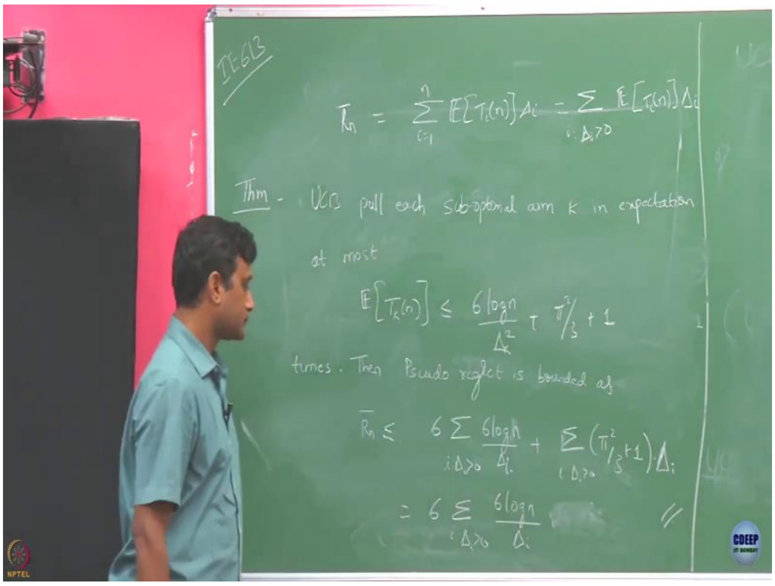
Bandit Algorithm (Online Machine Learning)
Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Lecture - 33
Upper Confidence Bound Algorithm

So, what is the performance guarantee of this algorithm? So, before we write this bound, we already discussed that the pseudo regret can be decomposed right. We had a decomposition result, we said that x the pseudo regret can be written as summation of expected number of place of arm times the sub optimality arm.

(Refer Slide Time: 00:47)



IE613

$$\bar{R}_n = \sum_{i=1}^{n} \mathbb{E}[T_i(n)]\Delta_i = \sum_{i:\Delta_i>0} \mathbb{E}[T_i(n)]\Delta_i$$

Thm - UCB pull each sub-optimal arm k in expectation at most

$$\mathbb{E}[T_k(n)] \leq \frac{6\log n}{\Delta_k^2} + \pi^2/3 + 1$$

times. Then Pseudo regret is bounded as

$$\bar{R}_n \leq 6\sum_{i:\Delta_i>0} \frac{6\log n}{\Delta_i} + \sum_{i:\Delta_i>0} (\pi^2/3+1)\Delta_i$$

$$= 6\sum_{i:\Delta_i>0} \frac{6\log n}{\Delta_i}$$

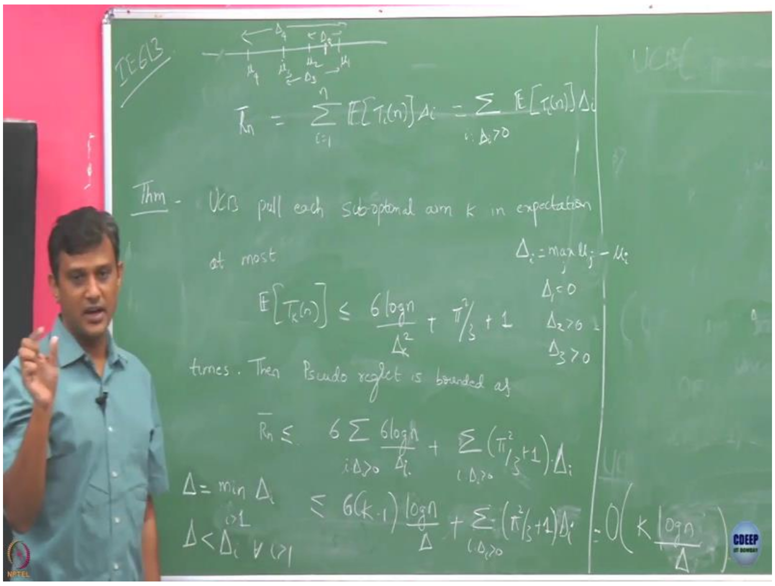
//

That is we said that this can be written as. So, then, in this case, it is enough to give a bound the expected number of pulls off each of these arm right. So, in this theorem, we are exactly going to do that at most times and then pseudo regret. So, if you just plug in this value in this, what are you going to get?

