CIS 5020 Analysis of Algorithms Fall 2024

Assignment 2

Due: September 27

Name: Yujie Xu, Richard Zhang

PennID: 51233809, 19331985

1 Question 1

1. We know that for any random variable X:

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

We can deduce that $\mathbb{E}[X]$ is the minimum mean squared deviation from any constant. Then, suppose c is a constant.

$$\mathrm{Var}[X] = \min_c \mathbb{E}[(X - c)^2]$$

This implies that for any constant c,

$$\mathrm{Var}[X] \leq \mathbb{E}[(X - c)^2]$$

Applying this equation to $X = f(Z)$, we can get:

$$\mathrm{Var}[f(Z)] \leq \mathbb{E}[(f(Z) - c)^2]$$

If $c = f(\mathbb{E}[Z])$, then the equation becomes:

$$\mathrm{Var}[f(Z)] \leq \mathbb{E}[(f(Z) - f(\mathbb{E}[Z]))^2]$$

Because f is 1-Lipschitz, for any $x, y \in \mathbb{R}$, we have:

$$|f(x) - f(y)| \leq |x - y|$$

If $x = Z$ and $y = \mathbb{E}[Z]$, then we can get:

$$|f(Z) - f(\mathbb{E}[Z])| \leq |Z - \mathbb{E}[Z]|$$

Squaring both sides and taking the expected values of both sides:

$$(f(Z) - f(\mathbb{E}[Z]))^2 \leq (Z - \mathbb{E}[Z])^2.$$

$$\mathbb{E}\left[(f(Z) - f(\mathbb{E}[Z]))^2\right] \leq \mathbb{E}[(Z - \mathbb{E}[Z])^2] = \mathrm{Var}[Z]$$

Because we know $\text{Var}[f(Z)] \leq \mathbb{E}[(f(Z) - f(\mathbb{E}[Z]))^2]$, the equation becomes:

$$\text{Var}[f(Z)] \leq \mathbb{E}[(f(Z) - f(\mathbb{E}[Z]))^2] \leq \mathbb{E}[(Z - \mathbb{E}[Z])^2] = \text{Var}[Z]$$

Thus, when f is a 1-Lipschitz function.

$$\text{Var}[f(Z)] \leq \text{Var}[Z]$$

2. Suppose W is a random variable. Based on the definition of variance, we can write the variance of W as:

$$\text{Var}(W) = E[W^2] - (E[W])^2$$

Based on the question, we let $W = X - \beta Y$, where β is a constant.

$$W^2 = (X - \beta Y)^2 = X^2 - 2\beta XY + \beta^2 Y^2$$

Taking the expected values of both sides, we get:

$$E[W^2] = E[X^2] - 2\beta E[XY] + \beta^2 E[Y^2]$$

Because variance of any random variable should be greater or equal to 0, we know $\text{Var}(W) \geq 0$. Thus, $E[W^2] - (E[W])^2 \geq 0$

Let us assume $E[W] = 0$ to simplify this question a bit. Then:

$$E[W^2] \geq 0$$

$$E[X^2] - 2\beta E[XY] + \beta^2 E[Y^2] \geq 0$$

Then we want to find a value of β such that $E[W^2]$ can be as small as possible (which is 0 in this case).

Taking the derivative with respect to β:

$$\frac{d}{d\beta}\left(E[X^2] - 2\beta \cdot E[XY] + \beta^2 \cdot E[Y^2]\right) = -2 \cdot E[XY] + 2\beta \cdot E[Y^2] = 0$$

$$\beta = \frac{E[XY]}{E[Y^2]}$$

Plugging in $\beta = \frac{E[XY]}{E[Y^2]}$ to the original inequality $E[X^2] - 2\beta E[XY] + \beta^2 E[Y^2] \geq 0$, we get:

$$E[X^2] - 2\left(\frac{E[XY]}{E[Y^2]}\right) E[XY] + \left(\frac{E[XY]}{E[Y^2]}\right)^2 E[Y^2] \geq 0$$

$$E[X^2] - \frac{E[XY]^2}{E[Y^2]} \geq 0$$

$$E[XY]^2 \leq E[X^2]E[Y^2]$$

Taking the square root of both sides

$$|E[XY]| \leq \sqrt{E[X^2]E[Y^2]}$$

Thus,

$$E[|X \cdot Y|] \leq \sqrt{E[X^2]E[Y^2]}$$

2 Question 2

1. Let $v_1, v_2 \cdots, v_m, m = |E|$ denotes all edges in $G = (V, E)$.

   Let w_i be the indicator of whether an edge is in cut(S). i.e.

$$w_i = \begin{cases} 1, & \text{if the i-th edge is in cut(S)} \\ 0, & \text{otherwise} \end{cases}$$

   Let random variable $X = |\text{cut}(S)|$ denotes the size of the cut.

