CS 537: *Randomness in Computing*
Boston University

Prof. Sofya Raskhodnikova
TF: Gavin Brown
February 20, 2020

Solutions to Homework 5

1. (Random hats) *Suppose n people come to a theater performance wearing hats. When they leave, they get back a uniformly random hat. That is, each of $n!$ assignments of n hats to n people is equally likely.*

   (a) *We say that a pair of people (let's call them Alice and Bob) exchanged their hats if Alice got Bob's hat and Bob got Alice's hat. Let X denote the number of pairs of people that exchanged their hats. Calculate the expectation of X.*

   Hint: *Write X as a sum of indicator random variables.*

   Let $X_{i,j}$ be the indicator random variable for the event that people i and j exchange their hats. Then $X = \sum_{i,j:1\le i<j\le n} X_{i,j}$. We get

$$\begin{aligned}
\mathbb{E}\left[X_{i,j}\right] &= \Pr\left[\text{people } i \text{ and } j \text{ exchange their hats}\right]\\
&= \Pr\left[i \text{ takes } j\text{'s hat and } j \text{ takes } i\text{'s hat}\right]\\
&= \Pr\left[i \text{ takes } j\text{'s hat}\right] \cdot \Pr\left[j \text{ takes } i\text{'s hat} \,\middle|\, i \text{ takes } j\text{'s hat}\right]\\
&= \frac{1}{n} \cdot \frac{1}{n-1}.
\end{aligned}$$

   By linearity of expectation,

$$\mathbb{E}[X] = \mathbb{E}\Big[\sum_{i,j:1\le i<j\le n} X_{i,j}\Big] = \sum_{i,j:1\le i<j\le n} \mathbb{E}\left[X_{i,j}\right] = \frac{n(n-1)}{2} \cdot \frac{1}{n(n-1)} = \frac{1}{2}.$$

   (b) *Calculate Var[X].*

   Hint: *Use the same sum as in part (a); remember that the indicators are not independent.*

   By definition of variance, $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. By part (a), $\mathbb{E}[X]^2 = 1/4$. It remains to compute $\mathbb{E}[X^2]$. We have,

$$X^2 = \sum_{i,j:1\le i<j\le n} X_{i,j}^2 + \sum_{i,j:1\le i<j\le n} \sum_{\substack{k,\ell:1\le k<\ell\le n;\\ (i,j)\neq(k,\ell)}} X_{i,j}X_{k,\ell}.$$

   For $i<j$, we have $\mathbb{E}[X_{i,j}^2] = \mathbb{E}[X_{i,j}] = \frac{1}{n(n-1)}$, since squaring an indicator random variable does not change its value. To evaluate the double summation, observe that if $\{i,j\}$ and $\{k,\ell\}$ overlap, then $X_{i,j} \cdot X_{k,\ell} = 0$, as three people cannot form two distinct pairs that both swap hats at the same time. Hence, we can assume that i,j,k,ℓ are all distinct. Then

$$\begin{aligned}
\mathbb{E}[X_{i,j}X_{k,\ell}] &= \Pr[X_{i,j} = 1 \cap X_{k,\ell} = 1] = \Pr[X_{i,j} = 1] \cdot \Pr[X_{k,\ell} = 1 \mid X_{i,j} = 1]\\
&= \frac{1}{n(n-1)} \cdot \frac{1}{(n-2)(n-3)}.
\end{aligned}$$

The last equality follows from the fact that conditioned on i and j having exchanged their hats, the probability of k picking ℓ's hat is $\frac{1}{n-2}$ and the probability of ℓ picking k's hat (conditioned also on k having picked ℓ's hat), is $\frac{1}{n-3}$. The number of pairs (i,j) and (k,ℓ) that satisfy all of the aforementioned conditions is $\binom{n}{4}\cdot\binom{4}{2}$. Thus,

$$\mathbb{E}[X^2] = \binom{n}{2}\cdot\frac{1}{n(n-1)} + \binom{n}{4}\cdot\binom{4}{2}\cdot\frac{1}{n(n-1)(n-2)(n-3)} = \frac{1}{2}+\frac{1}{4}.$$

Hence, $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{2}$.

