**NPTEL**

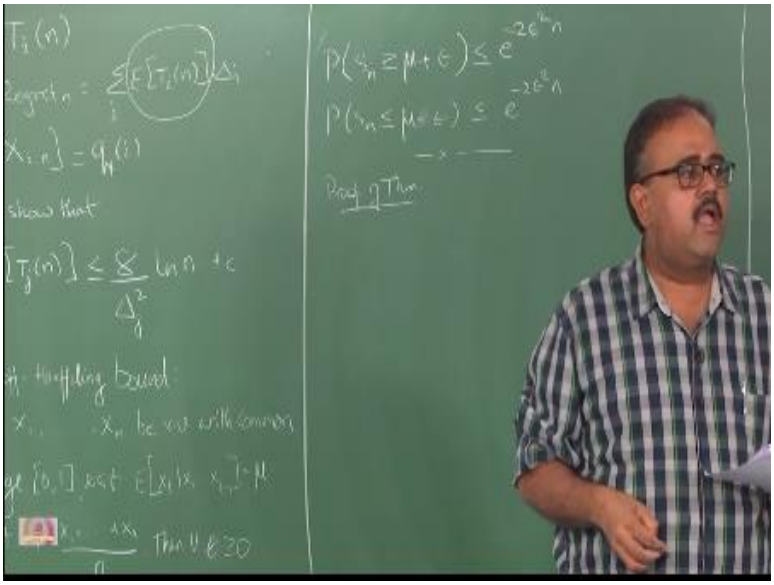
**NPTEL ONLINE COURSE**

**REINFORCEMENT LEARNING**

**Theorem 1 Proof (UCB1 Theorem)**

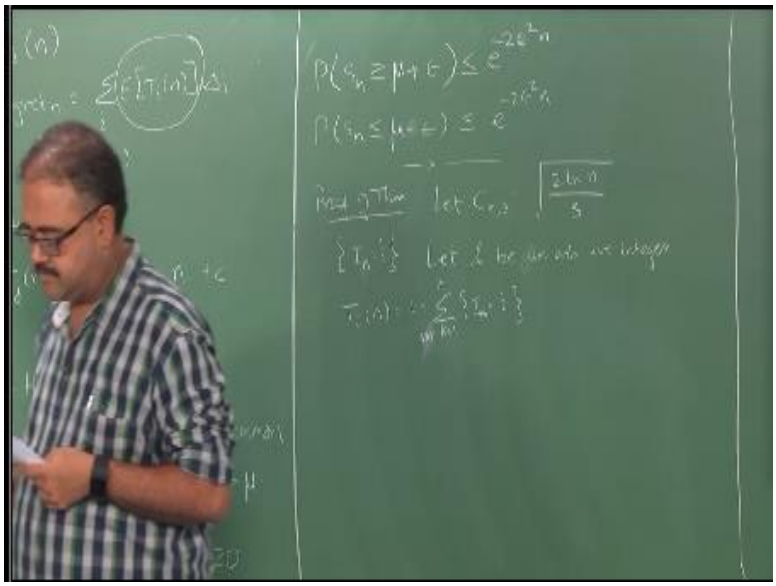
**Prof. Balaraman Ravindran**
**Department of Computer Science and Engineering**
**Indian Institute of Technology Madras.**

(Refer Slide Time: 00:23)




Now you remember but how many class we have for branded algorithms 4 you will try okay fine because the rate at which I am going I will probably need one more class just to finish the two proofs I wanted to do today so which should mean I would need one more extra class for finishing all of brand is did we do 4 class last year 6 yeah I think I we did we had visiting faculty you were talk okay. We have oh good fine then I am not too bad of okay.

(Refer Slide Time: 01:27)



Okay since of writing the compression 2ln n/s all over right so I am defining that as Cns okay then if I write Cn-1s it is ln of n-1/s and then I can do Cn, s + n so on forth right I can just like a function just substitute that make my life little easier right so the goal now as I mentioned earlier is to upper bound this expectation right so I will never play a suboptimal on more than this many times okay that is necessarily what we are try to show okay so one other notation I will just adopt the notation used in the proof here even though we introduced a different notation in the previous class where the same thing get makes it little easier for me.

