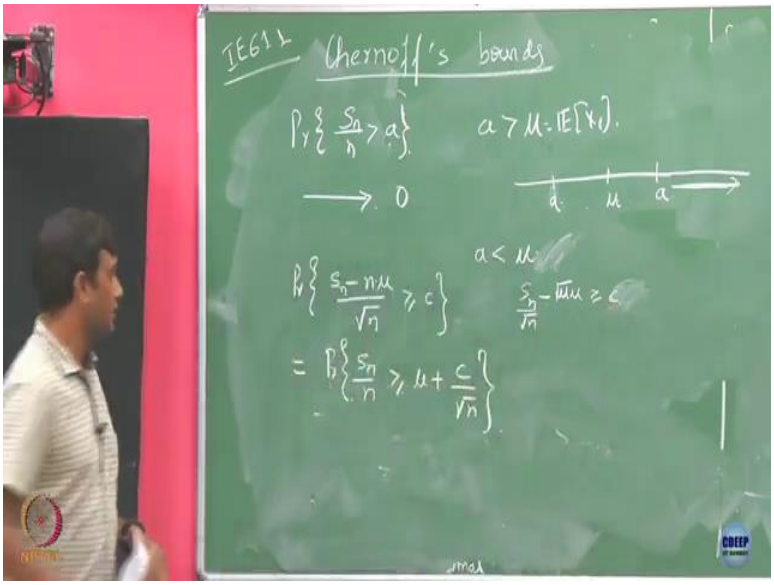
Introduction to Stochastic processes
Professor Manjesh Hanawal
Industrial Engineering & Operations Research
Indian Institute of Technology, Bombay
Lecture No. 28
Chernoff Bound

So, next thing we are going to study is something called Chernoff Bounds. We will just do a quick pass on that. We will not go much deeper into that.

(Refer Slide Time: 0:39)



Suppose, I am interested in I am interested in asking the question. Let us say I am henceforth I am assuming IID sequence. If I am not assuming that I will make it a special mention about this. If I am not making a mention by default I am assuming my sequence is IID. And this is my Sn is the sum of my n random variables and I am looking at their average.

I want to ask the question, what is the probability of that Sn by n is greater than a for larger. Now, what I know about Sn by n from my law of large numbers. So, Sn by n converges to mean, in probability and also almost (())(1:57) and this a let us say that this a is such that, this a is greater than mu. What is mu? Mu is my expectation. So, if my a is greater than mu. What do you think this should happen as n goes to infinity?

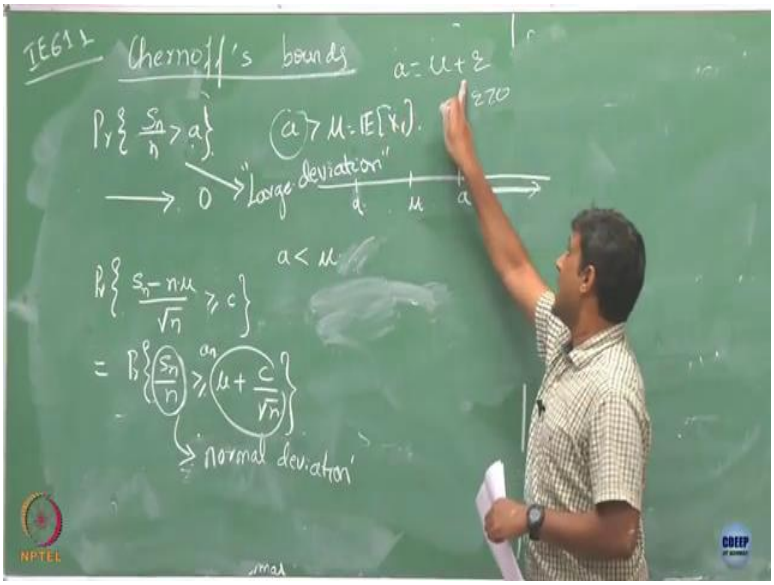
So, as n stands for infinity I know that Sn by n goes to mu. Now, what I am asking, it to take a value a, which is larger than mu that is not going to happen. Because Sn by itself is going to value mu so it cannot take, this is going to 0. And if I take a to be less than mu, what happens? Again 0, it is going to 1. Because if this a is going to be less than mu. I know that mu comes after that. So, I am going to take the value mu which is already greater then a. So its probability should be 1.

