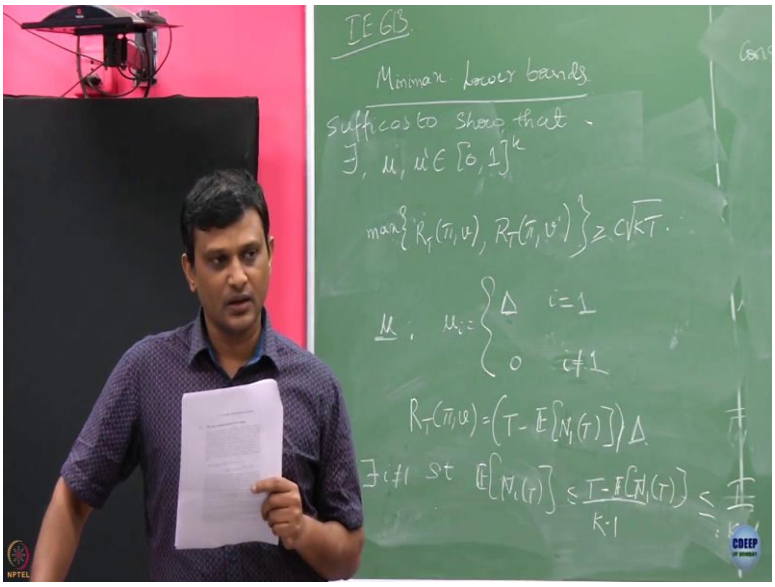
Bandit Algorithm (Online Machine Learning)
Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

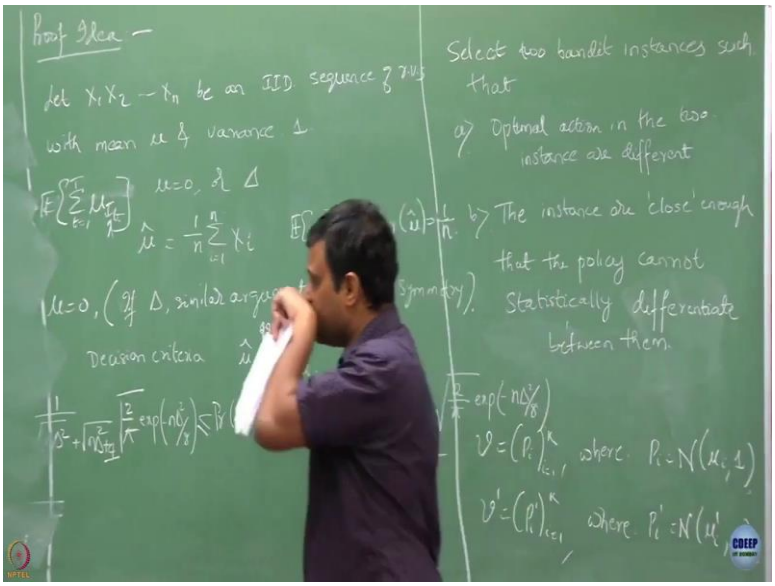
Lecture - 39
Proof Idea of Lower Bounds - 2

(Refer Slide Time: 00:19)



So, now, we are going to show that we show that there exist mu and mu prime; Suffices to show that. So, if I can show that there exist parameters mu and mu prime; such that, max. So, if I can show that on this instance either mu or mu prime.

(Refer Slide Time: 01:18)



So, when I say nu; this is collection of this Gaussian distribution and nu prime is collection of these distribution. If I can be able to choose this Gaussian distribution with this parameter certain parameters, so that the max of this regrets happens to be larger than C square root K T, then it kind of implies my regret lower bound right.

So, now, let us start looking into what are these parameters. So, for mu, I am going to set such that this mu i is going to be delta for i equals to 1 and it is going to be 0 otherwise. Yeah. So, this example we are running through Gaussian case.

Student: (Refer Time: 02:09).

Yeah, but we also said in the statement that it is sigma, sigma Gaussian it should be fine.

Student: Ok fine.

Ok

Student: Yeah.

Yeah, if it is it is fine for sigma sub Gaussian, it all automatically implied for bounded one right.

So, I am looking for one bandit instance now, where the first arm has mean delta.

Student: (Refer Time: 02:35).

And other are all have 0 mean ok. So, how I have chosen this delta? We have not yet specified which is. For time being I just assume that the delta is something which is positive, strictly positive and other arms has 0 mean.

