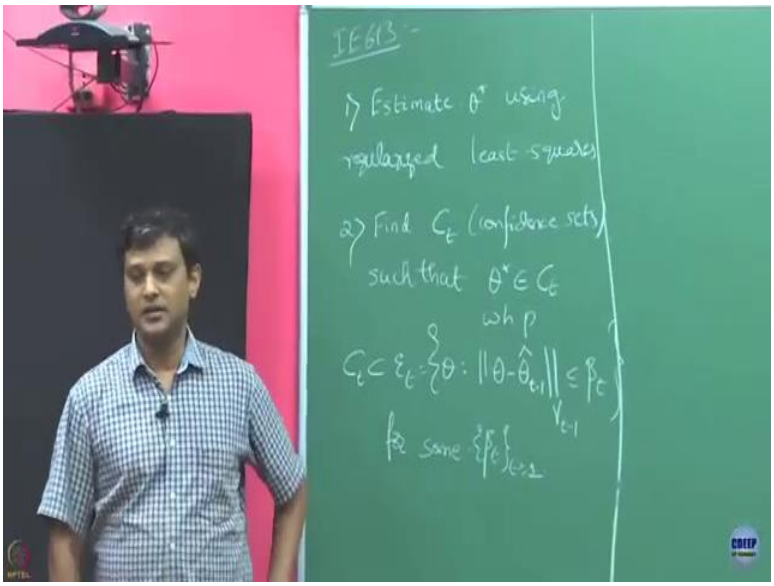
Bandit Algorithm (Online Machine Learning)
Professor Manjesh Hanawal
Industrial Engineering and Operation Research
Indian Institute of Technology, Bombay
Lecture 46
Regret Analysis of SLB - II

(Refer Slide Time: 00:30)



In the last class we have started like thinking about how to go ahead and solve this stochastic linear bandit. So, we just said, so we just said first thing we are going to do is first question is, how to estimate the parameter theta? So, what would say, how we are going to estimate it? What was the natural candidate for this estimation? How did we got an estimate for theta stars? We already discussed, we are going to just a regularized least square methods we are going to do this.

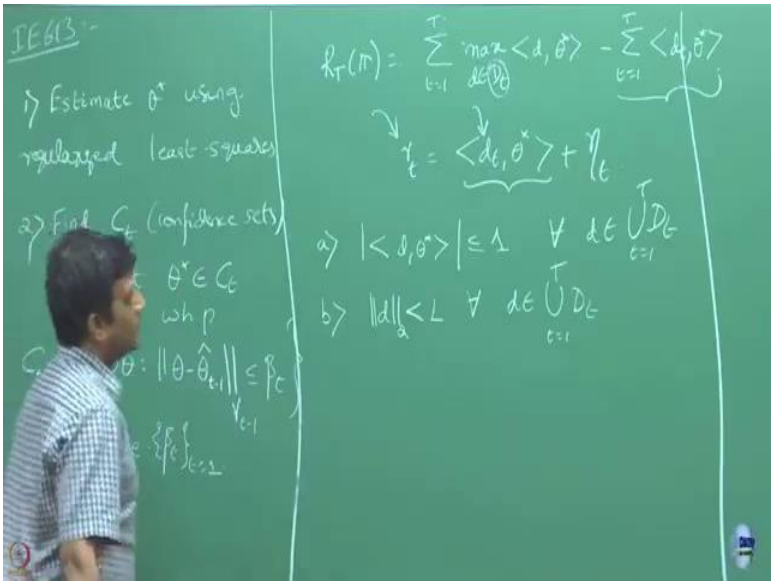
So, estimate regularized least squares and what is the second part we assume that we will be able to find confidence sets in every round so that my theta star belongs to that confidence set with high probability. It is, so now the first part is here we already know like given my observations how to find an estimate of theta. We will just do the regularized least square method and then the main question was here we said how to find this confidence interval so that my theta star belongs to this confidence set with high probabilities.

In the last class we said that we are going to write, let us assume my confidence set to be some subsets of the form such that so for some beta sequence. We said that suppose I will be able to

give subsets based on my estimate till round t minus 1 and if I construct such a balls let us say this ball includes my theta star with high probability.