So, I am looking at these two pictures. Here is mu, the first case I consider is a. So first I asked a, what is the probability of that. I take a value beyond this. I am not going to take any value in this. Because I will be always hitting here. Next, I am asking the question. Let us say a is here what is the probability that I take value greater than a, that is always going to be 1 right, because I am (always) always going to be take value mu which is always going to be greater than a.

Now, let us take the case. What does this goes to as n goes for infinity. What this can go to? So, what this will go to from CLT- Central Limit Theorem, Gaussian distribution. If any sufficiently large, let us say we are already approaching Gaussian distribution. Now, I am asking this Gaussian distribution being greater than equals to c. So, this is nothing but the complement of CDF of Gaussian distribution with 0 mean and variance sigma square, that will have some values.

So, let us, I will write to reorganize this. So, if you reorganize this what you are going to get is Sn by n. This should be greater than or equal to Mu plus c by square root n. Check this I am just reorganized this. Can you quickly verify this is indeed correct? Correct, so what I am going to do is, I will do this or do I draw it by square root.

(Refer Slide Time: 6:06)



Now, what I am asking what so, this is basically, this is the question basically central limit theorem is (ascering) answering as intense infinity. So, if I just put alternatively it is a. It is basically trying to ask the question what is the probability of Sn by n is going to be greater than equals to mu plus c by square root n fro sum c. And we know that this Sn by n as n goes to infinity convert this to this value mean.

Now, we are asking it to be away from mean by this quantity which is itself diminishing to 0. As n goes to 0 infinity where this quantity is going to mu. So often, now this type of distribution where we are trying to answer the question where Sn by n is trying to be away from its mean value by sum of this quantity this is we usually call normal deviations.

Because we just know that Sn is anyway going to converge to mu. So we are just asking this question to be Sn being just away from mu by some small neighborhood. Whereas questions of this form where a is arbitrary. All we are saying is some a which is greater than mu. It could be far away from a, from below or above wherever it is. But here this a, so you can treat this as an. Most of the time as n sufficiently large, this an is very close to mu.

So, this central limit theorem is basically trying to answer what is this normal deviation. And this Sn by n such questions usually these are referred to as large deviation and basically this large deviations and normal deviations, they try to answer the question. I know that this guy goes to 0

as n goes to infinity, but I want to know at what rate it goes to 0. So, basically I want to see, for a finite n how this quantity behaves. So, let us see whether we are able to get a bound on this. So, you understand the difference between this deviation and this deviation.

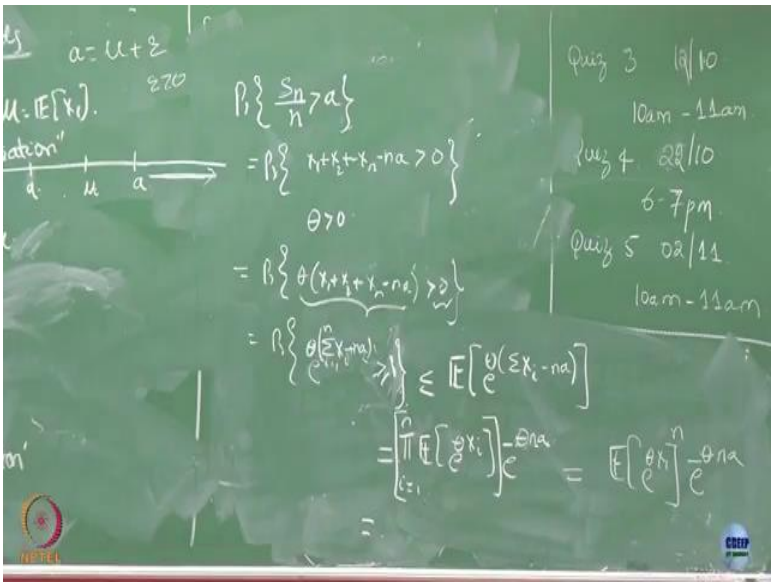
Now, we will come to Chernoff Bound. Chernoff bound gives the bound on such larger deviation. So this is, these bounds are very important when you are trying to do analysis of machine learning algorithms. So in machine learn algorithms, you we will often have large samples of data and from that you are trying to estimate some parameters. And now often those parameters will be related to the mean values of this random variable.