So, another thing is we know that if some arm i is optimal right. For that arm i, del i is going to be 0 by definition ok. So, above all these del i's, there will be a one del i i for which this del i value will be 0. So, I will just because of that, I do not need to consider that, I am only going to consider all this i's for which del i's going to be positive and then, write it ok. So, because of this now I am going to plug this value in this. So, this is going to be all i plus ok.

So, we are going to get this is upper bounded by this is sum 6 is the constant summation or all arm i's, where we sub optimally take up is positive 6 log n by del i plus summation over i. This is the constant here right. So, I can write this as well ok. Let me write just; sorry, I do not need to. So, there is a del i value anyway.

(Refer Slide Time: 05:28)



$$\bar{R}_n = \sum_{i=1}^{n} \mathbb{E}[T_i(n)]\Delta_i = \sum_{i:\Delta_i>0} \mathbb{E}[T_i(n)]\Delta_i$$

Thm - UCB pull each sub-optimal arm k in expectation at most

$$\mathbb{E}[T_k(n)] \le \frac{6\log n}{\Delta_k^2} + \pi^2/3 + 1$$

times. Then Pseudo regret is bounded as

$$\bar{R}_n \le 6\sum_{i:\Delta_i>0} \frac{6\log n}{\Delta_i} + \sum_{i:\Delta_i>0} (\pi^2/3+1)\Delta_i$$

$$\le 6(k-1)\frac{\log n}{\Delta} + \sum_{i:\Delta_i>0} (\pi^2/3+1)\Delta_i = O\left(k\frac{\log n}{\Delta}\right)$$

$\Delta_i := \max_j \mu_j - \mu_i$

$\Delta_1 = 0$

$\Delta_2 > 0$

$\Delta_3 > 0$

$\Delta = \min_{i>1} \Delta_i$

$\Delta < \Delta_i \ \forall i > 1$

And we define this del i as max of mu i let mu j for all j and as mu i right and now, and our convention was already we said that arm 1 is optimal right. We said that let us assume arm 1 is optimal that algorithm does not know; but let us assume arm 1 is optimal because of that this del 1 is going to be 0. There will be some value of del 2, del 3 like that.

So, delta 2 will be positive; del 3 will be also positive. What I am going to now define another term as delta to the minimum value of all this delta i's where del all i where delta i is positive or maybe in our case I am just going to take it as delta 2 delta i, where i is greater than 1. So, what I am doing? I know because arm 1 is optimal delta 1 is 0; delta 2, delta 3s are all positive.

I am interested in among a positive value, what is the smallest one ok? So, what is this, I am saying? Suppose, let say my mean is ok, this is my highest mean mu 1; let us say somewhere mu 2 is here; mu 3 is here; mu 4 is here ok. Let us say this is one example. I called this gap as delta 1. From here to here as delta 2 and here to here as delta 3.

Sorry, this one I called it as delta 2 right and I call this as delta 3 and from here to here I call it as delta 4 ok. What now I saying is among this delta 2, delta 3, delta 4 take the smallest one and that I am going to denote as a delta ok. So, this is for my convention, I have written this as mu 1 here, mu 2 here, mu 3, mu 4, it could be anywhere like you this could be I could has been mu 2 and this could have been mu 3.

This is just like labeling right whatever. What where I am dealt, what is this delta is telling is the gap between the highest mean value and the next highest mean value. So, this is the highest value here right. What is the next highest value? In this case, mu 2 the gap between them is what delta is denoting ok.

So, this delta is; so by definition this delta is smaller than delta i, for all i greater than 1 right; this is by definition. So, it is basically saying that what is the gap between the highest mean and the next highest mean right. This delta is exactly capturing good.

So, now, I just plug in this here. So, I am going to replace this delta i by this deltas and since this delta is a lower bound on this delta i, if I plug in I am going to get an upper bound here ok. Now, if you look into this bound, is this a sub linear bound?

