NPTEL

NPTEL ONLINE COURSE

REINFORCEMENT LEARNING

Concentration Bounds

Prof.Balaraman Ravindran

Department of Computer Science and Engineering

Indian Institute of Technology Madras

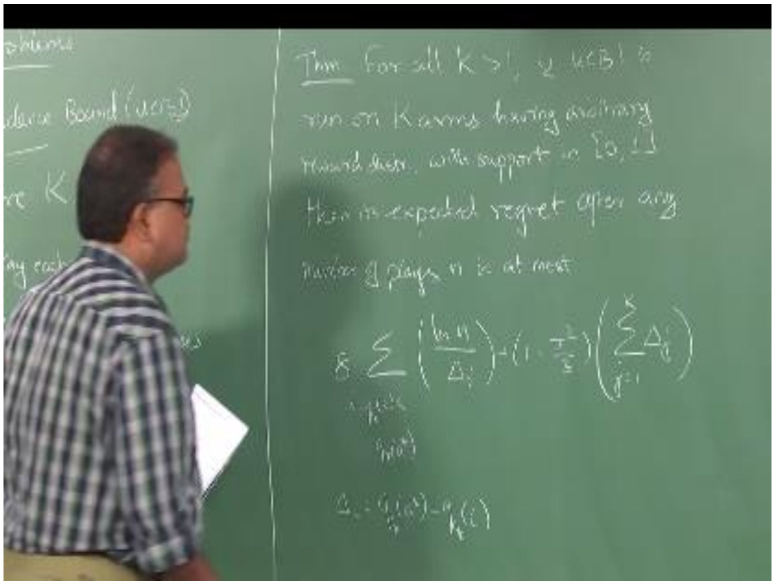
So this is interesting because we can have a nice theorem which says that, so this particular form of UCB is actually called UCB 1 right, and in the literature if you generally see you see be without any qualifier Center to it usually means UCB1 okay, there are many variants on it we are not going to look at all the variants that is like normal UCB then you see be revisited you see be improved UCB2 right so, there are many many variations of you see be right and each one gives you a slightly better critical results than the one that I am going to write down right, and sometimes involving exponentially harder analysis.

(Refer Slide Time: 00:35)

ndit Problems
per Confidence Bound (UCB)
here are K arms
it Play each arm once
p Play arm j that maximizes
$Q(j) + \sqrt{\frac{2 \ln n}{n_j}}$
Thm For all K > 1, if UCB

UCB1 is fairly trivial compared to the other algorithm set that were proposed later and just want to give you a flavor of the kind of work that is done in this area right, so I am NOT going to get into the other things but we will certainly put all those papers up on Moodle, if people want to read they can read those. Another thing about UCB1 is that we can have arbitrary reward distributions so, some of the other algorithms that are out there assume say Bernoulli rewards and if for showing their results and so on so forth the right thing about UCB 1 is arbitrary well.

(Refer Slide Time: 04:16)



Thm For all K > 1, if UCB1 is run on K arms having arbitrary reward distn. with support in [0, 1] then its expected regret after any number of plays n is at most

$$8 \sum_{i : q_*(i) < q_*(a^*)} \left(\frac{\ln n}{\Delta_i}\right) + \left(1 + \frac{\pi^2}{3}\right)\left(\sum_{j=1}^{K} \Delta_j\right)$$

$$\Delta_i = q_*(a^*) - q_*(i)$$

So capital ∆I is Q *A* - Q* essentially this means the loss that you will incur if ,you play I is expected loss that will incur if you play I instead of playing A*, remember what A* is the max over A of Q * a right so, that is a star max A of Q *A is a A* right s, they will always denote by a start the optimal action right so, what this is tell you so the regret right this upper bounded by eight times ln n by ∆I the summation running over all the non optimal arms ,why all the non optimal ops? Because ∆I is 0 for optimal arm, so I do not really want that to figure in the summation okay, plus 1+ л^2/3 that is a magic constant okay, times summation over all j ∆j right.

