E)$. This gives us $2m/3 \leq 2m/3 - 1 + \epsilon + \epsilon m/3 \implies 1 \leq \epsilon(1 + m/3) < \frac{1}{m}\left(1 + \frac{m}{3}\right) = \frac{1}{m} + \frac{1}{3}$ which is a contradiction. Therefore, $Pr(X \geq E) \geq \frac{1}{m}$.

Question 5: Birthdays and applications ..... **[11]**

Suppose we have n people, each of whom has their birthday on some random day of the year. Suppose there are m days in the year, and let us pretend that this is some parameter.

(a) **[5]** What is the expected *number of pairs* (i, j) with $i < j$ such that person i and person j have the same birthday? For what value of n (as a function of m) does this number become 1?

Describe a general solution with the right structure	3 points
Correct solution	2 point

Answer: For any pair of people (i, j), we can see that the probability that the (i, j) have the same birthday is $\frac{1}{m}$. Let $Y_{(i,j)} = 1$ if and only if (i, j) have the same birthday (0 otherwise). Note that we can now write the number of people with the same birthday, X as $X = \sum_{(i,j)|i<j} Y_{(i,j)}$. We can see that $E(X) = \sum_{(i,j)|i<j} E(Y_{(i,j)}) = \sum_{(i,j)|i<j} \frac{1}{m}$. We can see that there are $\frac{n(n-1)}{2}$ (i, j) pairs such that $i < j$. Therefore, $E(X) = \frac{n(n-1)}{2m}$.

For this value to be 1, we can see that $\frac{n(n-1)}{2m} = 1 \implies n^2 - n - 2m = 0$ and therefore, $n = \frac{1 \pm \sqrt{1+8m}}{2}$ and since n has to be positive this gives us $n = \frac{1+\sqrt{1+8m}}{2} \approx \sqrt{2m}$.

(b) **[6]** This idea has some nice applications in CS, one of which is in estimating the "support" of a distribution. Suppose we have a radio station that claims to have a library of one million songs, and suppose that the radio station plays these songs by picking, at each step a uniformly random song from its library (with replacement), playing it, then picking the next song, and so on.

Suppose we have a listener who started listening when the station began, and noticed that among the first 200 songs, there was a repetition (i.e., a song played twice). Prove that with probability > 0.9, the station's claim of having a million song database is false.

[One recent application of this idea was in proving that "GANs", a recent ML technique to produce realistic data such as images, typically have a pretty small support size.]

Describe a general solution with the right structure	3 points
Correct solution	3 point

Answer: Assume we have m songs in the library. Assuming we play n songs, the probability that there is no repitition (let this event be A), $Pr(A) = \prod_{i=0}^{n-1} \frac{m-i}{m} > \left(1 - \frac{n}{m}\right)^n$. Note that in our case, we have observed 1 repititon after $n = 200$ plays with a library of size $m = 1000000$. But with those parameters, we can see that $Pr(A) > \left(1 - \frac{2\times10^2}{10^6}\right)^{200} \approx 0.9608$. Which means that with probability > 0.96 there should be no repitition with $n = 200$ and $m = 1000000$. Therefore, with probability > 0.9, the claim made by the radio station is false.