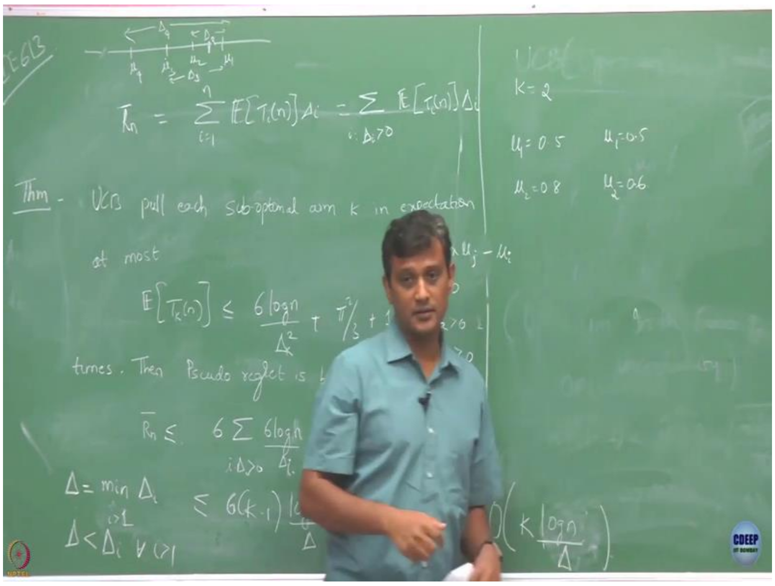
So, how does this regret go and end? It goes logarithmically end right because of that it sub linear. But if you look into the order wise quantity, what is the order wise regret? It will like order wise K log n by delta right. So, this is K minus 1, but this is constant right; I will just take it as K, this log n, this whole divided by.

So, this it should be (Refer Time: 11:24). So, the expected regret of my UCB algorithms K is like K log n divided by delta and here, this delta depends on the problem instance right. So, the problem if I have been given the problem instance which has mu mu 1, mu 2, mu 3, mu 4; this delta is basically capturing the gap between the best mean and the next best mean. So, I basically representing this complexity in terms of delta, but not all these delta i’s because delta actually matters more.

Suppose, now assume that you have been given mu 1, mu 2; these are the things and you are you are goal is to identify this mu 1. As long as there is a good separation between mu 1 and mu 2, maybe this is slightly easier. But if this mu 1 and mu 2 are very close to each other, its slightly becomes complicated right. I am saying that if instead of mu 2 was instead of this much separation my mu 2 was here.

It is very far from the optimal one ok. So, the gap between mu 1 and mu 2 is very small. Will my problem is becomes easier or harder? Suppose, let us say only take the case of 2 arms ok.

(Refer Slide Time: 13:26)


UCB
K=2
μ1=0.5 μ1=0.5
μ2=0.8 μ2=0.6
Thm - UCB pull each sub-optimal arm k in expectation
at most
times. Then Pseudo regret is
Δ = min Δi
Δ < Δi ∀ i>1

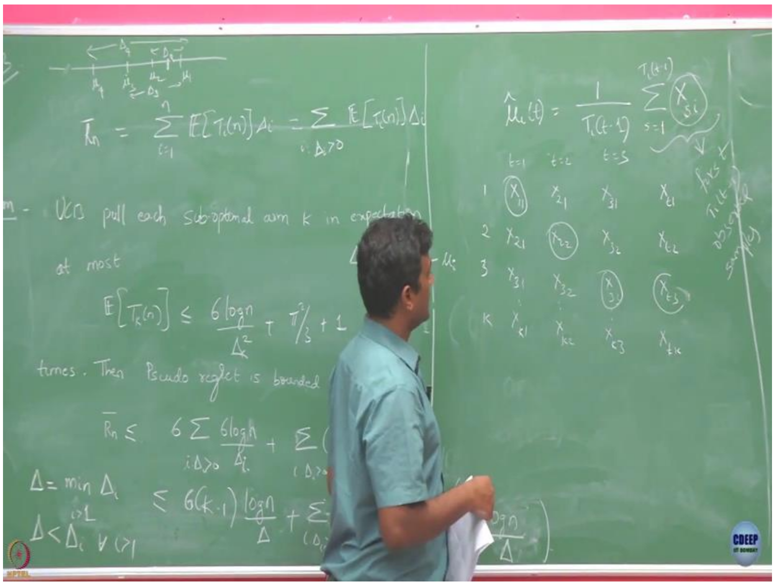
Let us take in one case, I will give an instance; where, mu 1 happens to be 0.5 and mu 2 happens to be 0.8. I ask you to identify an arm which has the highest mean in this case? Let us say you did something and identified and now, I what I do is I will keep the first one same. But make this 0.6. So, now, the gap between them has reduced right. Will you think it will be easier here to identify which one is the best or it will be easier to identify which is best here? Here, why?.

