CS265, Winter 2021-2022

Class 4: Agenda, Questions, and Links

1 Warm-Up

Suppose you flip a p-biased coin n times. What does Markov's inequality tell you about the probability that you see more than $2pn$ heads? What does Chebyshev's inequality tell you about that probability?

Group Work: Solutions

Let X be the number of heads you see. We have $\mathbb{E}X = pn$, and $\text{Var}(X) = np(1-p)$. (This second thing is because we can write $X = \sum_i X_i$, where X_i is 1 iff coin i is heads; then we have $\text{Var}(X) = \sum_i \text{Var}(X_i)$ by independence, and $\text{Var}(X_i) = p(1-p)$ since it's a Bernoulli-p random variable).

Thus, Markov says:

$$\Pr[X \geq 2pn] \leq \frac{pn}{2pn} = \frac{1}{2}$$

Chebyshev says:

$$\Pr[X \geq 2pn] \leq \Pr[|X - pn| \geq pn] \leq \frac{np(1-p)}{n^2p^2} = \frac{1-p}{np}.$$

2 Announcements

- HW1 due tomorrow!

3 Questions?

Any questions from the minilectures or warmup? (Markov and Chebyshev's inequalities).

4 Sampling-Based Median

Sampling-Based Median Algorithm

Median:(A list S of n distinct numbers, where n is odd):

1. Let $t = n^{3/4}$. Sample $R = \{r_1, \ldots, r_t\} \subseteq S$ by drawing r_i uniformly at random, independently.

2. Sort R in time $O(t \log t)$. Henceforth, assume that $r_1 \leq r_2 \leq \cdots \leq r_t$.
3. Let $a = r_{t/2-\sqrt{n}}$, $b = r_{t/2+\sqrt{n}}$.
4. Let $N_{<a}$ and $N_{>b}$ denote the number of elements in S less than a and greater than b respectively.
5. Let $T = \{x \in S : a \leq x \leq b\}$. Construct T, and compute $N_{<a}$ and $N_{>b}$, in time $O(n)$.
6. If $|T| < 4t$, sort T in time $O(t \log t)$; otherwise output FAIL.
7. If $N_{<a}, N_{>b} \leq n/2$ (aka, $median(S) \in T$):
   - Return the i'th smallest element of T, where $i = (n+1)/2 - N_{<a}$.
8. Otherwise, output FAIL.

[We'll see an example on a slide; this slide is posted on the course website.]

[**Note:** Above there should be some floors or ceilings or something. Don't worry about it, and ignore off-by-one errors throughout this class.]

5 Analyzing the sampling-based median algorithm

You will analyze this algorithm in group work.

Group Work

Note: Throughout this group work, don't worry about $\leq$ vs $<$, or whether or not something is true up to ± 1, or anything small like that.

1. Make sure that you all understand the algorithm. Pseudo-code is above, and the one-slide example is available on the course website (cs265.stanford.edu), in the class-by-class resources for Class 4. Ask/answer any questions that you have amongst yourselves, and flag down a member of the course staff if you still have questions.
2. Suppose that you could show that:
   - with probability ≥ 0.9, the median of S is in the list T; and
   - with probability ≥ 0.9, $|T| < 4t$.

   Explain (to each other) why these two things would imply that the algorithm returns the correct answer with probability ≥ 0.8. And if it does not return the median then it returns FAIL.
3. Convince yourself that this algorithm uses at most $O(n)$ operations. What is the leading constant in this big-Oh notation? (Assuming that "sample a random element of S", and comparing two numbers are each single operations).

4. In the following parts, you will show that the median of S is in T, with probability at least 0.9. Let m be the median of S. Consider two events:
   - $|\{r_i \in R : r_i < m\}| < \frac{t}{2} + \sqrt{n}$
   - $|\{r_i \in R : r_i > m\}| < \frac{t}{2} + \sqrt{n}$

   (a) Explain why, if both of these events hold, then $median(S) \in T$.

   (b) Use Chebyshev's inequality to bound the probability that the first event does not hold. (Hint: let X_i be the indicator random variable that is 1 iff $r_i \leq m$, and consider $\sum_i X_i$).

   (c) Convince yourself that the same argument will work for the second event, and write a statement of the form:

$$\Pr[median(S) \in T] \geq 1 - \_\_\_\_.$$

5. Now, we turn our attention to the probability that $|T| < 4t$.

   (a) Explain why it is sufficient to show that a is *not* one of the smallest $n/2 - 2t$ elements of S, and b is *not* one of the largest $n/2 + 2t$ elements of S.

