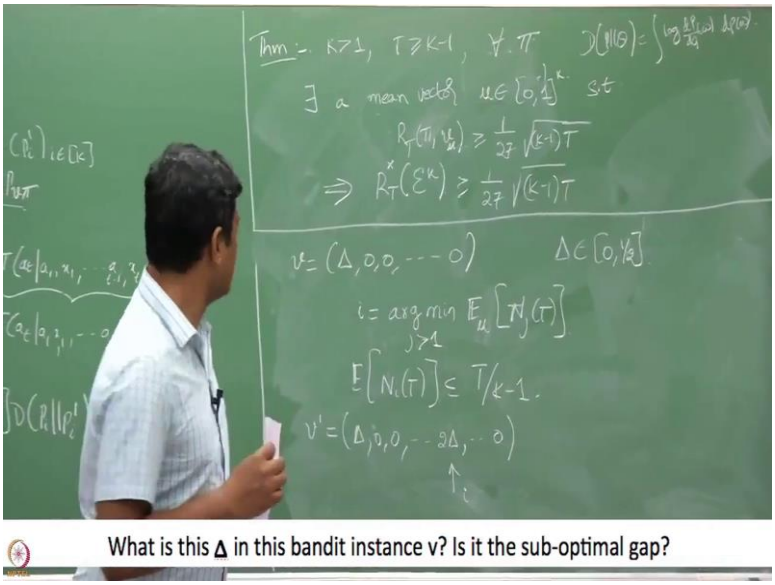
Bandit Algorithm (Online Machine Learning)
Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Lecture - 41
Proof of Lower Bound - 2

(Refer Slide Time: 00:21)



Like we discussed in the last class, let us have an environment nu which has this mean value. So, it has a mean value delta where I am going to assume that this delta is to 0 or maybe like I will assume it to be between zero and half and we are going to choose select it later. If you want to select delta later, but let us say I have one environment which has this one. And I have I already know that when I am going to run my policy on this bandit or this instance there exist one arm which would have played in expectation less than T upon K by 1 number of rounds. I know at least one such bandit exist right.

Now, let us call that as i to be the 1. So, let us call it i and I know that this i would have been played less than T times K by K minus 1 number of rounds ok. Now that is now for this I am going to change the mean rewards and I am going to construct a new bandit instance which is delta 0 0 and this is 2 delta to 0 and this is exactly at the ith part. Now, I know that.

Student: (Refer Time: 03:07) I is not equal to 1 for that delta will depends (Refer Time: 03:14).

So, in this I am assuming that first one is the optimal arm right and now I am looking for the arms which are other than the optimal arm.

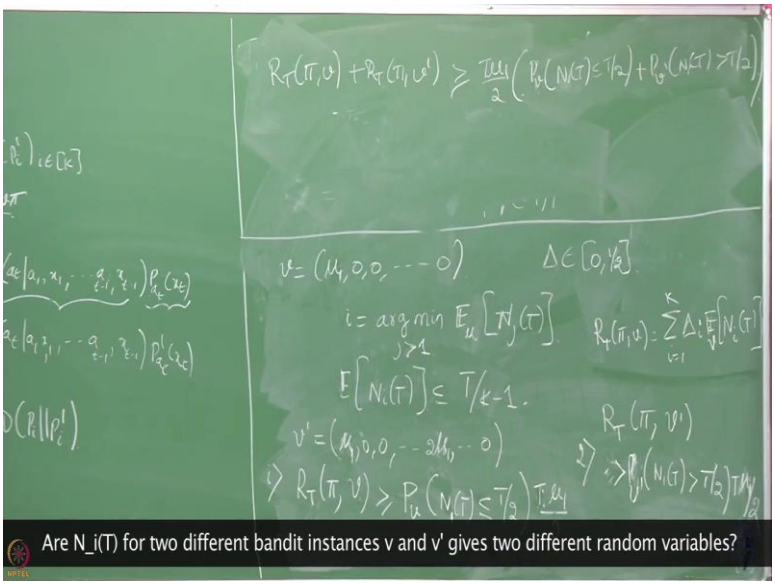
Student: But for optimal arm delta will be 0 right, that is the gap between nu star and.

So, do not confuse this delta with that. So, here I am telling this to be ok, maybe I should just tell this as mu, it is a mean value.

Student: Ok.

And this is mu and mu; again do not confuse again do not confuse. I initially said mu to be a vector. Maybe let us say let us say mu 1; mu 1 is the mean of the first one and then this is a mu 1, but this one I am going to make it as mu 1 then we are going to choose that mu 1 later ok.

