Bandit Algorithm (Online Machine Learning)

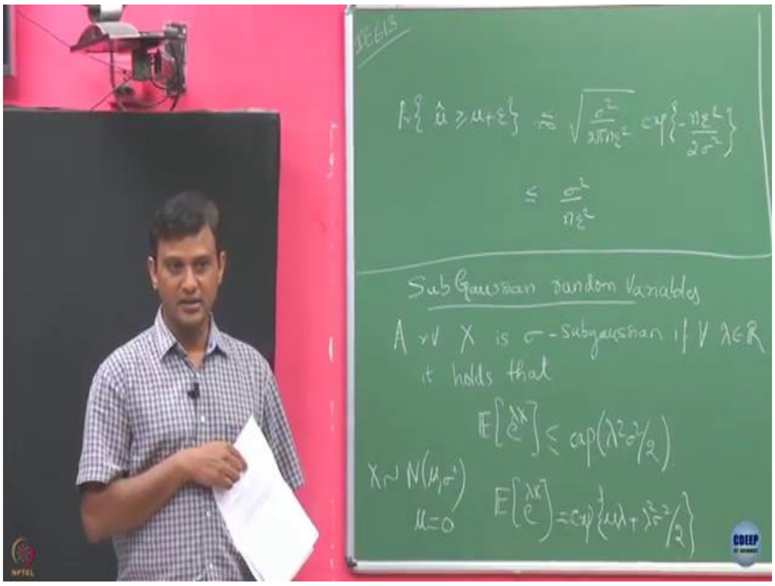
Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Lecture – 28
Sub-Gaussian Random Variable

So, in the last class we started talking about Stochastic Bandits. And, in the last class, we talked about the stochastic bandit setting that we will be interested in; then we introduced the different bandit environments, right? We said that my arms distribution can come from a set of class and that class is going to define my environment class and if I pick one particular set of distributions and assign it to one to each arm that is going to define a bandit instance.

After that, now we are started looking into how to estimate and one natural estimator we considered is simply take the average of all the samples. And then we are interested is how is this sample average is away from the true mean value. Ultimately our interest was to identify an arm which has the highest mean, right. So, we want to identify among the arms which we can play which has the highest mean.

(Refer Slide Time: 01:47)


IE613
$P_r\{\hat{\mu} \geq \mu+\varepsilon\} \approx \sqrt{\frac{\sigma^2}{2\pi n\varepsilon^2}} \exp\left\{-\frac{n\varepsilon^2}{2\sigma^2}\right\}$
$\leq \frac{\sigma^2}{n\varepsilon^2}$
Sub Gaussian random Variables
A r.v X is σ-Subgaussian if $\forall \lambda \in \mathbb{R}$
it holds that
$\mathbb{E}[e^{\lambda X}] \leq \exp(\lambda^2\sigma^2/2)$
$X \sim N(\mu,\sigma^2)$
$\mu = 0$
$E[e^{\lambda X}] = \exp\{\mu\lambda + \lambda^2\sigma^2/2\}$

We discussed both this Markov inequality and Chernoff inequality sorry, Chebyshev and then using our central limit theorem we are able we argued that the estimator being away from this. This is upper bounded by and again this was like an approximations for sufficiently enlarge we argued that that is like bounded like this. So, this is the boundary got using central limit theorem for n sufficiently large and also we had another bound right based on our Chebyshev inequality, what was that? It was sigma square divided by n epsilon square.

So, this bound here decrease inversely in n whereas, this bound here it kinds of decrease exponentially in n. So, this is a tighter bound, but we this bound we got it only approximately assuming that n sufficiently large. But, this gave us intuition that possibly this bound may be weaker; it is not like it is falling inversely in n, possibilities falling exponential in n. So, for that we will now further look into how to get this can we get a bound which are decaying exponentially n, but the exact the bounds of exact ones not like in approximately not necessarily n has to be sufficiently large, whatever n number of samples we have can we express our bond in similar to this ok.

