CMPSCI 590D: Algorithms for Data Science
Homework#1

October 3, 2017

1 Question 1

The probability of getting a sequence of 'proof' starting at position i is $\left(\frac{1}{26}\right)^5$ when $i \leq 999996$.

Consider a set of indicator random variable X_i's as follows :

$$X_i = \begin{cases} 1 & \text{if 'proof' starts at position i} \\ 0 & \text{if 'proof' doesnt start at position i} \end{cases}$$

To find : $E[\text{Number of occurences of 'proof'}] = E[\sum_{i=0}^{1000000} X_i]$

Using linearity of expectation, $E[\sum_{i=0}^{1000000} X_i] = \sum_{i=0}^{1000000} E[X_i]$

For $1 \leq i \leq 999996$, $E[X_i] = \left(\frac{1}{26}\right)^5$ and for $x > 999996$, $E[X_i] = 0$ because 'proof' needs atleast 5 characters.

Hence,

E[Number of occurences of 'proof'] $= 999996 \left(\frac{1}{26}\right)^5 \approx 0.084$

2 Question 2

Let X_i denote indicator random variable such that $X_i = 1$ only if $\pi(i) = i$, otherwise $X_i = 0$.

The number of fixed points for a permutation π is given by $X = \Sigma_{i=1}^{n} X_i$

From Linearity of Expectation, $E[X] = \Sigma_i \; E[X_i]$

$E[X_i] = 1.Pr[X_i = 1] + 0.Pr[X_i = 0] = Pr[X_i = 1]$

$Pr[X_i = 1] = Pr[\pi(i) = i] = \frac{(n-1)!}{n!} = \frac{1}{n}$

Because of symmetry, $E[X_1] = E[X_2] = \cdots = E[X_n]$.

$E[X] = n\frac{1}{n} = 1$

3 Question 3

3.1 Part (a)

Consider a set of random variables as $X_i = \begin{cases} 1 & \text{if coin toss is a heads} \\ -1 & \text{if coin toss is tails} \end{cases}$

It can be observed that expected payoff $= E[\sum_{i=1}^{100} X_i]$.

Hence

$$\begin{aligned} \text{Expected payoff} &= \sum_{i=0}^{100} E[X_i] & (1) \\ &= \sum_{i=0}^{100} 1 * P(\text{Heads}) + -1 * P(\text{tails}) & (2) \\ &= \sum_{i=0}^{100} 1 * 1/2 + -1 * 1/2 & (3) \\ &= 0 & (4) \end{aligned}$$

In expectation the payoff with an unbiased coin is 0$

3.2 Part (b)

From above it can be observed that

$$\begin{aligned} \text{Expected Payoff} &= \sum_{i=0}^{100} 1 * P(\text{Heads}) + -1 * P(\text{tails}) & (5) \\ &= \sum_{i=0}^{100} 1 * 0.3 + -1 * 0.7 & (6) \\ &= -40 & (7) \end{aligned}$$

Payoff with a biased coin is $-40$$

3.3 Part(c)

Suppose we get x tails out of 100 turns. Then the payoff = x - (100-x) = 2x-100 which is 50 for $x = 75$
For the friend to get more than 50$, there should be atleast 75 tails because when there are 75 tails, the friend earns 50$.
To find : P($X \geq 75$) where X = Number of tails.

E[X] = 0.7 * 100 = 70

Using Markov inequality,
$P(X \geq 75) = \frac{E[X]}{75} = \frac{70}{75} \approx 0.93$

Note: Here we cannot apply Markov inequality directly on $\sum X_i$ considered in parts (a) and (b) because Markov inequality can only be applied when the random variable is non negative.

4 Question 4

X = Sum of numbers appearing over 100 rolls.
Consider X_i be the number appearing on the i th roll of the die.
Hence $X = \sum_{i=1}^{100} X_i$.

$$
\begin{aligned}
E[X] &= E[\sum_{i=1}^{100} X_i] && (8)\\
&= \sum_{i=1}^{100} E[X_i] && (9)\\
&= \sum_{i=1}^{100} \left(1.\frac{1}{6} + 2.\frac{1}{6} + 3.\frac{1}{6} + 4.\frac{1}{6} + 5.\frac{1}{6} + 6.\frac{1}{6}\right) && (10)\\
&= \sum_{i=1}^{100} 3.5 && (11)\\
&= 350 && (12)
\end{aligned}
$$

