CS265/CME309, Fall 2019. Instructor: Gregory Valiant

Problem Set 3

Electronic submission via Gradescope due **11:59pm Tuesday 10/15**. You are strongly encouraged to submit a homework with a partner—that is, submit one homework with both of your names.

[You may discuss these problems with classmates. Feel free to look at wikipedia, course notes, etc. for reference material, but do not try to specifically search online for solutions to the problems. ***Your submission must be the original work of you and your partner, and you must understand everything that is written on your submission.*** *We strongly suggest that you write solutions using LaTex—see the course website for a latex solution template.]*

1. Chernoff Bound applications. For each question, be sure to specify which Chernoff bound you use:

   (a) (4 points) Suppose you are conducting a poll to forecast the upcoming election. Assume you randomly sample 10,000 likely voters, what is the probability that the percentage who support candidate X that you estimate differs from the true population value (the expected value) by more than an additive $\pm 2\%$? What about differing by more than 5%. What about more than 10%?

   (b) (4 points) How many random voters must you sample if you want to guarantee that with probability at least 0.95 (over the randomness in the choice of who you poll), you end up with an estimate of the percent of people who will vote for candidate X that is accurate to within an additive $\pm 1\%$? How does this change if you know that the fraction will vote for X is very small, say at most 5%?

2. In class, we saw a sampling-based randomized algorithm for computing the median of a set S of n (distinct) numbers. The algorithm succeeds with high probability (probability $> 1 - o(1)$) and, provided it succeeds, will compare $\frac{3}{2}n + O(n^{3/4} \log n)$ pairs of numbers. There is a different randomized algorithm that many of you might have seen, that resembles quick-sort with a random pivot. In this problem, you will show that this algorithm is significantly worse than the sampling-based algorithm we saw in class. The algorithm is as follows:

   - Choose a uniformly random element $x \leftarrow S$, and form the set $S_1 = \{y \in S : y < x\}$ and $S_2 = \{y \in S : y > x\}$ by comparing every element of S to x.
   - Based on the sizes of S_1 and S_2, determine which set the median lies in and recurse the algorithm on that set (keeping in mind the sizes of the partitions S_1,S_2 from earlier iterations of the algorithm in order to calculate which set the median of S belongs.

   (a) (4 points) What is the expected number of comparisons that the above algorithm performs on a list of n numbers? (The answer will be of the form $cn + o(n)$; figure out the leading constant, c).

   (b) (4 points) Show that with constant probability, the algorithm will exceed its expectation by at least n comparisons (and hence the runtime is not very strongly concentrated about its expectation.)

(c) DOUBLE BONUS You might wonder whether the leading terms of $3n/2$ in the number of comparisons that the randomized sampling based median algorithm uses is optimal. Prove that any randomized algorithm that finds the median of a set of n distinct numbers (with probability at least 0.9), must, in expectation, compare at least $3n/2$ pairs of numbers. [This is more of a research problem/food for thought, and we will not grade it.]

3. (4 points) For a random variable X distributed according to a Poisson distribution of expectation $\lambda \geq 0$, for all integers $k \geq 0$, by definition $\Pr[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}$. For independent random variables, X, Y, with X distributed according to a Poisson distribution of expectation λ_1, and Y distributed according to a Poisson distribution of expectation λ_2, show that $X + Y$ is distributed according to a Poisson distribution of expectation $\lambda_1 + \lambda_2$ by computing the moment generating function of the Poisson distribution, and then arguing that the moment generating function of $X + Y$ is equal to that of the Poisson distribution of expectation $\lambda_1 + \lambda_2$ (and hence, these two distributions must be equal).

4. Suppose a class has n students. On a given day, assume that if the ith student is asked "What is the temperature today", their response is drawn from a Gaussian (Normal) distribution whose mean is the actual temperature, and variance is σ_i^2 (and each student's guess is independent from the guesses of the other students).

   (a) (4 points) If $\sigma_i = 1$ for all i, how accurate will the average of the n guesses be? [Feel free to use the fact that the sum of independent Gaussians is Gaussian: namely for X_i distributed according to $N(\mu_i, \sigma_i^2)$, if the X_i are independent, then $\sum_i X_i$ is distributed according to $N(\sum_i \mu_i, \sum_i \sigma_i^2)$. This fact can be proved by looking at the moment generating function of X, as in the previous problem.] In this part, and subsequent parts, it is up to you to quantify what notion of accuracy you want—reasonable options include 1) bounding the expected error, 2) bounding the probability that the returned answer deviates from the truth by more than a certain amount.

   (b) (4 points) If $\sigma_i = 1$ for all i, how accurate can we expect the *median* of the guesses to be? Specifically, find some function $d(n)$ for which with probability at least 0.9, the median will be within distance $d(n)$ from the true temperature.

   (c) (4 points) Suppose $\sigma_i^2 = 1$ for $i \leq n/2$, and $\sigma_i^2 = n$ for $i > n/2$. Are we better-off returning the mean of the guesses, or the median? Support your claim with a Chernoff bound or two and a brief discussion.

   (d) BONUS (2 points): Suppose you are given the list of σ_i^2's, and hence you know the variance of each of the n guesses. What is the "best" estimate of the temperature, where, for example "best" could be defined as minimizing the expected square of the error of the guess.

   (e) DOUBLE BONUS: Suppose you are given the *set* of σ_i^2's, but are not told which student has each variance. What is the "best" estimate of the temperature? [This is more of a research problem/food for thought, and we will not grade it.]

5. Thresholds in random spatial networks: Suppose we have a square room (with side length 1) with n people that are positioned uniformly at random. Assume that each person will befriend the k nearest people.

(a) (4 points) Prove that there is some constant c such that if $k \geq c \log n$, then with high probability, the resulting friend network is connected (for every two people, there exists some chain of friend's friend's friends, etc. connecting them). [Hint: Divide the room into $\frac{n}{\log n}$ equal-sized squares, and argue that with high probability, every square has at least one person. Next, show via Chernoff bounds that with high probability, for each person, the number of people within distance $10\sqrt{\frac{\log n}{n}}$ is at most $k = c \log n$ for some c, and hence each person will be friends with everyone in the 8 neighboring squares, and hence the friend network is connected.]

(b) BONUS (2 points): Prove that there exists some constant c_2 such that for $k = c_2 \log n$, the probability that the resulting network is connected goes to zero as $n \to \infty$, and hence $\Theta(\log n)$ is the connectivity threshold in this network model. [Hint: to overcome issues of dependencies, consider analyzing the setting of a Poisson Point Process, namely where the number of people is chosen from a Poisson distribution of expectation n, and each person's location is chosen randomly, as before. In this setting, the number of people in any two disjoint regions are independent. Then note that because the probability that a Poisson random variable of expectation n is actually equal to n is large—at least $\Theta(1/\sqrt{n})$— one can translate the result from the Poisson setting to the setting of exactly n people.]