CS5330: Randomized Algorithms

Problem Set 4—Solutions

Due: February 19, 6:30pm

Instructions. The *exercises* at the beginning of the problem set do not have to be submitted—though you may. There are two problems to submit (one of which requires only a single sentence), along with your Final Project proposal (via the Google Form on the web page).

- Please submit the problem set on IVLE in the appropriate folder. (Typing the solution using latex is recommended.) If you want to do the problem set by hand, please submit it at the beginning of class.
- Start each problem on a separate page.
- If you submit the problem set on paper, make sure your name is on each sheet of paper (and legible).
- If you submit the problem set on paper, staple the pages together.

Collaboration Policy. The submitted solution must be your own unique work. You may discuss your high-level approach and strategy with others, but you must then: (i) destroy any notes; (ii) spend 30 minutes on facebook or some other non-technical activity; (iii) write up the solution on your own; (iv) list all your collaborators. Similarly, you may use the internet to learn basic material, but do not search for answers to the problem set questions. You may not use any solutions that you find elsewhere, e.g. on the internet. Any similarity to other students' submissions will be treated as cheating.

Exercises and Review

Exercise 1. Assume you have n balls and you place each ball in bin A with probability p (and discard the ball otherwise). Let Z be the random variable representing the number of balls in the bin. We showed in class that $\mathrm{E}[Z] = np$, and using the fourth moment method we showed that $\Pr[|Z - np| \geq 2np] \leq O(1/(np)^2)$. Using that sixth moment method, can you show that $\Pr[|Z - np| \geq 2np] \leq O(1/(np)^3)$?

Exercise 2. Given n independent binary random variables $X_1, \ldots, X_n$ where $X_j \in [0, 1]$, $\mathrm{E}[X_j] = p$, $X = \sum X_j$, and $\mu = \mathrm{E}[X] = np$. Let's see if we can prove the Hoeffding Bound!

In this proof, you can use (as a black box) the Hoeffding Lemma, which says that if $Z \in [-1, 1]$ is a random variable where $\mathrm{E}[Z] = 0$, then $\mathrm{E}\left[e^{tZ}\right] \leq e^{s^2/4}$.

Let $t > 0$ be some arbitrary constant. Explain why the following facts are true:

Step one:

$$\Pr[X - \mu \geq \delta] \leq \frac{\mathrm{E}\left[e^{t(X-\mu)}\right]}{e^{t\delta}} \tag{1}$$

Step two:

$$\mathrm{E}\left[e^{t(X_j - p)}\right] \leq e^{t^2/4} \tag{2}$$

Step three:

$$\mathrm{E}\left[e^{t(X-\mu)}\right] \leq e^{nt^2/4} \tag{3}$$

Step four: fix $t = 2\delta/n$ and show the final result:

$$\Pr[X - \mu \geq \delta] \leq e^{-\delta^2/n} \tag{4}$$

Exercise 3. In class (or in the previous exercise) we described the following Chernoff Bound. Given n independent binary indicator random variables $X_1, \ldots, X_n$ where $\Pr[X_j = 1] = p$, where $\mu = np$, then the following two bounds hold. First, for all $\delta > 0$:

$$\Pr\left[\sum_{i=1}^{n} X_j \geq (1+\delta)\mu\right] \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu} \tag{5}$$

Second, for all $0 < \delta < 1$:

$$\Pr\left[\sum_{i=1}^{n} X_j \leq (1-\delta)\mu\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} \tag{6}$$

Recall (or prove!) that $\ln(1+x) \geq \frac{x}{1+x/2}$ and $\ln(1-x) \geq -x + x^2/2$. Prove that:

$$\Pr\left[\sum_{i=1}^{n} X_j \geq (1+\delta)\mu\right] \leq e^{-\frac{\delta^2}{2+\delta}\mu} \tag{7}$$

From this, you can conclude that if $\delta \leq 1$, then

$$\Pr\left[\sum_{i=1}^{n} X_j \geq (1+\delta)\mu\right] \leq e^{-\frac{\mu\delta^2}{3}} \tag{8}$$

And also prove that:

$$\Pr\left[\sum_{i=1}^{n} X_j \geq (1+\delta)\mu\right] \leq e^{-\frac{\mu\delta^2}{2}} \tag{9}$$

Exercise 4.

In class, we showed that linear probing ensures that each access to the hash table has expected $O(1)$ cost. Our goal in this problem is to show that, with high probability, *every* access has $O(\log n)$ cost. To do that, use the Chernoff bound that we discussed in class:

$$\Pr\left[\sum_{i=1}^{n} X_j \geq e\mu\right] \leq e^{-\mu} \tag{10}$$

Assume that we are inserting n keys in our hash table of size $m \geq 4n$.

1. As in class, build a binary tree over the table, and define a node u at level ℓ to be crowded if there are more than $(3/4) \cdot 2^\ell$ hashed elements in the portion of the array below node u. (This is the same definition of crowded as from class.) Show the following: for some fixed value L, with probability at least $(1 - 1/n)$, every node of height $\geq L$ is not crowded. (What value of L did you choose?)