And here i am being little sloppy i have to exclude the optimal arm here also but i do not because it is 0 anyway, ∆j 0 anyway so, I mean symmetrically just as I have the summation ER I should have the summation there, but I can be a little sloppy and I can write it like that because, i do not have worried about the additional 0 terms I'm adding it okay, great. How did we get from here to there or more importantly how do we get this expression? I so, before we go on so I want to introduce a little bit of notation here that is a term that tells you the number of times it is a random variable that represents the number of times I have played arm I in the first n trials okay.

Number of times I have played out my in the first ten trials that makes sense, so then i can write me regret after n trails as summation over i expected value of that times ∆i, people get that so they expected the regret that I expect to have after n time steps is I will take each action, look at the number of times I, would have played that action till that time right, multiply it by the regret of playing that action right. So if i play the optimal action then the regulator is 0,so if i play some sub optimal action then i will be adding that ∆i times the number of times i have played that sub optimal action I right.

So this is essentially my definition of regret so i have a nova formal way of writing down my definition of regret, this is clear any questions? Okay. so one more notation is one thing which I yeah so one more notation i will have here, it is a random variable that denotes the reward that i obtained for playing action I at time n, some n ok some arbitrary n time n if I play action I write what I see the word that I get right? since we are assuming everything is stationary right so the expectation of this will be, this will tell me people are following it or not, so what is Xin the reward I get for playing are my a time n right what is the expected value of that somebody say something.

One of you put your hand up and say something, so that i can very hard for me to make out in that murmur and hand, Q*I, people see why since we are assuming things are stationary this expectation is really not going to depend on n, so then it becomes expected reward I get for pulling our my which is essentially Q*I okay, great. So anything else that I need to do notation wise so, essentially to show this result right, what I am going to try and show is that, this guy is bounded and I am going to show that that k is bounded so what I am going to show? We're going to show that the expected value of Tj(n)or Ti(n) is bounded by 8/ Δi^2 ln n so, since i multiply by another delta here to get the regret so that will get me the first term in the regret and some constant here which will get me the second term the regret. Expectations of the number of times plane and multiplying it by some value Δi which is calculated at the end stage right, it's not calculated at the n^{th} stage it's calculated by using their true values that's Q* you do not know the values but it's just a definition is a quantity that's the expected regret that you are going to get for playing arm I, this is the expected loss you would accrue for playing arm I.

If I had played the best arm over and over again right, if I played the best term over and over again I would get Q*A* if you play arm I am going to get Q*A right, so the loss i am going to get this Q*A*- Q*A right, so this is what I am denoting by Δi this is the loss that they expect a crew for playing arm I, let us think that one norm right and if we keep playing the arm repeatedly the total loss i am going to accrue is Δi a times the number of times i will pay i write n dice on this overall the arms. So one of these arms in the summation or one or more of these arms and we do not know one or more of these arms in the summation would be optimal.

And for those terms delta i will be 0, so the regret require the contribution to the regret will be 0 and all the other sub optimal arms will contribute Δi to the regret. Because I am assuming the reward distributions are stationary, I do remember I said you are deciding the by tossing a coin but you do not change the coin right because you do not change the coin does not matter when they toss the coin the probability of it coming up heads will be the same whether it toss it the first time or they toss it at the tenth time I pull they are the probability of it coming heads will be the same and that is the expectation, right so the expectation will be the same right.

