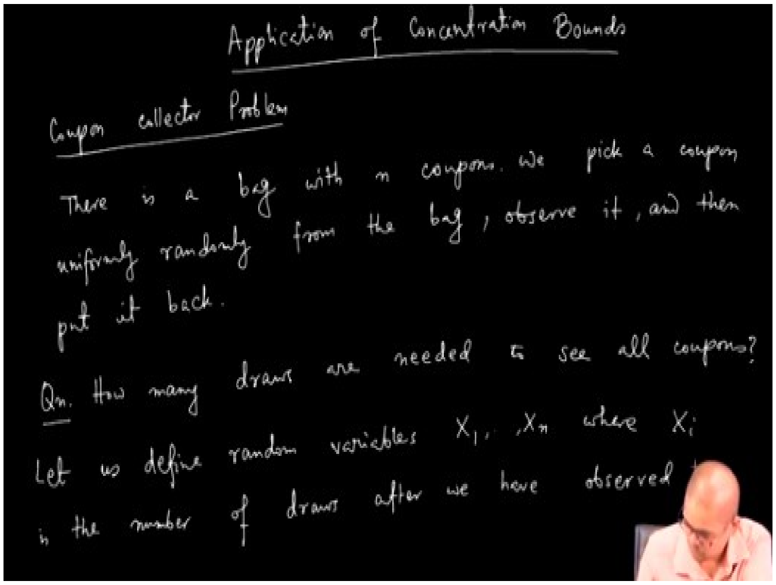
Selected Topics in Algorithm
Prof. Palash Dey
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Module No # 04
Lecture No # 16
Coupon Collector Problem

Welcome in the last class we have seen the proof of channel bound and see how we can use it to get a tight concentration on estimating the probability of how with what a coin comes up head and also the print in the winner of election.

(Refer Slide Time: 00:51)



Application of Concentration Bounds

Coupon collector Problem

There is a bag with n coupons. We pick a coupon uniformly randomly from the bag, observe it, and then put it back.

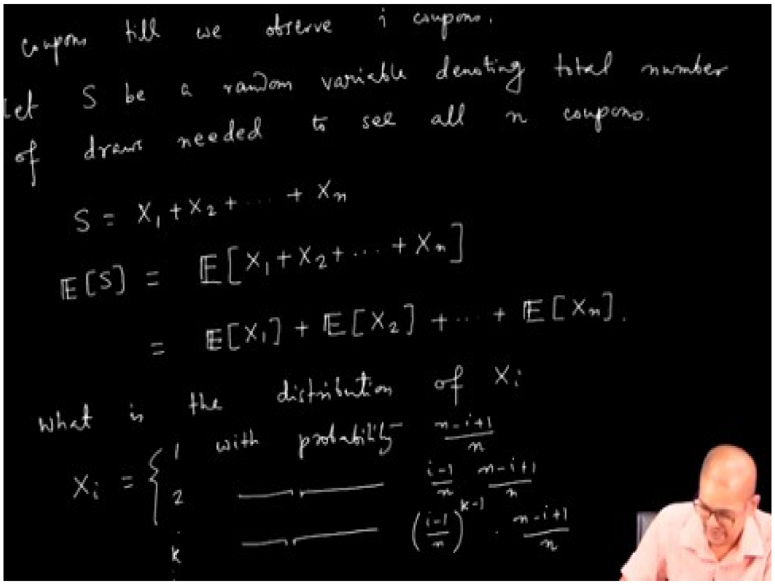
Qn. How many draws are needed to see all coupons?

Let us define random variables $X_1, .., X_n$ where X_i is the number of draws after we have observed

So in today's plus also we; will continue that discussion of application of concentration bounds. So our first problem is the coupon collector problem. So what is the setting? Setting is that there is a bag with in coupons we pick a coupon uniformly randomly from the bag observe it. We see what is that coupon observe it and then put it back. Question how many draws are needed to see all coupons? Of course we need at least n draws but because you are picking any random coupon and we are and this sampling is with replacement.

