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 $c \in (0, 1)$.]

Ans)

We know that f(n,k) represents the number of boxes we need to buy to see all stickers(n), stating that we have seen k distinct stickers.

Let us define an event X: We see a new sticker, given that we have seen k distinct stickers.
f(n,k) would consist of two parts: 1) We are seeing new stickers. 2) We are seeing old stickers.

Probability of seeing new stickers would be $\frac{n-k}{n}$

Therefore,
$\frac{n-k}{n} * (1 + f(n, k+1))$

Expected number of new boxes would be $1 + f(n, k+1)$ where 1 represents opening a box and $f(n, k+1)$ represents boxes we would be opening to get n stickers when we have seen k+1 distinct stickers.

Probability of seeing old(not new) stickers would be $\frac{k}{n}$

Expected new number of boxes would be $1+f(n,k)$ where 1 represents opening a box and $f(n, k+1)$ represents boxes we would be opening to get n stickers when we have seen k+1 distinct stickers.
Therefore,
$\frac{k}{n} * f(n, k+1)$

Therefore by law of conditional expectations.

$f(n,k) = \frac{n-k}{n} * (1 + f(n, k+1)) + \frac{k}{n} * f(n, k+1)$
Proved.
$f(n, 0) = \frac{n}{n}(1 + f(n, 1)) + 0$
$f(n, 0) = 1 + f(n, 1)$
Now we have to calculate f(n,1)
$f(n, 1) = \frac{n-1}{n} + \frac{n-1}{n} * f(n, 2) + \frac{1}{n} + \frac{1}{n} f(n, 1)$
$f(n, 1)(\frac{n-1}{n}) = 1 + \frac{n-1}{1} * f(n, 2)$
Dividing by $\frac{n-1}{n}$

$f(n-1) = \frac{n}{n-1} + f(n-2)$

Therefore we see a pattern

$f(n,n-1) = \frac{n}{1} + f(n,n) = \frac{n}{1}$ as f(n,n)=0

$f(n,0) = \frac{n}{n} + \frac{n}{n-1} + \dots \frac{n}{1}$
$= n(\frac{1}{n} + \frac{1}{n-1} + \dots \frac{1}{1})$
Using the identity.
$= n(\log n + c)$
$n \log n + cn$

(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 $\leq 1/4$.
Ans)
We are going to apply Markov's inequality.
$Pr[X \geq t.E[X]] \leq \frac{1}{t}$
let t=8
$Pr[X \geq 8.E[X]] \leq \frac{1}{8}$

Since $Pr[X \geq 8.E[X]]$ is less than $\frac{1}{8}$, it will be less than $\frac{1}{4}$ too.
Hence proved

(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.]

Ans)
Probability of seeing a particular sticker=$\frac{1}{n}$
Probability of not seeing a particular sticker=$1 - \frac{1}{n}$
This happens for $8n \log n - 1$ times.
$= (1 - \frac{1}{n})^{8n \log n - 1}$
We will use the hint $(1 - x) \leq e^{-x}$
We can write that as
$(1 - x)^k \leq e^{-kx}$, considering k as positive constant.
$x = \frac{1}{n}$ and $k = 8n \log n - 1$
$\leq e^{-(8n \log n - 1) * \frac{1}{n}}$
$\leq e^{(n \log n^{-8} + 1) * \frac{1}{n}}$
$\leq e^{(\log n^{-8} + n^{-1})}$
$\leq e^{(\log n^{-8}}) * e^{n^{-1}}$
Using log properties
$\leq \frac{1}{n^8} * e^{\frac{1}{n}}$
We notice that as n keeps on getting bigger, $e^{\frac{1}{n}}$ becomes very small so we ignore it.
Therefore,
$\leq \frac{1}{n^8}$
Therefore
$\leq \frac{1}{n^4}$

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 $\mathbf{E}[X_t]$.
Ans)
$E[X_t] =$ Probability at (s+1) and probability at s-1

$E[X_t] = \frac{1}{2} * (s+1) + \frac{1}{2} * (s-1)$
$E[X_t] = s$

Since we are at the origin at t=0. s would also be 0.

$E[X_t] = 0$

(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 $|X_t| \leq 2\sqrt{t}$.

Ans)
$X_t = X_{t-1} + S_t$
Here S_t is defined as a parameter which has two values
+1 meaning it moves forward with probability of $\frac{1}{2}$
-1 meaning it moves backward with probability of $\frac{1}{2}$
Var(X_t)=E($X_t - E(X_t)^2$)
We know $E(X_t)$ is 0.
$E(X_t^2) = \sigma^2$
Squaring the original equation.
$X_t^2 = X_{t-1}^2 + S_t^2 + 2 * X_{t-1} * S_t$
taking expectation both sides
$E[X_t^2] = E[X_{t-1}^2] + E[S_t^2] + E[2 * X_{t-1} * S_t]$
We can separate $E[X_{t-1} * S_t]$ into $E[X_{t-1}] * E[S_t]$
$E[X_{t-1}]$ and $E[S_t]$ both are 0.
Final equation
$E[X_t^2] = E[X_{t-1}^2] + 1$
We get a recurrence relation.
Let us assume a base case.
$E[X_o^2] = 0$
Let us write this as a normal recurrence.
$T(n) = T(n-1) + 1$
$T(n-1) = T(n-2) = 1$
$T(n) = T(n-2) + 2$
Say this happens for k steps
$T(n) = T(n-k) + k$
Let n-k=0
n=k
Therefore
$E[X_t^2] = t = \sigma^2$
$\sigma = \sqrt{t}$
Chebyshev's theorem
$P[X_t - E(X_t)] \geq k\sigma \leq \frac{1}{k^2}$
Let k=2
$P[X_t] \geq 2 * \sqrt{t} \leq \frac{1}{4}$
Taking complement
$P[X_t \leq 2 * \sqrt{t}] \geq 1 - \frac{1}{4}$
$P[X_t \leq 2 * \sqrt{t}] \geq \frac{3}{4}$

(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.

