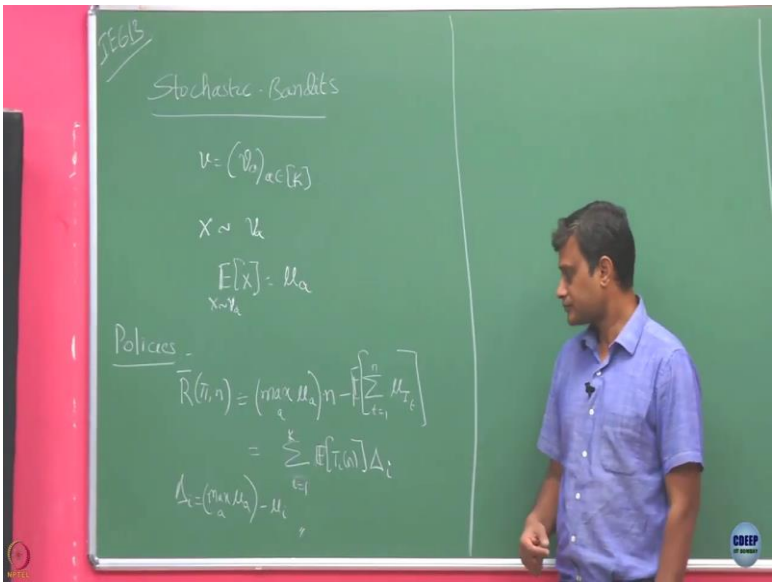
Bandit Algorithm (Online Machine Learning)
Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Lecture - 30
Explore and Commit (ETC) Algorithm

So, after talking about this concentration inequality, we returned back to the stochastic bandit setting. We define what is the regret would be interested in the stochastic bandit setting and then defined the pseudo regret which we want to take it as a performance measure and we want to give performance guarantee on that. And we concluded by giving the in the last class regret decomposition result right.

Now, we will move on and try to see what is the best bound or what kind of bounds we can give on the pseudo regret ok.

(Refer Slide Time: 01:06)



So, we know that we let us say we have a bandit instance, where we have this set of K distributions one distribution assigned to each of the arms. And we said that the sample when you are going to play an arm a, the samples are drawn. The random variable X that corresponds to the sample, that are drawn from arm a that will be distributed according to distribution a right. And we said that expected value of X in this case let us say we denote it as μ_a ok, where this guy X is drawn according to this distribution μ_a.

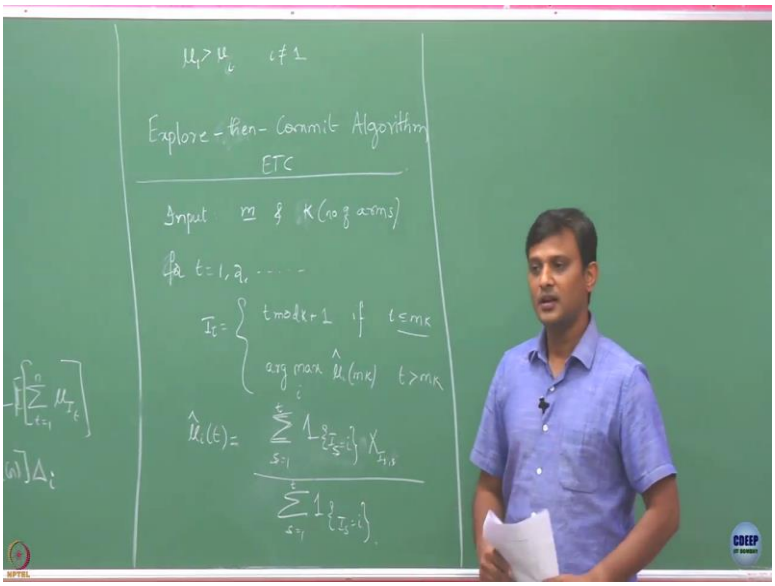
Now, our goal is to basically our goal boils down. If you want to minimize the regret our goals boils down to what? Identify an arm which has the largest value of this mean value right. And what would we say in the last time? Through our discussion on concentration inequality we said that if we have many samples from a particular arm, how to estimate this parameter μ_a and we said that if you are going to use sample mean as my estimator, how far the estimated we mean will be from that true mean after a certain number of rounds. We gave a bound on that right. We specifically focused on the cases at the case where this mu as are all sub Gaussian distributed and for that we gave this concentration bounds ok.