So, I have not yet specified how this beta star are defined. Of course there will maybe depend on this theta itself, sorry delta itself with what confidence you want your theta star to belong to this in addition this beta t may also depend on all the observations you have made, all the arms you have paid, played and the corresponding rewards you have observed so far. So, how we will get this we will look into that later. But now let us say you are able to do this. Now, what is the method to find out how you go about finding regret of my stochastic linear bandit?

(Refer Slide Time: 04:07)



So, how did we define regret of my stochastic linear bandit, it is R_T of pi we said this as summation minus what are the (())(04:46) this (())(04:49) theta (())(04:54). So, this is what you are going to obtain in round T. So, if you are going to play arm d_t in round T this is the remain divide you would have got. Actually what you have observed in round T is we have denoted it as the reward you are going to, reward sample you are going to observe is this quantity plus eta t which we say conditionally Sub-Gaussian.

So, this is the noisy reward you are going to observe in round T and we have defined your regret to be like this and here we want to bound this quantity and also the expected value of this quantity. So, here it is still this d_t's the arm you are going to play in every round could be random.

Because the choice of this d_t depends on what you have observed so far and that depends on what was the sets d_t that has been selected so far. So, now how to go about this? Before we start proving that we are going to make couple of more assumptions about this setup. First thing we are going to assume is, we are going to assume that the reward, the mean reward, so this is the mean reward in round T. We know that this is a Sub-Gaussian so this is a noise with mean 0.

So, the mean reward you are going to get in round T is this we are going to say that this is going to be always less than or equals to 1 for all d coming from D_t t equal to 1 to T. What does this mean? In round T this is D_t is going to be a decision set. Consider all union of all this decision set it is saying that you play any element from this decision all possible decisions. The mean reward for that is always bounded by 1.

So, this is same as saying that earlier the mean value of my bandit of each arms is less than 1. So, the analogue version of the mean here is this quantity here. It is just we are saying that mean is going to which irrespective of which arm here there going to pay, here arms are nothing but that decision vectors, decision points, which are you are going to pay for that you are going to get the mean reward, which is strictly less than or equals to 1.

The second we are going to assume that elements d are all bounded with L_2 norm again for all d belongs to.

Student: At the maximum (())(08:30).

Professor: No, the maximum reward is this, so in round T in round T, d_t is your decision set.

Student: Yes.

Professor: If you have whatever that d that maximises this if you play that d that is the maximum reward you are going to do.

Student: That d will also belong to the union of all decision sets.

Professor: No, this d is coming, you are going to choose this d only from d_t.

Student: But that d will also belong to the union of all these decision sets.

Professor: Yes, it belongs to.

Student: So, you said that mean is bounded by 1 but all those values are bounded by 1 actually.

Professor: See the reward.

Student: yeah.

Professor: Is given like that there reward is this part if you are going to play in round t some decision t plus noise. So, this is the, what is the mean reward you are going to get in round t, this is nothing but this part depending on which arm you played. I am just saying that whichever arm you are going to play in that round or in fact any round in for all the rounds the mean value should be less than or equals to 1.

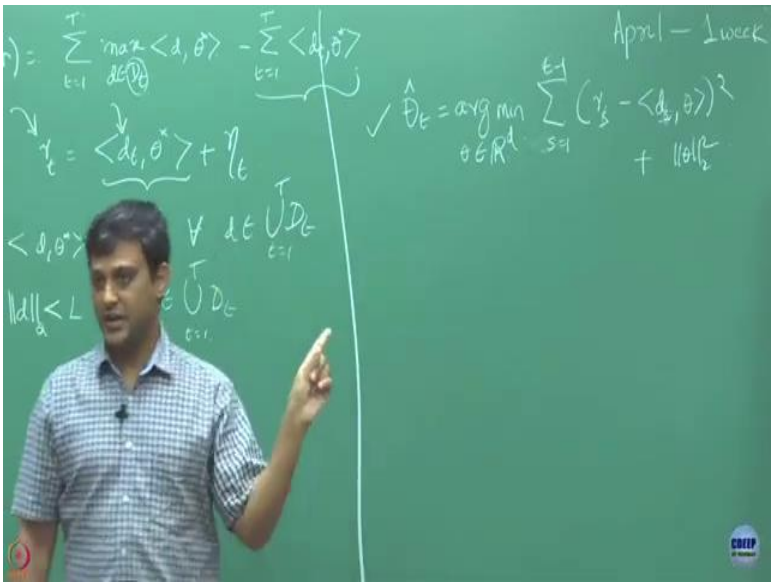
Student: (())(09:41).