That variable is around its expectation means whichever coupon we whichever sample we draw we put that sample again. So we can observe same coupon many number of times and what the expected number of draws that one needed? And so this is a random variable and we will see its expectation and we will I will see how concentrated that random. So first let us find out the expected number of cross towards that.

(Refer Slide Time: 05:11)



coupon till we observe i coupons.

let S be a random variable denoting total number of draws needed to see all n coupons.

$$S = X_1 + X_2 + \cdots + X_n$$

$$\mathbb{E}[S] = \mathbb{E}[X_1 + X_2 + \cdots + X_n]$$

$$= \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n].$$

What is the distribution of X_i

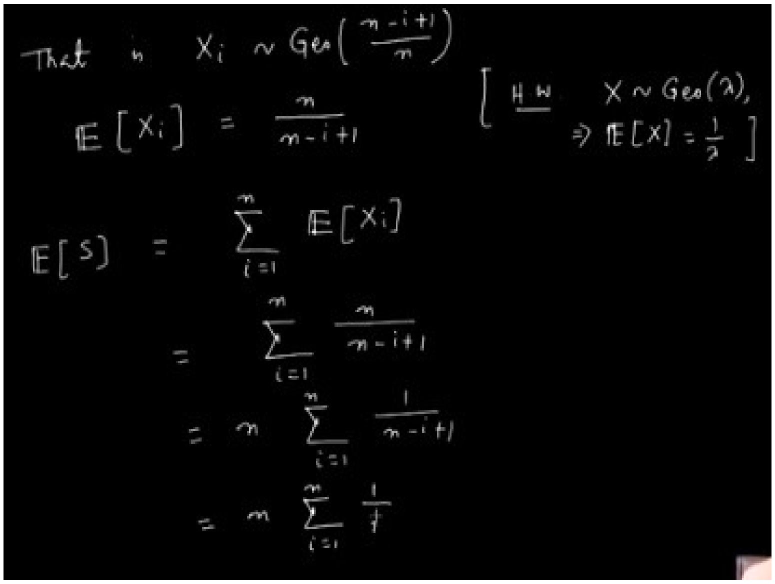
$$X_i = \begin{cases} 1 & \text{with probability } \frac{n-i+1}{n} \\ 2 & \text{——"——} \quad \frac{i-1}{n} \cdot \frac{n-i+1}{n} \\ \vdots & \\ k & \text{——"——} \quad \left(\frac{i-1}{n}\right)^{k-1} \cdot \frac{n-i+1}{n} \end{cases}$$

Let us define random variables $X_1, \ldots, X_n$ where X_i is the number of draws after we have observed i -1 coupons till we observe I coupons. So and let S, be a random variable denoting total numbered of draws needed to see all in coupons. So S is sum of $X_1, \ldots, X_n$ and hence expectation of S is expectation of $X_1 + X_2 + \ldots + X_n$ this is expectation of X_1 + expectation of X_2 + dot expectation of X_n. So to compute expectation of S it is enough to compute expectation of X_i now see that what is the how the random variable X_i is distributed.

What is the distribution of X_i? So X_i takes value 1 with probability you know i -1 coupons have been already seen and so there are n - i +1 many coupons which are not seen yet. And the next draw itself I pick 1 of such coupon is $\frac{n-i+1}{n}$. So this much probability X_i will take value 1; X_i take value 2 if this is with probability the far the fast draw I see a repeated coupon I see the coupon which I have already seen that happens with probability i - 1 of n and in the second row I see a new coupon which happens with probability $\frac{n-i+1}{n}$.

Similarly reasoning this way X_i take value k with probability the first k -1 time a coupon which has been already seen that coupon is picked and the k-th time and unseen coupon is picked. so this sort of random variables is called geometric random variable.