This notation indicates a I mean it denotes a random variable corresponding to that indicator right so this is a random variable whose value 1 if $I_T$ = I whose value sis 0 if $I_T$ = not equal to I okay so whatever is this I will put in some kind of events of here later right so if $I_{T=}$ I then what does it mean it means that okay fine I apologies for the confusion this yeah so when I say In = I okay it means that at time n I played MI right so if this the hole expression within curly braces will be one if at time at n I played MI it will be 0 if at time n I played arm not I okay so there are couple of things which we can observe about this so at any instant I In = I will be 1 for some I for everything it will be .

Right because at very time instant I can play only arm and at every time instance In I will play at least 1 arm so it will be 1 for 1 of the arms so everything it will be 0 okay this yeah the paper is not really the most accessible of papers out there right so if you are going to read it on your own you will have to do some work okay so these are not I mean they just assume that to your are all really smart people reading the paper so since some of the gabs that you will have to fill in yourself right I am just doing one of the proofs in the papers so if you want to read the rest of the paper and understand all the theorems and all the algorithms that they describe in the paper.

So you will need to some additional work okay so L is some arbiter positive integer so I start by saying this t as a running variable okay so the number of times I played arm I up till the $n^{th}$ instant will be one which I get from the initialization + k+1 obviously I could have played that ram again in the first K times right in the first k times I will play each arm only once right so am I also I would have also played once so in the first skate trails the $\Sigma$ will be just 1 so I just have that 1 there and from the k + trail till the nth trail I add the number of times this variable is 1 what does it mean number of times I have taken arm I rights so I M – I will be 1 when ever take arm I so I just sum this up so that so that the number of times I played arm I is what I get right.

Is it fine this expression is fine okay this is the only directly inductive path that we have here soon after this whatever you do or these are standard tricks that people play to bound random variables okay so to actually tell you about the general techniques all of this I recommend you take the ATTCO course because about third of the course is all about how to do proofs like this right how many of you have ATTCO already advanced techniques in theoretical computer science theory of computation.

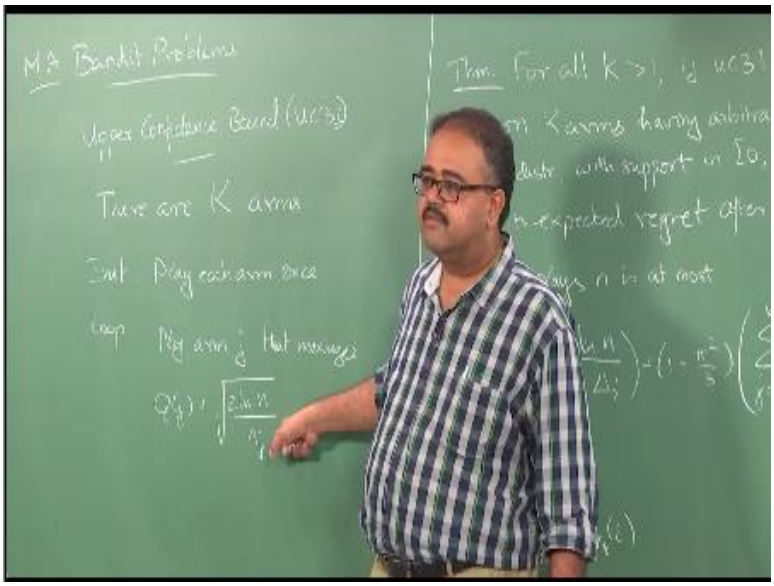
Okay that is a course that Jayala Srama teaches in the CS department so it is very relevant to lot of ml reaches because more lot, machine learning has become like randomized algorithms right so if people teach you how to analysis randomize algorithms you can take the same ideas and you algorithms machine learning algorithm as well so lot of the tools that are developed there are useful here and many of the things that we do here right essentially have their roots in the kind of analysis that people do for randomized algorithms okay.

Any way we so we will start by doing that so I have an orbiter number l okay I say that okay fine let us assume that I played that arm L time okay then let us start counting right so assuming that up till the M- 1 step okay I have L or greater than L right only then I will count this sample so at some point I have discovered the I have pulled arm I okay should I add it to my count or not what I do I go head and look at the cumulative count so far if the cumulative count is L or higher I Add it to my count if it is below L I do not add it to my count because I am any way starting count from L okay.

So this is slightly different way of doing that this I will star the count from some arbitrary number give that I have taken that arbitrary number already right so later on we will since now are going to bound this expression assuming that at least L time I have taken the arm right and then we will drive a bound for L so that we get whatever probability we want okay so that is idea behind doing this so this I can now derive a bound for this assuming I have done at least L time right then I will go back and find what is that value of L so that I get whatever probabilities I want right whatever grantees I want I will go and find that L okay that essentially we are going to do here.