Professor: What is a, what you are saying dt is why it is.

Student: (())(09:53).

Professor: Why it is probabilistic, because this is what you are going to observe in every round. So, let us say till some point time you have observed, you know which is the arm you are going to play and you know what is the corresponding reward you observed. So, this is the information you have and any estimate you are going to make it is going to depend, going to depend on these two quantities. It has to depend on the observation you have made that tell, that is noisy and because of that any next decisions you are going to make so.

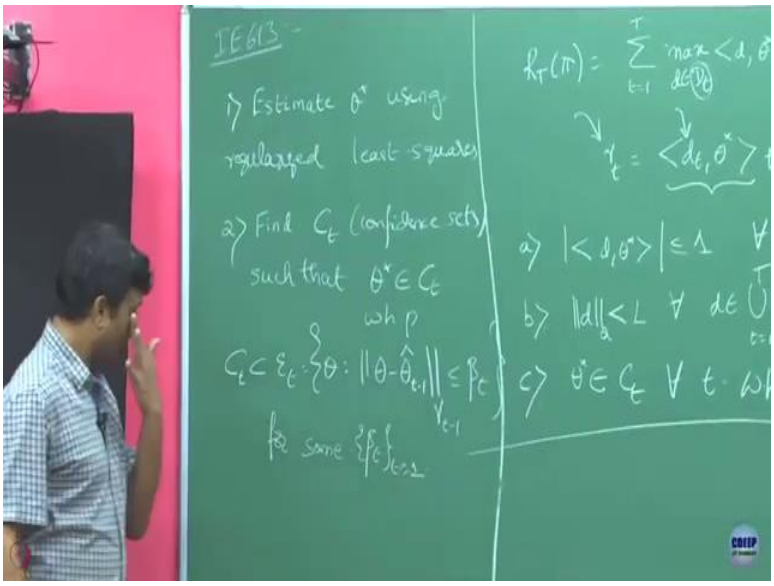
(Refer Slide Time: 10:42)



So, let us say how did we find theta t hat we said that this is nothing but arg mean of this is my r_t minus and then what did we say what d_t of theta square plus norm of theta square at so this is what is you see this is what, this is what our regularized least square regression, regularize the least square estimate. Whatever theta t hat you are going to find this should be maybe I should write it as s, s equals to 1 to t minus 1.

So, till around t you have observed all the samples from s to t minus 1. You have observed this reward and the corresponding for this corresponding arm ds you (())(11;56) this. Now, that decision which arm you are going to play in the next round has to depend on this, based on this estimates I am going to make a decision which I am going to play in the next round. So, I will come to that.

(Refer Slide Time: 12:25)