2. (MaxCut) *In the problem* MAXCUT, *we are given an undirected graph $G = (V, E)$ and asked to find a cut of maximum size in G. (Recall that a cut in G is a partition of the vertex set V into two parts; the size of the cut is the number of edges with one endpoint in each part of the partition.) In contrast to the seemingly very similar problem* MINCUT *discussed in class (recall Karger's algorithm),* MAXCUT *is a famous NP-hard problem, so we do not expect to find an efficient algorithm that solves it exactly. Here is a very simple linear-time randomized algorithm that gives a pretty good approximation:*

- *Randomly and independently color each vertex $v \in V$ red or blue with probability $1/2$.*
- *Output the cut defined by the red/blue partition of vertices.*

(a) *Let random variable X denote the size of the cut output by the algorithm. Compute $E[X]$ as a function of the number of edges in G, and deduce that $E[X] \geq OPT/2$, where OPT is the size of a maximum cut in G.*

Hint: *Write X as a sum of indicator random variables.*

Let m be the number of edges in the graph G. Note that $\mathrm{OPT} \leq m$. For any (undirected) edge $(u,v) \in E$, let X_{uv} be the indicator random variable for the event that the edge (u,v) is part of the cut (i.e., u and v have different colors). Then, $X = \sum_{(u,v)\in E} X_{uv}$. For every $(u,v) \in E$,

$$E[X_{uv}] = \Pr[X_{uv} = 1] = \frac{1}{2}.$$

This is computed noting that the edge (u,v) is part of the cut iff v is colored a different color than u, which happens with probability $1/2$. By the linearity of expectation,

$$E[X] = E\left[\sum_{(u,v)\in E} X_{uv}\right] = \sum_{(u,v)\in E} E\left[X_{uv}\right] = \sum_{(u,v)\in E}\frac{1}{2} = \frac{m}{2} \geq \frac{\mathrm{OPT}}{2}.$$

(b) *Let p denote the probability that the cut output by the algorithm has size at least $0.49\ OPT$. Show that $p \geq 1/51$.*

Hint: *Applying Markov's inequality to X will not work here. Try applying Markov's inequality to a different random variable.*

We need to show $\Pr[X \geq 0.49OPT] \geq 1/51$. Since $\Pr[X \geq 0.49OPT] + \Pr[X < 0.49OPT] = 1$, the statement we have to prove is equivalent to $\Pr[X < 0.49OPT] \leq 50/51$. To show this, we apply Markov's inequality to the random variable $OPT - X$. Since, $X \leq OPT$, this random variable is nonnegative. By Markov's inequality, the linearity of expectation, and the fact that $E[X] \geq OPT/2$, we get

$$\begin{aligned}\Pr[X < 0.49OPT] &= \Pr[OPT - X > 0.51OPT] \\ &\leq \frac{E[OPT - X]}{0.51\ OPT} = \frac{OPT - E[X]}{0.51\ OPT} \leq \frac{0.5\ OPT}{0.51\ OPT} = \frac{50}{51}.\end{aligned}$$

(c) *Now compute the variance Var[X].*
Hint: *Again write X as the sum of indicators, as in part (a).*
For any two different edges $(u, v), (y, z) \in E$,

$$\mathbb{E}[X_{uv}X_{yz}] = \Pr[X_{uv} = 1 \cap X_{yz} = 1] = \Pr[X_{uv} = 1] \cdot \Pr[X_{yz} = 1] = \frac{1}{2} \cdot \frac{1}{2}.$$

We use the fact that X_{uv} and X_{yx} are independent. To see this, note that if (u, v) and (y, z) are disjoint then X_{uv} and X_{yx} are clearly independent. Otherwise, suppose without loss of generality that $u = y$. But no matter how u is colored, with probability $1/2$, v takes the other color[1].