   Solution: Fix some node u of level ℓ. Let $X_j = 1$ if ball j lands in node u and 0 otherwise. We know that $\Pr[X_j = 1] = 2^\ell/n$. Let $X = \sum(X_j)$. We define $\mu \mathrm{E}[X] = n(2^\ell/n) = 2^\ell$. Node u is crowded if it has $3\dot{2}^\ell$ keys that hash to it, i.e., if $X > 3\mu$. We want to show that $\Pr[X > 3\mu] \leq 1/n^2$. Using the Chernoff Bound stated above, we know that:

   $$\Pr[X > 3\mu] \leq \Pr[X > e\mu] \leq e^{-\mu} \leq e^{-2^\ell} \tag{11}$$

   Choose $L = \log\ln(n^2)$ and assume $\ell \geq L$. We then conclude that:

   $$\Pr[X > 3\mu] \leq e^{-2^\ell} \leq e^{-2^{\log\ln(n^2)}} \leq e^{-\ln(n^2)} \leq 1/n^2 \tag{12}$$

   Since there are at most n nodes of neight at least L, by a union bound, we conclude that the probability that *any* of the nodes of height at least L is crowded is at most $n/n^2 = 1/n$. Hence with probability at least $1 - 1/n$, every node of height at least L is not crowded.

2. Show that with probability at least $1 - 1/n$, every operation on the hash table has cost $O(\log n)$.

   Solution: First, we know (from the previous part) that every node in the hash table tree of height at least $L = \log \ln(n^2)$ is not crowded. Let S be a cluster in the table of size at least 2^ℓ. Recall that if there is a cluster of length at least 2^ℓ, then there must be some crowded node at level $\ell - 2$. However, we know that there are no crowded nodes at level L or above, so we conclude that $\ell - 2 < L$, i.e., $\ell < L + 2$. From this we conclude that there is no cluster in the hash table of length:

   $$2^{L+2} = 2^{\log \ln(n^2)+2} = 8\ln(n) \ .$$

   Thus we conclude that with probability at least $1 - 1/n$, there are no operations on the hash table that cost more than $8\ln(n)$.

Problem 1. Imagine that Alice is trying to send Bob a message containing k bits over a noisy channel: with probability 1/2, each bit is corrupted and lost. Both Alice and Bob can detect that the bit was lost, and Alice resends the missing bit. (Of course the transmission might also be corrupted.) For a message of size k, how many bits does Alice have to send to ensure that Bob receives the entire message? Prove (using a Chernoff Bound) that in total, Alice needs to send $O(k + \log(1/\epsilon)$ bits to ensure that all k bits are received with probability at least $1 - \epsilon$.

Solution: Let X_i be an indicator random variable where $X_i = 1$ if the ith bit is transmitted successfully. Notice that after k successful transmission, Alice will have transmitted all the bits. Let $n = 4k + 16\ln(1/\epsilon))$. Let $X = \sum_{i=1}^{n} X_i$. Let $\mu = \mathrm{E}[X]$, and notice that $\mu = 2k + 8\ln(1/\epsilon)$. If we can show that $\Pr[X \le \mu/2] \le 1/n$, then we know that with probability at least $1 - 1/n$, there are at least $\mu/2 \ge k$ successful transmissions. Since all the transmissions are independent, we can apply a Chernoff Bound:

$$\begin{aligned}
\Pr[X \le (1 - 1/2)\mu] &\le e^{-\mu(1/2)^2(1/2)} \\
&\le e^{-\mu/8} \\
&\le e^{-(2k+8\ln(1/\epsilon))/8} \\
&\le e^{-\ln(1/\epsilon)} \\
&\le \epsilon
\end{aligned}$$

Problem 2. Consider the following balls-and-bins problem: we throw n balls in n bins, and want to know how many bins contain at least 2 balls. Mr. Smith proposes the following solution:

> `Let` $X_i = 1$ `if there are at least 2 balls in bin` i`, and let` $X = \sum X_i$ `be the number of bins with at least two balls. Let` $p = \binom{n}{2}(1/n)^2 = (1/2)(1 - 1/n) \ge 1/4$ `be the probability of there being at least two balls in bin` i`. Thus` $\mu = \mathrm{E}[\sum X_i] \ge n/4$`. We now use a Chernoff Bound to prove that there are at least` $n/8$ `bins that contain 2 balls, with high probability:`
>
> $$\begin{aligned}
\texttt{Pr}[X \le n/8] &\le \texttt{Pr}[X \le (1 - 1/2)\mu] \\
&\le e^{-\mu(1/2)^2(1/2)} \\
&\le e^{-n/32} \\
&\le 1/n
\end{aligned}$$

What are the problem(s) with Mr. Smith's proof?

Solution: The X_i are not independent! If one bin has two balls in it, then other bins are less likely to have two balls in it. In fact, the X_i are negatively correlated (or negatively associated), and there is (in fact) a version of a Chernoff Bound you can use in this case. But you cannot simply use a regular Chernoff Bound and ignore the dependence among the random variables.