(Refer Slide Time: 04:10)



Now, we already know that what is the regret. Regret of a policy pi on environment nu is nothing but delta I times expected number of pulls of that arm I, i= 1 to k and here this expectation is with respect to in whatever the distribution that is induced by this environment nu. Similarly, you can write it when what happens when you replace nu by nu prime, fine.

Now, I am going to write these two things. Let me tell whether these are correct. The regret I am going to incur from policy pi on this environment nu is going to be at least

probability under this environment nu times $N_1(T)$ less than or equals to T by 2 times n mu 1 by 2. Is this correct?

So, I am basically saying the number of pulls of the optimal arm has been less than or equals to T by 2, it has been played less than T by 2 rounds. So, if this is the event I am looking at then it must be the case that the suboptimals arm should have been played more than T by 2 rounds and whenever on those T by 2 rounds I should have incurred a regret of at least on each round I should have incurred a regret of mu 1 right, because whenever I am going to play a suboptimal arm, I am going to incur mu minus 0 regret in that round, is that clear?

And now, I am taking this probability. So, that event happening whenever this event happens I am going to get a regret of T mu 1 by 2 at least this much regret I am going to get ok.

So, I will have this one. Similarly, on this nu prime, so, here I am already assuming that $N_1(T)$ has happened less than played these many rounds in that case regret is going to be at least this much, but that event happening is this probability right.

Student: Right.

So, that is why I am taking yeah this is of course, this is an expectation and this quantity is going to be at least this much ok. So, just write down that is why I wrote this you have this just manipulate this you are going to get that.

And similarly this quantity being is going to be at least now I am going to take the event that my in the second environment I am looking for an event that I have played it I have played the first term more than T by 2 rounds. If I have played T by if I played the arm 1 in my second bandit more than 3 by 1 rounds, how much of regret I should have incurred? T mu.

Student: (Refer Time: 08:23).

T mu 1 by 2 right because in that case also regret is every round I am going to play a suboptimal if I play mu 1 the regret I am going to incur is 2 mu 1 minus mu 1 right?

Student: Yes.

And if I have played this guy itself more than T 1 T by 2 number of rounds I should have also incurred at least this much of T mu 1 by 2 also regret, but that probability is this is with nu prime $N_1(T)$ being greater than T by 2 into T delta by 2, is that clear? Now how these 2 lower bounds, yeah.

Student: (Refer Time: 09:19).

T mu 1 by 2 ok. So, now, what I have is I will rub this, now what I will just take the sum of these. 2 R_T pi nu plus R_T pi nu prime this should be greater than or equals to T mu 1 by 2 times these two probabilities.

Student: (Refer Time: 10:30) probably write this N1 dash because (Refer Time: 10:36). Policies?

Yeah, but policies the same, but I am already saying this is under the nu environment right.

Student: Two different.

They are the two different random variables.

Student: Yes.

So, what I am talking here is let us say in the first environment number of pulls is less than this, this is the number of pulls of the same arm happens to be greater than T by 2, this its distribution is now going to be governed by this nu prime, but I am still going to call that number of pulls of that arm to be just N by 1 N_1 (T) only right. N_1 (T) is just indicating number of pulls of that arm how this is distributed is going to be governed by the underlying environment. It is the same event like see number of arm pulls either in this environment or this environment, I am just the same thing the I am just counting the number of pulls of that arm that.

Student: But, values will be different.

Values will be different.

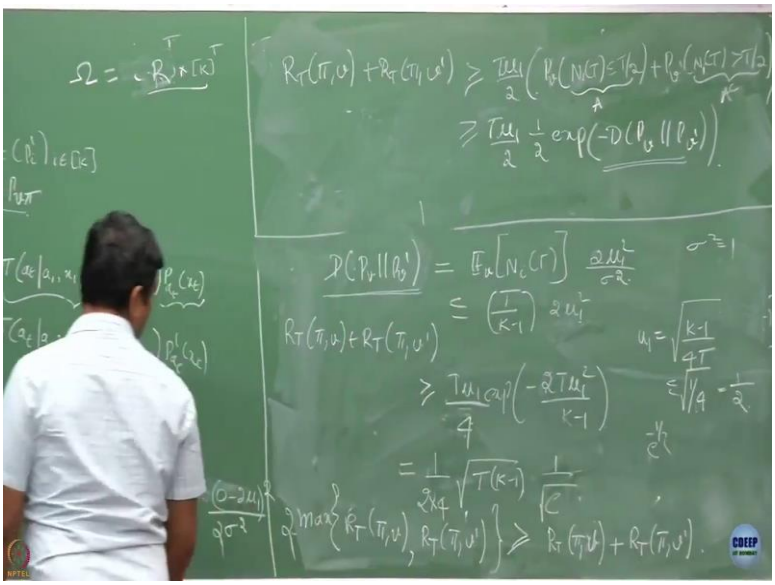
Student: And the probability distribution will (Refer Time: 12:05).