(Refer Slide Time: 09:29)



That is $X_i \sim Geo\left(\frac{n-i+1}{n}\right)$

$E[X_i] = \frac{n}{n-i+1}$

[H.W $X \sim Geo(\lambda)$, $\Rightarrow E[X] = \frac{1}{\lambda}$]

$$E[S] = \sum_{i=1}^{n} E[X_i]$$

$$= \sum_{i=1}^{n} \frac{n}{n-i+1}$$

$$= n \sum_{i=1}^{n} \frac{1}{n-i+1}$$

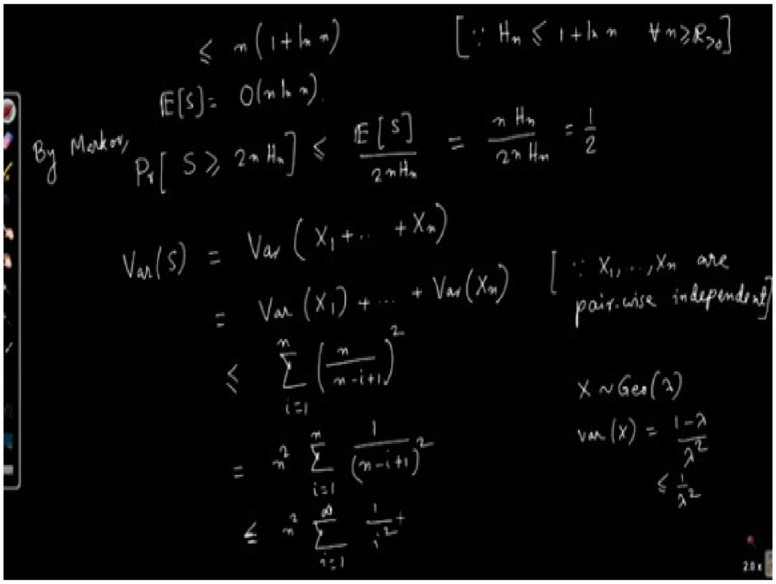
$$= n \sum_{i=1}^{n} \frac{1}{i}$$

So that; is X_i is distributed geometric distribution with parameter is probability of success that means in this case it is n -i +1. So these sorts of random variables are applicable where you know we are keep trying and probability of success in each try is the parameter of the geometric random variable. And we keep on trying till we succeed in the first which still we succeed and the random variable indicates how many times we have tried till success.

So we know that expectation of a geometric random variable is 1 over that probability of success so from that it follows this $\frac{n}{n-i+1}$. So you can if you do not know you can take this as a homework that if X is distributed geometrically with parameter λ then expectation of X is $\frac{1}{\lambda}$. And which is intuitive also it is like if you are if trying something where probability of success is say 10% then on expectation on average I need to try 10 times.

So now using this; what is the expectation of is recall what expectation of s was summation expectation of X_i = 1 to n this is summation i = 1 to $\frac{n}{n-i+1}$. n goes outside $\sum_{i=1}^{n}\frac{1}{n-i+1}$. Now indexing with n - i you know then this by renaming the index it can be written as $\sum_{i=1}^{n}\frac{1}{i}$.

(Refer Slide Time: 12:23)



$$\leq n(1+\ln n) \qquad [\because H_n \leq 1+\ln n \quad \forall n \geq R_{>0}]$$

$$E[S] = O(n\ln n).$$

By Markov,

$$\Pr[S \geq 2nH_n] \leq \frac{E[S]}{2nH_n} = \frac{nH_n}{2nH_n} = \frac{1}{2}$$

$$\begin{aligned} Var(S) &= Var(X_1 + \dots + X_n) \\ &= Var(X_1) + \dots + Var(X_n) \quad [\because X_1, \dots, X_n \text{ are pair-wise independent}] \\ &\leq \sum_{i=1}^{n}\left(\frac{n}{n-i+1}\right)^2 \\ &= n^2\sum_{i=1}^{n}\frac{1}{(n-i+1)^2} \\ &\leq n^2\sum_{i=1}^{\infty}\frac{1}{i^2} + \end{aligned}$$

$$X \sim Geo(\lambda)$$

$$var(X) = \frac{1-\lambda}{\lambda^2} \leq \frac{1}{\lambda^2}$$

Just like this sum is called the harmonic sum is nH_n which is less than equal to times H_n is less than equal to $1+\ln n$ for every n for all real numbers positive. Next now we have found the expectation so this is like O of so expectation of S is at most $n\ln n$ say $O(n\log n)$. And now let us try to bound this let us try to see how s is concentrated around its mean.

