GA HW 05

Piyush Hinduja

November 2023

Question 1

a)

Law of conditional expectations state that,

$$E[X] = P(F) * E[X|F] + (1 - P(F)) * E[X|\bar{F}]$$

In this problem, we need to find out the expected number of boxes we need to buy given that we have got 'k' stickers out of 'n', we are denoting it as $f(n, k)$.
Probability of choosing a box having one of the already seen 'k' stickers is $\frac{k}{n}$.
And the probability of choosing a box having other than those k seen stickers is, $1 - \frac{k}{n} = \frac{n-k}{n}$.
Now, expected number of boxes we will have to buy if we get a box with unseen sticker in the current box will be $f(n, k + 1)$, whereas, if we get a box with a seen sticker, our expected number of boxes will remain the same $f(n, k)$.
Using definition of conditional expectation,
Expected # of boxes (to buy) = Prob of getting a box with unseen sticker * Expected number of boxes after getting new sticker + Prob of getting a box with already seen sticker * Expected number of boxes after getting a seen sticker

$$f(n, k) = \frac{n-k}{n} * (1 + f(n, k + 1)) + \frac{k}{n} * (1 + f(n, k))$$

In the above equation, '1+' in the both cases of expected boxes denote the box bought in current step.

Now, we have to simplify this equation.

$$f(n, k) = \frac{n-k}{n} * (1 + f(n, k + 1)) + \frac{k}{n} * (1 + f(n, k))$$

$$f(n, k) = \frac{n-k}{n} + \frac{k}{n} + \frac{n-k}{n} * f(n, k + 1) + \frac{k}{n} * f(n, k)$$

$$f(n, k) = 1 + \frac{n-k}{n} * f(n, k + 1) + \frac{k}{n} * f(n, k)$$

$$f(n,k) - \frac{k}{n} * f(n,k) = 1 + \frac{n-k}{n} * f(n,k+1)$$

$$(1 - \frac{k}{n}) * f(n,k) = 1 + \frac{n-k}{n} * f(n,k+1)$$

$$\frac{n-k}{n} * f(n,k) = \frac{n + (n-k) * f(n,k+1)}{n}$$

$$(n-k) * f(n,k) = n + (n-k) * f(n,k+1)$$

$$f(n,k) = \frac{n + (n-k) * f(n,k+1)}{n-k}$$

$$f(n,k) = \frac{n}{n-k} + f(n,k+1)$$

Now we have to evaluate $f(n,0)$ using this equation,

$$f(n,0) = \frac{n}{n} + f(n,1)$$

$$f(n,1) = \frac{n}{n-1} + f(n,2)$$

$$.....$$

$$f(n,n-1) = \frac{n}{1} + f(n,n)$$

Hence,

$$f(n,0) = \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + ... + \frac{n}{2} + \frac{n}{1} + f(n,n)$$

We are given $f(n,n) = 0$,

$$f(n,0) = \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + ... + \frac{n}{2} + \frac{n}{1}$$

$$f(n,0) = n * (\frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + ... + \frac{1}{2} + \frac{1}{1})$$

$$f(n,0) = n * (1 + \frac{1}{2} + ... + \frac{1}{n-1} + \frac{1}{n})$$

$$f(n,0) = n * (logn + c)$$

$$f(n,0) = nlogn + c$$

This mean expected number of boxes we need to buy in order to collect all 'n' stickers will be $n * logn$.

b)

From part(a), we have got E[X] = nlogn because f(n, 0) is nothing but expected number of boxes to buy given that we have got 0 stickers out of n.

Using Markov's inequality,

$$Prob(X \geq 8nlogn) \leq \frac{1}{8}$$

And, $\frac{1}{8} < \frac{1}{4}$. So we can write above inequality as,

$$Prob(X \geq 8nlogn) \leq \frac{1}{4}$$

c)

We need to bound the Prob(X$\geq 8 * nlogn$) by $\frac{1}{n^4}$.

Probability of not seeing a sticker till 8nlogn-1 boxes is $(1 - \frac{1}{n})^{8nlogn-1}$.

