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

Lecture - 40
Proof of Lower Bound - 1

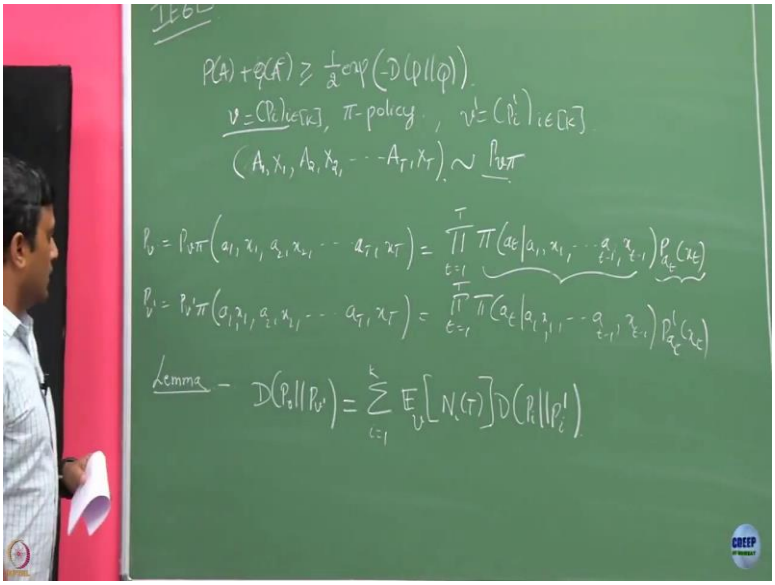
So, ok. So, let us start what we are discussing last time. So, last class, we just broadly discussed about the idea for Lower Bound Proof right. We said we stated that every for any stochastic k arm bandit where the rewards are either bounded or some sigma sub Gaussian. We said that the min max regret has to scale like some constant times square root tk. And last time we just discussed that idea. Let us now try to make that formal in today's class.

So, recall that at the end of the last class, I had told 1 result which said that for any measures P and this is greater than what?

Student: Psi (Refer Time: 01:30).

So, we are going to crucially exploit this result in proving this.

(Refer Slide Time: 01:10)



For the proof of the result, I need slight notation to make so, let me define those notations. So, there is an environment, there is a learner. So, let us fix a policy of a

learner let us call that pi. When you are going to interact with the environment, you are going to generate like you are going that is going to induce as distribution on the way actions and rewards are drawn.

So, what is that I mean? You remember if you play in round 1 let us say action A_1, you are going to observe the corresponding reward X_1, then you are going to play A_2, X_2 then that till round T A_T, X_T. X_1, X_2,.., X_T these are random samples you have observed by playing a corresponding arm and A_1, A_2 these are the arms you have pulled and maybe like this arm pull itself is random right; because this arm pull is going to be induced by what you have observed so far right.

So, now if we are going to look at the distribution of this quantity which consists of actions and this distribution on this will be induced by the environment and your corresponding policy right. So, let me nu is your policy sorry nu is your environment and pi is your policy. This sequence we are going to observe by interacting by the interaction between the learner and the environment this is going to be randomly distributed right and this distribution will be induced by this quantities nu and pi. Let me call that induced distribution to be nu pi is that ok? So, this quantity has to be randomly random which is has to be induced by nu and pi right whatever it is let us call this.

Now, we also know this right how to write this for example, nu pi of a particular realization let us call this a_1, x_1, a_2, x_2 all the way up to a capital T, x_T. How is this probability is going to be defined as? It is going to be product of and then what and then P of a_t of x_t this is the policy and this is the probability of observing that sample x_t when you are going to play arm a_t. So, this p a_t x_t is coming from your environment.

So, whatever policy you are going to use, maybe your policy is going to play action a_t based on the observation you have made so far time that is going to be random and then based and then, you are going to play a_t and then after you play action a_t, you are going to observe a corresponding sample a_t drawn from that arm a_t. So, it has introduced this. Right now, we have kind of assume that.

Student: Is that arm the mean of the arm will remain same.

The same throughout right; if you are going to every time you are going to play this action a_t, your sample is going to be coming from the same distribution. So, that is where that underline lying environment is there ok.

Now, we want to understand now similarly, let us say there is an another environment let me call that as nu prime that we will I will enter the same policy pi that nu environment will also induce an another distribution which I am going to call it as P nu prime pi. So, that is going to be what? Again that is a_1, x_1, a_2, x_2,.., a_T, x_T this is everything is going to remain only thing it changes is here pi x_t because my environment has changed.