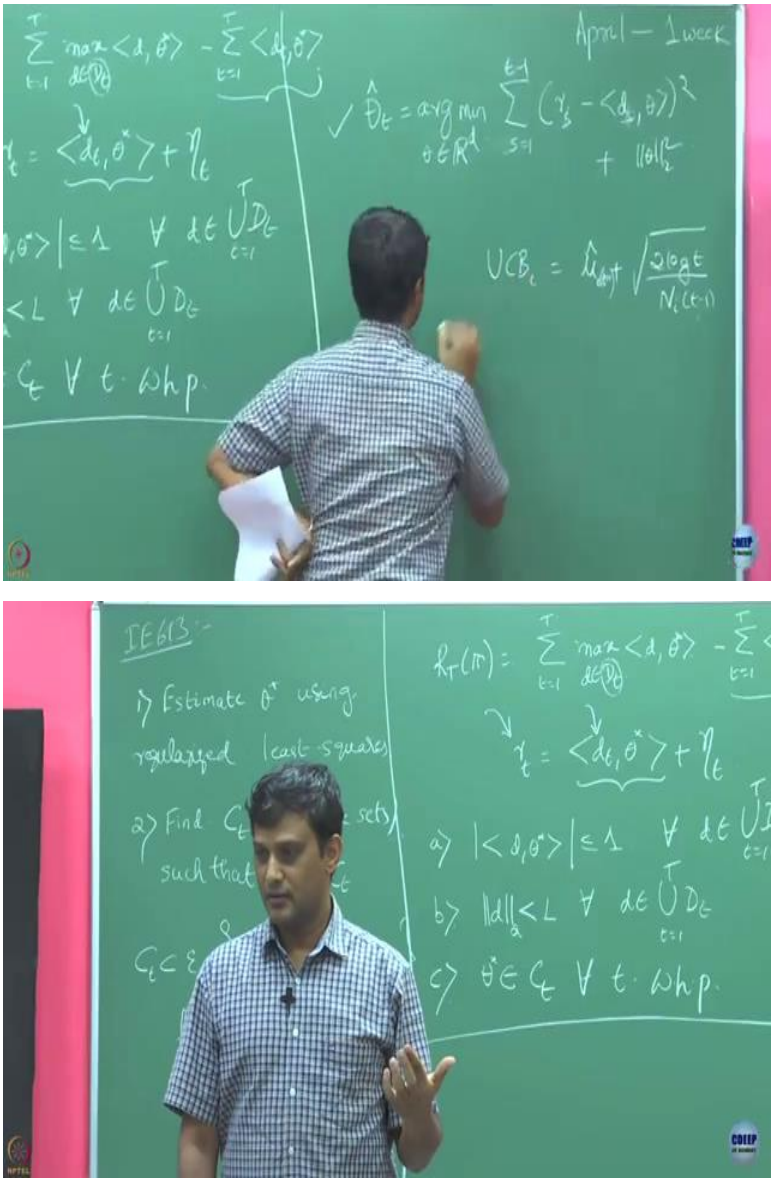


Let first me, let me write this conditions under which I am going to prove this. So, this is one assumption, this is assumption and then the third assumption is simply this like in every round so this theta star belongs to C_t for all t with high probability this is another assumption I am going to make, this is the setup.

Now, I have to give you algorithm how the generic algorithm looks like and then I am going to say for that algorithm how the regret bounds are going to look like. You estimate your theta in round t based on your past observations like this. So, this theta hat is a random quantity. So, now based on, using this I have maybe like in round t based on your previous estimate you are going to whatever in round you are going to construct a confidence set like this.

Now, let us see how to find, now how to make a decision in every round based on this information. So, now I have to make a decision. So, this in every round I have to see a d_t set is revealed to me looking into that d_t set I have to now decide which is the small d_t from that set I have to, I am going to play. How you are going to do that? Suppose, if I can assign a value to each element in the decision set d_t, then what I will do. So, let us go back to my multi-armed bandit setting the original one, what we did for each arm based on its estimate plus the confidence term I defined a value for each arm.

(Refer Slide Time 14:42)

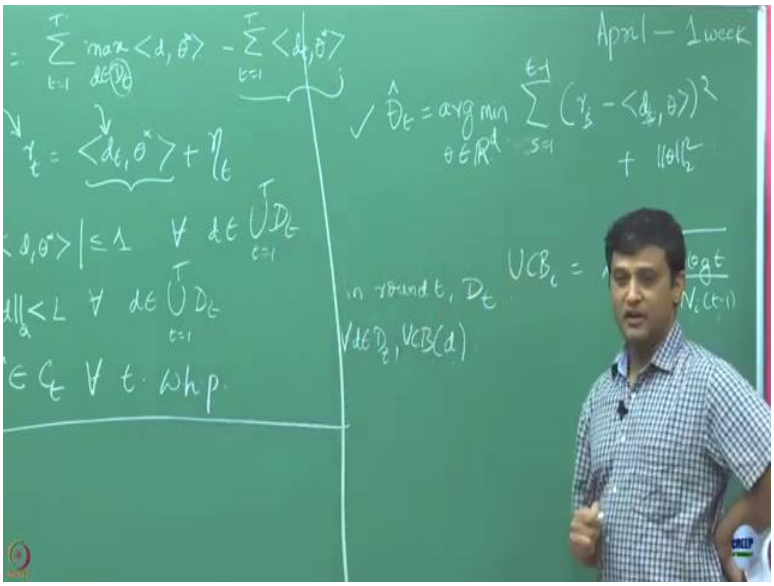


So, in my standard, I have mu t hat plus 2 log t divided by N so this is my estimate of arm i till round let us say t minus 1 and this is my number of pulls of this. So, I have defined a term like this for arm i and what I did, I played an arm which has this highest value. So, this is kind of a index value that I assigned to arm i. Now, for me the arms are this entire set. So, first suppose if I can assign a index value to each of the elements in this set d_t then which is the arm you are going to play in that round t.