That will be yes, yes of course, that will be governed by this nu environment that distribution its distribution. It is how many times I am going to pull yes it is going to be having a different distribution. It is just that I am looking at the event whether this arm has been pulled more than this numbers or not, I am just looking at the event in this case I am looking at the event that whether that arm has been pulled less than T number of T by 2 number of rounds and here I am looking whether it has been pulled T by 2 more than T by 2 number of rounds, yes the distributions are different.

Student: So, the (Refer Time: 12:47) depends on nu.

Yeah, it depends on nu prime; it is with respect to this distribution.

(Refer Slide Time: 13:01)



Now, this is an event A, this is an event A compliment, even though I am looking at their probability in different environments. So, that I am looking at this A and A compliment in with respect to two different probability measures ok.

Now, this is like what? Can I use my this result? I have two events A and A compliment there. The event A corresponding to number of pulls of arm A being less than or equals to T by 2, that P there is governed by my environment nu Q is governed by my environment nu prime. So, then this is what? T mu 1 by 2 times half of exp now divergence between?

Student: Minus.

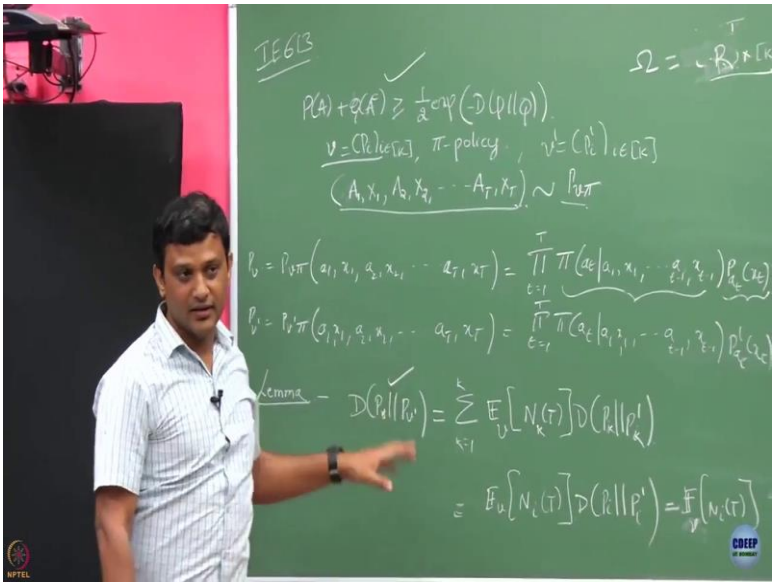
nu P divided by nu P prime.

Student: nu prime (Refer Time: 14:15) same (Refer Time: 14:18).

Yeah because the underlying sample space is the same for both right, you are looking at the same set of arms and the rewards are coming from the same region even though distributions are different.

Student: Yeah.

(Refer Slide Time: 14:41)



So, what is the underlying sample space? Underlying sample space is 0 to 1 let us say [0, 1]^ T * K^T times and then yes this rewards we have assumed to be Gaussian. So, let us say this could be real number.

Student: Ok.

So, this is the sample space right, this is my sample space over which both the distributions are defined. So, I am looking at T rewards and T pulls of my arms. So, this is what I mean like you understand this notation.

Student: Yeah.

Student: 1 2 3 4.

Yeah, 1 2 3 4 K but.

Student: (Refer Time: 15:50).

Cartesian product taken T times ok, because I said this is nothing but this right these samples are coming from this underlying sample space and both of them are defined on the same space. Now, we have this, now we want to invoke this guy here. So, we have invoke this guy we have let us invoke this guy.

Now, what is happening? The way we are constructed nu and nu prime they are same at every for each arm, but they only differ at ith location ok, only that i where they differ this divergence is going to be nonzero for all others it is going to be 0 right, because all other distributions are same, I have only changed at one for one arm ok.

So, then this is going to be nothing but what is that let us say. So, just for clarity I am just making this initially K. Now for that whatever ith one this is nothing but this value is simply going to be mu i of T time D (Pi||Pi'). So, this value is just one term remains everything is 0 here, because for all others the distributions are same and hence the divergence is going to be 0 for there. So, mean same does not mean they have the same distribution right ok.