Because they are well separated right like if let us say in a class, there is only one good guy and others are all bad guys. It is easier to identify a good guy right like and I mean if you are to assign only one top grade and there is only one good guy and others are all bad guy. So, you know whom to assign this thing and if there are too many good guys, maybe that is the case we have and you want to identify whom to assign the top grade. Then, only one top grade, then it become bit challenging right.

So, the problem is all about identifying the best from the rest ok. So, if there are too many I mean there are too many people, who are close to the best; then, its separating the best guy from the rest is slightly challenging right. So, the problem becomes hard. So, that why that is why, this gap matters, what is the gap between the best and the second best the mean values of the arms ok.

So, fine. Now, the question is how to show this ok. So, for that, we will introduce some notation may be and next time we will continue it. I just said it for K equals to 2, but you can imagine right; this should hold for any number of arms. If there are too many guys, who are close to the best guy then separating them, it becomes bit not so easy task. So, and this boundary is exactly capturing that. If the gap becomes very small, your bound is also going to be large accordingly.

(Refer Slide Time: 16:14)



In the algorithm, I wrote mu i hat of t right. What does that meant? This is the estimate i for arm i at time t and what was that? Like we said that this is nothing but the sample average of all the samples, I have gotten for arm i till that point. So, this is actually T i of t minus 1 summation s equals to 1 to T i.

Because I am going to till when I am trying to estimate the value at time t, I am going to use this many samples right T i t minus 1 because till round t, I have observed my arm i, this many times right and what is this? This is going to be X times i. So, thus, how we are going to interpret? We are going to have the average of T i of t minus 1 samples to get this mean value.

But then, what is this? The way we are going to interpret it is in the these are the; so, we have arms right like let us say; let us say arm 1, arm 2, arm 3 like let us say arm 4 we have. Let this is let us say this is X_{11}; this is X_{21}; X_{31} and X_{k1}. So, this is in round t = 1 and this is round t = 2. This is going to be X_{21}, X_{22}, X_{32},.. X_{k2}. And let us say t equals to 3; X_{31}, X_{32},.., X_{k3}. This K ok. Like this you can continue like I am going now write some for arbitrary t, this is going to be X_{t1}, X_{t2}, X_{t3},.., X_{tk}. So, like this.

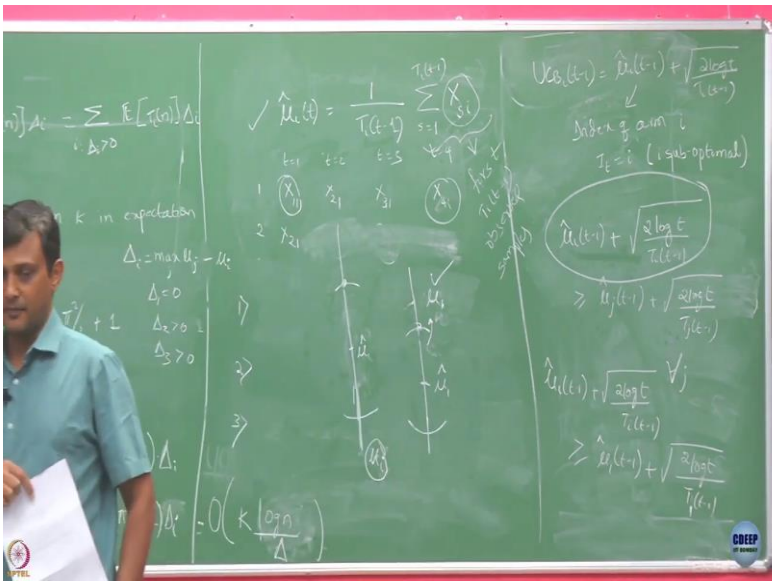
So, in had you played arm 1 all the time in round 1? Let us say the sample value, you observed is this. In round 2, sample value observed is this. In round 3, sample value observed is this and in round 3, sample observed like 2. Had you applied t all the arm number 2, this is the sample value observed.

But you are choosing right, suppose let us say in round 1, you played this; round 2, you played this; round 3, you played this. So, you have got this sample and similarly, depending on which arm you played, let us say you played arm 3 in round t, you got these samples ok.

So, it is not that you are getting samples from each arm in every round right. You are going to if you happen to play that arm in a particular round, then only you are getting samples from that. So, when I write this, it may happen that for s equals to 2, s to i, I mean I may not have played I term in the second round, that sample is not there. But what is our interpretation of this is where whichever slot you played, I do not care; but suppose, let us say you have these many samples, you have just take them and average.