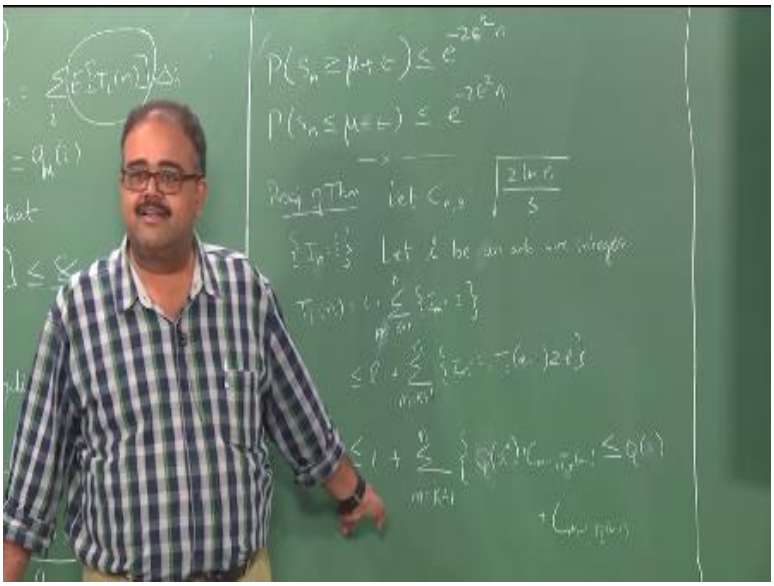
Okay I just need more notation here write this let me try this any way so when am I going to take arm I right when will I take arm I when arm I has the highest right arm I has the highest value for this quantity right when will I take arm I when in that as the highest value.

(Refer Slide Time: 12:58)




For this quantity among all the arms specifically I will take arm I only when Q+ C so that root term is now C right that root there is what I have written as C so I will take arm I only when Q+ that C term is greater than Q ( A* + the C term corresponding to S term.

(Refer Slide Time: 13:30)



Yes that is why I have lesser than or equal to here see at especially this should be 2 right in fact it as larger than everything else but at least it has to be larger than S * right so it has been every other action but like he lightly pointed outer so this is the larger set there when I will actually be taking action but that is okay because I am trying to upper bound it right I am trying to give an exact number because trying to give exact number make it two involve right I am just trying to give an upper bound so I'm saying very specifically I am looking at the best arm I am saying that this should be better than the best arm.

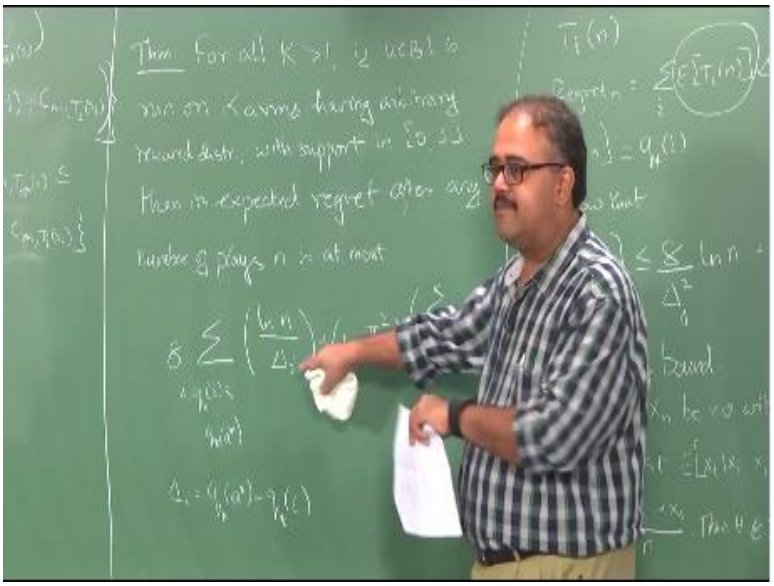
At least right so anything else we need have will still have this right that condition is also there so will but that conation back in and what does this curly bracket mean it is the curly bracket is a random variable that indicates the value of 1 and the expression within the curly bracket is true it will have value of 0 and expression with the curly bracket is falls so essentially this summation counts how may time this in equality become true after I as been taken L times nice so that I what we did here right so this counting is whatever we are doing here is after L times I has been taken because L I have added here as a count so like vise I am counting the number of times is in equality becomes true after I has been taken L times.