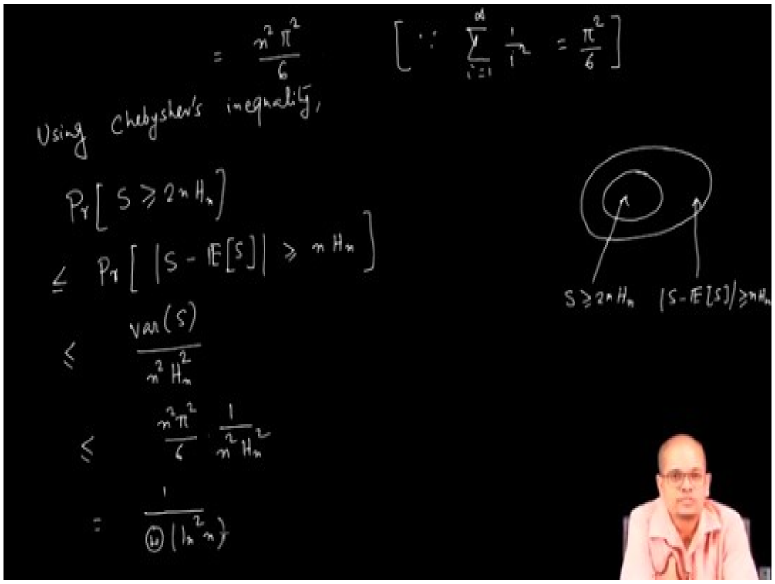
So probability that S is greater than equal to twice in H_n this is less than equal to expectation office so now by Markov. See each concentration inequality what bound it gives so his expectation of S by twice nH_n an expectation of S is less than equal to nH_n and this is should be equality is half. So this is the bound we get using Markov let us see what bound we get using Chebyshev and for Chebyshev we need to find expected variance of a square so variance of S is variance of X_1 +...+ X_n.

Now each X_i is independent so this becomes variance of X_1 + ...+ variance of X_n. They are not identically distributed but they are independent like linearity of expectation unlike linearity of expectation variation does not go linear there are some covariance terms that we need to subtract unless they are independent they are pairwise independent. But here they are fully independent and for independent pairwise independent random variables covariance is 0 this is variance goes inside.

So let me write here since $X_1, \ldots, X_n$ are pairwise independent. And what is the variance of variance of this? So again if X is geometric λ then variance of X is $\frac{(1-\lambda)}{\lambda^2}$. So in particular this is less than equal to $\frac{1}{\lambda^2}$. So that we will use so is less than equal to $\sum_{i=1}^{n} \frac{1}{\lambda^2}$ and the parameter here in this case is this $\left(\frac{n}{n-i+1}\right)^2$.

Again let us do the reindexing instead of i index with n - i and let us call that I this is $\sum_{i=1}^{n} \frac{n^2}{i^2}$. This is actually it can be easier if we make this sum to infinity and write it less than equal to infinity and this sum summation over 1 over i square is a convergent sum and convergent series and this is $\frac{\pi^2}{6}$.

(Refer Slide Time: 18:39)



$$= \frac{n^2 \pi^2}{6} \qquad \left[\because \sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6} \right]$$

Using Chebyshev's inequality,

$$\begin{aligned} & \Pr\left[S \geq 2n H_n \right] \\ \leq\ & \Pr\left[\left| S - \mathbb{E}[S] \right| \geq n H_n \right] \\ \leq\ & \frac{\text{var}(S)}{n^2 H_n^2} \\ \leq\ & \frac{n^2 \pi^2}{6} \cdot \frac{1}{n^2 H_n^2} \\ =\ & \frac{1}{\Theta(\ln^2 n)} \end{aligned}$$