Now, let us start thinking about how to estimate this means, but our goal is to not just estimate this, but to quickly estimate this so, that I do not incur regret after some time. Now, what could be our strategy? Now, we are interested in to come up with the policies for this right. So, what are the policies? Policies are algorithm. What so, any thoughts on how we will go about this? Let us say you have been asked to we have been given n rounds and you have been your goal is to come up with a policy π such that you want to was this which in the last class we discussed that this can be given also as until the there are K arms.

So, this is nothing but expected value of right, where we defined as delta i equals to and also we said the last time suppose some particular arm is optimal. So, throughout we will assume without loss of generality that my arm 1 is optimal that is the mean which has the highest value is the one with arm 1 ok. So, we know this, but we the algorithm does not know this ok. So, this is just for our analysis point of view.

So, because of this if arm 1 is the optimal one delta 1 is going to be 0 and delta 2 will be positive and delta 3 will be positive. And also we will assume for time being that the optimal arm is the unique right.

(Refer Slide Time: 06:25)



So, if μ_1 is the highest mean all the other means μ_2 and all mu to all others are going to be strictly smaller than the mean of arm that is why the arm 1 is the optimal arm.

So, we know that this delta 1 is going to be 0, but this delta 2 and all are going to be positive. So, if I want to minimize my regret, I need to ensure that the expected number of plays of my arm 2 and others should be as small as possible or I want to keep them low ok.

So, now I am going to use the terminology that arm 1 since it has the highest mean I am going to call it as optimal arm and all other arms are suboptimal arms ok. Any thought on how to do this? Just like some random ideas ok. So, first we will start with something called this explore and commit algorithm. So, one obvious is you sample each of these arms for a certain number of rounds and then find the means of sample means of each of these arms and then maybe after that you just play the one which has the highest mean ok.

But here the question is how much time I should sample each one of them so, that I get the good estimates all of them if I can. If I get once I am sufficiently confident these estimates are good maybe then I do not need to explore any more further then just play the one which has the highest mean after that wards right.

So, let us that one natural policy let us we call explore and commit; explore call ETC. Now, all the algorithm looks in this case. So, for this algorithm we need to tell how many rounds it need to explore. So, explore let us say we have to tell the algorithm, sample each of these arms these many times that has to go as an input to this algorithm right.

So, let us say input is that number m and of course, and k; k is the number of arms. And then what this algorithm does? The algorithm is going to play arm 1 m times arm 2 again m times arm 3 m times. So, there are k arms. In the first k m number of rounds it will sample each of the arms m number of rounds. After that it finds the estimates of all of them and commits to the one with the highest mean. I am just going to write that.

So, all of you understand. What is this? This is t mod k and till the first m k number of rounds what it is going to do? It is going to do t mod k. So, suppose let us say t equals to 1, it is going to be what? 1 mod k 1. So, what it is going to play? It is going to play 1 mod k right. Let us say k some greater than 1, it is going to be this term why it is going to be 1?

Student: (Refer Time: 11:37).

Yeah, plus 1 there is right. It is going to be 2 in that case. So, like that when t equals to 2 it is going to play arm 3. When t equals to 4 sorry, t equals to 3 it is going to arm play and when t becomes k, it becomes 0 and it is going to play arm 1 and it continues right. So, it is going to play the arms in a Round Robin fashion. It starts with actually 2, 3, 4, 5 up to k and then with comes with 1, 2, 3, 4, 5, like that until that is it is going to do m k number of rounds..

And for t greater than m k, so, what is mu hat k here? So, we are going to define μ_i hat at time t; that means, the notation here is the estimates of arm i I have at round t and this is going to be; so, this is the number of samples averaged till time t right.

So, what is a numerator doing? It is; so, first focus on the denominator. It is counting on the number of plays of arm i till round t. And what is numerator doing? Numerator is taking the sum of all the samples that comes from arm i till round t right. So, this notation is saying that whenever i's in the S round if S equals to i then only this term is retained. So, it is returning only those samples that you have collected from arm i till round t.

