NYU CS-GY 6763: Algorithmic Machine Learning and Data Science Spring 2025

Homework Solution key

Name: Solution key

Problem 1

1. $\mathbb{E}[X] = 1/2$, so by Markov's $\Pr[X \geq 7/8] \leq \frac{1/2}{7/8} = .571$.

2. $\mathbb{E}[(X - \mathbb{E}\,X)^2] = \int_{0^1} (x - .5)^2 dx = .08333\ldots$ via Wolfram Alpha. So by Chebyshev's $\Pr[|X - .5| \geq (7/8 - .5)] \leq \frac{.08333}{(7/8-.5)^2} = .593$.

3. $\mathbb{E}[X^2] = \int_0^1 x^2 dx = \frac{1}{3}$, so by Markov's $\Pr[X^2 \geq (7/8)^2] \leq \frac{1/3}{(7/8)^2} = .435$. So the uncentered moment gives a better bound.

4. The general equation for an upper bound is:

$$\frac{1/(q+1)}{(7/8)^q}.$$

For $q = 3, 4, \ldots, 10$ we get values:

$$\begin{bmatrix} .373 & .341 & .325 & .318 & .318 & .323 & .333 & .346 \end{bmatrix}$$

So using higher moments at first improves our bound, but then eventually starts giving a weaker bound. The tighted bound is obtained at $q = 6$ and $q = 6$.

5. Let g be the step function which is 0 for $X < 7/8$ and 1 for $X \geq 7/8$. Then we have $\Pr[X \geq 7/8] = \Pr[g(X) \geq 1]$. We have $\mathbb{E}[g(X)] = \frac{1}{8}$, so $\Pr[g(X) \geq 1] \leq \frac{1}{8}$ by Markov's, which gives the tight bound.

Problem 2

Let $Y_1, \ldots, Y_r$ be indicator random variables with $Y_i = 1$ if the i^{th} run of the algorithm failed to output an answer $\mathcal{A}(X)_i$ with $|\mathcal{A}(X)_i - f(X)| \leq \epsilon$. Otherwise, let $Y_i = 0$ if the i^{th} run of the algorithm succeed.

The key observation is that, if $|M - f(X)| \geq \epsilon$ then it must be that at least $\frac{r}{2}$ of the random variables $Y_1, \ldots, Y_r$ equal 1. In particular, suppose $M \geq f(X) + \epsilon$ (so the median is a bad overestimate). The algorithm return an equal or higher value to M at least half the time, so it thus returned an overestimate (and failed) half the time. On the other hand, suppose $M \leq f(X) - \epsilon$ (so the median is a bad underestimate). The algorithm return an equal or smaller value to M at least half the time, so it thus returned an underestimate (and failed) half the time.

So, this problem boils down to proving that:

$$\Pr\left[\sum_i^r Y_i \geq r/2\right] \leq \delta.$$

If $\sum_i^r Y_i \leq r/2$, then by the argument above, it cannot be the $|M - f(X)| \geq \epsilon$, so providing this shows that we obtain error $\leq \epsilon$ with probability $\geq 1 - \delta$.

This is a sum of independent Bernoilli random variables, so let's apply a Chernoff bound! To do so, we will use the fact that the algorithm fails with probability $\leq 1/3$. Indeed, $\Pr\left[\sum_i^r Y_i \leq r/2\right]$ is maximized when $\Pr[Y_i] = 1/3$ exactly for all i. So, assume this is the case. Then we have $\mathbb{E}[\sum_i^r Y_i] = 1r/3$. Plugging into Chernoff bound:

$$\Pr\left[\sum_i^r Y_i \geq (1+\epsilon)r/3\right] \leq e^{-\frac{\epsilon^2 r/3}{(2+\epsilon)}}.$$

Setting $\epsilon = 1/2$, we have

$$\Pr\left[\sum_i^r Y_i \geq r/2\right] \leq e^{-.05333r}.$$

Setting $r = 2\log(1/\delta)$, we have $e^{-.05333r} \leq \delta$ as desired.

Problem 3

1. Once there are $n+1$ servers in this setup, the expected number of items on the $(n+1)^{\text{st}}$ server is $\frac{m}{n+1}$, by symmetry. All of these items (and only these items) must have been relocated when the $(n+1)^{\text{st}}$ server was added. So the expected number of items that move is $\frac{m}{n+1}$.

2. For a server S to own more than a $c\log n/n$ fraction of the interval , it would need to be that *no other server* falls within distance $c\log n/n$ to the left of the server. We can choose the random location of server S first. Then the probability of any one server landing within distance $c\log n/n$ from S's left is $c\log n/n$. So the probability *no servers* land that close is:

$$(1 - c\log n/n)^{n-1} \leq \frac{1}{10n},$$

as long as we choose c to be a large enough constant (same analysis as homework 1). By a *union bound*, we thus have that no server owns more than an $O(\log n/n)$ fraction of the interval with probability $\geq 1 - n\frac{1}{10n} = \frac{9}{10}$ which proves the claim.

3. From Part 2, we could have equivalently proven that no server owns more than a $c\log n/n$ fraction of the interval with probability $19/20$ (by choosing c larger). For the rest of the problem, assume that this event happening.

For servers $S_1, \ldots, S_n$ let $Y_i^{(j)}$ be the indicator random variable that item j lands within distance $c\log n/n$ to S_i's left. Let X_i equal $X_i = \sum_{j=1}^m Y_i^{(j)}$. Since we assumed that no server owns more than a $c \log n/n$ fraction of the interval, X_i is an *upper bound* on the number of items assigned to server i. So it suffices to show that X_i is not too large for all i.

To do so, note that, for a fixed i, $Y_i^{(1)}, Y_i^{(2)}, \ldots, Y_i^{(m)}$ are an independent $\{0,1\}$ random variables, where each is 1 with probability exactly $c \log n/n$. So they are just biased coin flips!

Let $c > 2$ be a sufficiently large constant. Using the Chernoff bound from class with $\epsilon = c$, we get that:

$$\Pr[X_i \geq 2c \cdot \frac{m \log n}{n}] \leq e^{\frac{-c^2 m \log n/n}{2+c}} \leq e^{\frac{-c \log n}{2}} \leq \frac{1}{20n},$$

for large enough c. The last inequality uses that $m > n$ (as specified in the problem).

We conclude via a union bound that no server is assigned more than $O(m \log n/n)$ items with probability $\frac{19}{20}$.

There's one last step – we needed two events to hold for our proof to go through: 1) no server owns more than a $c \log n/n$ fraction of the interval and 2) no server was assigned two many items. Since each holds with probability 19/20, by another union bound, both hold with probability 9/10.