Now, for this bandit instance clearly arm one is the optimal arm ok. Now, what is the regret is going to be? Its regret is going to be. So, number of rounds in that you just pull whatever the number of times you have played the optimal arms.

So, this is the regret you are going to get for this instance. Is that correct? So, let us say whatever is, $N_1(T)$ denotes what? Number of place of the arm one.

So, these are the one which are these many in expectations these many rounds, do not give any regret. Other than this place, all the place T minus these many rounds. They are going to you are you are going to incur regret. And each one of them is going to add a regret of delta. So, this is the total regret if you have this problem instance.

Now, we have total in number of rounds to be capital T let us say. Then the claim is there should be at least one arm which would have been played less than T by K number of rounds. Is that correct?

Student: Yeah.

Yeah no, by what principle?

Student: Pigeonholes.

Pigeonholes principle. So, we are just going to apply that. So, one of the arms which are like not the optimal arm, there should be one arm which is going to be at least played less than T by K number of rounds, there exists i such that expected number of. So, if this is not true the total number of place is going to be greater than T.

Now, based on this arm. So, whichever is that i not equals to 1 we are going to. Now.

Student: (Refer Time: 05:42).

What?

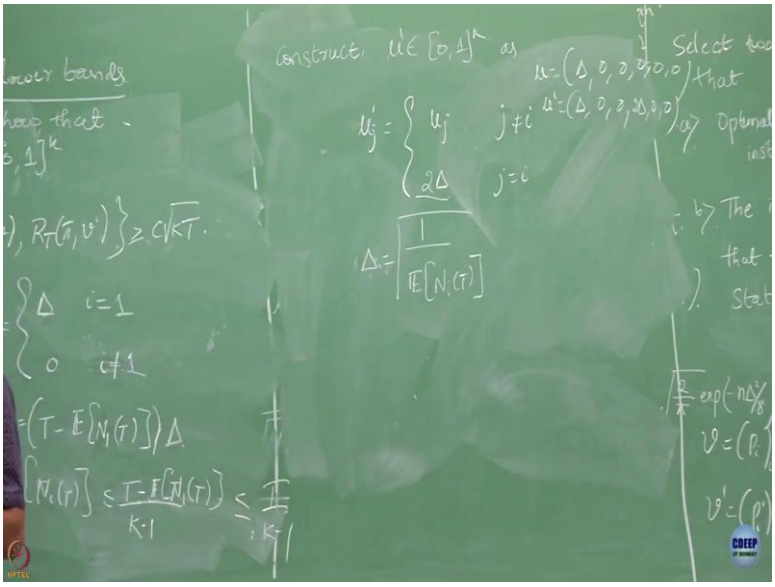
Student: (Refer Time: 05:45).

Yeah ok. So, let us say that is right. I have excluded i equals to 1. If i I have included everything, this is fine ok.

So, if I say; this should be correct right. So, just forget the optimal arm. So, this is the remaining number of place you have. And from that you are going to play and there are from the remaining K minus 1 arms.

So, I am just going to be upper bound it further and just take it K minus 1. Earlier one was correct, but it was loose fine ok.

(Refer Slide Time: 07:01)



Now, we are going to now construct the second bandit instance as following. We are going to say now, construct mu prime as. So, what I am going to do is mu j prime.

So, the new set of parameters I am going to define like, mu j for j not equals to i, and then I am going to define 2delta for j equals to i. So, what (Refer Time: 08:00) by just saying is, so there should be at least 1 arm i not equals to 1 for which this upper bound expected number of pulls has to be T K minus 1.

Let us say for that arm, I am going to make now the mean reward to be 2 delta, others I am keeping the same. Now, what has changed? Earlier the mu, my mu look like delta 0 0 0 0 0 right. And now, my mu prime looks like this delta 0 0. Let us say this i th one is the

this one whatever I got. Let me call it as 2 delta like this. So, what is the optimal arm in the first bandit instance

Student: (Refer Time: 09:00).

and what is that in the second bandit instance?

Student: i th arm.

i th arm right. It has.