So, what is this so we said that this is nothing but UCB of arm i and what I did I basically played an arm which has the highest value of this UCB quantity, I am going to do possibly similar thing

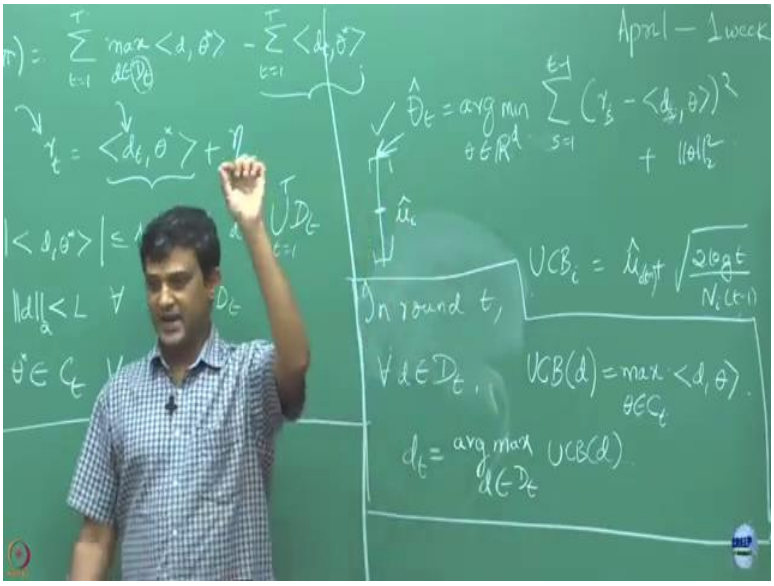
maybe when I said d_t is revealed to you do a similar thing for every point in that define a UCB index for that and then go and find out which is the point arm in that which has the highest UCB Index. So, now let us worry about how to find this UCB index, one. Now, this is for multi-bandit.

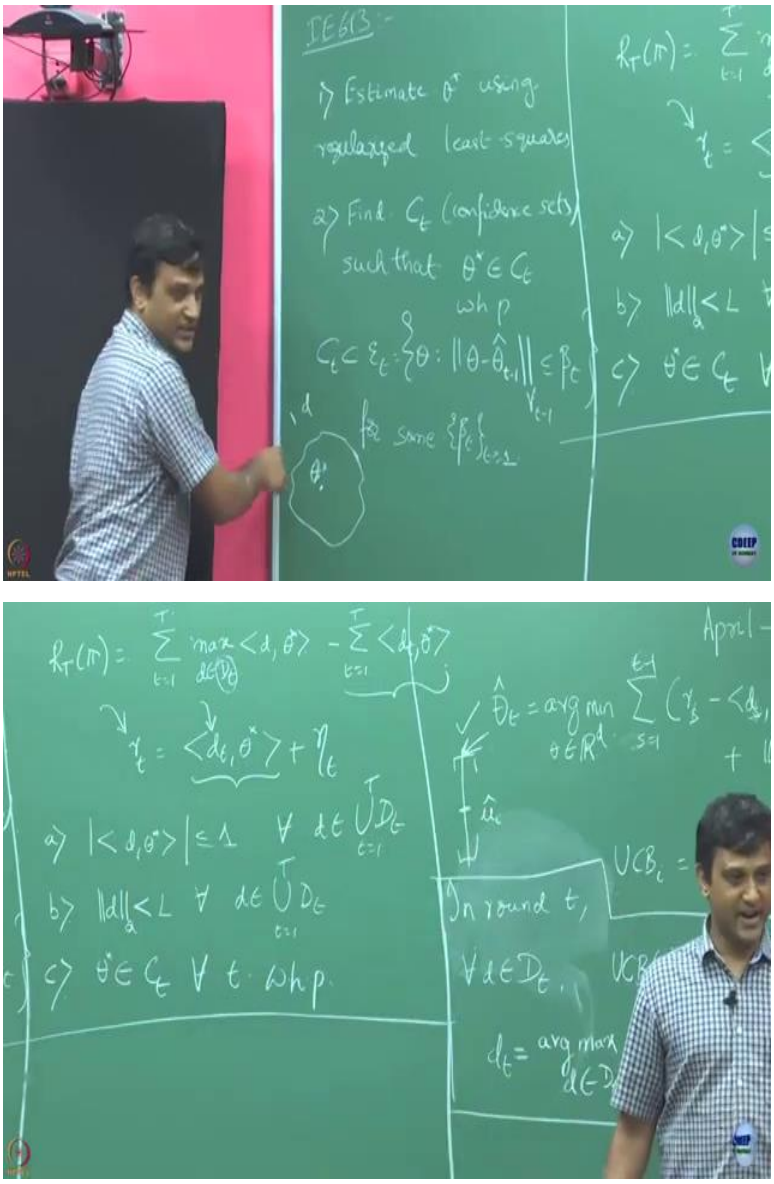
(Refer Slide Time: 16:42)