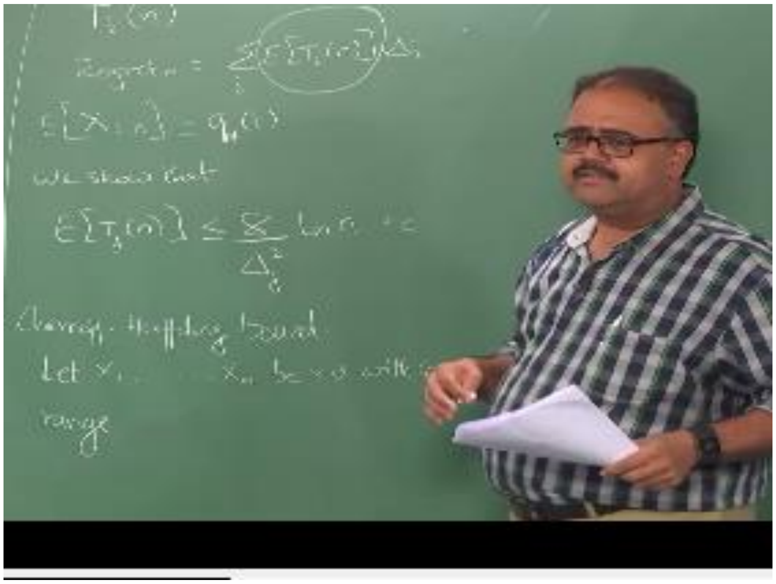
so one way to think about the X this expectation is right so, I do multiple experiments and I do several experiments every time starting from time 0 right, I do i do like the millions of experiments every time I start from time 0 and I keep pulling arms for some number of time steps and then i reset my whole system start from time 0 and pull for another say 100 time steps or something like that right when i do this experiments like this in the 10th time step if I pull on three right so in this million experiments some number of times I would have pulled on three in the 10thtime step okay, then I'll take all the rewards I saw then I'll take the average of that right that is one way of thinking about what this expectation means.

So now you understand it does not matter whether I pulled on three at 10th step or whether a pull on three at the thousandth time step if i take this expectation across these million trails, it will be the same. So that is essentially what we're saying here okay, is it is it clear so what we are trying to show is this right so if you are able to show this then we are all done as long as this constant here evaluates to 1+л^{2}/3 here proof is all done okay. Now we have to now go and start counting that okay, so here things start get getting interesting so you want to complete the proof so, before I move on with the proof i will need one more i need one more fact, so they're on this whole set of results in probability theory called concentration inequalities or concentration bounds.

They are sometimes called large deviation bounds okay, these are essentially results that relate the true expectations of distributions okay, with estimated values of those expectations from samples not this expectations in different kinds of statistics you have different bounds you have bounced on expectations you have bounds on variance right and so on so forth, right so essentially the idea behind these bounds is to kind of characterize right so, given a certain number of samples from which you are estimating a statistic and the statistics in this case will for the one that we are interested in is the expectation right.

So we are interested in the expectation as a statistic right, so I have some number of samples from which I makes estimating this expectation right and i have the true expectations right so as the number of samples increases how quickly or how slowly does this estimation approach my true expectation so there is something that we want to characterize so we will talk about one such bond which is a very popular plant which applied all over the place okay, so if you are going to do anything in machine learning forget about RL anything that has remotely shades of theoretical results in machine learning it's a good idea to know more about concentration bones but at least the least you should know as a bad thing sure enough bonds or sheriff of bonds.

(Refer Slide Time: 17:20)


$T_j(n)$
Regret n = $\sum_j E\{T_j(n)\}\Delta_j$
we show that
$E\{T_j(n)\} \le \frac{8}{\Delta_j^2} \ln n + c$
Chernoff-Hoeffding bound
let $x_1 \ldots x_n$ be r.v. with
range