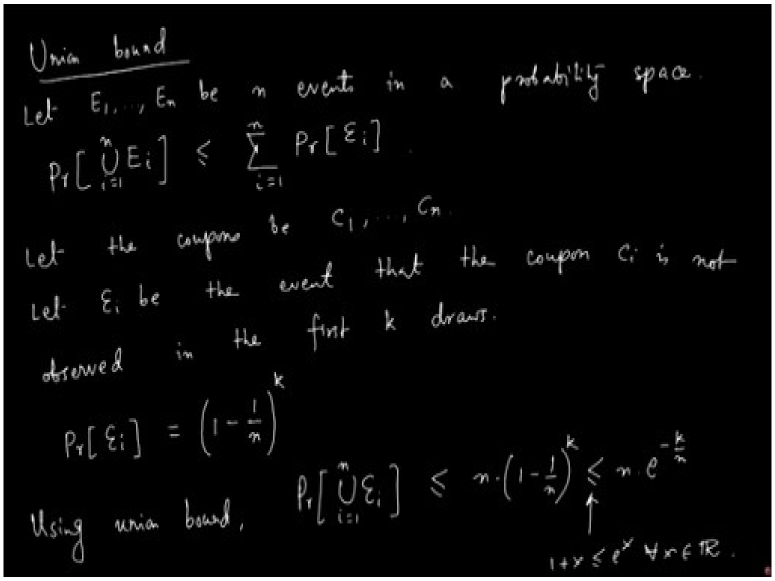
So this is $n^2 \frac{\pi^2}{6}$. This is a standard using standard technique of infinite series and sum you can you can get this. So now using Chebyshev's inequality what we get is probability mod of S minus or just a minute so your bounding probability is greater than equal to twice $n H_n$ this bound came out to be less than equal to half using Markov let us see how what we get using Chebyshev

So for Chebyshev we need to apply Chebyshev we need to bring it into Chebyshev's term or Chebyshev's form it means S minus expectation of S this is mod this is greater than $n H_n$ and this is less than equal to 0. Because mod S minus expectation of S greater than equal to H_n implies that n does not imply that S is greater than equal to twice $n H_n$ but if S is greater than equal to twice $n H_n$ then it must be that S mod of S minus expectation of S is greater than equal to $n H_n$.

It is basically this is a larger event so this is a suppose this is a sample space of S minus expectation of S greater than equal to $n H_n$. and in a subspace it holds that is greater than equal to twice in h n. and now apply Chebyshev's bound is less than equal to variation of S by $n^2 H_n^2$ and variations of S is at most n square phi square by 6 this is $n^2 \frac{\pi^2}{6}$ times $\frac{1}{n^2 H_n^2}$.

So this is $\frac{1}{\Theta(\ln^2 n)}$ in which is a much tighter bound than what we got using Markov's inequality. But for this coupon collector problem there is another way to get an even tighter bound using simple union bound so what is union bound? Let us introduce union bound.

(Refer Slide Time: 22:26)



Union bound

Let $E_1, \dots, E_n$ be n events in a probability space.

$$Pr\left[\bigcup_{i=1}^{n} E_i\right] \leqslant \sum_{i=1}^{n} Pr[E_i].$$

Let the coupons be $C_1, \dots, C_n$.

Let E_i be the event that the coupon C_i is not observed in the first k draws.

$$Pr[E_i] = \left(1 - \frac{1}{n}\right)^k$$

Using union bound, $Pr\left[\bigcup_{i=1}^{n} E_i\right] \leqslant n \cdot \left(1 - \frac{1}{n}\right)^k \leqslant n \cdot e^{-\frac{k}{n}}$