And now you want to see for let us say you took you have n samples. When you take the average of those n samples, Mu if n is sufficiently large, you understand that that will be the correct value of the parameter, the mean value. But now if n is not so you want to see like if you do not have enough samples enough data points that you will be being away from that mean value, what is that probability.

So let us say you want to, you have lot of data points from let us say weather prediction or something, the average temperature of a city or something and then you want to estimate what is the average temperature of the city. You let us say from past 10 or 20 years you have collected the samples. So, these are like daily recorded values. On this day this temperature this day this temperature this day this is the temperature and some let us say you have 10 -20 years samples, you take that value and then average it.

Then fine you know that it is not exactly if even take the average of this infinite number of samples it will not give me the exact value. But if at all, if it is away from that mean value and if it is going to be larger than by some margins. So, let us take this a to be some mu plus epsilon, where epsilon is some positive quantity and you want to say that will it be other from epsilon margin of true value. So, this will try to answer it with what probability this is going to happen. And you will see that this probability actually fall exponentially in the number of samples.

(Refer Slide Time: 11:54)



So, what we want to show is equals to a is bound to bound this, how to bound this. What we will do is? We will try to simplify this by taking X1 plus X2, all the way up to Xn minus na greater than 0. This are same, I just took an a there and replace Sn by their sum. This is what I will get. Now, let us take a parameter theta which is strictly positive and then multiply this both sides. So since this parameter theta is strictly positive. If I multiply it on both sides ordering will not change.

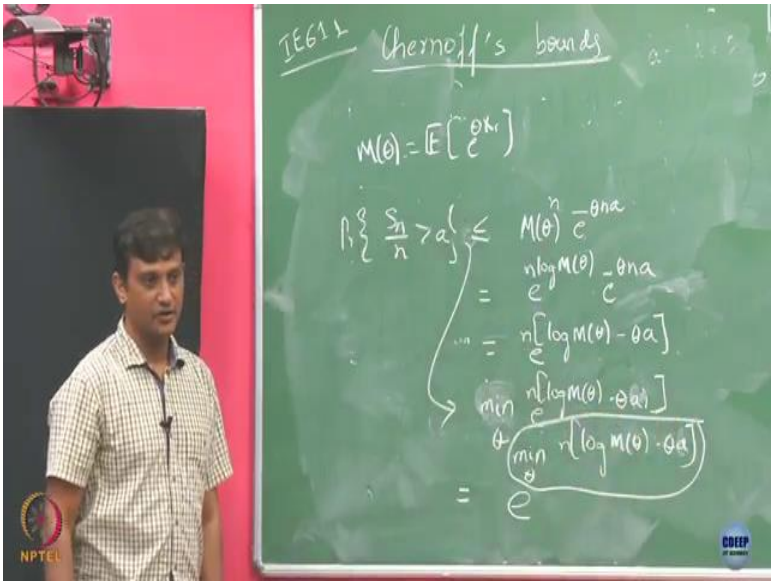
Now, let us exponentiate both this sides. If I exponentiate this, what I am going to get, minus n here. Now, because I have exponentiated. Why did I exponentiate here, any guess? So now, when you have this, what is the theorem, what is the result you want to apply? So, to apply Markov's Inequality what I want that random variable to be non-negative and 1, I have exponentiated 1 that becomes non-negative. So this the left side is non-negative random variable. I can treat this now as entire random variable and this is going to greater than equals to 1.

