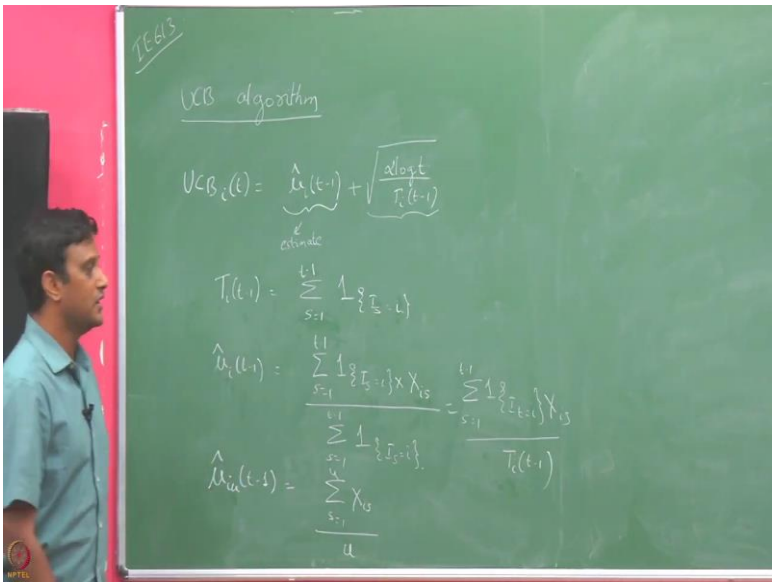
Bandit Algorithm (Online Machine Learning)
Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Lecture - 34
Regret Analysis of UCB

Hello, so, we have been discussing UCB algorithm, we introduced it briefly last time. So, this is one of the algorithm based on optimism in the face of uncertainty.

(Refer Slide Time: 00:37)



So, the UCB algorithm is said that it combines both exploration and exploitation together. Unlike, ETC algorithm which first did exploration and then did exploitation separately in a subsequent rounds, it is it combines exploration and exploitation in each step by looking into upper confidence bounds of the arms in each round.

So, in UCB algorithm we are going to so, like I said last time UCB is a like index based algorithm in which we are going to assign an index to each arm and based on the value of these indices, we are going to pick an arm. So, we are going to pick the arm with the highest index in this.

So, what is the index in the UCB algorithm? The index of the UCB algorithm for arm i at time t is defined to be mu hat of i t minus 1 plus alpha log t divided by T_i of t minus 1.

So, what is this? This we said as this is the estimate based on the number of samples of arm i, I have observed till time t minus 1 and here this is the confidence term here, which basically determines the confidence width of my estimates. And, here how did we define T i t minus 1 this is the number of pulls of arm i over, so maybe I should write it as s equals to 1 to.

So, recall that I_s here denotes the arm I am going to pull in round s and depending on how many times I have pulled this arm I over t minus n rounds this is going to give me T i of t minus 1 and based on this is just a recall of power notation; we said that this is nothing but, summation of all over s equals to t minus 1 times what is this? We said this is X i of s.

What is X i s? This is the sample observed from arm i, if you have pulled it in round s, it is not necessarily that you would have pulled it in round s if at all you pulled it that is the indicator here, then you are going to include that sample in the summation and whole of this divided by s equals to t minus 1 indicator of I of s equals to i or just to be t minus 1 and I_t equals i times X_{is} divided by Ti t minus 1.

I am also, so notice that when I said mu i hat till time t minus 1; that means, this is going to be a random quantity based on the Ti t minus 1 sample you have observe till round t minus 1 ok. So, a priori you do not know how many samples of arm i would have observed till t minus 1. So, Ti t minus 1 here is a random quantity.