For that we are going to now focus or introduce one set of random variables called sub Gaussian random variables. So, what is the sub Gaussian random variable? A random variable X is sigma sub Gaussian if for all lambda in R it holds that expected value of e to the power lambda X is upper bounded by exponential value of lambda square sigma square by 2.

So, we were just saying that if you have a random variable X, exponent it or basically take the what we call this as? Some moment functions, right? moment generating function and if that is upper bounded by like this, we are going to call it as sub Gaussian with parameter sigma ok. Just to see are there any random variable which satisfies this property, so, let us take a x to be Gaussian with mean μ and σ^2.

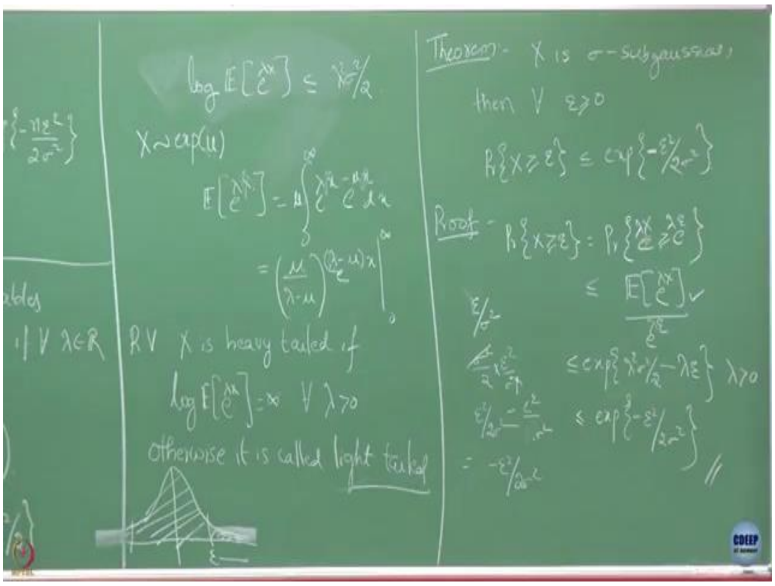
So, if you are going to compute its moment what is the bound look like? So, what do we know the characteristic function of a random variable is unique right? What is the Gaussian look like? exp($\mu\lambda$).

Student: (Refer Time: 06:19).

Lambda square σ^2 by 2 ok. This is what. But, suppose I assume that this μ I assume the distribution this Gaussian distribution have to have a 0 mean ok. So, if I have a 0 mean, then this guy is exactly exponential σ^2 lambda square by 2. This is what it is saying. So, at least the Gaussian random variable with 0 mean and sigma square variance is sigma sub Gaussian, have at least one random variable which is sub Gaussian.

Since this is like expressed in terms of moment generating functions, how did you define the moment generating function this or log of this?

(Refer Slide Time: 07:17)



$\{-\frac{\lambda \varepsilon^2}{2\sigma^2}\}$

ables

$\forall \lambda \in \mathbb{R}$

$\log E[e^{\lambda X}] \leq \lambda^2\sigma^2/2$

$X \sim \exp(\mu)$

$E[e^{\lambda X}] = \mu \int_0^\infty e^{\lambda x} e^{-\mu x} dx$

$= \left(\frac{\mu}{\lambda - \mu}\right) e^{(\lambda-\mu)x} \Big|_0^\infty$

RV X is heavy tailed if

$\log E[e^{\lambda X}] = \infty \quad \forall \lambda > 0$

Otherwise it is called light tailed

Theorem:- X is σ-subgaussian, then $\forall \varepsilon \geq 0$

$P_r\{X \geq \varepsilon\} \leq \exp\{-\varepsilon^2/2\sigma^2\}$

Proof:- $P_r\{X \geq \varepsilon\} = P_r\{e^{\lambda X} \geq e^{\lambda \varepsilon}\}$