Now we compute Variance of $X = \sum_{i=1}^{100} X_i$. Since X_i's are independent, variance of X can be written as the sum of variance of X_i's.

$$
\begin{aligned}
Var(X) &= Var(\sum_{i=1}^{100} X_i) && (13)\\
&= \sum_{i=1}^{100} Var(X_i) && (14)\\
&= \sum_{i=1}^{100} \left(E[X_i^2] - (E[X_i])^2\right) && (15)\\
&= \sum_{i=1}^{100} \left(E[X_i^2] - 3.5^2\right) && (16)\\
&= \sum_{i=1}^{100} \left(\left(1.\frac{1}{6} + 4.\frac{1}{6} + 9.\frac{1}{6} + 16.\frac{1}{6} + 25.\frac{1}{6} + 36.\frac{1}{6}\right) - 3.5^2\right) && (17)\\
&= \sum_{i=1}^{100} \left(\frac{91}{6} - \frac{49}{4}\right) && (18)\\
&= 100\left(\frac{35}{12}\right) = \frac{3500}{12} && (19)
\end{aligned}
$$

Using Chernoff bound,

$$
\begin{aligned}
P(|X - \mu| \geq t) &\leq \frac{Var(X)}{t^2} && (20)\\
P(|X - 350| \geq 50) &\leq \frac{Var(X)}{50^2} && (21)\\
&= \frac{3500}{12 * 2500} && (22)\\
&= \frac{7}{60} && (23)
\end{aligned}
$$

5 Question 5

Given: m balls and n bins such that $m \geq n$. B_i is the random variable that indicates the number of balls in bin i.

To find: $E[B_i]$

Let X_{ij} be an indicator random variable such that:

$X_{ij} = \begin{cases} 1 & \text{if ball } j \text{ falls in bin } i \\ 0 & \text{otherwise} \end{cases}$

For $i \in [1, m]$ and $j \in [1, n]$, we have

$$E[X_{ij}] = P(X_{ij} = 1) * 1 + P(X_{ij} = 0) * 0 = P(X_{ij} = 1) = \frac{1}{n} \tag{24}$$

Now $B_i = \sum_{j=1}^{m} X_{ij}$ (Since B_i is the total number of balls in bin i)

$\Rightarrow E[B_i] = E[\sum_{j=1}^{m} X_{ij}]$ (By Linearity of Expectation)

$= \sum_{j=1}^{m} E[X_{ij}]$

$$\Rightarrow E[B_i] = \sum_{j=1}^{m} (1/n) = m/n \tag{25}$$

using eq. 24

5.1 Soln. 5.1

Given: $m = 100n\text{ln}n$

To prove: $P(|B_i - E[B_i]| \leq 25 \text{ ln } n) \geq (1 - \frac{1}{n^2})$

Proof: In order to prove $P(|B_i - E[B_i]| \leq 25\text{ln } n) \geq (1 - \frac{1}{n^2})$ we shall need to prove that if

$$P(B_i - E[B_i] \geq 25 \ln n) \leq p_1 \tag{26}$$

and,

$$P(B_i - E[B_i] \leq -25 \ln n) \leq p_2 \tag{27}$$

then $p_1 + p_2 \leq \frac{1}{n^2}$

Using Chernoff bound we have,

$P(X \geq (1 + \delta)\mu) \leq e^{\frac{-\delta^2}{2+\delta}\mu}$ for all $\delta > 0$ and $\mu = E[X]$ and,

$P(X \leq (1 - \delta)\mu) \leq e^{\frac{-\delta^2}{2}\mu}$ for all $\delta \in [0, 1]$ and $\mu = E[X]$

$$\Rightarrow P(X - \mu \geq \delta\mu) \leq e^{\frac{-\delta^2}{2+\delta}\mu} \tag{28}$$

for all $\delta > 0$ and $\mu = E[X]$ and,

$$\Rightarrow P(X - \mu \leq -\delta\mu) \leq e^{\frac{-\delta^2}{2}\mu} \tag{29}$$

for all $\delta \in [0, 1]$ and $\mu = E[X]$

In our case $X = B_i$

$E[X] = E[B_i] = \frac{m}{n} = 100\text{ln}n$ (from eq(25) and since $m = 100n\text{ln}n$)

Putting $\delta\mu = 25\text{ln}n \Rightarrow \delta = \frac{1}{4}$ (equating equations (26),(28) and (27),(29))