So, let us state that. We are saying that we are keeping not disturbing the distribution of other arms except for the ith arm, which we are changing it and making it is to have a different mean which is 2 mu, rest of the distributions remain same and so their means also remains the same. That is where now we will be able to have the good expression for this divergence; otherwise in general we do not have a good expression for divergence right. It is for a Gaussian we have a good, I already told you in the last class right, what is the divergence between two Gaussians?

Student: (Refer Time: 18:28).

What was this value?

Student: mu 1 by root 2; 1 minus mu 2 whole square.

Yeah just a minute let me write this mu 1 minus.

Student: Mu 2 whole square.

So, let us put that values. So, mu 1 is this and mu 2 for us is.

Student: 0.

Not really right. So, I am going to compare distributions of ith component only. So, ith component here has mean what?

Student: 0.

0 and.

Student: 2 mu 2.

It is 2 mu. So, 0 minus 2 mu.

Student: By sigma square.

Divided by 2 sigma square or just sigma square.

Student: Sigma square.

Just sigma.

Student: 2 sigma.

What is the value?

Student: 2; 2 sigma square.

Yeah according to this it is should be 2 sigma square. So, right now we assume that the sigma square is fixed for all the arm variance is fixed only the means are changing ok. So, without loss of generality we can assume that sigma square is just 1 the variance is 1. What we know about this quantity for this i? We already know that under this environment nu this has been pulled no more than T divided by K minus 1 number of rounds. So, this divergence I have between two environments, it is nothing but this quantity I will just repeating this; no, but I know that this is nothing but T divided by K minus 1 times 2 mu 1 square sigma square ok.

So, let me I am going to assume sigma square to be 1 and just let me I am just going to ignore this denominator sigma square equals to 1. Now, I have an upper bound on

divergence. So, let us go back and plug in here, if I plug in the upper bound on divergence and this is with a minus sign I am going to get a further lower bound on this. So, I will have T pi nu R_T pi nu bar is equals to what? T mu 1 divided by 2 or 4 exp minus T 2 T mu 1 square divided by K minus 1, right.

Now, making use of both these results, I have been able to show that the sum of the regret under those two environments for the same policy has to be at least this much. Now, we will set mu 1. We are going to set mu 1 such that it is exactly equals to K minus 1 divided by 4 T. And by assumption in my theorem the hypothesis was that T is already larger than K minus 1, we have assumed that right; number of rounds is going to be at least that. So, K minus 1 by T is already less than 1. So, this quantity should be at least upper bounded by 1 by 4, this is like 1 by 2.

At least this quantity if I am going to set my mu 1 like that it is going to be at most half, no more than that. That is what I also want it right, initially my means should be in the interval 0, 1 and I have it here and I have also deliberately made that this mu 1 is going to be less than or equals to half, so that in my other environment mu prime, where I have 2 mu that is also between 0, 1. So, in both the environment all the mu's are between 0, 1 only ok.

Student: Sir.

Yeah.

Student: Sir, why did we require?

Mu 1.

Student: No, why do we require the means to be (Refer Time: 23:51)?

Because, I just want to explicitly give you an environment whose mu lies in the interval 0, 1 for which this holds.

Student: So, even if you are something else, let us say it is (Refer Time: 24:06) for the (Refer Time: 24:08).

Yes.

Student: (Refer Time: 24:09).

That should also be fine, but like to be consistent with whatever the theorem statement I have said right, where I will come up with a mean the environment, so, the. So, suppose like later you want to also make a similar the same claim for where the rewards are all bounded, bounded rewards. So, even though I am doing it for a Gaussian distribution here, but if you want to do it for rewards within some bounded interval and there let us say means are further bounded in some interval, you want to ensure that you are able to come up with an environment from that class where you any policy is going to suffer this regret. So, you have to like whatever the class you are going to claim you have to give me an environment from that class over which your policy is going to fail. So, that is why whatever the class I was looking at I am now trying to construct an environment from that class fine.

So, now, if I am going to set like this what I am going to get? I have set mu to be exactly like that. So, this is going to get me what, just simplify this. So, I am going to get mu is I am going to get like this square root T K minus 1 and what this will happen?

Student: Minus 1.

Minus 1.

Student: Half.

Half or like, I will simply write it as square root T right 1 upon e square.

Student: 1 by (Refer time: 26:10).