$\leq \frac{E[e^{\lambda X}]}{e^{\lambda \varepsilon}}$ ✓

$\leq \exp\{\lambda^2\sigma^2/2 - \lambda\varepsilon\} \quad \lambda > 0$

$\leq \exp\{-\varepsilon^2/2\sigma^2\}$ //

ε/σ^2

$\frac{\sigma^2}{2} \frac{\varepsilon^2}{\sigma^4}$

$\varepsilon^2/2\sigma^2 - \frac{\varepsilon^2}{\sigma^2}$

$= -\varepsilon^2/2\sigma^2$

So, alternative characterization of whatever we have here is you should take log on both sides. It just says that number square a sigma square by 2. So, they are saying that if a random variable X satisfies this this can bound then it is a sigma sub Gaussian. And, it is not necessary that every random variable we have need to be sub Gaussian ok. Here we have shown that if I have a random variable, which is Gaussian with mean 0 and variance sigma square that is sigma Gaussian.

But, if you take arbitrary random variable it need not be it need not satisfies this condition. For example, if you take x to be exponential distributed with parameter μ, what is the bound we are going to get? Just compute it now. So, just take X is exponential lambda. So, what is this value is going to be?

So, how we are going to compute this? This is going so, sorry you take it as μ for some positive μ. Well, this is going to be e to the power lambda X e to the power μ x and there is also μ here dx 0 to infinity, right. What is this value is? So, this is going to be μ upon lambda minus μ into e to the power lambda minus μ into x between 0 to infinity.

So, if I am going to choose this lambda to be greater than μ, what is this quantity? If I choose lambda to be greater than μ, this is anyway positive quantity this whole term, but when I put the upper limit infinity this is going to become infinity ok. And, what happens when this lambda is less than μ? It becomes negative quantity and becomes 0 right like we will not be able to what for no value of whatever sigma you are going to say you will not be able to come up with such a characterization here because of this random variable is not going to be sub Gaussian here for whatever parameter. It is not going to be sigma sub Gaussian for whatever sigma you are going to take.

So, we will see bit more of this examples later and some in the assignments what kind of distribution satisfy sub Gaussianity and what does not. So, another based on this sub Gaussianity one can also defined something called heavy tailed and light tailed. This not important for us, but just I will make a remark. This is maybe this will be useful to you some in other cases random variable X is called heavy tailed if for all. So, otherwise it is called light tailed.

So, basically we want this quantity to be are log of to be upper bounded by some finite number here. So, that is why we are interested in distributions which are kind of light tailed ok. So, we will be focusing on light tail distribution fine. So, we have this at least we have shown that one Gaussian random variable that is Gaussian with 0 mean and variance sigma square. It has sigma sub Gaussian.

What is the support of this Gaussian random variable here? So, it is a entire are real line right minus infinity to plus infinity. So, it support is infinity, it is not the random variable need not be bounded to have a sub Gaussianity property; even unbounded random variable can have sub Gaussian properties. The only thing we required is it requires that it has to be somehow light tailed. It means that as you go along so, how does the Gaussian look something like inverted bell right, but after some time as we go further this probability becomes very small.

So, if you. So, most of the probabilities will be in this region, and if you go beyond this so, this probabilities will be small. So, that is why in essence that is light tailed like if you go look at the tail the probability content in this region is very small.

So, we will be now going to look into the sub Gaussian; that means, our random variable need not our distribution need not have finite support, it can have unbounded support.

Now, this is the first theorem. If X is sigma sub Gaussian then for all epsilon greater than r equals to 0 it is tail probability is upper bounded by minus epsilon square by 2 sigma square.

So, what we are interested in? We are interested in the random variable taking value beyond epsilon. So, here in this case suppose this is epsilon, we are interested in my random variable taking probability in this part ok. We are saying that that is upper bounded like this. Now, why this bound holds true? So, the simple proof follows from the Markov inequality.

So, suppose we are interested in probability that X is greater than or equals to epsilon right. So, this X need not be positive random variable. It can be a positive and negative random variable, but what I will do we have already done this when we proved Chebyshev inequality we will I will explain both sides by X right. This should still hold, the inequality still hold and the probability holds with equality.

Now, e to the power X is? Is a positive valued random variable right; if X is any random variable, but e to the power X is a positive value random variable. Now, I am asking this positive value random variable being greater than e to the power epsilon. So, it see this is a positive random valued random variable, I know Markov inequality can I mean I can apply Markov inequality here right. So, what is that gives me?