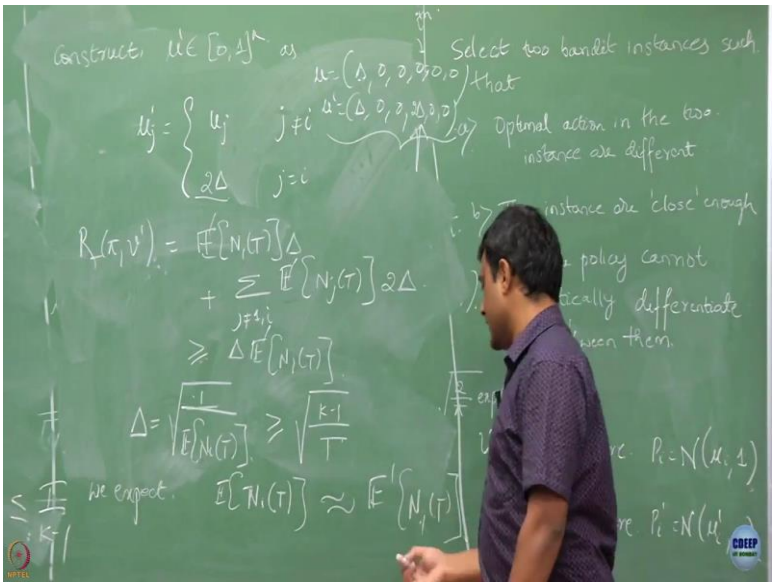
Now, the question is, the two instances differ only in the i th arm. The two optimal arms in the two scenarios are different. Now, the question is, can I set my delta in such a way that even if my bandit instance is happens to be 2, but my algorithm thinks it is still bandit instance 1 and ends up selecting my arm 1 most of the time. If I can do this then I have made my algorithm make err right. I will make it err most number of the times. So, now, let us see what will be a good choice of delta here. So, here I have little bit too much hand waving. Now, this is just like we will make the thing bit more formal.

So, right now, assume that uh. So, just you notice that in the upper and lower bounds they have a factor N delta square right, N into delta square. Suppose I set my delta N, so I had this delta such that. So, before I write this let me.

So, when I wrote this expectation here? So, this expectation is induced by the interaction between my policy and the environment ok. That is going to induce the environment here is defined by nu.

So, the interaction between nu and my policy resulted in that induced this expectation here. Now, for this bandit instance.

(Refer Slide Time: 11:47)



What is the expected regret? Expected regret suppose, if I continue to apply the same policy on this also. First arm is no more optimal, but that is going to cost me a regret of delta factor right, because it is off by delta.

The optimal is 2 delta arm 1 is delta. So, the regret by the place of arm 1 is going to be this much. And then, there is no regret from the i th arm in this bandit instance, because that is the optimal. And other than arm 1 and arm j, everybody is going to cost me a regret.

And what is that amount? j not equals to 1 and i expected value of $N_j(T)$ and 2 delta. And this quantity has to be at least delta times expectation of $N_1(T)$. So, I have just written the expression for regret, that I am going to incur on the bandit instant.

Now, suppose, if I set this delta to be let us say, 1 upon expected number of $N_i(T)$, what is this expected number of place? So, and here what is E here this is I am going to put E prime here, because this expectation is induced by interaction my policy pi with environment nu prime. Whereas this expectation was induced by interaction of my policy pi with environment nu.

Now, suppose whatever be the expected number of pulls I had with respect to my original or initial bandit instance right. Let us say, this was the number of expected number of pulls. I know that this quantity has to be upper bounded by T upon K minus 1.

So, this quantity let us say. So, what I have basically made is suppose, if you ignore the expectation here, just do some hand waving and just take this to be simply the number of pulls of i th arm. Then, what I am basically doing is delta square times number of pulls of i th arm. I am just setting it to be 1, that is I am making the factor N delta square big enough that is close to 1 ok.

Now, if I can do this by choosing it in such a fashion. Now, you can again go and see that on even this instance, I can come up with a similar upper and lower bounds. I had earlier on the sample mean of a Gaussian random variable. And you can see that my algorithm will fail to identify this as the optimal arm with this. And it may still end up choosing on this instance this as. And it will confuse this instance as this instance and think this is the arm one is the optimal arm most of the times.

So, that again will bit make it more formal, but that is the idea like if we can formally argue that. By the choice of this delta even on this instance, it is likely that my algorithm will still continue to think that it is this instance and place arm 1 most of the times. That means it is making a wrong choice right in terms of the optimal arm. And even if that is the case.