From equations (26),(28) and values of δ and μ we get,

$$P(Bi - E[B_i] \geq 25lnn) \leq e^{\frac{-\delta^2}{2+\delta}\mu} = e^{\frac{-25}{9}lnn} = p_1 \tag{30}$$

Similarly, from equations (26),(29) and values of δ and μ we get,

$$P(B_i - E[B_i] \leq -25lnn) \leq e^{\frac{-\delta^2}{2}\mu} = e^{\frac{-25}{8}lnn} = p_2 \tag{31}$$

Now

$$\begin{aligned} p_1 + p_2 &= e^{\frac{-25}{9}lnn} + e^{\frac{-25}{8}lnn} && (32)\\ &= n^{\frac{-25}{9}} + n^{\frac{-25}{8}} && (33)\\ &\leq n^{-2}(n^{-7/9} + n^{-9/8}) && (34)\\ &\leq n^{-2} \text{ for } n > 2, \text{ as it is a decreasing function} && (35) \end{aligned}$$

Hence proved.

Now we try to analyse a situation when we can get the ratio of maximum to minimum bin as a constant.

Let R denotes the range:$[E[B_i] - 25\ln n, E[B_i] + 25\text{ln}n]$. We have proved above that $P(B_i \in R) \geq 1 - \frac{1}{n^2}$

$\Rightarrow P(B_i \notin R) \leq \frac{1}{n^2}$

$\Rightarrow P($ atleast one of the estimates lie outside $R) \leq \sum_{i=1}^{n} P(B_i \notin R)$

$\Rightarrow P($ atleast one of the estimates lie outside $R) \leq \frac{n}{n^2}$

$\Rightarrow P($ all the estimates lie in $R) \geq 1 - \frac{1}{n}$

Hence, probability that all the estimates lie in range R is atleast$1 - 1/n$. If this holds true, then
maximum load possible in a bin $= E[B_i] + 25\text{ln}n = 125\text{ln}n$ and,
minimum load possible in a bin $= E[B_i] - 25\text{ln}n = 75\text{ln}n$
Therefore, ratio of maximum to minimum load $= \frac{5}{3}$ (= constant) with a probability greater equal to $1 - 1/n$

5.2 Soln 5.2

Given: $m = \Omega(n\text{ln}n)$

To prove: $P($ all estimates lie in range R $= \left(\frac{m}{n} - O(\sqrt{\frac{m}{n}lnn}, \frac{m}{n} + O(\sqrt{\frac{m}{n}lnn})\right) \geq (1 - 1/n)$

Similar to the **Soln 2.1** if we prove that $P\left(\frac{m}{n} - O(\sqrt{\frac{m}{n}lnn} \leq B_i \leq \frac{m}{n} + O(\sqrt{\frac{m}{n}lnn})\right) \geq (1 - \frac{1}{n^2})$, then we can prove that $P($ all estimates lie in range R $= [(\frac{m}{n} - O(\sqrt{\frac{m}{n}lnn}, \frac{m}{n} + O(\sqrt{\frac{m}{n}lnn})]) \geq (1 - \frac{1}{n})$

Proof:
In order to prove this, we can split above probability equation as:

$P(B_i \leq \frac{m}{n} + O(\sqrt{\frac{m}{n}lnn})) \geq p1(say)$

$\Rightarrow P(B_i - \frac{m}{n} \leq O(\sqrt{\frac{m}{n} lnn})) \geq p1$

$$\Rightarrow P(B_i - \frac{m}{n} \geq O(\sqrt{\frac{m}{n} lnn})) \leq p1 \tag{36}$$

$P(B_i \geq \frac{m}{n} - O(\sqrt{\frac{m}{n} lnn})) \geq p2$

$\Rightarrow P(B_i - \frac{m}{n} \geq -O(\sqrt{(\frac{m}{n} \text{ln} n)})) \geq p2$

$$\Rightarrow P(B_i - \frac{m}{n} \leq -O(\sqrt{(\frac{m}{n} \ln n)})) \leq p2 \tag{37}$$

We need to prove that $p1 + p2 \leq 1/n^2$.