So, because of that, with this the way we have to interpret this the first observed samples; is that fine; are you lost? All, I am saying is; so let us say I want to ok, let us only take only 2 arm case.

(Refer Slide Time: 21:25)



And you happen to play arm 2 in the second and third round and you happen to play, let us take this to be t equals to 4 and you happen to play arm 1 in the first round and in the fourth round. So, after this for arm 1, you have two samples and arm 2 also you has two samples. So, now, you have to you can only average for arm 1, only these 2 samples right.

So, we are saying that you are going to take this sample and this sample and averages even though, we have written that s is running from 1 to t; it is not that this sample and this sample, it is this sample and the one which you have observed till that time and similarly, here it is going to be average of these two samples here whichever you observed till that time.

Now, so that is the meaning of the estimate mu i t here. Like you have just taking average of whatever the number of samples, you have observed for that and the number of samples you have observed for arm, we are T i t minus 1. This what we have denoted as fine.

Now, how to go about the proof ok? We will just write it and we will continue discussion. Suppose, assume that in round t some arm i is played and assume that this is not one, that is not the optimal one; some suboptimal arm is played. What could be the reason that suboptimal arm could have played have been played? Obviously, it is played because the mu i hat t minus 1 plus; what is the confidence term we had here? 2 log t divided by Ti t minus 1; this happens to be larger than rest everybody.

That is why you are able to you played it. This is happens for all j. Why this should happen? Yes, because this has happened you played arm i in the ith round that is what the algorithm said right. Then, why this should happen; in what cases this could have happened?

This could have happened because or let us say and this is the case and in particular, this should be also the case that this guy is the larger than the optimal arm. So, that is mu i t and s 1 plus 2 log t T i t minus 1 should be wherever equals to mu 1 the first arm.

So, you happen to play arm i because it is better than everybody else. So, by the way, we are going to call this to be the index of arm i and in the. So, earlier, we defined UCB t minus 1 to be. So, we are going to call this index of arm i. So, it is called UCB index in this case, but in general what are the value we are going to assign to that particular arm in that round that is going to call the index of that arm.

So, what UCB algorithm is doing is basically it is finding an arm which has the highest index and it is playing that right. So, because of this what of course, the UCB index of if I_t happens to be i, its index should be larger than everybody else and in particular, its index should be larger than that of the optimal arm itself in that round. That is why and assume that this i is suboptimal.

Suppose, in some particular round let us say you did not play optimal arm and you happen to play an arm i which is a suboptimal 1. Now, it must be the case that its index is also going to be larger than that of the optimal arm in that round right.

So, this can happen. It is so happen that in that round the mu is index is kind of overestimated right and the index of optimal arm is kind of underestimated or it may happen that the exploitation term dominated for the ith term because of this happened because the ith term was not played sufficiently many times ok.

So, just to we will write it formally next time. So, what could have happened? So, suppose let us say this is ith term and let us say this is the one with optimal arm. So, their estimates are like this. It has so happen that the true value of mu 1 is somewhere outside, it is not in the confidence term or because of this and what is this; so, I have to make this, mu i star and it confidence term is going to be.

This one is this and I want I want it to be flashy and what about this mu 1 star, I have and star the confidence term is lower, mu i minus this somewhere this happened ok. So obviously, if I had pick ith arm in round t right, it must be the case that its upper confidence come here which is this value should have been larger than this point right. That is the only reason I the if I have picked ith arm, it must be happen that this should be larger than this quantity this is the upper confidence term for the ith one.

Now, because of this in this, if this has happened further, some other cases can arise right. It may happen that the true value of the mean this is my interval, where I am expecting my mean to lie. It may happen that it may not have lied in this; but it would have lined something outside this and also, the mean value of I which I expected into be in this interval that it did not lie in this interval. But it lied somewhere below even though this mu i, we are saying that this is the optimal arm.

So, its value is going to be higher than this value, but it so happened that when I estimated I estimated this to be in this interval, so it is not capturing it correctly and when I wanted to estimate mu 1, I wanted interval somewhere here; but it has so happened that it is this value has come below.

So, because of that it is so happening that this guy is exceeding this guy and I may end up playing ith arm ok. So, I am just giving you pictorially what could be the possible reason, I term has picked instead of first term ok. So, fine. We will formalize in the next class and write it as a three possible conditions.