Since squaring an indicator random variable does not change its value, $\mathbb{E}[X_{uv}^2] = \frac{1}{2}$ for all $(u, v) \in E$. By definition of variance,

$$\begin{aligned}
\operatorname{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \frac{m^2}{4} \\
&= \mathbb{E}\Big[\Big(\sum_{(u,v)\in E} X_{uv}\Big)^2\Big] - \frac{m^2}{4} \\
&= \sum_{(u,v)\in E} \mathbb{E}[X_{uv}^2] + \sum_{(u,v)\in E} \sum_{\substack{(y,z)\in E:\\ (u,v)\neq(y,z)}} \mathbb{E}[X_{uv}X_{yz}] - \frac{m^2}{4} \\
&= \frac{m}{2} + \frac{m(m-1)}{4} - \frac{m^2}{4} = \frac{m}{4}.
\end{aligned}$$

(d) *Let p be the probability defined in part (b). Use Chebyshevs inequality together with part (c) to show that $p = 1 - O(1/|E|)$.*
(Note how Chebyshev's inequality gives us a better bound here than Markov's.)
Chebyshev's inequality states that for any random variable X,

$$\Pr[|X - \mathbb{E}[X]| \geq k] \leq \frac{\operatorname{Var}(X)}{k^2}.$$

As before, let X denote the random variable for the number of edges crossing the cut. We first use the fact that $OPT \leq m$ and then apply Chebyshev to X,

$$\begin{aligned}
\Pr[X < 0.49 \text{ OPT}] &\leq \Pr[X < 0.49m] \\
&= \Pr[0.5m - X > 0.01m] \\
&\leq \Pr[|X - 0.5m| > 0.01m] \\
&\leq \Pr[|X - \mathbb{E}[X]| > 0.01m] \\
&\leq \frac{\operatorname{Var(X)}}{(0.01m)^2} = \frac{m/4}{0.0001m^2} = \frac{2500}{m}.
\end{aligned}$$

Thus, $p = \Pr[X \geq 0.49\text{OPT}] \geq 1 - \frac{2500}{m} = 1 - O(1/m)$, as desired.

3. (Generalization of Randomized Median Algorithm) *In this problem, you are asked to generalize the randomized median-finding algorithm from class (Section 3.5 of the MU book), so that it finds an element of rank k (that is, the kth largest element) in an array of n distinct elements, for any given $k \in [4n^{3/4}, n - 4n^{3/4}]$. You may ignore rounding issues in your algorithm and analysis.*

[1] The random variables $\{X_{uv}\}_{(u,v)\in E}$ are only *pairwise independent*, not mutually independent. This can be observed from the fact that the random variables X_{uv}, X_{vw}, X_{wu} corresponding to the three edges of a triangle are dependent.

(a) *Explain how to modify lines 3, 4, 6 and 8 of the algorithm in the book.*

Line 3: Let d be the $(\frac{k}{n} \cdot n^{3/4} - \sqrt{n})$th smallest element in R.

Line 4: Let d be the $(\frac{k}{n} \cdot n^{3/4} + \sqrt{n})$th smallest element in R.

Line 6: If $\ell_d > k - 1$ or $\ell_u > n - k + 1$ then FAIL.

Line 8: Output the $(k - \ell_d)$th element in the sorted order of C.

(b) *Analyze the running time of the modified algorithm.*

All the steps that we changed take a constant time to execute. Computing the set C and numbers ℓ_d and ℓ_u continue to be the bottleneck steps even in the modified algorithm. Thus, the modified algorithm runs in time linear in the size of the input.