Now what is if I of divided by that is 1. I will just leave it like this. So now this quantity here, I can write it as expectation of e of theta i product i equals to 1 to n times e to the power minus na. Why I could write this as like this? Yeah, minus theta na. So, why I could write a product like this? These are independent random variables. And now, why I have multiplied theta because we want to later optimize it over theta. Right now, what theta, I said is theta is greater some greater

than 0. So what I am got is because I am at upper bound. And this upper bound holds for any theta. And now later as you will see. I want to tighten this upper bound.

How can I do that? Look for a theta which minimizes this upper bound. I have upper bound I want this value of upper bound to be as small as possible. If this is small then this is going to be a good upper bound for me. How can I do that? By minimizing it our theta. So, let us try to do that. So this can be further written as expectation of e of theta x one to be power n times e to the power theta n.

(Refer Slide Time: 16:59)



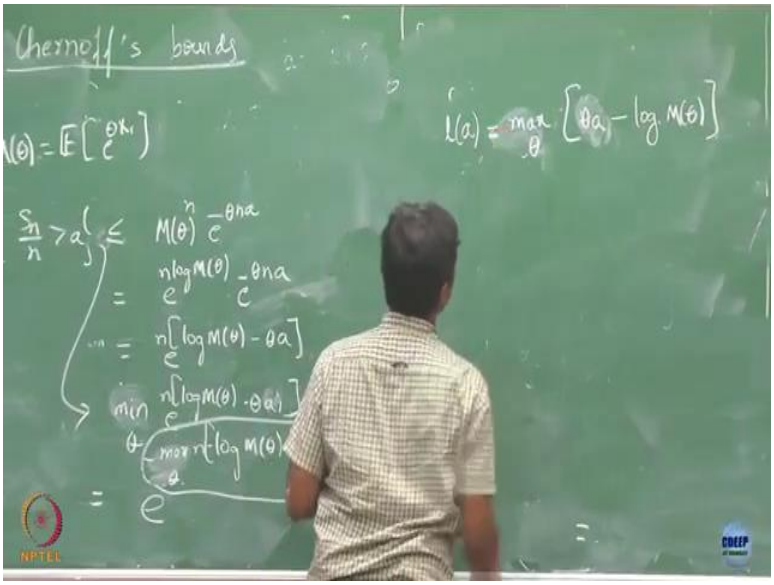
So often this here expectation e of theta 1. This is called moment generating function of X1. So, what is the difference between this moment generating function and characteristic function? There is no just i here. And often this is denoted as M of theta. So what we have now, finally putting all these things together. Probability that Sn by n is M of theta n times e to the power theta na, less than or equals to. Now, what I will do is. I am going to write it in a exponential format. This is going to be then n of log M of theta e to the power theta na. This is correct. This is log to the base e here.

Now, further this is going to be e times n log of M theta minus theta of a. So, now I have this quantity here which is a function of theta. Now, what I want do is? I want to and this is true for any theta which is greater than 0. I have just arbitrarily chosen theta which is strictly positive.

Now, I want to find a theta which makes this bound smallest. How can I do that? By minimizing this quantity. So and that is if I just do this if I so this is true for all the things. So here instead of this if I am going to write e to the power min over or I could still write min over theta of e to the power n log of M of theta minus theta a.