So, before applying any Markov inequality here what I will do is I will exponent both sides again by some number lambda positive ok. So, this inequality still holds right, both sides I am exponenting by lambda. So, if I do that now my Markov inequality tells that e to the power lambda x divided by e to the power lambda epsilon. Is it this or whole square of this that is just this right, it is just this.

Now, I invoke the assumption that this X is sigma sub Gaussian. So, if this random variable X is sigma sub Gaussian I know the this is upper bounded by this quantity here. So, which I and I will take also it in the numerator are this will give me this quantity, right and this is true for any lambda positive. So, now because this is true for any lambda positive, and this exponent here this is like a convex function in lambda. What I will now look for is the smallest value of this lambda or smallest value of this exponent. So, can you just optimize it over lambda and tell me what is the value of lambda that minimizes this exponent?

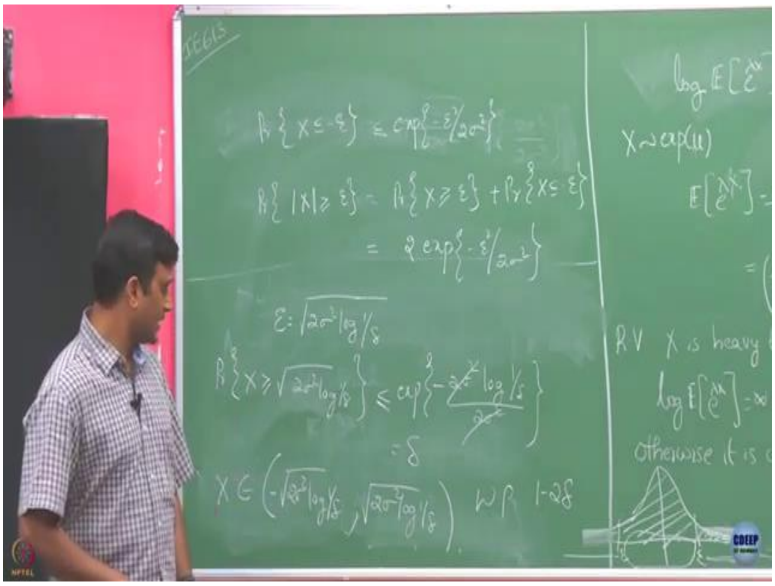
So, it has to be what 2 epsilon sigma square or these 2 will be there simply epsilon by sigma square right. Now, you just plug in that value here, that optimal value. So, if you plug in that what you will get? So, what are you getting? So, if I just plug in this is epsilon square, but this is only sigma square here. So, if I do this, this is going to give me what? minus epsilon square to sigma square and it is exactly what we wanted.

X is positive, why where did I use that X is positive?

Student: (Refer Time: 19:37).

This one? Why? No, it is whatever this value whatever this value it is there just both side you are exponenting. So, this is for X greater than or equals to epsilon. You can also have a similar bound which says that X is less than or equals to epsilon ok. So, instead of you asking epsilon is being greater than here, you ask this epsilon is less than or equals to minus epsilon ok, instead of that then you will also get the same thing.

(Refer Slide Time: 20:19)


IE615
$\Pr\{X \le -\epsilon\} \le \exp\{-\epsilon^2/2\sigma^2\}$
$\Pr\{|X| \ge \epsilon\} = \Pr\{X \ge \epsilon\} + \Pr\{X \le \epsilon\}$
$= 2\exp\{-\epsilon^2/2\sigma^2\}$
$\epsilon = \sqrt{2\sigma^2 \log 1/\delta}$
$\Pr\{X \ge \sqrt{2\sigma^2 \log 1/\delta}\} \le \exp\{-\frac{2\sigma^2 \log 1/\delta}{2\sigma^2}\}$
$= \delta$
$X \in (-\sqrt{2\sigma^2 \log 1/\delta}, \sqrt{2\sigma^2 \log 1/\delta})$ WP $1-2\delta$
$\log E[e^{\lambda X}$
$X \sim \exp(u)$
$E[e^{\lambda X}] =$
$=$
RV X is heavy
$\log E[e^{\lambda X}] = \infty$
otherwise it is
CDEEP

