CS 6150: HW 5 – Randomized algorithms

Submission date: Wednesday, November 20, 2019, 11:59 PM

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

Question	Points	Score
Collecting coupons	14	
Brownian motion	10	
Trade-offs in sampling	6	
Satisfying ordering constraints	10	
Birthdays and applications	10	
Total:	50	

Question 1: Collecting coupons . . . . **[14]**

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

(b) [**2**] 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 $\leq 1/4$.

(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.]

Question 2: Brownian motion . . . . **[10]**

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

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

(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.

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?

Question 4: Satisfying ordering constraints . . . . **[10]**

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.]

(b) [**4**] 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 $\Pr[X \geq E] \geq 1/m$.

Question 5: Birthdays and applications .... **[10]**

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) [**4**] 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?

(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.]