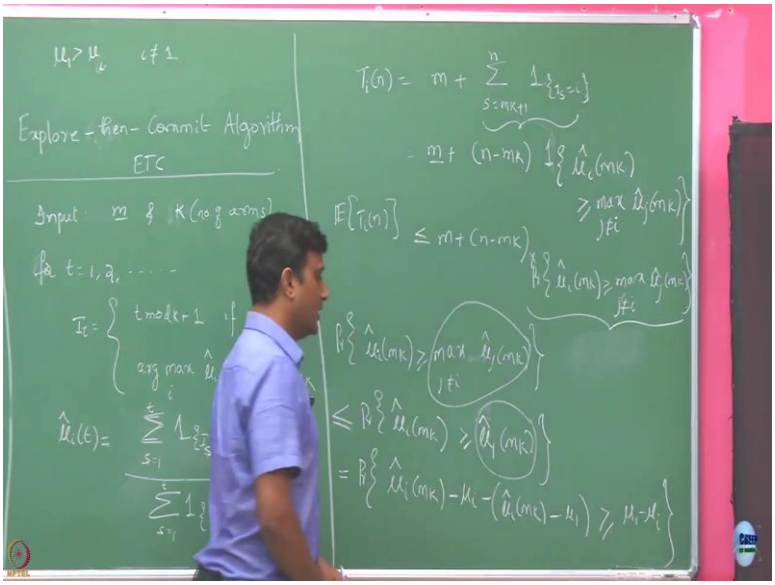
Student: (Refer Time: 14:41).

Which one? This has to be i's yes ok.

So, once you hit m k number of rounds ok, what you will do? You will at that round you will see which arm has the highest mean and after that that is it, you keep on playing the same arm which has the highest mean. You are not going to change the estimates after that. Now the, so, that is why we are saying that we have explored till the first mk rounds and after mk rounds we are just committing ourselves to the one which have which has the highest empirical mean in that till that round ok.

Now, if I am doing going to do like this, what is the performance I am going to get ok?

(Refer Slide Time: 15:56)



So, now, let us try to bound this ok. Now, let us count T_i (n). So, how many rounds? I have played arm i till n rounds, right. You know that for the first m rounds you have played arm i that has been guaranteed by this first part right ok. And then for S equals to m k onwards you do not know whether which arm is played that depends on which one has the highest mean.

So, let us right and now from this round m k plus 1 and n whether you have played arm i or not depending on in the mk-th after the mk-th round whether the arm ith term happened to be the best one or not right.

So, let us try to do this. So, what is this is going to be then? This is going to be so, this depends whether I am going to play or not. Only depends whether on the m kth round I happen to be the best or not. So, that is going to be n minus m k times indicator. So, I am saying that so, now, after this m k plus 1 to n rounds I will be playing arm i only if μ_i hat of k this guy happens to be larger than the other arms right; all computed at m kth round. So, at the end of this m kth round I have computed the empirical means of all of them and I know that after t is going to be greater than mk I am going to play arm i only if its empirical mean happens to be larger than others right, fine, ok.

Now, so, this is a constant that has been given to me as an input to the algorithm. So, n is also let us assume that is given to me the number of rounds for a given number of rounds. So, what will be the expected number of plays of T_i (n)? So, this is going to be m plus n minus m k and now if I take the expected value of this indicator it is going to be probability of right.

So, now the problem boils down to if how can I bound this probability? What is the probability that at the end of m k rounds, now i arm happens to be the one with the highest mean highest empirical mean ok? So, now, let us try to focus on this term. So, now, for time being let us assume that this I am interested in this arm i which is other than the optimal arm that is i is not equals to 1 ok.

Now, I am interested in a bad event right. Like, I am I want to basically see that I assumed my arm 1 is optimal, but what is the probability that I will end up some arm i to be empirically best right. So, if that happens if the arm i which is not 1 happens to have a highest empirical mean compared to the arm 1 then that is a bad thing for me right that is going to cause regret for me.