So, probability that X is less than or equals to minus epsilon we will get. And, now, if you want to use what is the probability that your X; mod of X being greater than or equals to epsilon; that means, what I am asking here? I want it to be both can I write it like this probability that X is greater than or equals to sorry, this is mod of X epsilon greater than epsilon. So, then this is going to be greater than epsilon and plus probability that X is less than or equals to minus epsilon. So, if you just plug this quantity we are going to get 2 times exponential of minus epsilon by 2 square.

Now, this bound here I am going to choose a specific value of epsilon; let us say let me choose this epsilon to be some quantity 2 sigma square log 1 by delta. So, I am now interested in knowing what is the probability that my X is going to be larger than this epsilon? So, if you are plug in that value what is the probability that X is going to be greater than or equals to 2 sigma square log 1 by delta.

Now, just substitute this value of sigma there what you are going to get exponential of plus 2 lambda square log 1 by delta this whole divided by 2 sigma square and this simply becomes log delta and this is going to be delta. So, what it says is, the probability that if you want this X to be larger than this quantity that is not going to be more than delta ok.

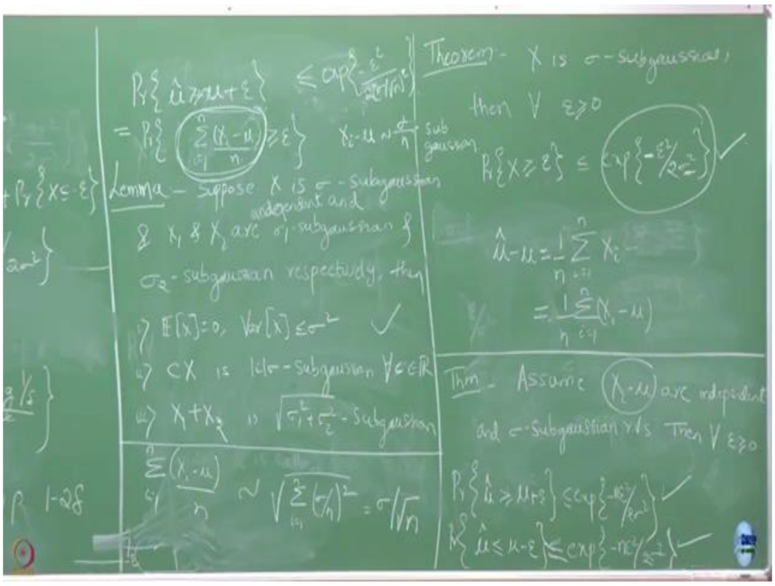
So, is it then in that case can I say that for any given delta if I want my X to be in this interval one is. So, if I want my X to be in this interval right so, what I am basically asking? I know what is the probability that my X being away from this epsilon and what is the probability that X is below this quantity here that is minus epsilon. So, I have set epsilon to be like this and when epsilon is set like this I know that my random variable going beyond this quantity is going to be upper bounded by lambda.

Now, I am asking what is the probability that your random variable lies in this interval? So, that is just the going to be the compliment of this right. What is that? This is going to happen with probability 1 minus 2 delta ok. So, with probability delta it is can be away from this with probability another delta it can be below this and if you remove this then this is the probability that it is going to be in this region.

So, what that is telling us is, if you are interested to know that your random variable x is going to take value in this interval, then that is going to be given with probability 1 minus delta. Suppose, let us say your delta is very small, so that if delta is very small you want; that means, basically you want with very high probability you want in this interval. So, if delta becomes very small, what we expect this these two terms to be like? They are going to be large right. If delta is small, then 1 minus delta is going to be large and because of that this these points move away from each other fine.

Now, this is just one thing. What I was finally, interested in what happens to my estimators like when I have n samples and if I average them, how far that will be from the true mean right that is what of my interest.

