CSE 565: Algorithm Design and Analysis **Due 2022-12-08T23:59**

Problem Set 6 Solutions

Instructor: Dr. Antonio Blanca
TA: Jeremy Huang

Release Date: 2022-11-29

Notice: Type your answers using LaTeX and make sure to upload the answer file on Gradescope before the deadline. Recall that for any problem or part of a problem, you can use the "I'll take 20%" option. For more details and the instructions read the syllabus.

Problem 1. MAX 4COLOR

ou are given an undirected graph $G = (V, E)$, and you need to color each vertex with one of the given four colors. We say an edge $(u, v) \in E$ is satisfied if u and v are assigned different colors. Given $G = (V, E)$, the problem is to color all vertices so that the number of satisfied edges is maximized.

Design a randomized 4/3-approximation algorithm for this problem, that is, the expected number of satisfied edges returned by your algorithm should be at least 3/4 fraction of the number of the satisfied edges in the optimal solution. Prove your algorithm can achieve such randomized approximation ratio. Your algorithm should run in polynomial-time.

Solution

The algorithm independently assigns one of the 4 colors with probability of 1/4 for each vertex. We now show its expected performance. Let Z be the random variable indicates the total number of satisfied edges. Let Z_e be the binary random variable indicates whether edge $e \in E$ is satisfied. We have that $Z = \sum_{e \in E} Z_e$ and therefore $E(Z) = \sum_{e \in E} E(Z_e)$. Further, edge $e = (u, v)$ is satisfied if and only if u and v are colored differently, and with probability of $4 \cdot 1/4 \cdot 1/4$, they are colored the same. Hence, $\Pr(Z_e = 1) = 1 - \Pr(Z_e = 0) = 1 - 1/4 = 3/4$. Combined, $E(Z) = \sum_{e \in E} E(Z_e) = 3 \cdot |E|/4$. As the optimal solution satisfies at most |E|edges, this algorithm is a randomized 4/3-approximation algorithm.

Problem 2. Close Games

Consider a balls-and-bins experiment with $2n$ balls but only two bins. As usual, each ball independently selects one of the two bins, both bins equally likely. The expected number of balls in each bin is n. In this problem, we explore the question of how big their difference is likely to be. Let the random variables X_1 and X_2 denote the number of balls in the two bins, respectively. Prove that for any $\varepsilon > 0$ there is a constant $c > 0$ such that the probability $\Pr[X_1 - X_2 \geq c\sqrt{n}] \leq \varepsilon$.

Hint: use Chebyshev's inequality.

Solution

X_1 and X_2 are binomially distributed, with $2n$ trials and $p = 0.5$. Their variance is $2np(1-p) = \frac{n}{2}$. $X_1 - X_2$ is also a random variable, with mean $\mu = E[X_1] - E[X_2] = n - n = 0$ and variance $\sigma^2 = \mathrm{Var}(X_1 - X_2) = \mathrm{Var}(X_1 - (2n - X_1)) = \mathrm{Var}(2X_1) = 4\,\mathrm{Var}(X_1) = 8np(1-p) = 2n$. We can now apply Chebyshev's inequality using a given ε to find an expression for c which makes the given equality true. Chebyshev's inequality is $\Pr[|(X_1 - X_2) - \mu| \geq k\sigma] \leq \frac{1}{k^2}$. Set

$\frac{1}{k^2} = \varepsilon$ (and thus $k = \frac{1}{\sqrt{\varepsilon}}$) to get the following derivation:

$$\begin{aligned} \Pr[|(X_1 - X_2) - \mu| \geq k\sigma] &\leq \frac{1}{k^2} \\ \Pr\left[|(X_1 - X_2) - 0| \geq k\sqrt{2n}\right] &\leq \frac{1}{k^2} \\ \Pr\left[|X_1 - X_2| \geq k\sqrt{2n}\right] &\leq \frac{1}{k^2} \end{aligned}$$