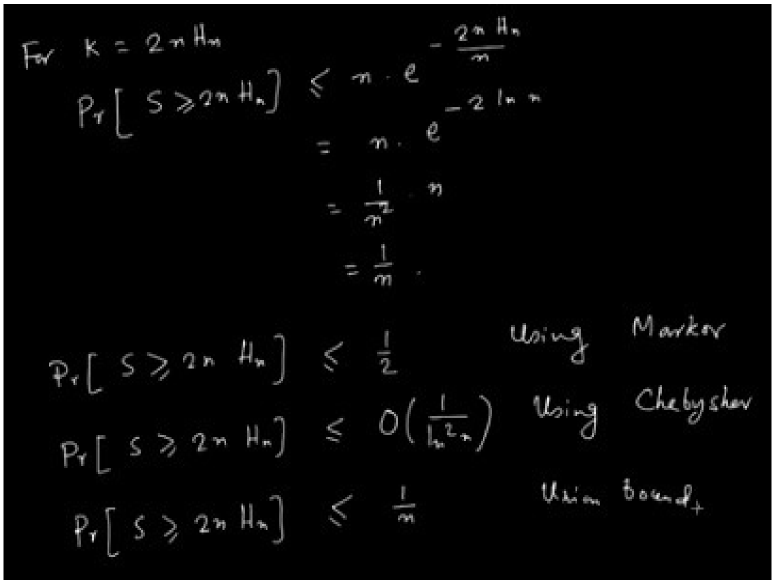
$\uparrow$

$1 + x \leqslant e^x \ \forall x \in \mathbb{R}$.

So let no $E_1, \dots, E_n$ events in sample space in a probability space then probability of $\cup_{i=1}^{n} E_i$ this is less than equal to $\sum_{i=1}^{n} Pr[E_i]$ this union bound using the simple union bound we will get an even tighter bound for coupon collector problem. So let the coupons be let us just give some name these are the coupons $C_1, \dots, C_n$ and let us define an event let E_i be the event that the coupon C_i is not observed in the first k draws.

So probability of E_i so there are n coupons so in 1 draw this particular coupon C_i is not drawn is with probability $1 - \frac{1}{n}$. And each draw is independent so the probability that i'th coupon is not drawn at all is $\left(1 - \frac{1}{n}\right)^k$. Now using union bound using union bound probability of $\cup_{i=1}^{n} E_i$ is less than equal to $n\left(1 - \frac{1}{n}\right)^k$. And this is less than equal to $n e^{\frac{-k}{n}}$ this follows from $1 + x \leq e^x$ this holds for all real number $x \in \mathbb{R}$.

(Refer Slide Time: 26:47)



For $k = 2nH_n$

$$\Pr[S \geq 2nH_n] \leq n \cdot e^{-\frac{2nH_n}{n}}$$
$$= n \cdot e^{-2\ln n}$$
$$= \frac{1}{n^2} \cdot n$$
$$= \frac{1}{n}.$$

$\Pr[S \geq 2n H_n] \leq \frac{1}{2}$ Using Markov

$\Pr[S \geq 2n H_n] \leq O\left(\frac{1}{\ln^2 n}\right)$ Using Chebyshev

$\Pr[S \geq 2n H_n] \leq \frac{1}{n}$ Union bound+

And now let us put for k equal to twice nH_n. So this is the probability that you know all coupons are not observed in the fast twice in H_n many draws. This is probability that this is s greater than equal to twice nH_n; this is less than equal to this is $ne^{\frac{-k}{n}}$ and in, k I will put twice nH_n times e to the power minus twice nH_n by n. So and this is like $ne^{-2\ln n}$ is like $\frac{1}{n}$.

So you see that using this Machinery we get the tightest possible bound so in summary using Markov bound we get probability of S greater than equal to twice nH_n less than equal to half this is using Markov. There is a weakest bound we get using Chebyshev we get probability S greater than equal to twice in H_n this is $O\left(\frac{1}{\ln^2 n}\right)$ this is using Chebyshev. And using union bound we get that probability is greater than equal to twice nH_n this is less than equal to $\frac{1}{n}$ this is union bond.

Observe that the charm of bound is not directly applicable because X_i's are not you know 0-1 random variable. And not only that you know there are versions of channel bound which works for not 0-1 random variables but they need not they need to be at least bounded random variable. Which are not in this case these are unbounded so chart of bound is not directly applicable here. Of course we apply to what is called a sub Gaussian class of random variables but these are not suitable for applying channel bond so we will stop here today thank you.