Sorry 1 by root T; what is that I have. So, yeah, now, what we have? So, now, you can just see this 27 whatever I have write in the theorem 1 upon 27 that is coming from 8 into square root e, what this will happen, turn out to be?

Student: This should be 16 into 12, because now there are this is sum of (Refer Time: 26:38) that one of the.

Yeah. So, this is just the sum of this we are noted there fine. So, why is the how is the final claim that my regret is going to be at least this. I know that the regret I am going to

incur is going to be the max of these two regrets and this max of these two regrets is nothing but is going to be at least.

Student: Half.

Half of this sum. So, let me write that. So, max of the regret I am going to incur is going to be larger than this. So, 2 times this is going to be larger than this sum, so.

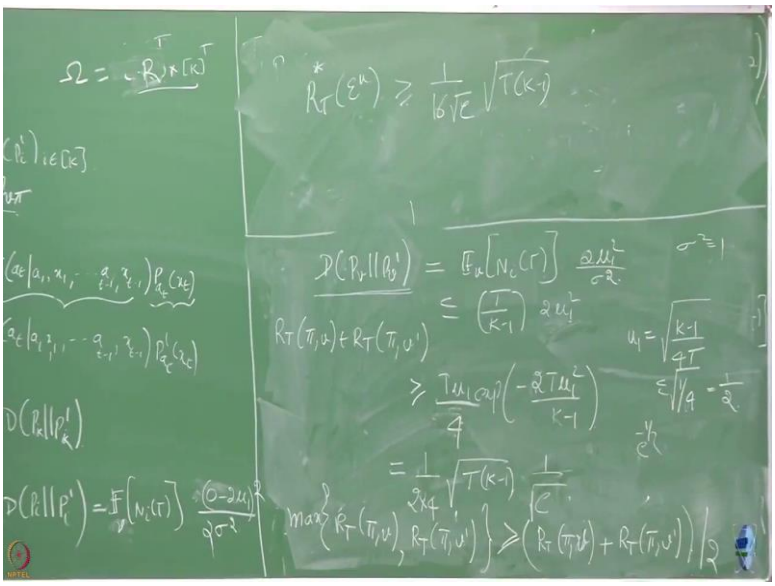
Student: Is this is the regret in R_T pi some mu.

Right, I am going to incurring R_T pi mu. So, when I am going to apply a policy pi right. My regret is R_T pi mu, now what I am saying this the max of these 2 regret is going to be larger than this either you are look at regret in this environment or this environment, it is going to be larger than the sum they are read by these two if I am going to take the two other side let us square.

Student: Possibly into any of these.

I could be facing any of this environment right. I that is not in my control, I should be I could be facing nu or I could be facing nu prime. So, one of this case, I am going to be doing worse and that I am going to guarantee that, that worse is going to be at least this much. You may be lucky that on one nu environment you might do be extremely well, but on nu prime you may take a hit, because nu, nu prime you may confuse it as nu and you are going to incur a large regret on that and it may be otherwise you may do extremely well at nu prime, but you may get confused when you are playing with on environment nu and do extremely bad.

(Refer Slide Time: 29:26)



So, my regret has to be in that case is going to be maximum of this. Now just do this. So, my whatever my R_T pi star whatever I have right like on that particular environment I have here, now it should be larger than this sum divided by 2 which is 1 upon 16 square root T times T K minus 1 and this 16 square root T is approximately 27 I guess.

So, yes, when we went through the proof fine like you have some nice results from how to relate to event and its complement under a different measures you are able to show this, but if I just tell you to prove this how to get all this intuition like ok, I said how I am going to do all these things right. I mean this is hard like in general, in general this lower bounds are very hard to come up with the instances and get an intuition that under what instances your algorithm can potentially get confused and think one for the other and make mistakes it is hard, but somebody has done it and he has and these results take a quite efforts.

And you see that it they need to use a nice tricks and so, the thing is like, but I think in general to prove the lower bound these are the tricks. There are I do not think there are you can there are many tricks available to prove these tricks, for some reason somebody got hold of this tricks and is able to do this, but I think if you understand all these things if you want to prove lower bounds in other case if you generally follow these tricks you should be able to prove, the lower bounds for that case also ok.

So, the main thing is this result and this divergence result ok. So, I mean it is not possible that like other things like algorithms we can think of ok, this is how I am going to do, but for proving a lower bound it has to be you have to make any algorithm fail right and you want to construct such an environment. So, and it may not be always easy, but that is fine like we just know some idea now what could be a potential tools or what could be potential tricks to use to get a feel of what the lower bound look like ok. So, let us stop here.