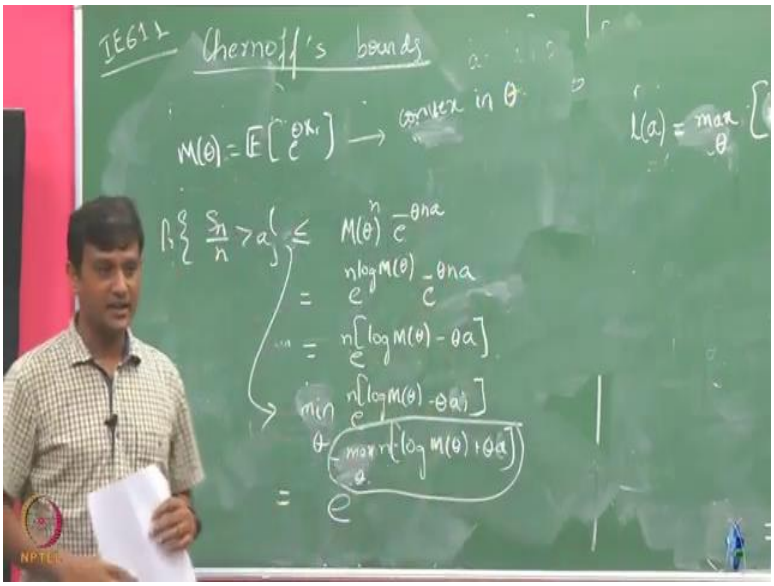
If I do this, you still think that this quantity is upper bound on this. It is still an upper bound on this. Because this is true for anyway for all theta. What I am just doing is look for a theta which makes this smallest. And is it true that this I can write it as e to the power, I can take the min upper, is this true? I just did this. And then optimizing, now my optimation problem just want to minimizing this quantity here. So, let us say so let us do this minimization.

(Refer Slide Time: 20:28)



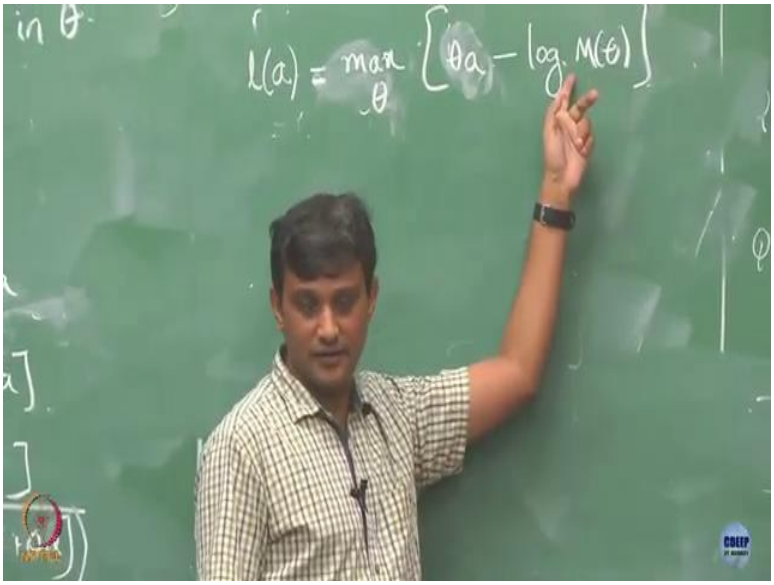
So, I can pull this n outside also because this n is independent of theta. Now what I will get is. I will also do one thing I just take minus inside outside this and just do. This so I am going to get theta of a minus log of M theta. But now if I am going to take this minus entirely outside. It becomes max over theta. I just took minus outside my min. Then I will have, that is correct. And often this is denoted as l of a.