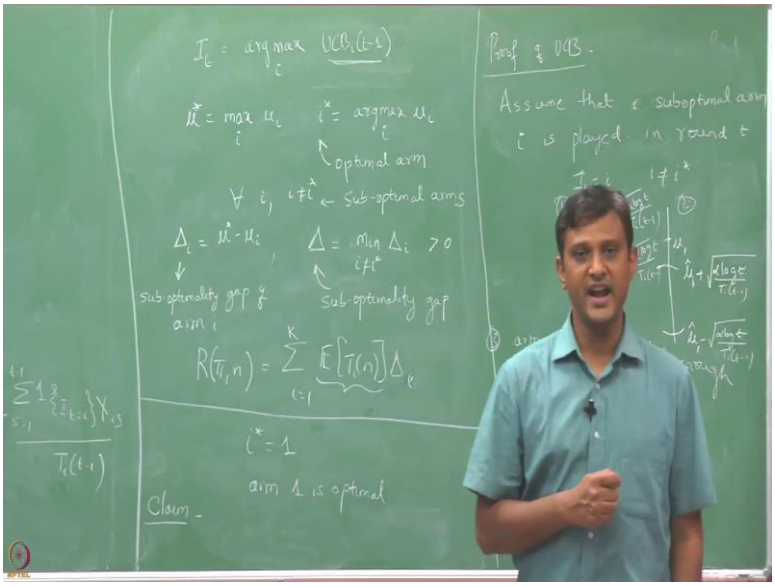
But, suppose I fixed this number, suppose let us say till round t minus 1 I said that I have observed exactly certain number of samples, then for that I am going to use this notation, may be u I will use. So, when I use this I mean to say that, I have exactly u number of samples till round t minus 1 and I am going to use them to find this estimate.

So, when I wrote this notation this is whatever be the number of samples I have observed till t minus 1 that was T t minus 1, I am going to use this to find the average, but somehow if I fix this number of samples to have been observed till t minus 1 to be u, then I am going to use this notation.

And, in this case it is simply going to be right. So, this u is this u which is I am taking as number of samples of arm i .

Now, our goal is to come up with what will be the regret of my UCB algorithm which works which selects arm like this in each node ok.

(Refer Slide Time: 06:30)



Now, to prove the regret bound of this we are going to use the regret decomposition result that we have shown last time let us recall that, we said that mu star is the mean of the highest mean we have across all the arms and we are going to denote i star to be arg max of mu i over i.

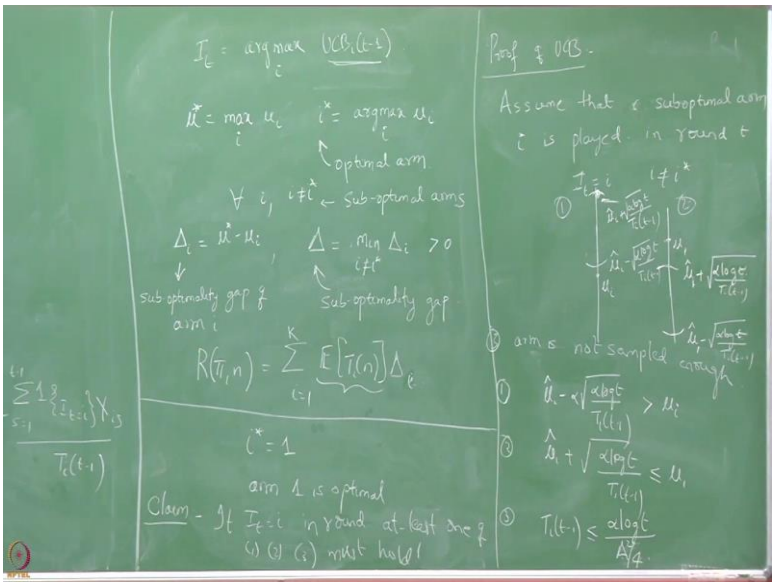
So, this i star we are going to call it as optimal arm and all other i which are not i star we are going to call them as sub optimal arms for all i, such that i we are going to call them as sub optimal arms.

Last time we also denoted this del i which said the gap between the means of the i-th arm and the optimal arm mu star minus mu i right and we also said delta to be minimum value of i not equals to i star of delta i ok. And, this term we are going to call it as sub optimality gap of arm i and this we will simply call it as sub optimality gap.

So, we also going to assume that the optimal arm is unique, because of that we are this guy is going to be positive in our case. Fine, this was the notation and with this notation we have defined that regret of any policy over n round can be given as expected number of pulls times plus a i where i running from 1 to k right.