$X_1 - X_2$ is symmetric about 0 since both X_1 and X_2 are symmetric about 0, so $\Pr\left[|X_1 - X_2| \geq k\sqrt{2n}\right] = 2\Pr\left[X_1 - X_2 \geq k\sqrt{2n}\right]$.

$$\Pr\left[X_1 - X_2 \geq k\sqrt{2n}\right] \leq \frac{1}{2k^2}$$

We can then set $\frac{1}{2k^2} = \varepsilon$ (giving $k = \frac{1}{\sqrt{2\varepsilon}}$) to get the inequality requested by the question.

$$\Pr\left[X_1 - X_2 \geq \frac{1}{\sqrt{2\varepsilon}}\sqrt{2n}\right] \leq \varepsilon$$

Clearly the requested inequality always holds for $c = \frac{1}{\sqrt{2\varepsilon}}$, and $\frac{1}{\sqrt{2\varepsilon}} > 0$ when $\varepsilon > 0$ because inverse, doubling, and square root all preserve positivity; so this proves that for any $\varepsilon > 0$ there always exists a $c > 0$ such that the inequality holds.

Problem 3. Speeding on the Grass

Consider the following problem: given an unsorted array A of n elements, we wish to sample from the middle half of the array. i.e. we want a procedure that returns an element a of A such that $a = A'[i]$ where A' is a sorted version of A and $i \in \mathbb{N}$ where $n/4 \leq i \leq 3n/4$.

Consider the following algorithm for this problem: choose $k = 10\log n$ elements from A u.a.r., sort them, and return the median of the sorted elements. What is the time complexity of this algorithm? What is the probability that one iteration of this algorithm returns an element that is not from the middle half of the array?

You can use these inequalities: $\binom{k}{k/2} \leq \frac{1}{2} \times 4^{k/2}$, $\sum_{i=k/2}^{k}(1/3)^i \leq (1/3)^{k/2}\frac{3}{2}$, $(3/4)^{k/2} \leq (1/2)^{k/5}$. Assume that it takes $O(\log n)$ time to compare two elements and $O(k\log n)$ time to select k random elements from an array of length n. Note that for the error probability an exact calculation is difficult and a reasonable upper bound is fine.

Solution

Error Probability:

The return value of the algorithm is erroneous iff the returned element is from the bottom or top quarter of the array. WLOG we assume an erroneous element from the bottom quarter was returned. Since the algorithm returns the median of the size-k subset, the size-k subset must contain at least $k/2$ elements from the bottom quarter.

The probability of selecting an element from the bottom quarter is 1/4, so the probability of selecting i elements from the bottom quarter in k selections is $\binom{k}{i}\left(\frac{1}{4}\right)^i\left(\frac{3}{4}\right)^{k-i}$. So the probability of the algorithm choosing at least k/2 elements

from the bottom quarter is

$$\begin{aligned}
\sum_{i=k/2}^{k} \binom{k}{i} \left(\frac{1}{4}\right)^i \left(\frac{3}{4}\right)^{k-i} &\leq \binom{k}{k/2} \sum_{i=k/2}^{k} \left(\frac{1}{4}\right)^i \left(\frac{3}{4}\right)^{k-i} \\
&= \binom{k}{k/2} \left(\frac{3}{4}\right)^k \sum_{i=k/2}^{k} \left(\frac{1}{4}\right)^i \left(\frac{4}{3}\right)^i \\
&= \binom{k}{k/2} \left(\frac{3}{4}\right)^k \sum_{i=k/2}^{k} \left(\frac{1}{3}\right)^i \\
&\leq \binom{k}{k/2} \left(\frac{3}{4}\right)^k \left(\frac{1}{3}\right)^{k/2} \frac{3}{2} \\
&\leq 4^{k/2} \left(\frac{3}{4}\right)^k \left(\frac{1}{3}\right)^{k/2} \\
&= (3/4)^{k/2} \\
&\leq (1/2)^{k/5} \\
&= (1/2)^{(10/5)\log n} \\
&= 1/(2^{\log n^2}) \\
&= n^{-2}
\end{aligned}$$

So the probability of the algorithm selecting at least $k/2$ element from the bottom quarter is $\frac{1}{n^2}$. By symmetry, the probability for the top quarter is the same, so the probability of an erroneous result is $\frac{1}{n^2} + \frac{1}{n^2} = \frac{2}{n^2}$.