Using the formula, $1 - x \leq e^{-x}$,

Prob(not seeing a sticker till 8nlogn-1 boxes) $\leq e^{-\frac{8nlogn-1}{n}}$

Now, Not seeing a sticker i is an independent event from not seeing a sticker j, hence we can calculate probability of not seeing any sticker (out of n) after buying 8nlogn-1 boxes using union bound,
Prob(not seeing at least one sticker after buying 8nlogn-1 boxes) $\leq n * e^{-\frac{8nlogn-1}{n}}$

Hence, probability of seeing remaining sticker(s) in $8nlogn^{th}$ or later boxes will be less than $n * e^{-\frac{8nlogn}{n}}$, that is,

$$Prob(X \geq 8nlogn) \leq n * e^{-\frac{8nlogn-1}{n}}$$

$$Prob(X \geq 8nlogn) \leq n * e^{-8logn+\frac{1}{n}}$$

$$Prob(X \geq 8nlogn) \leq n * e^{-8logn} * e^{\frac{1}{n}}$$

$$Prob(X \geq 8nlogn) \leq \frac{n}{n^8} * e^{\frac{1}{n}}$$

$$Prob(X \geq 8nlogn) \leq \frac{e^{\frac{1}{n}}}{n^7}$$

And, $\frac{e^{\frac{1}{n}}}{n^7} < \frac{1}{n^4}$, so,

$$Prob(X \geq 8nlogn) \leq \frac{1}{n^4}$$

Hence Proved.

Question 2

a)

We are given,

$$E[X_t] = \frac{1}{2}(X_{t-1} + 1) + \frac{1}{2}(X_{t-1} - 1)$$

That is,

$$E[X_{t-1}] = \frac{1}{2}(X_{t-2} + 1) + \frac{1}{2}(X_{t-2} - 1)$$

.....

$$E[X_1] = \frac{1}{2}(X_0 + 1) + \frac{1}{2}(X_0 - 1)$$

where X_0 is the position of particle at time 0 and we are given that $X_0 = 0$.

$$E[X_1] = \frac{1}{2} * (1) + \frac{1}{2} * (-1) = 0$$

When we keep substituting the values in above equations, we will get,

$$E[X_1] = E[X_2] = ... = E[X_t] = 0$$

Hence we get, $E[X_t] = 0$, for all t$\geq$1.

b)

Now for the first part of this sub-question, we will calculate $E[X_t^2]$ using Var(X_t),

$$Var(X_t) = E[X_t^2] - (E[X_t])^2$$

We know $E[X_t] = 0$, so lets calculate $E[X_t^2]$,

$$E[X_t^2] = \frac{1}{2}(X_{t-1} + 1)^2 + \frac{1}{2}(X_{t-1} - 1)^2$$

.....

$$E[X_1^2] = \frac{1}{2}(X_0 + 1)^2 + \frac{1}{2}(X_0 - 1)^2$$

$$E[X_1^2] = \frac{1}{2}(X_0^2 + 2 * X_0 + 1) + \frac{1}{2}(X_0^2 - 2 * X_0 + 1)$$

$$E[X_1^2] = X_0^2 + 1$$

We know, $X_0 = 0$,

$$E[X_1^2] = 1$$

Now,

$$E[X_2^2] = \frac{1}{2}(X_1 + 1)^2 + \frac{1}{2}(X_1 - 1)^2$$

$$E[X_2^2] = X_1^2 + 1 = 2$$

.....

$E[(X_t)^2] = t$, for all $t \geq 1$
Finally, we get, $Var(x) = E[X_t^2] = t$

Now in the second part, we will use the Chebyshev's inequality to get the desired proof.
Chebyshev's inequality states that,

$$Prob[|X_t - E[X_t]| \geq k * \sqrt{Var(x)}] \leq \frac{1}{k^2}$$

$$Prob[|X_t - 0| \geq k * \sqrt{t} \leq \frac{1}{k^2}]$$

Putting $k = 2$

$$Prob[|X_t| \geq 2\sqrt{t}] \leq \frac{1}{4}$$

That is,

$$Prob[|X_t| \leq 2\sqrt{t}] \geq \frac{3}{4}$$

Hence Proved.

c)

For t=40000, element crosses origin **164.78** times on an average of 50 runs.

For t=90000, element crosses origin **257.62** times on an average of 50 runs.

For t=160000, element crosses origin **360.78** times on an average of 50 runs.

Link to code

Question 3

Probabilities on different sample size (Sampling is done WITH REPLACEMENT):

# of samples 20: **0.64**

# of samples 100: **0.71**

# of samples 400: **0.8**

In order to get a probability of 0.9, we will need a sample size of **900**.