Now, to bound the regret of our UCB algorithm, we will just try to bound what is the expected number of pulls of each arms and in particular expected number of pulls of the sub optimal arms. So, if you have a bound on this we will directly get a bound on the regret of my UCB algorithm.

(Refer Slide Time: 10:02)



So, now how to go about this? Let us start the proof of UCB algorithm. Assume that in round t an arm i is played suboptimal arm is played and let in round t suboptimal arm i is played in round t. What I am saying is I_t is equals to some i and we have this i is not i star.

Now, what could be the reason that this sub optimal arm has been selected in round i? One possibility is this estimate of this the UCB index of this sub optimal arm happened to be very high in this; obviously, if it is selected in this round i it must be the case that it is the arm i should be the one which has the highest UCB index.

Then the question is what is the reason that this arm happened to have the highest UCB index; obviously, it could be because somehow my estimates are not good in this case, let us visualize this let us say this is my mu i and this is my mu 1 by the way we are going to assume that without loss of generality arm 1 is the optimal one. So, we are going to take this i star is 1 that is arm 1 is optimal.

So, let us say this is the mean of optimal arm is here and true mean of i-th arm is here. So, now it might happen that, when we construct confidence intervals of this arms; the confidence intervals may or may not include this bounds and this confidence intervals may overshoot or undershoot this true values ok.

Suppose, this guy has been selected in round i and let us say and optimal arm is not selected in that, one possibility could be it has so, happened that this guy happen to lie my confidence bounds whatever let us say my UCB my confidence bounds let me just call it as confidence term, let me directly write the confidence term we have already discussed the confidence term we are going to use this part here right.

Whatever this term happened to be and like this and for arm it happened to be like they poor estimated is in this the confidence term of this happened to be like mu i hat plus alpha log t here alpha log t divided by T i t minus 1 and this lower term happened to be mu i hat of minus alpha log t T i t minus 1.

Because of this you see that; obviously, this arm gets a preference over this arm because it is UCB index is this. So, this is one possibility, that is the intervals for the sub optimal arm actually overshoot and they over estimated its value and it is telling that my intervals are telling that the true value of mu i's are here, but it is actually below this whereas, for the optimal arm the true value of the mean happens to be here but, my confidence interval said that that is going to line here because of this bad case I may end up missing my optimal arm.

Other possibilities that I have not sampled my arm i sufficiently enough because of this my T i for the i-th arm is smaller and this term happens to be larger and made it dominate the made its UCB index dominate over the others.

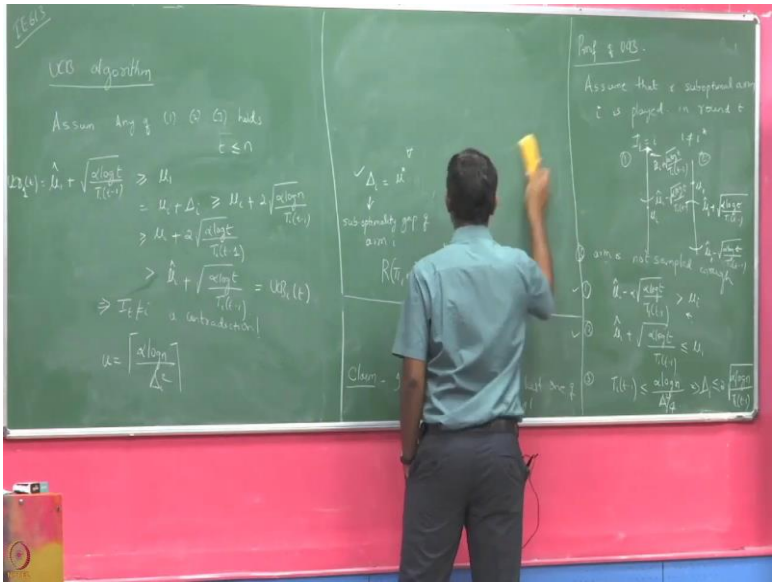
So, other possibility is that the T i at round 1 is small, is arm i is not sampled enough. So, because of this results may be let us call 1, 2, 3 we ended up possibly choosing the arm i which is sub optimal in this round.