Time Complexity:

It takes $O(k \log n) = O(\log^2 n)$ time to select k elements and $O(k) = O(\log n)$ time to copy them into a new array.

It takes $O(k \log k) = O(\log n \log \log n)$ comparisons to sort the new array. Each comparison takes $O(\log n)$ time, so overall it takes $O(\log^2 n \log \log n)$ time to sort the new array.

It takes $O(1)$ time to return the median of the new sorted array (just return the element at index $k/2$).

These three steps happen sequentially, so overall the algorithm has time complexity $O(\log^2 n \log \log n)$.

Problem 4. Cliche Cliques

Let $G \sim G(n,p)$ and let X be the random variable corresponding to the number of cliques of size 4 in $G = (V,E)$. Let Ω_4 be the set of all the subsets of size 4 of V. That is, $\Omega_4 = \{C \subset V : |C| = 4\}$ and so $|\Omega_4| = \binom{n}{4}$. Here $G(n,p)$ is a distribution of n node graphs generated by including each possible edge independently with probability p.

1. Show that if $pn^{2/3} \to \infty$, then $E[X] \to \infty$ and that if $pn^{2/3} \to 0$, then $E[X] \to 0$.
   Hint: Observe that $X = \sum_{C \in \Omega_4} X_C$ where X_C is the $0/1$ random variable for whether C is a clique or not in G.
2. Use part (a) and Markov's inequality to show that $\Pr[G \text{ has a 4 clique}] \to 0$ when $pn^{2/3} \to 0$ (here $n \to 0$).
3. Show that the variance of X satisfies:
   $$\mathrm{Var}(X) = \sum_{C \in \Omega_4} \mathrm{Var}(X_C) + \sum_{C,D \in \Omega_4 : D \neq C} \mathrm{Cov}(X_C, X_D),$$
   where the covariance is defined as $\mathrm{Cov}(X_C, X_D) = E[X_C X_D] - E[X_C]E[X_D]$.
4. Show that $\sum_{C \in \Omega_4} \mathrm{Var}(X_C) = O(n^4 p^6)$.

5. Show that $\sum_{C,D\in\Omega_4:D\neq C}\text{Cov}(X_C,X_D) = O(n^6p^{11}) + O(n^5p^9)$.

6. Use part 4., 5. and Chebyshev's inequality to show that $\Pr[G \text{ has a 4 clique}] \to 1$ when $pn^{2/3} \to \infty$ (here $n \to \infty$). *Hint: Use Chebyshev's inequality to prove that it is sufficient that* $\frac{\text{Var}(X)}{E[X]^2} \to 0$.

Solution

1. Let $C \in \Omega_4$, and let X_C be the indicator variable for whether C is a clique in G. Since C is a clique when all 6 edges induced on C are present, $X_C \sim Bernoulli(p^6)$. By linearity of expectation,

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{C\in\Omega_4} X_C\right] = \sum_{C\in\Omega_4}\mathbb{E}[X_C] = |\Omega_4|\cdot p^6 = \binom{n}{4}p^6 = \Theta(p^6n^4).$$

   Naturally, if $pn^{2/3} \to \infty$, $\mathbb{E}[X] \to \infty$ and if $pn^{2/3} \to 0$, then $\mathbb{E}[X] \to 0$.