In that case by this choice, let us say then we expect, because my algorithm got confuse between these two instance. The number of times it is it would have pulled is going to be close as expect number of times that would have pulled under bandit with the second instance. So, in this case it is clear that, let us assume that the algorithm is making eventually I figuring out arm 1 as the best arm. Now, with the appropriate choice of this delta there is will a possibility, that even in this instance this still continue to make a bad yeah.

Student: (Refer Time: 17:45) it mean the (Refer Time: 17:47).

No no no.

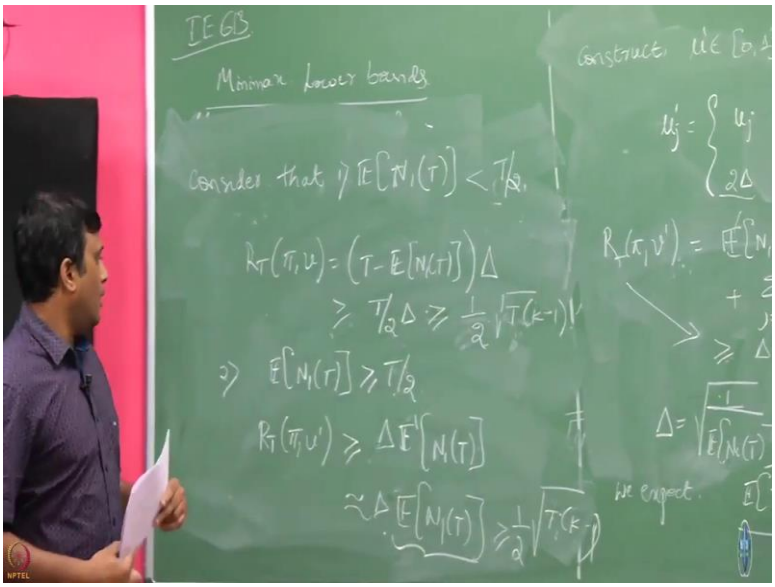
Student: Again.

Again it is like I just restarted and gave you a new bandit instance.

Once I start algorithm I am not going to change the bandit instance ok. In this case, because even in this instance my algorithm got confused thinking it as this instance then, it is going to spend most of the time on the first arm.

We are going to expect the expected number of pulls on bandit. The expected number of pulls that I am going to see with the second instance is almost going to be same as that I am going to have it in the first bandit instance.

(Refer Slide Time: 18:33)



Now, I think we are more or less done ok. Now, first we are going to consider now consider a case. Expected number of pulls is going to be N by 2 T by 2 if this is the case. So, earlier we had demonstrated that R T of pi nu is what is equals to T plus expected number of pulls of.

Now, in if this is the case then this is nothing, but what?

Student: 2.

This is T by 2 and delta. And what is my delta we have set it as? K minus 1 by 2. So, this is like some yeah K minus 1 root K T by sorry T times K minus 1 right ok. Now, consider the second case where has happened to be (Refer Time: 20:07) 2.

Then we have also argued that this guy R_T time nu prime. We have had a lower bound here this lower bound here; we had shown this is nothing but delta expected value of

$N_1(T)$, but then using this approximation here. This is nothing but expected value of $N_1(T)$. And this will give us and this will give us again the same thing right. Again it is going to be half of. So, this quantity this one I am going to replace it by T by 2 and delta from that, this is going to be K T times K minus 1.

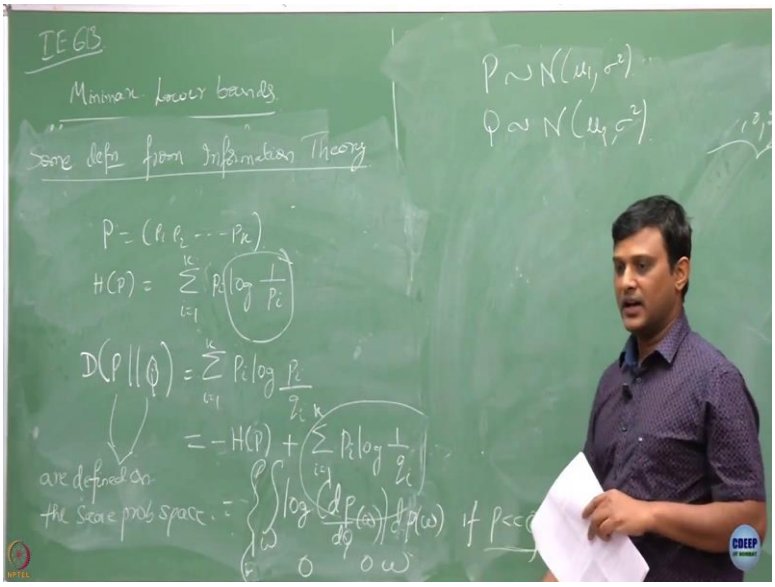
So, for both the cases we have this lower bound. And now, we have our. Now, if we are going to take max over these two instances. Now, we have that the regret is going to be order square root T K minus 1 and there is an half factor here. So, this is just like very heuristic top level arguments we have.