So, now I have to find UCB for at d in round t. So, I have d_t now for all d belongs to D_t I need to find UCB of d. So, let me write it more clearly.

(Refer Slide Time: 17:07)





So, one natural candidate is to find an assigning a UCB value to each element in this is to look for how. Let me define this and then we will discuss is equals to max over theta belongs to C_t times d, theta. What I have define this C_t is something which involves my true parameter with high probability in a range this C_t is some ball which involves my d star, sorry my theta star with high probability. Now, what I am doing for my arm d I am looking for, if is suppose if I have to play this arm d. Now, I am looking at the best reward I would have got if the parameter theta is drawn from this set C_t.

So, what I am basically assuming is my true parameter is somewhere in this all I know is it is somewhere in this, all I am trying to do is, I am trying to take this UCB index, the best I would

value I would have obtained if I played so the best that could have happened to me if I played my arm d. So, this is what like I am basically choosing the reward for d optimistically. So, how we did it in the UCB, UCB we have a mu hat and then around this mu hat we have constructed the confidence interval and what would I do, I always went and choose this as that true index.

This has the index for my arm and then I defined the UCB index to be, so I am trying to do similar thing here, in this ball I am just trying to see so if I will just compute take my d just see if with this d, what is the theta that gives me maximum value in the set C_t that I am going to define the UCB index of that arm d.

That is just like thinking that with my current information of confidence set I am just assuming that optimistically the parameter is the best possible one when I am going to I mean, when I going to, if I am going to play arm d, I am going to see that the parameter that is unknown to me is the best that could have come from C_t. I do not know what is that parameter but from this C_t, I am just going to do this maximization and assign UCB to be this value.

So, do you agree that this is like assigning the value to d optimistically here from the set C_t. Earlier I was doing this optimistically by just adding this upper confidence bond. Here also, in a way analogue say I am doing some kind of upper confidence bound. The best I could have got from this set d if I have to play arm d. So, this is how we are going to define UCB index of arm d. Once I have defined UCB index of arm d like this, then what I am going to do?