(Refer Slide Time: 25:46)



$Pr\{\hat{\mu} \geq \mu + \varepsilon\} \leq \exp\{-\frac{\varepsilon^2}{2(\sigma/\sqrt{n})^2}\}$
$= Pr\{\sum_{i=1}^{n} \frac{(X_i - \mu)}{n} \geq \varepsilon\}$
$X_i - \mu \sim \frac{\sigma}{n}$ sub gaussian
Lemma - Suppose X is σ - subgaussian
& X_1 & X_2 are independent and σ_1-subgaussian & σ_2-subgaussian respectively, then
i) $\mathbb{E}[X] = 0$, $Var[X] \leq \sigma^2$ ✓
ii) CX is |c|σ - subgaussian ∀ c ∈ ℝ
iii) $X_1 + X_2$ is $\sqrt{\sigma_1^2 + \sigma_2^2}$ - Subgaussian
$\sum_{i=1}^{n} \frac{(X_i - \mu)}{n} \sim \sqrt{\sum_{i=1}^{n} (\sigma/n)^2} = \sigma/\sqrt{n}$
Theorem - X is σ - subgaussian, then ∀ ε ≥ 0
$Pr\{X \geq \varepsilon\} \leq \exp\{-\varepsilon^2/2\sigma^2\}$ ✓
$\hat{\mu} - \mu = \frac{1}{n}\sum_{i=1}^{n} X_i - \mu$
$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)$
Thm - Assume $(X_i - \mu)$ are independent and σ - Subgaussian r.v.s. Then ∀ ε ≥ 0
$Pr\{\hat{\mu} \geq \mu + \varepsilon\} \leq \exp\{-n\varepsilon^2/2\sigma^2\}$ ✓
$Pr\{\hat{\mu} \leq \mu - \varepsilon\} \leq \exp\{-n\varepsilon^2/2\sigma^2\}$ ✓
$+ Pr\{X \leq -\varepsilon\}$
1-2δ

Now, let us see how. So, now, finally, I am interested in asking the question if my random variables are Gaussian, what is this probability is going to look like and what is μ hat here? μ hat is this is the average of n IID samples ok. So, now before we conclude what is the bound we are going to get using this result, we need some properties of sub Gaussian random variable which I am going to list now.

This is called a lemma. Suppose, X is sigma sub Gaussian and X_1 and X_2 are sigma 1 sub Gaussian and sigma 2 sub Gaussian respectively then the following properties holds. 1) is if X is a sub Gaussian it must be the case that it is mean value is going to be 0. This is the property of a sub Gaussian random variable and its variance will be less than or equals to sigma square that sub Gaussianity parameter sigma.

And, 2) if you multiply your random variable by some constant C, it could be positive or negative they are saying that is going to be mod C of sigma sub Gaussian, for all sigma belonging to R C belonging to R. So, if you multiply a sub Gaussian random variable by a constant, it is going to be a new sub Gaussian random variable with a parameter mod of C times sigma. So, suppose if this C is a constant, then the CX will be simply C sigma sub Gaussian. So, a sub Gaussianity parameter got scaled by C.

The 3) one is $X_1 + X_2$. So, we are saying that X_1 is sigma 1 sub Gaussian, X_2 is sigma 2 sub Gaussian, then if we are going to add $X_1 + X_2$ then it is going to be still sub Gaussian, but with the parameter sigma 1 square plus sigma 2 square ok. Now, let us see if we can I am going to leave the proof to you just this is application of a definitions, work it out yourselves.

Now, how can we get what we want using this result and this lemma ok. First of all notice that what is this μ hat minus μ? By definition it is nothing, but which I will write it as right. So what we will show try to show is before I write. Now, we are going to show that assume X_i - μ are independent and sigma sub Gaussian random variables, then for all epsilon greater than or equals to 0.

Our claim is that probability that μ hat greater than or equals to μ plus epsilon for bounded by exp and epsilon square by 2 sigma square and probability that μ hat less than or equals to μ minus epsilon is again bounded by the n epsilon by 2 sigma square ok. Now, why this is true? We are saying that if this X_i - μ are independent sigma sub Gaussian random variable then for any epsilon this is going to happen.