Let $m = cn\text{ln}n$ for some positive constant c

Since $E[B_i] = \frac{m}{m}$ (from (2)) we can use Chernoff bound here. From equations (21),(36) and (22),(37) we have,

$\delta\mu = O(\sqrt{(\frac{m}{n} \text{ln} n)}) = \alpha\sqrt{(\frac{m}{n} \text{ln} n)}$ (say) for some positive constant α

$\Rightarrow \delta = \alpha\sqrt{(\frac{n}{m} \text{ln} n)} = \frac{\alpha}{\sqrt{c}}$ (substituting $\mu = \frac{m}{n}$)

From (21) and (36) using values of δ and μ we get

$P(B_i - \frac{m}{n} \geq O(\sqrt{(\frac{m}{n} \text{ln} n)})) \leq e^{\frac{-\delta^2}{2+\delta}\mu}$

$\Rightarrow p1 \leq^{\frac{-\delta^2}{2+\delta}\mu} = e^{\frac{-\alpha^2}{2+\frac{\alpha}{\sqrt{c}}} lnn}$ (substituting the values of δ and m from above)

Similarly, from (22) and (37) using values of δ and μ we get

$P(B_i - \frac{m}{n} \leq -O(\sqrt{(\frac{m}{n} \text{ln} n)})) \leq e^{\frac{-\delta^2}{2}\mu}$

$\Rightarrow p2 \leq^{\frac{-\delta^2}{2}\mu} = e^{\frac{-\alpha^2}{2} lnn}$ (substituting the values of δ and m from above)

$\Rightarrow p1 + p2 \leq e^{\frac{-\alpha^2}{2+\frac{\alpha}{\sqrt{c}}} lnn} + e^{\frac{-\alpha^2}{2} lnn} = n^{\frac{-\alpha^2}{2+\frac{\alpha}{\sqrt{c}}}} + n^{\frac{-\alpha^2}{2}}$

$\Rightarrow n^{\frac{-\alpha^2}{2+\frac{\alpha}{\sqrt{c}}}} + n^{\frac{-\alpha^2}{2}} \leq n^{-2}$
For $c = 40$ and $k > 2$ above constraint holds True as LHS is a decreasing function.
Now, we have proved above that $P(B_i \in R) \geq 1 - \frac{1}{n^2}$

$\Rightarrow P(B_i \notin R) \leq \frac{1}{n^2}$

$\Rightarrow P($ atleast one of the estimates lie outside $R) \leq \sum\limits_{i=1}^{n} P(B_i \notin R)$

$\Rightarrow P($ atleast one of the estimates lie outside $R) \leq \frac{n}{n^2}$

$\Rightarrow P($ all the estimates lie in $R) \geq 1 - \frac{1}{n}$

Hence, probability that all the estimates lie in range R is atleast$1 - 1/n$.
Hence proved.

5.3 Soln 5.3

Given: $m = n$

To Prove: Height of the heaviest bin is $O(\frac{logn}{loglogn})$ with probability $1 - o(1)$

Proof: Probability that the bin i has atleast t balls will be atmost

$$\binom{n}{t}\left(\frac{1}{n}\right)^t \leq \frac{n^t}{t!}.\frac{1}{n^t} \leq \frac{1}{t!} \leq \frac{1}{t^{\frac{t}{2}}} = p(say) \tag{38}$$

For $t = 8\frac{\ln n}{\ln \ln n}$, we try to simplify p.

It can be observed that $t \geq \sqrt{\ln n}$, making $t^{t/2} \geq (\sqrt{\ln n})^{4 \ln n/ \ln \ln n} \geq e^{2 \ln n} = n^2$

So $p \leq 1/n^2$

Now, the probability that some bin has more than t balls is $\leq t*$Prob(bin i has atleast t balls) $\leq tp \leq 1/n$.
So we have shown that the heaviest bin has more than t balls with a probability $\geq 1 - 1/n = 1 - o(1)$ because $1/n = o(1)$ as $1/n \to 0$ as $n \to \infty$, by definition of $o()$.

6 Question 6

Consider an indicator random variable, X_{ij} which is 1 if ith item is chosen in jth run and 0 otherwise

P$(X_{ij} = 1) = 1/100$
Hence $E[\sum_j X_{ij}] = t/100$

Using Chernoff bound, $Pr(|X_i - t/100| \geq t/300) \leq 2e^{\frac{-t}{2700}}$

Taking union bound over all i, we get the probability is less than $200e^{\frac{-t}{2700}}$ which should be less than 1 - 0.99 = 0.01.

Hence we need to evaluate $200e^{\frac{-t}{2700}} \geq 0.01$ which gives $t \geq 2700ln(20000) = 26740$