   Then the size of the size equals to the sum of indicators.

$$X = |\text{cut}(S)| = \sum_{i=1}^{m} w_i$$

   When $w = 1$, the 2 vertices of the edge belongs to different vertex set S and $\overline{S}$. Also note that the chance of 1 vertex to be chosen in S is random and have a possibility of 1/2. Thus,

$$\mathbb{E}(w_i) = \sum Pr(w_i = 1) = \frac{1}{2}$$

$$\text{Var}(w_i) = \mathbb{E}(w_i^2) - \mathbb{E}(w_i)^2 = \sum Pr(w_i^2 = 1) - \frac{1}{4} = \sum Pr(w_i = 1) - \frac{1}{4} = \frac{1}{4}$$

   Note that the indicator w is independent from each other.

$$\mathbb{E}(X) = \mathbb{E}(\sum_{i=1}^{m} w_i) = \sum_{i=1}^{m} \mathbb{E}(w_i) = \sum_{i=1}^{m} \frac{1}{2} = \frac{m}{2}$$

$$\mathrm{Var}(X) = \mathrm{Var}(\sum_{i=1}^{m} w_i) = \sum_{i=1}^{m} \mathrm{Var}(w_i) = \sum_{i=1}^{m} \frac{1}{4} = \frac{m}{4}$$

Conclusion: $\mathbb{E}(|\mathrm{cut}(S)|) = \frac{|E|}{2}, \mathrm{Var}(|\mathrm{cut}(S)|) = \frac{|E|}{4}$

2. Let $p = Pr(|\mathrm{cut}(S)| \geq \frac{|E|}{2}) = Pr(X \geq \frac{m}{2})$.
   Recall the expectation of X:

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x} x \cdot Pr(X = x) \\ &= \sum_{x < \frac{m}{2}} x \cdot Pr(X = x) + \sum_{x \geq \frac{m}{2}} x \cdot Pr(X = x)\end{aligned}$$

   Note that the probability in single point is less equal than cumulative probability.

$$\begin{aligned}&\max_{x < \frac{m}{2}} Pr(X = x) \leq Pr(x < \frac{m}{2}) = 1 - p \\ &\max_{x \geq \frac{m}{2}} Pr(X = x) \leq Pr(x \geq \frac{m}{2}) = p\end{aligned}$$

   Also note that:

$$\begin{aligned}&\max_{x < \frac{m}{2}} \ x = \frac{m}{2} - 1 \\ &\max_{x \geq \frac{m}{2}} \ x = m\end{aligned}$$

   Lemma. $\sum_i a_i \cdot b_i \leq \max_i(a_i) \cdot \max_i(b_i)$
   Proof. $\sum_i a_i \cdot b_i \leq \sum_i a_i \max_i(b_i) \leq \max_i(a_i) \cdot \max_i(b_i)$
   Thus,

$$\begin{aligned}&\mathbb{E}(X) = \\ &\frac{m}{2} = \sum_{x < \frac{m}{2}} x \cdot Pr(X = x) + \sum_{x \geq \frac{m}{2}} x \cdot Pr(X = x) \\ &\quad \leq \max_{x < \frac{m}{2}} Pr(X = x) \max_{x < \frac{m}{2}} \ x + \max_{x \geq \frac{m}{2}} Pr(X = x) \max_{x \geq \frac{m}{2}} \ x \\ &\quad \leq (1 - p)(\frac{m}{2} - 1) + p \cdot m \\ &\quad \leq \frac{m}{2} + \frac{mp}{2} + p - 1\end{aligned}$$

Therefore,

$$p \geq \frac{1}{\frac{m}{2}+1}$$

Assume existing a positive constant c,

$$p \geq \frac{1}{\frac{m}{2}+1} \geq \frac{c}{m}$$

$$0 < c \leq \frac{2}{3} \leq \frac{2}{1+\frac{2}{m}},\ m = |E| \geq 1$$

Therefore, there must exists a positive constant $0 < c \leq \frac{2}{3}$,

s.t. $p = Pr(|\text{cut}(S)| \geq \frac{|E|}{2}) \geq \frac{c}{|E|}$

3 Question 3

1. If we assign M items to N bins uniformly and independently, we can easily get the following conclusion:

   Suppose there are two items x and y $\in$ M. The possibility that x would collide with y is $1/N$.