So, what is happening is. We are able to come up with when we are just arguing that if at all we have this instance, where my algorithm is going to confuse it with other instance; then, it is possible that my regret is going to be order square root T K on these two instances ok.

Now, we have to make this bit more formal ok. So, for this we need some information theoretic quantities and some bounds based on them. I am just going to introduce them in today's class and in the next class we will try to go through the steps formally.

So, how many of you know already entropy? You know or no? Half? You know half ok.

(Refer Slide Time: 22:54)



Now, we are just going to some definitions. So, suppose let us say if we have a probability distribution P on K alphabets. We are going to define entropy of this distribution as expected value of Pi log 1 by Pi.

So, those who know this entropy, can they tell what is the meaning what is the operational meaning of entropy? I mean we know that the amount of information contained is inversely proportional to its probability right like.

Suppose, if a event is going to happen with probability one. Is there any information contained in that you already know that is going to happen right. If a event is very less likely then may be like information contained in its more right. Because if something rare happens like that is the big news right.

So, in that way the amount of information contained is like inversely proportional to this probability and also. And that way; I mean you can also give a interpretation that, if something is more frequent and if you want to assign, let us say some codes to it. You are going to assign larger code or shorter codes the things which are going to appear occur very frequently.

Student Shorter codes.

You want to assign shorter right, because it is appearing more and more times. The one which are going to appear less frequently you are going to. You may have you will be forced to assign larger one, because you need to distinguish right.

If all the small length codes are taken by the quantities which are happening frequently then, what will remains is the larger one you are gone. In a way, that if you want to like encode something you want to encode something which is more frequently with less code lengths and the one which are happening rarely with larger code lengths.

So, in that way if you are going to think this log 1 P as the length to code a message. Let us say then this is going to give you basically the expected length of the code. And information tells that this is the minimum average code length you need, if you want to record the message correctly ok.

So, fine this is for a given entropy. Of course, like I am assuming that all these quantities P i's have positive mass, like if some P i is 0, it is 1 by P i is not well defined right. So, I

am assuming that all the alphabets I have here positive mass. Then, there is a quantity called divergence. This also we defined in the last class.

Let us now, here I am assuming that both P and Q are distribution on the same probability space ok. Let us say P and Q are defined on the same probability space, then divergence between these two quantities defined as

Student: (Refer Time: 27:05).

Yeah.

Student: Sigma p_i Log p_i by q i.

Log p_i by q_i . Can I write it in terms of entropy? So, in that case it is like if I have to write it in terms of entropy, it is like minus

Student: H P.

H P then.

Student: Log H P q i.

Log.

Student: (Refer Time: 27:43).

In a way.

So, if you are just going to give the same interpretation that we just gave here. So, if you are feeling that let us say some messages are generated according to probability distribution p_i but you misinterpret that and you assume that generated according to distribution q_i right. The true generation is p_i the messages. But you are you somehow got confused and think that they are generated according to another distribution q_i. So, then the length you are going to assign is log 1 by q_i and but the true one happens to be p_i right.

So, this is still the expected length you are going to code those symbols. And this is anyway minus the best you could have done right. This is you are not doing that good

right. So, this is the best you could have done, but this is not the best you could have done. What do you think, this difference has to be positive or negative?

Student: It has to be positive.