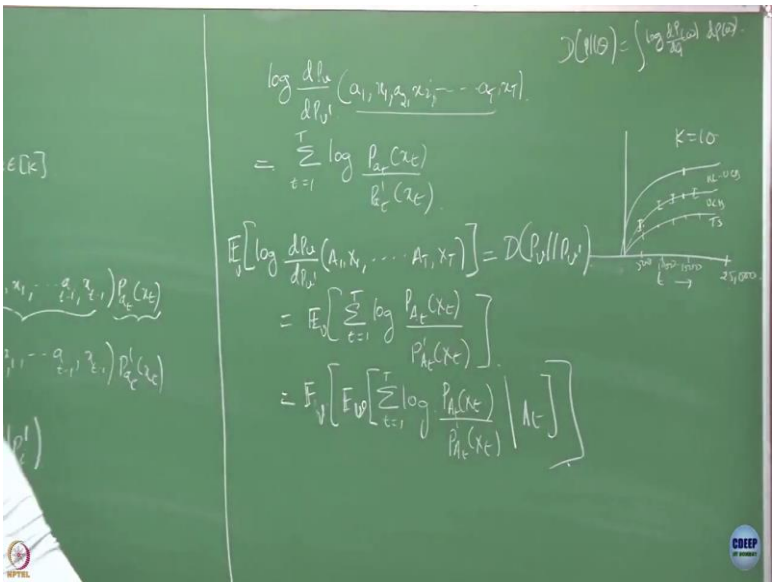
Our first claim is a lemma D P v. So, since I am going to fix a policy pi right, I am just going to and only my environment I have changed in this and this is my P prime. So, I have fixed a policy pi and now I am considering two environments one is nu and another is nu prime. So, I am just denoting by them and now I am going to consider divergence between these two quantities.

We are just going to argue that this divergence is going to be nothing but and nu prime here is another distribution which is P i prime and what is this expectation here? This expectation is taken with respect to the underlying environment mu whatever that distribution that environment mu indices this expectation is with respect to that environment.

Now, why this is true? See now when I am looking at this induce distribution P nu pi right here, this quantities a_1, a_2,.., a_T these are nothing but actions right. This actions are coming from one of the k actions. But this x_1, x_2,.., x_T these are the reward samples correspondingly coming from different arms, but there is are continuous valued ok. So, here I am looking on this distributions nu pi on this quantity where some components can be continuous right. So, this distribution is defined on a vector where some components can be continuous, and some components are discrete.

So, for this I have to define an appropriate divergence between them what is that? So, that divergence whatever it is we already defined for a continuous random variable right. How we have defined the divergence between two continuous distributions? In terms of I wrote it in the last class, if one distribution is absolutely continuous with the other, then we wrote it in terms of the integral.

(Refer Slide Time: 10:53)



So, now that integral has this term log d P v d P v'. So, this has meaning like what we mean by d P v this is I also said in the last class this is Radon-Nikodyn derivative. But let us not get into that we will just think it as like just a P by P prime I mean P nu divided by P nu prime and now calculated at the corresponding points. Let us compute this quantity.

And just imagine that these are nothing but the distribution value taken by P nu at this quantity divided by P nu hat taken at this quantity that is the meaning of this. And now if I just plug in these two quantities here, right this is nothing but d nu is nothing but d nu pi this quantity divided by d nu prime is nothing but these three quantities. If you take the log there and then simplify what you are going to simply get is t equals to 1 or capital T log P of a_t by x_t divided by P of a t prime divided by x_t. The factor corresponding to the policy cancel out; because I am using the same policy in both the bandit environment.

Now, this is the our one realization. If I look at the expected value of this quantity, now expectation with respect to what? My nu the underlying environment nu this is going to be nothing but expectation of this t equals to 1 to T log of P A_t X_t divided by P prime A_t X_t. But this quantity here this is nothing, but I have taking expectation with respect to the distribution that is induced by environment nu. This is exactly equals to this quantity in the numerator is exactly equals to the divergence between nu by P nu hat.

So, what we have defined the divergence between two quantities you recall? I defined it as and then, we have defined it has what d of P omega right. So, this integrate this expectation is nothing, but exactly that, but with these two distribution P nu and P nu prime. So, that is why this is divergence and this now a divergence is nothing, but this quantity expectation of summation of the logarithms of these two ratios.

Now, in this both X_t is the random quantity, A_t is the random quantity. Now what you could do is this in this case, now I am going to write this expectation or two parts the first one I am going to condition on the A_t and then, I am going to take the expectation over the other part.