So, now, what we are interested in this guy here right what we are what we are saying is this μ hat minus μ this can be expressed like this. So, this is nothing, but. So, now, if you look into this, so, this X_i - μ is sigma sub Gaussian and we have addition of n random variables here. If you just look into this summation here now instead of that let us take this n also inside. If X - μ is sigma Gaussian, what is X - μ by n? Is it the sub Gaussian? With what parameter? It is going to be what parameter?

It is going to apply this. I know if our random variable is multiplied by C, then this is going to be still a sub Gaussian with mod C term sigma right. So, here X - μ sigma Gaussian say X - μ divided by n is 1 minus n sigma Gaussian right 1 minus n sigma sub Gaussian. So, is what is it? This is 1 by n sigma sub Gaussian and now we are saying we are now taking sum of n random variables right. Each one of them is sigma by n sub Gaussian. Now, what is this entire thing will be? It is sub Gaussian with what parameter? With the sigma. Why is that?

Student: (Refer Time: 34:14).

Root sigma, root of now we have to apply this result, right?

So, now we have to; now we have to say that this i equals to 1 to n, X_i - μ by n. This is going to be what? Now, this is summation each of them is sigma by n. So, this is going to be sigma by n whole square and we have to add it for how many times? n times. So, what will this will give you? sigma by root n ok. So, this whole thing here is going to be sigma by square root n sub Gaussian.

Now, treat this as entire one random variable, now asking that this random variable is gain greater than epsilon. If this is a sum sub sub Gaussian value, we being greater than epsilon then just apply this. Replace whatever the sub Gaussianity parameter by the sub Gaussianity parameter of this random variable. What is the sub Gaussianity parameter of parameter of this sum here? We have just showed that that is sigma by square root n right. So, to get a bound on this all you need to do is replace the sigma by square root of n.

So, if you do that that is what I got this exact bound. I just replace sigma by square root n. So, if you got that you get an extra factor n in the numerator. So, that is how we get. So, this is done. This guy here is nothing but; now this guy is nothing but by applying this exp minus epsilon square by 2 by 2 by square root n whole square and you get exactly what you want.

So, notice that again, now, we got this bound here finally, which is decaying exponentially in n, right. If you increase n this bound is also decaying exponentially in n. So, finally, we and this bounds are not for only enlarge, this is true for any n right. So, now, we are have a bound which exponentially decreases in n and that holds for any n. Like when we have applied central limit theorem we got also exponential decay, but that we will hold for only large n, but here when I use the sub Gaussianity property, I get this bound which holds true for any n ok.

Student: (Refer Time: 37:26).

I did not make any such assumption right. Let me see if I need that assumption. Yeah, they are independent. Sub Gaussian are sub Gaussian are independent and so, we need to be they need to be independent ok. Just before that I just take one more minute. So, here maybe.

(Refer Slide Time: 38:30)

When we had this Chebyshev inequality right, when we applied that what is the bound we got? We got that μ hat being greater than μ by epsilon. This is the Chebyshev inequality I was talk I am talking about we got this to be equal upper bounded by what? Some sigma square by n epsilon square right when we applied Chebyshev inequality, but whereas, when I apply this sub use this expert the sub Gaussianity, I got this quantity to be right. This is what I got in epsilon square 2n square.

Now, of course, this is tighter right this is going to be this value is going to be tighter because this is decaying exponentially and where as this is decaying inversely in n. Now, actually this bound here we can show that. So, we know that exponential of e can be written as upper bounded by 1 upon e to the power x for all x greater than or equals to 0. x if x is greater than 0 e to the power x is 1 divided by e times x.

So, if we apply this property on this treating this quantity as x, what we are going to get? Upper bounded by 1 upon e times 2 sigma square divided by epsilon square by n. So, you see that finally, what we have gotten is. So, and also 2 by e is; is the ratio 2 by e is less than 1 or greater than 1? It is going to be less than 1 right because e is 2 points have. So, finally, we are going to get sigma square by epsilon square n. You see that what are the bound we got this is of course, this is going to be much tighter than this. Ok.