So, let us consider that event and see how if we can capture that. So, what we want is in this I want the when this happens to be larger than μ_1 hat of k that is the mean empirical mean of the i-th arm happen to be greater than or equals to the empirical arm of the arm 1 that is the best one right.

So, here I was looking at max over all the arms. Here I am just trying to replace it by the one with arm 1. So, what can I say the relation between this and this? So, this is basically saying μ_1 hat is greater than maximum of several terms and μ_1 is one of the terms here in the max ok. So, if I am just going to return one of the terms here, so, what is the relation

between this probability and this probability? So, which is going to be larger; this one is going to be larger or this is going to be larger?

Student: (Refer Time: 23:31).

First one is going to be larger right and anyway left hand side is the same.

So, now I am saying that μ_1 being greater than equals to some larger quantity, now I am asking μ_1 to be greater than somewhat smaller quantity. Which event implies which event?

Student: (Refer Time: 23:50).

So, if this happens this automatically happens right. So, which one should have a large larger probability, if this A implies B?

Student: (Refer Time: 24:09).

First one will have larger probability?

Student: (Refer Time: 24:15).

So, A implies B?

Student: (Refer Time: 24:26).

So, A is this event if whenever A happens, we are saying that B happens right. Which should be correct?

Student: (Refer Time: 24:43).

Yes of course, this is larger.

Student: (Refer Time: 24:49).

Student: (Refer Time: 24:51).

Yeah, this one is going to be more stringent right.

Student: (Refer Time: 24:57).

And this is happening this is automatically implied.

Student: (Refer Time: 25:01).

So, no power is more stringent. So, because of that you can check that if this is more stringent this probability should be smaller than the other one right. Or like maybe like we are basically saying that if this means basically you are saying otherwise this were saying that is contained in.

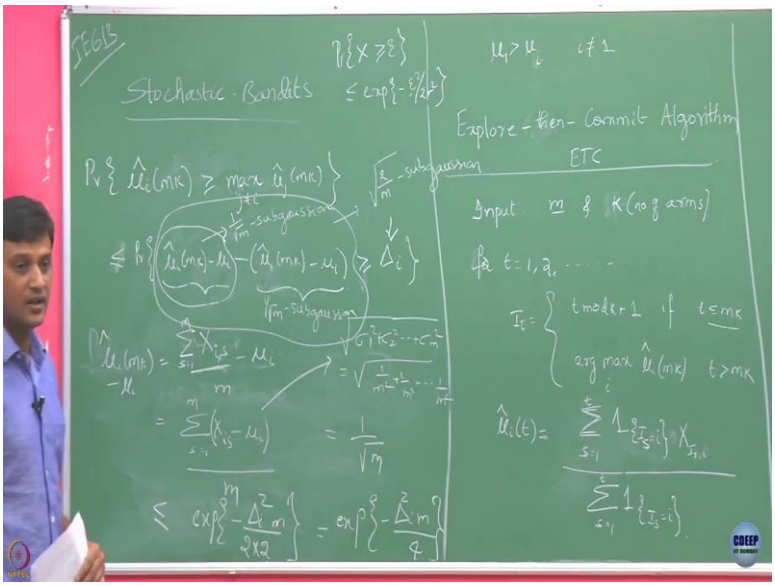
Student: (Refer Time: 25:30).

So, this is just like this guy is asking it to be larger than some quantity and now you want it to be smaller than in this quantity to be larger than somewhat smaller quantity, so, this is all automatically implied but, the otherwise not true right. Like, suppose μ_1 hat is greater than let us say μ_1 hat that does not mean that μ_1 hat is going to be greater than or equal to max over all of this or is that automatically implied. Actually, so, we should think it like this.

Suppose let us say, μ_1 hat is going to be greater than or equals to μ_1 hat and now I am looking at max. In this max μ_1 hat is already content, ok. So, now, this quantity is going to be larger than this quantity right. So, you are basically asking this to be more stringent. You want more than what is already guaranteed in this. So, that also implies that, this quantity has to be less than or equals to this quantity ok.