So, what I will do is this expectation is this expectation of this quantity t equals to 1 to T given A_t can I do like this? This expectation I can write it in I can do it or two steps first, I conditioning on A_t and then, find it expectation and then, I will take the expectation of that quantity again. What is that?

Student: It is outside (Refer Time: 16:21).

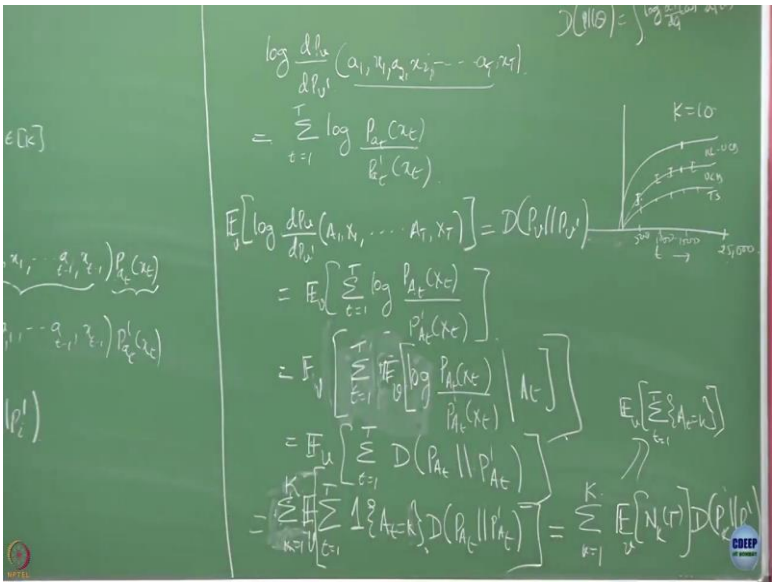
This is same, this is nu again nu whatever this nu is going to induce the distribution on this underlying environment.

Student: (Refer Time: 16:36).

Yeah?

Student: (Refer Time: 16:38).

(Refer Slide Time: 16:53)



It does not make much difference right like let us see this if I just to do so, this anyway there are only finitely many terms right? I can also write it as summation t equals to 1 to T expectation of nu over this log and then, you can write this as this expectation over here. So, in this case, we are only conditioning each terms.

Student: Ha.

But yeah this is fine. So, far every t you are conditioning that A_t part there. Earlier also whatever you did you could have just taken that inside that summation inside.

Now, what I am doing? Conditioned on this A_t, I am looking for expectation of this quantity. Once I condition A_t, what is the randomness there.

Student: X_t.

only X_t and what is X_t now? Conditioned on A_t, these X_t are coming from that particular arm right now what I am doing now what is the randomness here? It is only going to be corresponding the randomness is due to the corresponding armed right whatever that arm you are conditioned upon. Now, these are this is corresponding to the distribution of that arm under nu and this one is the corresponding distribution for the same arm under this nu arm nu prime and now you are doing the expectation with respect to nu that is with P distribution. What is this quantity is going to be? This quantity is going to be nothing but divergence between.

Student: (Refer Time: 18:54).

The distribution corresponding to arm A in the first environment and the distribution corresponding to another arm the same arm in the other environment right. So, this is nothing, but.

So, now, this distributions are completely absorbed the randomness in the samples X_t. Now this nu is only talking about how is this distribution over now it only remains so, this quantity are completely absorbed that distribution in the reward samples and now, this expectation is only over now the randomness in the arms pull A_t.

Now, another step. So, now, I am just going to simply rewrite this in this fashion. Let me know if it is correct ok. So, now, initially it was this is over sum going from all rounds. Now, I am splitting them in over different arms. Now first I am taking the summation over k equals to 1 to K and then and this summation I am only looking for those terms which when A_t equals to i and then, I am going to do it for each arm. So, then, this sum should be equals to the same as this sum right.

Student: A_t equals to k.

A_t equals to k yes. I am just looking at let us fix A_t to be one arm, then look at the divergence with respect to the distributions in the two distributions corresponding to the term and then do it over all arms and I did it by bringing in that indicator term there inside the summation over t equals 1 to infinity.

Now, we are done whatever we want to do right. Now, we have just now, k equals to 1 to K keep the outer summation just like that and if you just take the inner summation inside, now you are looking for a particular arm k.

Student: (Refer Time: 22:43) P k.

Yeah P k, but now then you are going to add the number of pulls of that arm right and what is that going to be? That going to be exactly number of pulls of that arm. So, this is now going to be expected pulls of N k T times divergence between P k and P k prime and that is what our claim was. That is what we have basically expressed through this lemma is the divergence between the induced the distributions in the two environment by

my policy can be expressed, decomposed into the corresponding divergence of the individual arms

Student: (Refer Time: 23:38).

distributions in this fashion yeah.