   If we select two items from M, then the number of possible combinations is:

   $$\binom{M}{2} = \frac{M(M-1)}{2}$$

   Thus,

   $$E[C] = \frac{1}{N} \cdot \frac{M(M-1)}{2} = \frac{M(M-1)}{2N}$$

   When $M \leq 0.01\sqrt{N}$, we can rewrite $E[C]$ as follows:

   $$E[C] \leq \frac{(0.01\sqrt{N})(0.01\sqrt{N}-1)}{2N} \approx \frac{0.0001N}{2N} = \frac{0.0001}{2}$$

   $$E[C] \leq \frac{0.0001}{2} = 0.00005$$

   Therefore, the probability that every item gets its own bin is at least (using Markov's inequality):

   $$P(C=0) = 1 - P(C \geq 1) \geq 1 - 0.00005 = 0.99995 \geq \frac{9}{10}$$

   Thus, when $M \leq 0.01\sqrt{N}$, every item gets its own bin with probability at least $\dfrac{9}{10}$.

2. To compute $E[C^2]$, we start by expressing C in terms of indicator random variables. Let X_{ij} be the indicator that two items i and j are assigned to the same bin. Then,

$$C = \sum_{1 \leq i < j \leq M} X_{ij}.$$

$$E[C] = \sum_{i<j} E[X_{ij}] = \binom{M}{2} \cdot \frac{1}{N} = \frac{M(M-1)}{2N}$$

Now, we need to compute $E[C^2]$. Let X_{kl} be another indicator that two items k and l are assigned to the same bin:

$$C^2 = (\sum_{i<j} X_{ij})^2 = \sum_{i<j} X_{ij}^2 + \sum_{i<j,\, k<l} X_{ij} X_{kl}$$

We know: $1 \leq i < j \leq M$ and $1 \leq k < l \leq M$.

We know $X_{ij}^2 = X_{ij}$ since it's a indicator variable (it can either be 1 or 0). Then, we get the following equation:

$$E[C^2] = \sum_{i<j} E[X_{ij}] + \sum_{i<j,\, k<l} E[X_{ij} X_{kl}] = E[C] + \sum_{i<j,\, k<l} E[X_{ij} X_{kl}]$$

$$E[C^2] = \frac{M(M-1)}{2N} + \sum_{i<j,\, k<l} E[X_{ij} X_{kl}]$$

Then, we need to consider computing $E[X_{ij} X_{kl}]$. We will discuss it under three different scenarios.

(1)Assuming $(i,j) \neq (k,l)$: In other words, i,j,k,l are four different items. Because the two event are independent, we have:

$$E[X_{ij} X_{kl}] = E[X_{ij}] \cdot E[X_{kl}] = (\frac{1}{N})^2 = \frac{1}{N^2}$$

The number of possible terms under this scenario is $\binom{M}{2}\binom{M-2}{2}$ as we want to select 4 different points from M.

Thus,

$$\begin{aligned}\sum_{i<j,\, k<l} E[X_{ij} X_{kl}] &= \frac{M(M-1)}{2} \cdot \frac{(M-2)(M-3)}{2} \cdot \frac{1}{N^2} \\ &= \frac{M(M-1)(M-2)(M-3)}{4N^2}\end{aligned}$$

(2)Assuming there is one overlapping item:

When there is only one overlapping item, the probability that both pairs collide is the probability that the three items are all in the same bin (suppose i is the overlapping item):

$$E[X_{ij}X_{il}] = E[X_{ij}] \cdot E[X_{il}] = \frac{1}{N^2}$$

The number of possible terms is $M \cdot \binom{M-1}{2} \cdot 2$ as we want to find one overlapping item from all items first and then find the other two items from the remaining $M-1$ items.