(Refer Slide Time: 21:50)



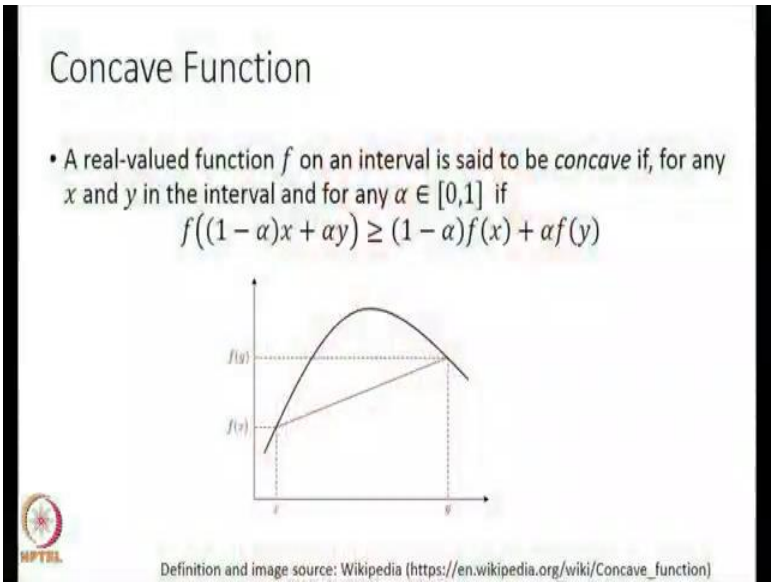
So, now let us understand this, this M of theta which I have defined here. This guide is going to be in general convex in theta. And just you can verify this like taking the double differentiation. You will see that as a function of theta this is going to be greater than or equals to 0 that is going to be convex function.

(Refer Slide Time: 22:25)



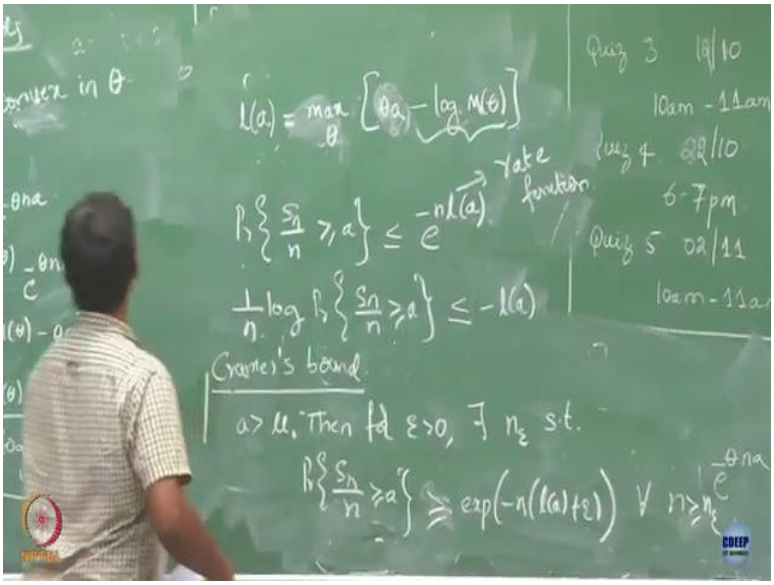
So, now what is negative of a convex function?

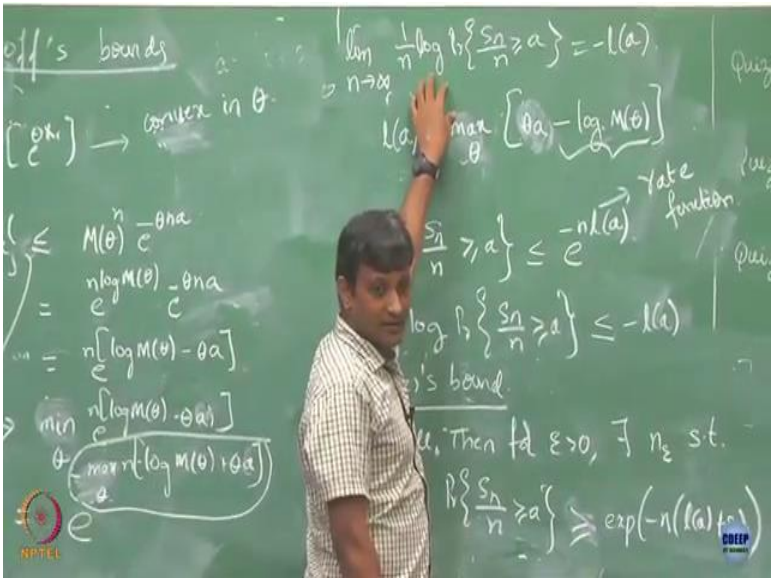
(Refer Slide Time: 22:30)



So you know what is concave function?