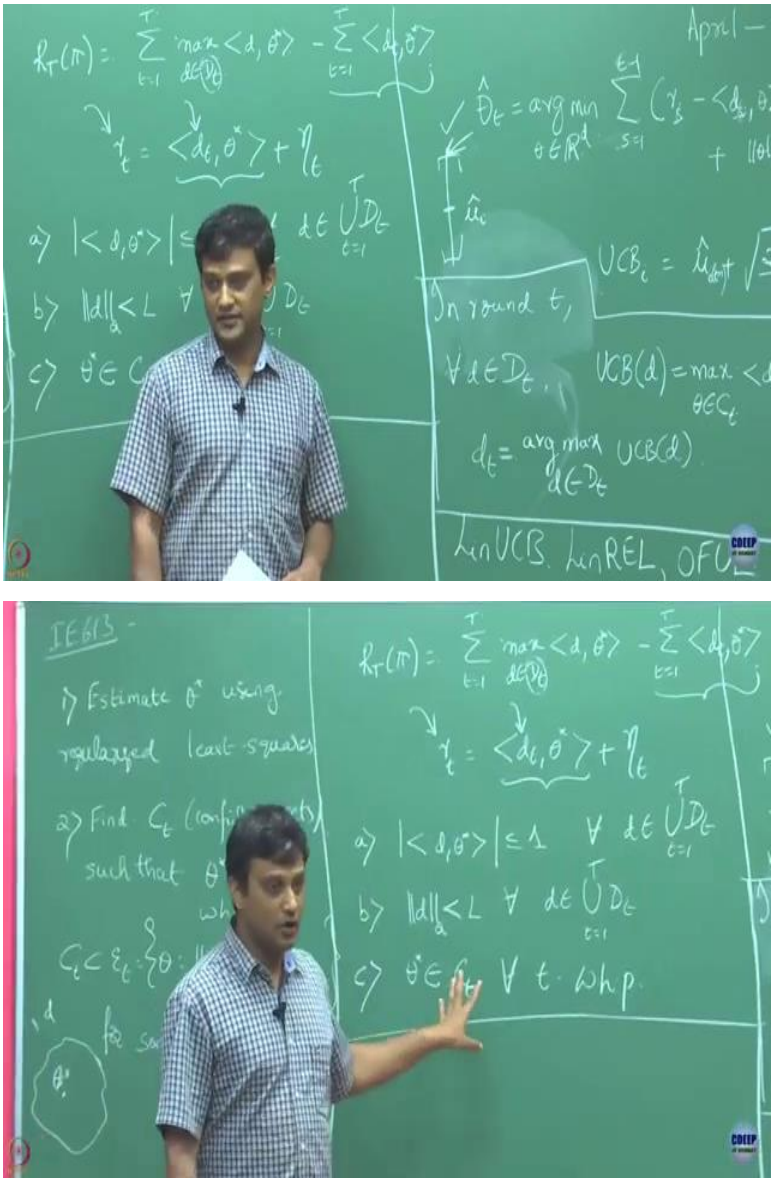
Student: Find the one (())(21:21).

Professor: Find one which is going to maximize that, then you are going to do play arg max over d of UCB of d and this is what we are going to call it as dt star dt. So, this is the one I am going to play. So, this is basically my algorithm. So, you see that this is nothing but again applying the UCB idea upper confidence bound idea to this contextual settings where I have not finitely many arms, but that arm set is now a complete set dt which could have uncountably many elements in this.

But now that I have that parameterized it. So, I have parameterize the rewards and now I am trying to estimate this parameter and building a confidence about that parameter using the C_t and now I have defined my confidence upper confidence for each my arm in this fashion. This is the broad idea that all the algorithms that work in, that apply on linear stochastic bandits work and

there are different different names like based on how they are going to come up with this confidence sets.

(Refer Slide Time: 23:06)



So, in literature, you will see many algorithms like Lin UCB, it is called linear UCB. There is another called I think Lin REL. So, this is likely Lin UCB stands for linear stochastic bandits UCB, Lin REL stands for I think linear reinforcement learning this is one of the earlier algorithm.

And there is another algorithm that is more recent it is called optimist optimism in the face of uncertainty L stands for linear. Now, people have this different different algorithms on this. So,

all of them are going to hinge on kind of this they just differ in terms of how they are going to construct these confidence sets.

Student: We need assumptions so I mean what was the complex (())(24:12) in a problem is simply taken away by the assumptions that we are making.

Professor: These assumptions.

Student: Yes.

Professor: No, in fact these assumption are not actually I mean bad. What is this this is just about the setup. What we are basically (())(24:26) we are just saying mean rewards are bounded by one. If they are not.

Student: I am, talking about C part theta star belong to C_t.

Professor: Yes, yes, of course this is the more complex part. Now, later we will see that how to construct such sets. Because right now you should have begun starting by how to construct that set, I mean we will lost the overall picture. So, right now I am saying that let us say we have such a sets then how to go about this algorithm. So, now you will see that even with this setup to bound by regret I need to have a slightly different, take a different approach than what I had then for the standard bandit algorithm.

So, how did we do? How did you prove? The regret for our multi-armed bandit in the standard set up we basically bounded the number of pulls are suboptimal arm and how we were able to do that we know that if you have to play total number of rounds t of course each one of has to be played less than that many rounds t. But here the number of arms itself could be countably many or uncountable the number of so. Because of this it may happen that you may not even end up playing some of the arms. So, what in the standard multi-armed bandit setting you played each arm at least once, that there were only finitely many arms and your time horizon you always took larger than to be than number of arm you got at least one samples.