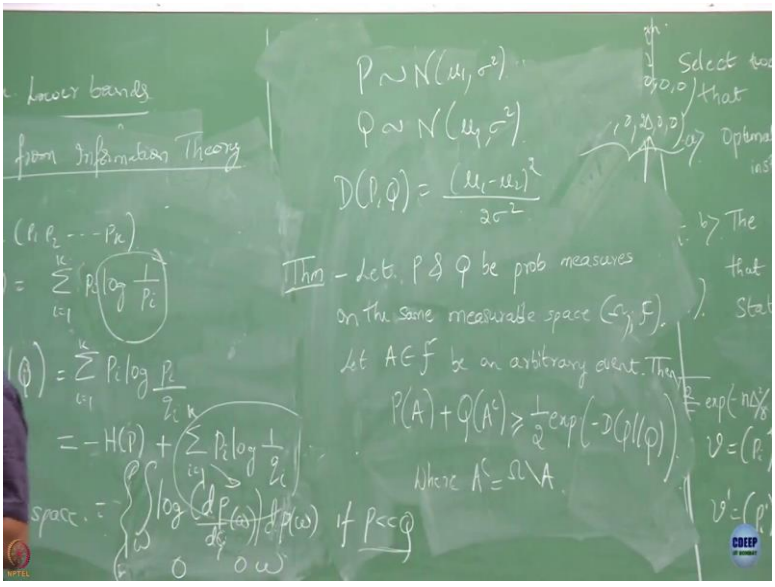
It has to be positive right, because this quantity is going to be larger than this.

So, this quantity divergence, we already discussed when we discussed the proof of KL UCB that using this for at least Bernoulli case, you can get a very tight upper bounds ok. And now, in general this is this divergence is have other names like Kullback - Leibler divergence, K L divergence in short. It has some nice properties.

It kinds of measures the distance between two distributions, even though it is not a true metric though. So, it is not a true metric, because it does not satisfy the transitive property ok.

Let us see now, I am now just going to state one result and then we will stop, which is going to be come handy to us.

(Refer Slide Time: 30:18)



When we are going to state our proof formally. By the way like, if P happens to be Gaussian normal with Gaussian with mean mu 1 and sigma square and Q happens to be another Gaussian with parameter mu 2 sigma square.

So, notice that the way I have defined here, this is assuming that these are discrete

Student: Probabilities.

Probabilities right, like there are these probabilities defined on discrete random variables. But this random this distribution could be on a continuous random variable, in which case I have to appropriately define this ok.

So, that definition is has something bit involved it depends on the Radon-Nikodyn derivative, but just let me write it here. This is defined as log of d P d Q if 0 otherwise.

So, it is defined as. This should be d P here sorry. So, this is defined as d P by d Q of omega d of P omega here.

Student: (Refer Time: 31:49).

This integration. So.

Student: Q.

This is right P is absolutely continuous with Q.

So, that means, I just want to make sure that this is well defined by making sure that this guy is in the denominator I do not end up with a 0's right. And whenever this guy and also like these are do not end up with 0 by 0 format in this. Yeah I just recall what do you mean by just absolute continuity.

Student: The lower (Refer Time: 32:29) P has to be wrong.

Whenever P is non-zero q is going to be non-zero ok.

Student: Sir (Refer Time: 32:40).

Yeah, so if it is q is p is 0 1 q is not zero, this is 0 by log 0 right log 0 is.

Student: Sir we have p i multiplied (Refer Time: 32:50) we will take 0 log 0 to be.

Yeah 0 log 0 is to be 0 fine. So, we want as long as q i to be positive in non-zero p i to be non-zero.

Student: (Refer Time: 33:05).

No no, there is a d this is ah what we have this Radon-Nikodyn derivative equal derivative this is d P by d Q into omega. Let us not going into that it is. Yeah

Student: (Refer Time: 33:19) cumulated (Refer Time: 33:21) density.

Yeah it is a density function density. Density,

Student: (Refer Time: 33:26) is the density function.

P and Q are the density functions, but this is something different ok. So, let us let us leave it like, this is another quantity we have to slightly interpret in a different way ok.

So, now if you have P and Q we are going to get it as D of P Q. You can just compute. This we only need this, that is why I am writing instead of working out all these. This is going to be mu 1 minus mu 2 whole square by 2 sigma square ok.

Next theorem. Let P and Q be probability measures on the same measurable space ok. Let A belongs to a compliment is omega (Refer Time: 35:05) A. So, this is we are going to not going to prove this. It says that if you have two measures or two.

Let us say, probability measures P and Q on the same space omega script f. Then, if you look at probably of A and probably of its compliment, but with respect to another measure Q, that is going to be lower bounded by exponential of minus divergence between these two measures ok.

So, what it is basically saying is. If you are interested in an event with respect to some measure P then that not happening with respect to another measure Q, that shows total sum is going to be lower bounded by like this. So, we will just going to use this result and later to prove that ok. So, let us stop here. So, we will continue it in the next class.