2. By Markov's inequality, when $pn^{2/3} \to 0$,

$$\mathbb{P}[G \text{ has a clique}] = \mathbb{P}[X \geq 1] \leq \mathbb{E}[X] \to 0.$$

3. We use properties of variance and expectation to manipulate the terms

$$\begin{aligned}
\text{var}[X] = \text{var}\left[\sum_{C\in\Omega_4} X_C\right] &= \mathbb{E}\left[(\sum_{C\in\Omega_4} X_C)^2\right] - \mathbb{E}\left[\sum_{C\in\Omega_4} X_C\right]^2 \\
&= \mathbb{E}\left[\sum_{C\in\Omega_4} X_C^2 + \sum_{C,D\in\Omega_4:C\neq D} X_CX_D\right] - \sum_{C,D\in\Omega_4}\mathbb{E}[X_C]\mathbb{E}[X_D] \\
&= \sum_{C\in\Omega_4}\mathbb{E}[X_C^2] - \mathbb{E}[X_C]^2 + \sum_{C,D\in\Omega_4:C\neq D}\mathbb{E}[X_CX_D] - \mathbb{E}[X_C]\mathbb{E}[X_D] \\
&= \sum_{C\in\Omega_4}\text{var}[X_C] + \sum_{C,D\in\Omega_4:C\neq D} Cov(X_C,X_D).
\end{aligned}$$

4. Recall $X_C \sim Bernoulli(p^6)$. Then

$$\sum_{C\in\Omega_4}\text{var}[X_C] = \sum_{C\in\Omega_4} p^6(1-p^6) \leq p^6\binom{n}{4} = O(n^4p^6).$$

5. Observe that $Cov(X_C,X_D)$ depends on $|C\cap D|$. If $|C\cap D| \leq 1$, C and D share no edge, so X_C and X_D are uncorrelated. If $|C\cap D| = 2$, then they share exactly 1 edge, and

$$Cov(X_C,X_D) = \mathbb{E}[X_CX_D] - \mathbb{E}[X_C]\mathbb{E}[X_D] = p^{11} - p^{12};$$

   if $|C\cap D| = 3$, then they share 3 edges, and $Cov(X_C,X_D) = \mathbb{E}[X_CX_D] - \mathbb{E}[X_C]\mathbb{E}[X_D] = p^9 - p^{12}$.

$$\begin{aligned}
\sum_{C,D\in\Omega_4:C\neq D} Cov(X_C,X_D) &= \sum_{C,D\in\Omega_4:|C\cap D|\leq 1} 0 + \sum_{C,D\in\Omega_4:|C\cap D|=2} p^{11} - p^{12} + \sum_{C,D\in\Omega_4:|C\cap D|=3} p^9 - p^{12} \\
&\leq p^{11}\binom{n}{6} + p^9\binom{n}{5} = O(n^6p^{11}) + O(n^5p^9).
\end{aligned}$$

6. Using part (d) and (e), we have

$$\text{var}(X) = O(n^4p^6) + O(n^6p^{11}) + O(n^5p^9).$$

Chebyshev's inequality implies

$$\mathbb{P}[X < 1] \leq \mathbb{P}\left[X < \frac{\mathbb{E}[X]}{2}\right] \leq \mathbb{P}\left[|X - \mathbb{E}[X]| > \frac{\mathbb{E}[X]}{2}\right] \leq \frac{\mathrm{var}[X]}{(\mathbb{E}[X])^2} \tag{0.1}$$

Hence, if $\frac{\mathrm{var}[X]}{(\mathbb{E}[X])^2} \to 0$, we have $\mathbb{P}[X \geq 1] \to 1$. When $pn^{2/3} \to \infty$,

$$\frac{\mathrm{var}[X]}{(\mathbb{E}[X])^2} = \frac{O(n^4p^6) + O(n^6p^{11}) + O(n^5p^9)}{\Theta(p^{12}n^8)} = \frac{1}{\Omega(n^4p^6)} + \frac{1}{\Omega(pn^{2/3} \cdot n^{4/3})} + \frac{1}{\Omega((pn^{2/3})^3 n)} \to 0.$$