But here once t is finite even though this set d_t let us say it is going to same in every round. It may happen that for some of the arms you will never get any observation. So, because of that like it is not clear that the same method we used to bound the number of rounds, number of place

of the suboptimal arm is going to work out here. So, that is why let us see if we have this what is the way to go and prove through the bound.

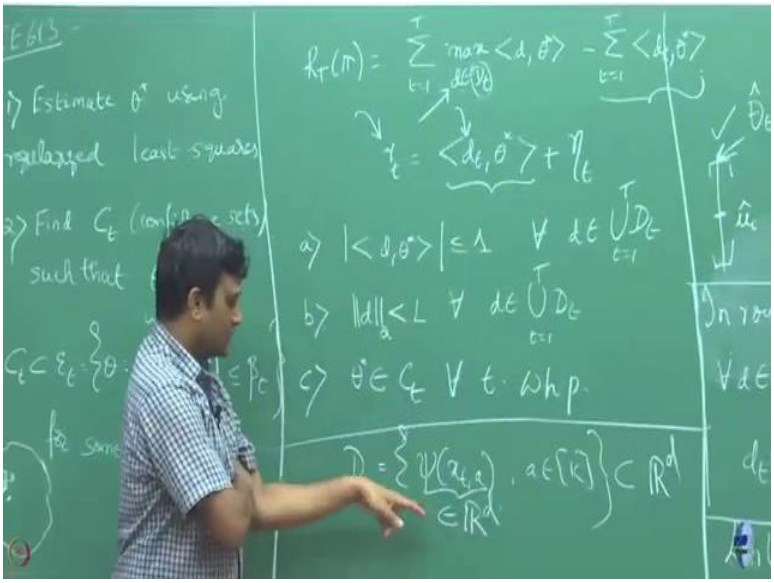
Student: Taken away the concept of arms in this (())(27:06).

Professor: Yeah, but now this that has been now got in to this set d. Now, this is what I am going to call as arms.

Student: All values that are possible in d_t.

Professor: Solve the values possible in d_t.

(Refer Slide Time: 27:28)



So, let us rewind this, what is D_t for us? Let us now go back to our initial problem from where we arrived at here. So, D_t for us is nothing but phi of x_t of a, a coming from K. So, in round t a feature x sorry, a context x_t was revealed and I had a map, feature map which said for each action what is this value.

Student: This was the contextual bandits for the linear bandits you generalized this even more to.

Professor: So, this is not like, this is the contextual bandit.

Student: Yes.

Professor: Now, I am just saying that this set d_t consist of these features.

Student: Cardinality is finite.

Professor: This cardinality is finite. So, now if by making it more general this C_t D_t could be any arbitrary, any arbitrary subset, bounded subset. The thing here is so where did this each of these feature remained in, this is feature vector this remained in some R^d space, now instead of that let us say why always talk about these specific features I am just saying in every round I am going to get a set dt which will consist of these feature vectors and those feature vectors I am going to call this arm.

Student: Dt right now is the close subset of R^d.

Professor: Is a subset of R^d as.

Student: With the currently half of the (())(29:08) right now the area dealing with (())(29:10) as a closed subset of R.

Professor: Yes.

Student: That is all.

Professor: Yeah, that is what like this D_t is a bounded subset of R^d. So, other way to think about this is. So, in this case, in this case, it is clear that there are only finitely many features right in this because one corresponding to each arm. So, let it the number of arms goes to infinity or like number of arms is could be uncountable then for each arm I have this feature space and that I am going to now call it as dt. But each feature vector is corresponding to one particular arm.

So, that is why now because we have done this abstraction I am now just saying that any element in this d_t is corresponds to an arm. So, when I have done this if I take a particular element here phi of x_t and a, that corresponded to that arm a. So, I have basically in that round t, I have mapped a arm to a feature.

So, here just like I have given so many features that means I am just thinking that they correspond to different different arms. We have, that is why I have thrown away that concept as actual physical arms and now we have these features of this decision space and again I am coming back and saying every element in that decision space is another arm, is just an arm.