Link to code

Question 4

a)

r.v. X: Number of constraints getting satisfied

$$X = X_1 + X_2 + ... + X_m$$

where X_i: 1, if i^{th} constraint is satisfied; 0, otherwise.
Using Linearity of expectation,

$$E[X] = E[X_1] + E[X_2] + ... + E[X_m]$$

Now, $E[X_t]$ = 1*Prob(constraint t is satisfied) + 0*Prob(constraint t is NOT satisfied).
In a uniformly random ordering scenario, lets say we have three elements placed in order as: *abc*.
For this ordering we will have 3! possible constraints (a,b,c), (a,c,b), (b,a,c), (b,c,a), (c,a,b), (c,b,a) out of which 4 ((a,b,c), (a,c,b), (c,a,b), (c,b,a)) of them are satisfied in this random ordering.

Hence we can say that given a random ordering and 'm' constraints, each constraint will have a probability of $\frac{2}{3}$ of getting satisfied.
$E[X_t] = \frac{2}{3}$
Therefore, $E[X] = 1 * \frac{2}{3} * m + 0 = \frac{2}{3} * m$

b)

We have E[X]=$\frac{2m}{3}$, hence, for r.v. Y: number of unsatisfied constraints, E[Y]=$\frac{m}{3}$.

Using Markov's Inequality,

$$Prob[Y \geq \frac{m+3}{3}] \leq \frac{E[Y]}{\frac{m+3}{3}}$$

$$Prob[Y \geq \frac{m+3}{3}] \leq \frac{\frac{m}{3}}{\frac{m+3}{3}}$$

$$Prob[Y \geq \frac{m+3}{3}] \leq \frac{m}{m+3}$$

Now we know that as X increases, Y decreases, that is,

$$Prob(X \geq \frac{2m}{3}) = Prob(Y \leq \frac{m}{3}) = 1 - Prob(Y > \frac{m}{3}) = 1 - Prob(Y \geq \frac{m}{3} + 1) = 1 - Prob(Y \geq \frac{m+3}{3})$$

$$Prob[X \geq \frac{2m}{3}] \geq 1 - \frac{m}{m+3}$$

$$Prob[X \geq \frac{2m}{3}] \geq \frac{3}{m+3}$$

Finally, $\frac{3}{m+3} > \frac{1}{m}$, we can say that,

$$Prob[X \geq \frac{2m}{3}] \geq \frac{1}{m}$$

Hence Proved.

Question 5

a)

r.v. X : Expected number of pairs (i, j) having birthday on the same day such that $i < j$.

$$X = X_1 + X_2 + ... + X_n$$

where X_k: 1, if pair k has same birthday; 0, otherwise.
Using Linearity of Expectation,

$$E[X] = E[X_1] + E[X_2] + ... + E[X_n]$$

$E[X_t]$ = 1*Prob(Pair t having birthday on same day) + 0*Prob(Pair t having birthday on different days)
And Probability of 2 people having birthday on the same day is: $\frac{m}{m} * \frac{1}{m} = \frac{1}{m}$.

Now we need to find the number of pairs such that $i < j$.
For n=2, we will have 1 pair; for n=3, we will have 3 pairs; for n=4, we will have 6 pairs.
Hence we will have a sum of series as: $1 + 3 + 6 + 10 + ... = \frac{n*(n-1)}{2}$.

Therefore, expected number of pairs (such that $i < j$) having birthday on the same day are, $E[X] = \frac{1}{m} + \frac{1}{m} + ...(\frac{n(n-1)}{2} times) = \frac{n(n-1)}{2} * \frac{1}{m} = \frac{n(n-1)}{2m}$.

Writing n as a fraction of m:

$$\frac{n(n-1)}{2m} = 1$$

$$n(n-1) = 2m$$

$$n^2 - n = 2m$$

Solving the quadratic equation, we get,
$n = \frac{1+\sqrt{1+8m}}{2}$ OR $n = \frac{1-\sqrt{1+8m}}{2}$
For m≥1, we will get the value of second root negative, which is not possible, hence,

$$n = \frac{1 + \sqrt{1 + 8m}}{2}$$

b)

The station claims of having 1 million songs, n = 1 million.

Probability of first song being repeated is 0; second song being repeated is $\frac{1}{1million}$; third song being repeated is $\frac{2}{1million}$ and so on.

So, Probability of i^{th} song being played before $= \frac{i-1}{1million}$.

Prob(of a song getting repeated in first 200 songs) $= \frac{0}{1million} + \frac{1}{1million} + ... + \frac{199}{1million}$

Prob(of a song getting repeated in first 200 songs) $= \frac{199*200}{2*1000000} = \frac{199}{10000} = 0.0199$

Prob(of no song getting repeated among first 200 songs) $= 0.9801$

After having such high probability of no song getting repeated, a song was still repeated.

Hence, we can say that the station's claim of having a 1 million song base is false with a very high probability.