So, putting these 3 conditions more formally. So, what is this condition is saying mu i hat minus alpha log t divided by T i t minus 1 happens to be larger than mu i and mu 1 hat plus alpha log t T i t minus 1 happens to be less than mu 1 and let us say and this guy T i I square by 4.

Now, the claim is if suboptimal arm i has been played in round t that it means it must be the case that at least one of these 3 condition must hold. If I t is equals to i in round t at least one of 1 2 3 must hold.

Now, let us see why this claim holds true. So, how to prove this claim? We are going to show that suppose, none of them holds true none of this condition hold true then we are going to argue that then it must be the case that I_t should not have played in this round t ok.

(Refer Slide Time: 18:24)



So, we are going to assume that neither a neither 1, 2, 3 are true, that is assumed then we will show that that means, I_t that is cannot be i in round t. So, to show that lets assume that first let us say this condition is violated; so that means, I have mu 1 hat plus alpha log t divided by T i t minus 1, so by the way this is not t here we will take it as n, n is what n is the number of rounds this algorithm is played for. This guy is taken to be mu 1 and now our definition of sub optimality gap this should be equals to mu i plus delta i right.

So, we have taken this case we are assuming that this does not hold. Now, let us take that this condition does not hold, because of this so, let us just simplify this what does this imply? So, if this condition here implies that, delta i is less than or equals to 2 times square root by T i t minus 1.

So, if this condition is violated we will take the opposite inequality and that will give me 2 times; so, that will give me a lower bound that is alpha log t divided by T i t minus 1.

Now, let us appeal to the third condition suppose, this also is violated then I am going to take the opposite inequality of this that will tell me, if I just take this guy on the other side it says basically mu i plus this quantity is lower bounded by mu i t. So, just appealing to that, this is going to give me mu i of hat plus alpha log t by t 1 t minus 1

So, what is this now? We have started with the left hand side which is UCB index of 1 in round t and now we have ended up with so, this is one of them we have to take it as strict inequality right and we have showed that this is equal to UCB of i n round t.

If none of this 3 condition holds what you have just demonstrated is the UCB of arm 1 is going to dominate that of UCB of index of i; that means, I_t i cannot be the I_t the arm played in round i cannot be i right, because anyway arm 1 is going to dominate that it would have been picked if nothing else right.

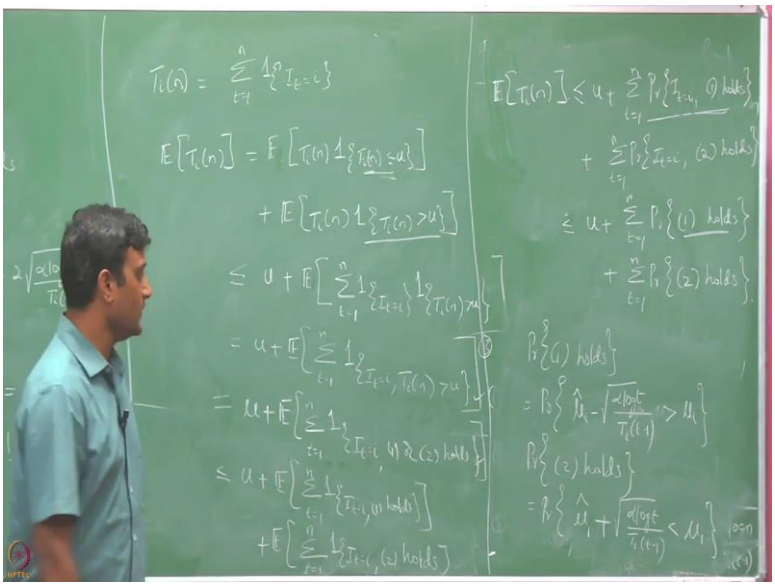
So, we do not know at least for sure we know that i cannot have been played in round i. So, this implies that I_t cannot be equals to i ok, so a contradiction.