Thus,

$$\begin{aligned}\sum_{i<j,\, k<l} E[X_{ij}X_{il}] &= M \cdot \binom{M-1}{2} \cdot 2 \cdot \frac{1}{N^2} \\ &= \frac{M(M-1)(M-2)}{N^2}\end{aligned}$$

(3)Assuming (i,j) is the same with (k,l):

This situation is the same with $\sum_{i<j} E[X_{ij}]$, so we won't consider it here.

Thus,

$$\begin{aligned}E[C^2] &= E[C] + \frac{M(M-1)(M-2)(M-3)}{4N^2} + \frac{M(M-1)(M-2)}{N^2} \\ &\approx \frac{M^2}{2N} + \frac{M^4}{4N^2} + \frac{M^3}{N^2}\end{aligned}$$

Computing $(E[C])^2$:

$$\begin{aligned}(E[C])^2 &\approx (\frac{M^2}{2N})^2 \\ &\approx \frac{M^4}{4N^2}\end{aligned}$$

Calculating $\frac{(E[C])^2}{E[C^2]}$:

$$\frac{(E[C])^2}{E[C^2]} = \frac{(E[C])^2}{\frac{M^2}{2N} + \frac{M^4}{4N^2} + \frac{M^3}{N^2}} = \frac{\frac{M^4}{4N^2}}{\frac{M^2}{2N} + \frac{M^4}{4N^2} + \frac{M^3}{N^2}}$$

The Paley-Zygmund inequality states for a non-negative random variable Z, $0 < \alpha < 1$:

$$Pr(Z > \alpha E[Z]) \geq (1-\alpha)^2 \frac{(E[Z])^2}{E[Z^2]}$$

Let $\alpha = \frac{1}{2}$, then:

$$
\begin{aligned}
Pr(C > \frac{E[C]}{2}) &\geq \left(1 - \frac{1}{2}\right)^2 \frac{(E[C])^2}{E[C^2]} \\
&\geq (\frac{1}{2})^2 \times \frac{(E[C])^2}{E[C^2]} \\
&\geq \frac{1}{4} \times \frac{(E[C])^2}{E[C^2]} \\
&\geq \frac{1}{4} \times \frac{\frac{M^4}{4N^2}}{\frac{M^2}{2N} + \frac{M^4}{4N^2} + \frac{M^3}{N^2}}
\end{aligned}
$$

We know C can only be non-negative integer, so $Pr(C \geq 1) = Pr(C > \frac{1}{2})$.
Since $M \geq 0.01\sqrt{N}$, we can set $M = \sqrt{2N}$. We know N is a positive integer:

$$
Pr(C > \frac{E[C]}{2}) = Pr(C > \frac{\frac{M^2}{2N}}{2}) = Pr(C > \frac{M^2}{4N}) = Pr(C > \frac{1}{2})
$$

$$
\begin{aligned}
Pr(C > \frac{1}{2}) &\geq \frac{1}{4} \times \frac{\frac{M^4}{4N^2}}{\frac{M^2}{2N} + \frac{M^4}{4N^2} + \frac{M^3}{N^2}} \\
&\geq \frac{\frac{4N^2}{4N^2}}{\frac{2N}{2N} + \frac{4N^2}{4N^2} + \frac{2N \cdot \sqrt{2N}}{N^2}} \cdot \frac{1}{4} \\
&\geq \frac{1}{1 + 1 + \frac{2N \cdot \sqrt{2N}}{N^2}} \cdot \frac{1}{4} \\
&\geq \frac{1}{2 + \frac{2N \cdot \sqrt{2N}}{N^2}} \cdot \frac{1}{4} \\
&\geq \frac{1}{8 + \frac{8N \cdot \sqrt{2N}}{N^2}}
\end{aligned}
$$

Since N is a positive integer, $\frac{1}{8 + \frac{8N \cdot \sqrt{2N}}{N^2}}$ is a positive constant.

Thus,

$$
Pr(C \geq 1) = Pr(C > \frac{1}{2}) \geq \frac{1}{8 + \frac{8N \cdot \sqrt{2N}}{N^2}} \geq 0
$$

Therefore, if $M \geq 0.01N$, with some constant probability $\gamma > 0$, there is a collision.