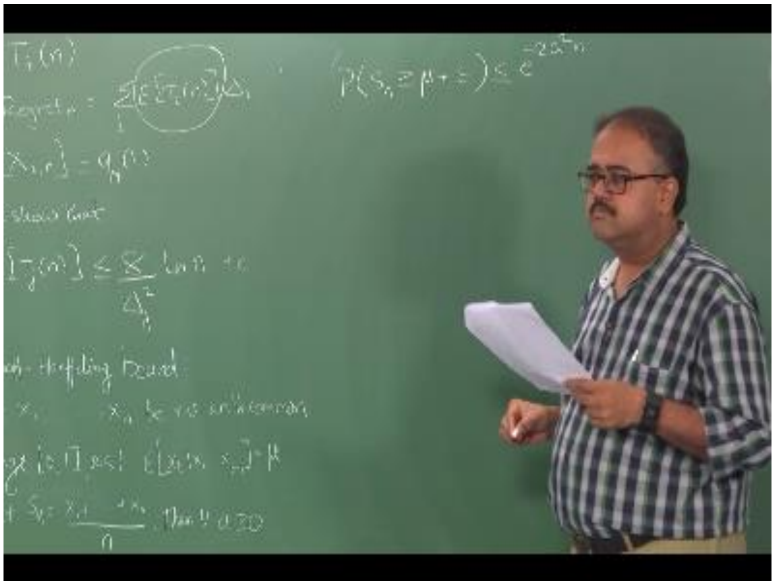
So this is a very specialized form of the bound I am writing at very narrow form of the bond i am writing specifically for this setting that we are interested in right the Shernoff bond is slightly more general than what i am stating now right so, you can look it up I mean so you can look up I don’t know your favorite online resource at Wikipedia or what a mathematical one of those things right so it will give you a more expanded bound. So you can see the setup here I am assuming that x1 to xn are random variables right each one with a range 0 to 1, that is kind of putting it here and if it is not in the range 0 to 1 you will have to do some kind of normalization in the bound right, but for the time being let's assume it's in the range 0 to 1.

And such that the expected value of XT right given all the variables that came before that is μ and then this holds for all T basically it, essentially means that all of the variables have a mean μ, all the variables each one independently x1 is a random variable with mean μ and the range 0 to 1 X 2 is a random variable with mean μ in the range 0 to 1 okay. So essentially that is what we are assuming and I am defining another random variable which we call SN which is essentially summing up all the X 1 to Xl and divide by n, okay so get that.

So what I am I trying to estimate us with SL you right and try to estimate them is μ so, one right now let us try to make this little concrete in our case so, that it becomes easier for you to understand right so x1 to xn or random variables that are corresponding to different times, I have pulled the same arm I right so I have pulled arm I n times right, so every time I pull the arm I am going to get some reward from the range 0 to 1 right, since I am assuming it is a stationary distribution all these rewards will have the same mean μ ,correct? And SL is the average right this SN is essentially make you j and if you think about it so, QJ is like one sample of my random variable SN, is it what I mean by saying QJ is a sample of SL because, they said is a random variable right what is QJ it is one realization of that random variable, because I have taken actual samples and taken the average it is one realization of that random variable right.

And what I am going to now tell you is okay given that i have taken n samples so, in our case it will be NJ samples right given that i have taken NJ samples, how can you relate QJ with what how can I relate QJ with Q *J okay, so given that I have taken NJ samples which is my n here how can relate QJ with Q *J okay. So this is how the hafting bound is applicable in our case right so i will write down the expression i am not going to prove it okay, so that will get little bit more involved write down the expression and after that we can apply it in there deriving those bonds okay.

(Refer Slide Time: 22:26)


$T_j(n)$
$P(S_n \geq \mu + \epsilon) \leq e^{-2\epsilon^2 n}$
shown that
$\frac{8}{\Delta_j^2} \ln n$
Hoeffding Bound

So essentially what i have here is now the probability that SN greater than or equal to μ plus some ε, where it will do a here but because a is action right sudden want to confuse A for action with something else yeah, this is what this is a problem with translating on the fly okay, fine so it should be n or n^2? o what does this mean the probability that my SN will be far away from my μ in one direction so $\mu+\varepsilon$ is some use a true mean, SN is estimated mean the probability that SN will be greater than at least ε from μ right is lesser than $e^{-2\varepsilon 2}$ right, so the smaller the ε , what happen? Smaller the probability or larger the probability larger the probability smaller the epsilon larger the probability right so larger the probability of me making an error, I mean if I want to be really really confident then I need more and more samples right.

If I want ε to be very small then n has to be very large only then I will get a high probability format right what about this it says in the other direction probability that the estimate I make will be far away from ε, basically below ε right is again bounded by the same expression.

IIT Madras Production

Funded by

Department of Higher Education

Ministry of Human Resource Development

Government of India

www.nptel.ac.in