So, now it is must be now we have just proved that, if arm i is played in round a sub optimal arm i is played in round t it must be the at least one of this must hold ok. Next, based on this we are going to bound now expected number of pulls of arm i, so before that I am going to define I am going to take u to be alpha log n divided by delta i square.

So, just so we have made a small mistake here right. When I applied this bound here, I should first get mu i times 2 alpha log n divided by T i t minus 1 here because, this delta i the way is defined here it depends on n here not t.

But, we know that we are considering this in round t and in round t is less than or equals to n; so, because of that this inequality also holds and rest of the things are fine here. Now, we defined u to be alpha log n by delta i square which is a constant right, like because alpha is some constant, n is the number of rounds which is again constant and delta i is a sub optimality gap that is a constant now. So, this quantity need not be integer, so we can take it to be simply the ceil of this, but I will not necessary light ceil all the time.

(Refer Slide Time: 25:02)



Now, how to bound the expected pulls of arm i? I am going to now write this by definition of this is going to be summation t equals to 1 to n indicator of I_t equals to i right. Anyway, I am interested in the expectation I am going to write my expectation of Ti of n; I split it into two parts: first part is expectation of Ti of n indicator that my of n is less than or equals to u plus and the other part is expectation that my Ti of n is greater than or equals to u.

Notice that, this when I said in round t if i term is played all at least one of these might hold and that round t I could arbitrarily take it any round right between 1 to n.

So, I can take that that t to be the last round here and now I am going to use that logic here, I know that because of this indicator here that Ti, n is less than or equals to n this first expectation I can upper bounded it by u right what is u here? The u is the one which I have defined here plus the remaining expectation, but in the remaining expectation let me substitute the terms for Ti n which is defined to be Ti 1 indicator I_t equals to i times this another indicator.

Notice that, this indicator here it does not depend on time here. So, simplifying this I am going to write it as expectation of so now, this is indicator consist of indication is based on this joint event here that is i i-th arm is played and number of pulls of i are till round n is greater than u.

Now at this point we are going to now appeal to this claim that we have made and proved. We are saying that, if at any point arm i is played right and T of n is greater than u, that is this condition is violated here for the nth round. So, as I said this is true for any n right. Now, we are saying that if this is the case it must be true that if this is condition is violated then other at least one of them other should hold right.

So, because of that it is going to be u plus expected value of integer 1 to n; integer I_t equals to i, if this is not this Ti n is greater than u, then at least this or this should hold right either 1 or 2 holds.

Now, applying union bound I am going to write it as u plus expectation now, equation t equals to 1 to n indicator that I_t equals to i comma 1 holds plus expectation that equation t equals to 1 to n indicator that I_t equals to i 2 holds.

So, I have basically simply applied the on this union bound on this indicators right. So, indicator of I_t equals to i and 1 or 2 holding is like I t and 1 holds and I t or I t and 2 holds.

Now, pulling per expectations inside the summation and applying this expectation on the indicator you will get that, expected number of pulls of Ti n upper bounded by u plus summation of t equals to 1 to n; probability that, I_t equals to i and 1 holds plus t equals to 1 to n probability that, I_t equals to i and 2 holds.

Now, further this is the joint event. So, if I just skip the I_t equals to i part here, I will further get an upper bound here with that summation t equals to 1 to n probability that 1 holds and plus t equals to 2, t equals to 1 to n that probability that, 2 holds.

Now, let us recall the definition of what was condition 1? So, condition 1 told us that, the confidence interval of the i-th arm it overshoot and or over estimated that mean value of i, so let us write it separately holds is same as saying that is probability that mu i hat minus alpha log t divided by t minus 1 is greater than mu i that round and probability that is equals to 2 holds is same as probability that mu i 1 hat plus alpha log t divided by t i t minus 1 less than mu 1.

Now, these probabilities we already know how to handle right, because this is nothing but, the estimate and if you are to going to take the mu i left hand side this is the difference between the estimate and the true parameter being away from this quantity.

And, similarly this quantity also we know how to bound this is for r mu i hat and this is the estimate in the error of mu 1 hat. So, both of them we know how to bound using our concentration inequality for our sub-Gaussian distribution. So, I am going to bound give a bound for this and similarly we are going to get a similar bound for this.