Student: Sir, so the last term how did you pull the divergence out of the.

So, now, we are if you have this indicator right k going from 1 to infinity. So, I have already said A_t is to be k. So, now, take that particular k; let us say that A_t equals to k so, if you just take now this is independent of t right, then what remains is summation of this indicators with an expectation term here.

Student: N i of T into divergence of.

Yes. So, this term basically I had just skipped this, this term came because there is an expectation of n when I put A_t equals to k this term was this.

Student: This term become independent.

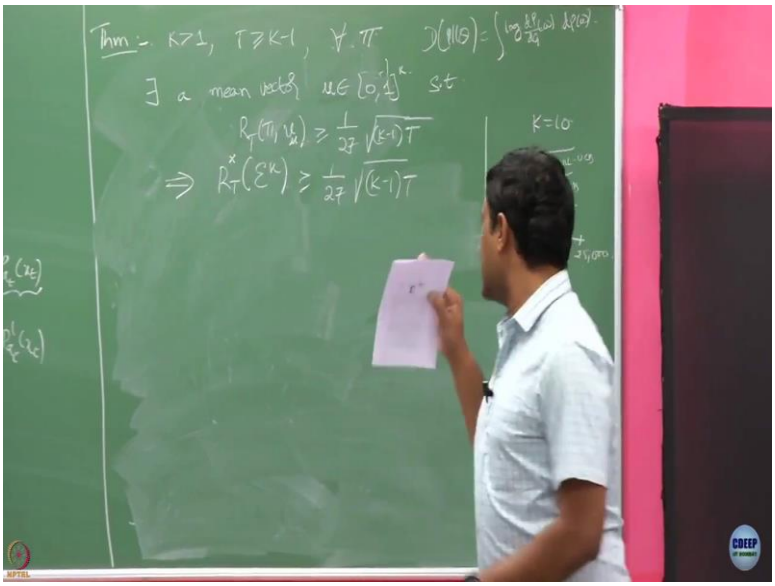
Of a time.

Student: Ha sir.

Now whatever this total number of rounds you have to going to play that particular arm over T ps here is nothing, but the expected number of pulls of the term under this environment nu fine. So, this is one result.

Now, the rest of it once we have established it now rest of the things is going to follow just exactly the same way we discussed in the previous class. We are going to construct two distribution P and P prime such that they differ at only one point, but still they have a different optimal arms and then, we are going to invoke our this results on this setup and we will be able to show that ok.

(Refer Slide Time: 25:40)



Just to recall what we wanted to show? We wanted to show that so, we assume let k equals to greater than 1 and number of rounds is T is greater than k minus 1, then we said there exists a policy pi no, we said that for all policy pi for all pi there exist what we are going to show is for all pi there exist a mean vector mu which is such that R_T pi, nu what is nu mu? This nu mu is equals to this, this is going to be 1 upon 27 square root k minus 1 times T.

What is nu mu here? The nu mu here nothing, but an environment whose distributions have this means coming from this mean vector mu. So, this mu is a vector with k components each quantity coming from 0, 1 which component coming from 0, 1 we are saying that there exist a such mu vector from which you can define an environment on which your regret is going to be at least this.

And now, once you have this, I have just demonstrated exist on such a policy right; that means, I can come up with a I have a class of environment over which your algorithm is going to make a mistake. I have so, what I have; what I am saying is through this theorem? I will be able to come up with one mean vector; that means, one environment for which for any policy we can its minimax regret is going to be at least this much;

Student: Yes.

That means, there exist a class of stochastic bandits right or environment such that on that this policy pi is going to incur at least this much of regret. We have just demonstrated like existence of one pi we are going to do that; that means, there exist a class such that this holds; on that class your algorithm is going to incur at least this much of regret or this implies basically what this implies is basically R_T star of epsilon k is going to be upon 27. So, thus nu k is set of all environments whose mean parameter are going to be in the interval 0, 1 ok.

So, now, earlier we have stated this result saying that this holds for any bounded environment where the rewards are taking in a particular interval or sub Gaussian any sigma sub Gaussian right. But now, the result we are going to show is what we are going to restrict ourselves to Gaussian distributions because we know that Gaussian distributions are also a sub Gaussian and for that we are going to show that as long as you can construct a Gaussian distribution whose mean vectors so, where all the arms have Gaussian distribution and the mean vectors are lying in the interval 0, 1. I wish if we can construct and we are going to show that such an environment with appropriately defined mu values is going to satisfy this result.