   (b) Use Chebyshev's inequality to bound the probability that a is not one of the smallest $n/2 - 2t$ elements of S. (Hint: Consider the indicator random variable Y_i that is 1 if r_i is in the smallest $n/2 - 2t$ elements of S. Argue that a is one of the smallest $n/2 - 2t$ elements of S iff $\sum_i Y_i \geq t/2 - \sqrt{n}$ (why?) and apply Chebyshev's inequality.)

   (c) Convince yourself that the analogous statement for b, and write a statement of the form:

$$\Pr[|T| < 4t] \geq 1 - \_\_\_\_\_.$$

Group Work: Solutions

Note: This median algorithm is also worked out in the lecture notes, so if you prefer to read the solutions with fewer bullet points and more complete sentences, check out those!

1. Understood!
2. If $|T| < 4t$, then the algorithm doesn't return Fail on that check. If the median is in T, then the algorithm doesn't return Fail on that check. And by construction, if the median is in T, and we don't return Fail, then the algorithm returns the $(n+1)/2 - N_{<a}$'th thing in T, which is the $(n+1)/2$'nd thing in S, which is the median of S.
3. The number of operations is $2n$, naively, or $\frac{3}{2}n$ if we are slightly crafty:

- $O(t) = o(n)$ to sample R (assuming that we can sample an item from S in time $O(1)$)
- $O(t \log t) = O(n^{3/4} \log(n)) = o(n)$ to sort R.
- $2n$ to compute T and $N_{<a}$ and $N_{>b}$ if we do it naively, but $\frac{3}{2}n + o(n)$ if we compare each element to a, and then only compare the elements that are greater than a to b.
- $O(t \log t) = o(n)$ to sort T.
- $O(1)$ to return the correct element of T.

4. (a) If the first event holds, then $m \leq b$. If the second event holds, then $m \geq a$. Thus if both hold, $a \leq m \leq b$, so m will appear in T.

   (b) Let X_i be as in the hint, so $\mathbb{E}[X_i] \leq \frac{1}{2}$. Then by Chebyshev's inequality,

$$\begin{aligned}\Pr[|\{r_i \in R : r_i < m\}| > \frac{t}{2} + \sqrt{n}] &= \Pr[\sum_i X_i > \frac{t}{2} + \sqrt{n}] \\ &\leq \Pr\left[\left|\sum_i (X_i - 1/2)\right| > \sqrt{n}\right] \\ &\leq \frac{t/4}{n} = O(n^{-1/4}),\end{aligned}$$

   using the fact that the X_i are independent, so $\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}\left(X_i\right) \leq t/4$.

   (c) The other claim is exactly the same, so we conclude that

$$\Pr[m \in T] \geq 1 - O(n^{-1/4}).$$

   In particular, when n is sufficiently large, this is at least 0.9.

5. (a) Suppose that a is not one of the smallest $n/2 - 2t$ elements of S, and b is not one of the largest $n/2 - 2t$ elements of S. Then the number of elements of S that are between a and b are at most $2t + 2t = 4t$, which is what we want to show.

   (b) Let Y_i be as in the hint. Notice that a is among the smallest $n/2 - 2t$ elements of S iff $\sum_i Y_i \geq t/2 - \sqrt{n}$. Indeed, if that's the case, then there are more than $t/2 - \sqrt{n}$ elements of R that are in the smallest $t/2 - \sqrt{n}$ elements of S, and so by the definition of a as the $t/2 - \sqrt{n}$'th smallest thing in R, a must be among them.

   Notice that

$$\mathbb{E}Y_i = \frac{1}{2} - \frac{2t}{n} = \frac{1}{2} - \frac{2}{n^{1/4}}.$$

Thus,

$$\begin{aligned}
\Pr[a \text{ is among the smallest } n/2-2t \text{ elements of } S] &= \Pr[\sum_i Y_i \geq t/2 - \sqrt{n}] \\
&\leq \Pr[\left|\sum_i Y_i - \mathbb{E}Y_i\right| \geq \frac{2t}{n^{1/4}} - \sqrt{n}] \\
&= \Pr[\left|\sum_i Y_i - \mathbb{E}Y_i\right| \geq \sqrt{n}] \\
&\leq \frac{\mathrm{Var}\left(\sum_i Y_i\right)}{n} \\
&\leq \frac{t}{4n} = \frac{1}{4} \cdot n^{-1/4},
\end{aligned}$$

using the fact that $\mathrm{Var}\left(\sum_i Y_i\right) = t\mathrm{Var}\left(Y_1\right)$ since the Y_i are independent, and then using the fact that $\mathrm{Var}\left(Y_i\right) \leq 1/4$.

(c) The same thing is true for the second event, and so we have

$$\Pr[|T| > 4t] = O(n^{-1/4}).$$