(c) *We will follow the same analysis outline as in class (and in the book). Change the definitions of events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_{3,1}$ and $\mathcal{E}_{3,2}$, so that they apply for general k.*

Let s_k denote the element of rank k in S.

i. $\mathcal{E}_1$: $Y_1 = |\{r \in R : r \leq s_k\}| < \frac{k}{n} \cdot n^{3/4} - \sqrt{n}$.
The event $\mathcal{E}_1$ is implied by the failure condition $\ell_d > k - 1$ of Line 6. If $\ell_d > k - 1$, the rank k element s_k is less than d. This implies that Y_1 is less than $\frac{k}{n} \cdot n^{3/4} - \sqrt{n}$, since d is the $(\frac{k}{n} \cdot n^{3/4} - \sqrt{n})$th element in R (assuming R is sorted).
Thus, to bound the probability that $\ell_d > k - 1$, it is enough to bound $\Pr[\mathcal{E}_1]$.

ii. $\mathcal{E}_2$: $Y_2 = |\{r \in R : r \geq s_k\}| < n^{3/4} \cdot \left(1 - \frac{k}{n}\right) - \sqrt{n}$.
If $\ell_u > n - k + 1$, the element s_k is greater than u. This implies that Y_2 is less than $n^{3/4} - \left(\frac{k}{n} \cdot n^{3/4} + \sqrt{n}\right) = n^{3/4} \cdot \left(1 - \frac{k}{n}\right) - \sqrt{n}$, since u is the $(\frac{k}{n} \cdot n^{3/4} + \sqrt{n})$th element in R (assuming R is sorted).
Thus, to bound the probability that $\ell_u > n - k + 1$, it is enough to bound $\Pr[\mathcal{E}_2]$.

iii. The event $\mathcal{E}_3$ is as defined in the textbook.

iv. $\mathcal{E}_{3,1}$: at least $2n^{3/4}$ elements of C are greater than s_k.
$\mathcal{E}_{3,2}$: at least $2n^{3/4}$ elements of C are smaller than s_k;

(d) *Write $\Pr[\mathcal{E}_1]$ and $\Pr[\mathcal{E}_{3,1}]$ as probability expressions involving the tail of a suitable binomial random variable.*

We know that, $\Pr[\mathcal{E}_1] \leq \Pr[Y_1 < \frac{k}{n} \cdot n^{3/4} - \sqrt{n}]$. We will now show that Y_1 is a binomial random variable. Let X_i be the indicator random variable for the ith sample in R being at most s_k, the element of rank k. Since each sample in R is chosen uniformly at random from the set of all elements in S, we have $\Pr[X_i] = \frac{k}{n}$. Then $Y_1 = \sum_{i \in [n^{3/4}]} X_i$ is a binomial random variable with parameters $n^{3/4}$ and $\frac{k}{n}$.

We will bound the probability that $\mathcal{E}_{3,1}$ occurs as the tail of a binomial random variable. If there are at least $2n^{3/4}$ elements above s_k in C, then the rank of u in S is at least $k + 2n^{3/4}$ and hence the set R has at least $n^{3/4}\left(1 - \frac{k}{n}\right) - \sqrt{n}$ samples among the $n - k - 2n^{3/4}$ largest elements in S.

Let X denote the number of samples in R that are from among the $n - k - 2n^{3/4}$ largest elements in S. It is clear from the preceding discussion that $\Pr[\mathcal{E}_{3,1}] \leq \Pr[X \geq n^{3/4}\left(1 - \frac{k}{n}\right) - \sqrt{n}]$. It remains to show that X is a binomial random variable and the mean of X is less than $n^{3/4}\left(1 - \frac{k}{n}\right) - \sqrt{n}$. Defining X_i to be the indicator random variable for the ith sample to be from among the $n - k - 2n^{3/4}$ largest elements in S, we have, $X = \sum_{i \in [n^{3/4}]} X_i$. Since the X_i's are i.i.d., X is a binomial r.v. with parameters $n^{3/4}$ and $1 - \frac{k}{n} - \frac{2}{n^{1/4}}$.