Now, we have this. Let us manipulate this. What I will do is, I want to ok; just I will just take both sides. μ_i is its μ_i minus; I am taking this quantity on the other side and now I have added minus μ_i here and plus μ_i. This is going to be what? Sorry, μ_1 here. Is going to be what? μ_1 minus μ_i; is this correct? I have just done a manipulation here. I have just taken this on the other side and there is minus I minus plus μ_1 here and I have added the same thing over here.

(Refer Slide Time: 28:15)



So, now continuing this line of inequality, μ_1 hat mk μ greater than or equals to μ_1. So, this is max over j not equal to i μ_j hat. So, now, this is going to be greater than or equals to probability that; and by definition what is this μ_1 minus μ_i for us? Delta i, that we have defined as delta i ok, sorry. So, this is what we have ok.

Now, recall, what was these quantities. This μ_i hat is nothing but average of m samples that are obtained from i curve right. So, that is how you have computed it. It is when it did it, we are compute taking t equals to m k. It is looking at all the samples that are drawn from arm i till m kth round till m kth round we have exactly m samples right and this is going to be m; because m the arm i has been played m number of rounds.

So, this is nothing, but this quantity is nothing, but average of m samples that has been drawn from arm i. And we know that arm i these samples are iid further the mean of arm i is μ_i right. Now, we argued last time if all my arms distributions are sub Gaussian what this quantity is going to be? Sub Gaussian. Sub Gaussian with what parameter? Suppose, assume that all my distributions are one sub Gaussian ok.

Student: (Refer Time: 30:59).

1 by m; why is that? So, there are so, if you simplify this.

Student: (Refer Time: 31:10).

It is 0 sub Gaussian?

Student: Under root of sigma (Refer Time: 31:16).

Under root of sigma?

Student: (Refer Time: 31:20).

1 square minus 1 square.

Student: (Refer Time: 31:25).

Right.

Student: (Refer Time: 31:33).

Yes.

So, what happens? So, in our case we said sigma 1 is 1 and sigma 2 is 2; also 1. It is, but there is also like now let us rewrite it. So, I know that for mk, the denominator becomes what? It becomes m right. So, this is m and what is this numerator is like? This is basically summation of?

Student: (Refer Time: 32:08).

This is going to be the summation of i samples which has drawn from arm i only that is what it is saying right. This is X_i, i's the sample you are going to observe in the s-th round, but I am interested in only those samples where i's equals to i right.

So, because of if you just ignore all the places where my i's is not equals to i, if you only return those where is equals to i; I have exactly m samples in the numerator right. So, if I just took that it is basically going to be what? X_i and so, this is basically going to be just i right, where i is s 1 to m. So, these are all the samples of the I even do not need to write this yes. This all these samples are coming from i-th arm and all these samples are one sub Gaussian.

So, now we have this and now we have also μ_i here right. Let us take this μ_i here ah. This is whole of this quantity divided by μ_i. We have already discussed this. So, S equals to 1

to m X is times μ_i this whole divided by again. So, there is some small mistake we made. We did not say that the samples are coming from one sub Gaussian.

We said that these samples when you subtract from μ the mean we said that this is one sub Gaussian that is what we said in the last class right. That the distributions are such that if you centre it is distribution that is if you subtracts the mean from the samples those samples are one sub Gaussian. So, that is the value right simplify this. So, I have to also not this I am taking this entire thing here. This is going to be like this and we know that this quantity is one sub Gaussian ok.

Now, if this is one sub Gaussian and if you going to sum this, what is this this is going to be? 1 by?

Student: (Refer Time: 34:45).

1 by root m; why is that? So, we already said that you said that it is going to be like sigma 1 square sigma 2 square like this right, we have m quantities here and first each one is like 1 by m square right. So, 1 by m square plus 1 by m square all the way to 1 by m square; there are m quantities like that and if you just going to m by m square, it becomes under square root. So, it is going to be 1 by m square. So, the entire quantity here is 1 by square root m sub Gaussian.

What about this quantity? This is μ_1 here. We are just saying that all the distributions are all the know is one sub Gaussian right even though they have different the centering value that all the centre distributions are one sub Gaussian. So, even for the optimal arm, when you centre it, that is going to be also the sub Gaussian with the same parameter ok.