(Refer Slide Time: 22:50) 2





So, now and we have this theta. This is linear this is of, this is a linear function in theta and we will see that we are already doing optimization course. Some of took concave functions here. So, this is. So if you add a linear function to convex function, it will remain convex. And if you add a linear function to concave function, it will remain concave. So, I have a concave function. I added a linear function. This is a concave function. And what I am trying to do is maximize a (conva) concave function. So will it have a unique maxima? So, this optimization is well defined.

So, I will this is the convex function so I am try to maximize a convex function sorry concave function. And then if I have this. Finally, what I have ended up is showing probability that Sn by n is greater than or equals to a is upper bounded by e to the power minus n of l a, where l a is this quantity here. Defined as this and this is some constant right, whatever it is it is some constant which depends on the distribution of my random variable through this momentum rating function. But this is some constant.

Now this l a is constant and now what we have done is we have bounded this probability in this manner and how this is decaying. And this is decaying exponentially. Because as you increase n it is decaying as e to the power minus n. So, the deviation of my average from a, from the mean. If you take this a to be which is strictly greater then mu is going to be exponentially decaying and that is going to be of this format. And often this l of a is known as rate function.

Now finally, what does this say, if you just take log on both side and do normalization like this that probability that Sn by n greater than or equals to a. What is this quantity is or if I just take, I can take minus on both sides and this becomes greater than equals to l a. So what we have is basically an upper bound on this. This is called Chernoff bounds and it so that Chernoff bound is actually tight. You can end up showing a similar result for the lower bound and that is called Cramer's bound. I will just show it.

And the Cramer's bound says that so let us say that a is greater than mu. Then then for epsilon greater than 0. There exist n epsilon such that probability that Sn by n greater than or equals to 0. Upper bounded by is lower bounded by exponential minus l a plus epsilon. So, like this we had a but this holds for, notice that this result here the upper bound this is true for any n. I did not make any assumption on what is the value of n here.

But for the lower bound, which is given by Cramer bound is that the lower bound holds almost like this but there is an extra factor of n epsilon in exponent here. But this is true only for some large n for that is for all greater than equals to n epsilon this is true only after some point. Now if you are going to combine this two results. I have an upper bound here and I have lower bound here but if you let. And this is true for any epsilon. I can make this epsilon arbitrarily small.

And because of that what we can conclude finally is, if you are going to look like n by n log probability that Sn by n equal to this and look at that limit n equals to infinity. What this limit is going to be? So, this guy is going to be upper bounded by like this. And as n goes to infinity. This guy is lower bounded by again exponential minus n l a but some arbitrary epsilon here. So because of that you can conclude that this is nothing but minus of l.

So the rate, so what is this? Rate is exponential growth rate. Normalize exponential growth rate of this deviation is a constant, which is given by minus l of a. Normalize exponential rate of growths. So, this is what this is, this is exponential this is the exponential rate of growth but 1 by we are taking because we are normalizing. So only then we will be left with minus l a. So this is the (expo) normalize exponential growth right and that l of a here.

So, this kinds of results will also be useful to, means these results are very-very basic very fundamental and they have very basic result and people how build much-much advance results

based on these ideas which are heavily used in areas like machine learning. If you want to understand deep into machine learning about how to analyze algorithms, how your algorithm performance on a given data set and all.

You need results of this flavor or much-much advanced versions of this result. So, this is just a given flavor like how based on our basic random variables and basic random processes how can we derive such results. Let us stop here and next class onwards we are going to start Markov Chains.