Ans)

For t=40000, average crossing is 156.04
For t=90000, average crossing is 206.2
For t=160000, average crossing is 338.64

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?

Ans)

a) When sample size is 20, P(+1 majority)=0.54
b) When sample size is 100, P(+1 majority)=0.58
c) When sample size is 400, P(+1 majority)=0.71
333 needs to be the sample size for getting 0.9 probability.

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

Suppose we have n elements, labelled $1, 2, \dots, n$, 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.]

Ans)

We define random variable as X_i with:

1, That means all the constraints are satisfied with the probability $\frac{4}{6}$ or $\frac{2}{3}$
0, That means all the constraints are not satisfied with the probability $\frac{2}{6}$ or $\frac{1}{3}$

Expected value is
$E[\sum_{i=0}^{m} X_i] = \frac{2}{3} * (1) + \frac{1}{3} * (0)$
$E[\sum_{i=0}^{m} \frac{2}{3}]$

$\frac{2m}{3}$

Since it is uniformly random ordering, it only can only be $\frac{2}{3}$ as all the items can be at any place.

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

Ans)

Let us consider Y as constraints that represents number of constraints not satisfied.

Y is complement of X
Y=m-X
Taking expectation both sides
$E[Y] = E[m] - E[X]$
$E[Y] = \frac{m}{3}$
We have
$Pr[X \geq E] \geq \alpha$
We take complement
$1 - Pr[X \geq E] \leq 1 - \alpha$
$Pr[X < E] \leq 1 - \alpha$
We saw in the first question, $E = \frac{2m}{3}$
Therefore
$X < \frac{2m}{3}$
Therefore
$X \leq \frac{2m}{3} - 1$
Therefore Y which is a complement would be
$Y \geq m - (\frac{2m}{3} - 1)$
$Y \geq m - E + 1)$
Changing the equation in terms of Y
$Pr[Y \geq m - E + 1] \leq 1 - \alpha$
Using Markov's inequality
$Pr[Y \geq m - E + 1] \leq \frac{E[Y]}{m-E+1}$
Therefore we get
$\frac{E[Y]}{m-E+1} = 1 - \alpha$
$\frac{\frac{m}{3}}{m-\frac{2m}{3}+1} = 1 - \alpha$
$\frac{m}{m+3} = 1 - \alpha$
$\alpha = \frac{3}{3+m}$
We know m is very large, so we ignore 3 in front of m.
$\alpha \approx \frac{3}{m}$
This is greater than $\frac{1}{m}$
$Pr[X \geq E] \geq \frac{1}{m}$
Proved

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?

Ans)
Probability for the 1st person to have distinct birthday=$\frac{m}{m} = 1$, since all days are available.
Probability for the 2nd person to have distinct birthday=$\frac{m-1}{m}$, since one day is gone.
Same birthday would be=1- distinct
$= 1 - \frac{m-1}{m} * 1$
$= \frac{1}{m}$
therefore
Selecting 2 people having same birthday would be:
$\binom{n}{2} * \frac{1}{m}$
It is given that this value is 1.
$\binom{n}{2} * \frac{1}{m} = 1$
$\frac{n!}{(n-2)!*2} * \frac{1}{m} = 1$
$\frac{n*(n-1)}{2} * \frac{1}{m} = 1$
$n * (n - 1) = 2m$

$n^2 - n - 2m = 0$
$n = \frac{1 \pm \sqrt{1-4(1)(-2m)}}{2}$
$n = \frac{1 \pm \sqrt{1+8m}}{2}$

(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.]

Ans)
Let us consider that we have m songs.
Probability that first song is unique is=$\frac{m}{m}$
Probability that second song is unique is=$\frac{m-1}{m}$
This will go on till 200 songs since we notice repetition then.
Therefore probability that last song is unique is=$\frac{m-199}{m}$
Total probability of distinct songs would be:
$\frac{m}{m} * \frac{m-1}{m} * \dots \frac{m-199}{m}$
This can be changed to this form:
$1 * (1 - \frac{1}{m}) * (1 - \frac{2}{m}) * \dots (1 - \frac{199}{m})$
Using $1 - x \leq e^{-x}$
$e^{\frac{-1}{m}} * e^{\frac{-2}{m}} * \dots * e^{\frac{-199}{m}}$
Therefore probability of song repetition would be 1-distinct
Therefore
$= 1 - e^{\frac{-1}{m}} * e^{\frac{-2}{m}} * \dots * e^{\frac{-199}{m}}$
$= 1 - e^{-1(\frac{1}{m} + \frac{2}{m} + \dots + \frac{199}{m})}$
$= 1 - e^{\frac{-1}{m}(1+2+\dots+199)}$
$= 1 - e^{\frac{-1}{m}(\frac{200*199}{2})}$
$= 1 - e^{\frac{-19900}{m}}$
We know this would be greater than 0.9
$0.9 \leq 1 - e^{\frac{-19900}{m}}$
$e^{\frac{-19900}{m}} \leq 0.1$
Taking log both sides
$\frac{-19900}{m} \leq -2.3$
$\frac{19900}{m} \geq 2.3$
$m \leq \frac{19900}{2.3}$
$m \leq 8652$
Therefore it is false that song database has 1 million songs.