So, now what about this? So, I know that this guy is one sub Gaussian 1 by m sub Gaussian this entire thing and this another guy is also 1 by square root m sub Gaussian, what about this difference ok? Before difference this entire thing is 1 by square root m sub Gaussian, what happens with a minus sign with a minus sign is what is this?

Student: (Refer Time: 37:02).

Why is that?

Student: (Refer Time: 37:05).

So, right, we have said that even if you scale it by some constant c whether positive negative does not matter it is going to give the same 1 by square root m with that constant and that that constant is 1 here ok.

Now, this is 1 by square root m sub Gaussian, this quantity is also 1 by square root m sub Gaussian and think of these are we are adding 2 sub Gaussian random variables and are they independent here? They are independent right. When what would we say? When I am going to try sample from an arm, it is going to be independent of the past all pulls from that arm and also independent from the pulls of other arms ok.

So, it has to be these two are independent and now what is the what is this then?

Student: (Refer Time: 38:01).

Then it is going to be under root 2 by m right because then it is like square root of. So, now, this entire thing here is under root 2 by m sub Gaussian. So, if this entire thing is under root square root 2 by m sub Gaussian, now I already know a result which we showed last time which we written as one of the lemma right. Can I apply that and find a bounds for this, what is this probability?

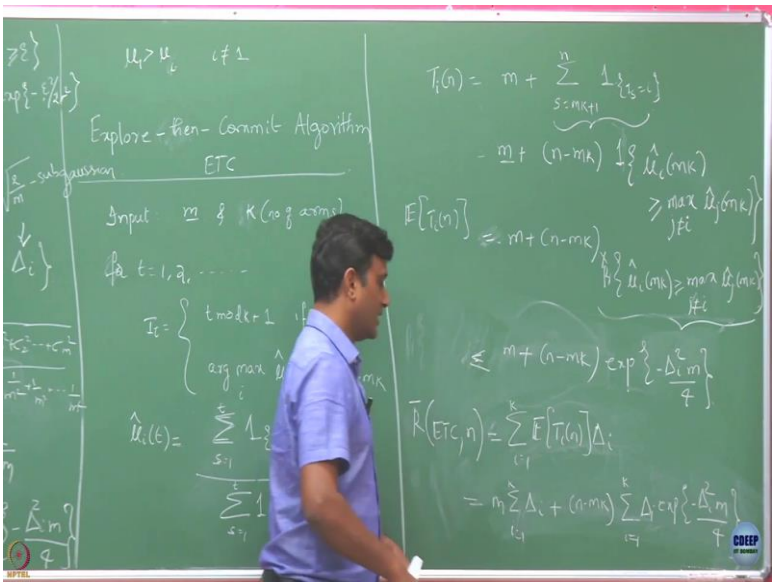
So, recall that, we said that probability that X is greater than or equals to epsilon, when X is a sigma sub Gaussian. What would you say this is upper bounded by exponentially exponential minus?

Student: (Refer Time: 39:05).

By 2?

What are sigma here? The sigma is the sub Gaussianity parameter of this X. Now, this X is nothing but this entire quantity right and so, replace this sigma by square root by 2 by m. Now, what you are going to get if you do that? And you are epsilon is now delta i. Now this X is the centre quantity. So, if you just apply that this probability is going to be upper bounded by exponential minus epsilon is delta i square 2 times 2 times this; sigma is going to be square root 2 by m. And if you are going to put it we are going to get after squaring 2 by m and this is going to be 2 times m right. So, this is nothing but exponential times minus delta i square m by 4.

(Refer Slide Time: 40:38)



So, what we finally, actually showed is ok, I will just erase this part I will rewritten there. m plus that the I am just m plus n minus mk and this whole quantity we have just argued that this is upper bounded by exponential equals to this ok.

Now, once we have I have bound on this expected pulls of an arm i right, now we can go and use our regret decomposition result to get a bound of my regret bound of my regret. So, what is that then? My pseudo regret is here it is given by expected value of $T_i(n)$ which I substituted this and this quantity is upper bounded by like this. So, this is going to be what? First term is going to be m k times sorry, so, if I substitute this I am going to get m times plus n minus m k times; i equals to 1 to k delta i must be. This is correct. I have just substituted bound on the expected number of pulls in this expression here.