So this is essentially what I have so far right. Back here this entire thing take another seat vacant here if you want to move as well you can see everything you guys are fine right no not to complain because I will tell you there is  row vacant here no cannot you see what I have warren or do not understand what I have written well sun scripts become small right so right then large then it becomes confusing so that is Cm- 1 for people who cannot read it I am reading it that Cm -1 , T A*M-1.

That is basically the number of times I have played A * until the M-1 nth point right note that I do not have to know this because I do not know what is A* right this is just the there but you know the there exists A * right and I know that my I as to have better whatever into that quantity then A* let so it is just that right I do not know I do not need to know which is A* okay just do not get confused by the fact that I'm using A* in the proof right did you get that CM-1 , TA*M-1 and this CM-1 , TI (m-1) no A* is a true.

Bets action which you do not know anything about right for reader we have to compare again so true best action right so we are comparing against the true best action right okay so far so good people can who could not see heared me right and people understand what is going on right so now we will I'm shifting over here.

(Refer Slide Time: 17:47)



Writing one line I am talking for 10 minutes man really we too have an additional index for the Q-function for us to say this things, right. What I write there is M-1 so the Q-function at step M-1 okay I really need that additional thing because Q keeps changing over time right so I need to know at what stage I am talking about the Q-function, so at that point I am talking about the Q-function at the M-1 stage, right. Remember that the Q-function at the M-1 stage can be the same as the Q-function at the M stage if I had not taken the action previous.

At the M-1 stage if I am not taking action A for action I then at the M stage also the Q-function of action A will be the same it will interchange, right just you have to be careful about how you interpret the index and I hope I also do not mess it up when I do this lecture. No so we do read the different interpretation for and I do need an additional notation I follow this for that because we need to what we need to insert into this bound here.

Is the number of time I have taken A* so far, right what I need to insert into this one, is the number of times I have taken A* so far which is TA* M-1 right and so it become a little combustion now this can be is this will be TA* or X okay fine we can get away with that. If it is

a little weird expression so what I am saying here is okay take the minimum of this expression okay overall time instant 0 to end.

Take the maximum of this expression right for all time instant L to M okay and look at the number of times one will be had on the other, right. This will be on over count of that event this is clear right but here I am talking about this quantity being higher than this quantity at that particular instant right I am talking about some time M-1 where the quantity for A* was lesser than the quantity for write.

Talking about sometime instant M-1 at which this happened here what I am talking about, okay at a time M-1 right I look at all the A* values I have had in the previous time step okay, I will take the least of that, right and I will take all the I values Qi values I have added in the previous time step I will take the maths of that okay.

If the maths of this is greater than the min of that then I will count that has 1 right so can you convince yourself that, that event will certainly be counted right because at that instant the values already greater so obviously the maths will be greater than the min so if this is not the Min right the Min should be lesser than this and if this is not the maths then the maths should be greater than this.

So if this value is greater than this value then the some quantity greater than this value will certainly be greater than some quantity lesser than this value right you can see t that right, so this is just the over count of that event, right why am I writing this over count? Because it makes it easier for me to sorry now that is the event right the curly brackets makes it easier for me to write a expression.

Yeah what is it, one second I will come let me finish yeah 0 <S<M L<= S less than because L because I am I have that condition there right I have to I have to take this atleast L times, so the first L time what happens I do not worry so I will remove that right so only after that I am worried about, but this will be an over count okay this will be an over count of that. So little tricky here so know the proof actually assumes that at this point, okay.

I have taken this at least S times not TES or S times okay the it is mainly because this is the notation translation problem but we will see we work with this and then I will probably adjust the results later, right. But we are reading the proof I am just telling you there is a mismatch in the notation that we have used so far, okay. But this is the neat trick that we have noted, right. We have taken rid of all the Min Max and everything, right,

So what are we doing here so assume that the Min occurs at some index right let us call that index I do not p1 let we assume that the Max occurs at some index say P2 right, so now what I am doing here is I am summing over all possible indices for A when I am summing over all possible indices for I, correct. So the combination P1, P2 will occur somewhere because I am summing over all possible indices, right.

At some point I would have use P1 for S and P2 for SA, correct because I am summing over all possible indices right so what was P1, P1 was the one for which this minimum was achieved P2 was the one for which this maximum was achieved, so when I am summing over S at some point S will have P1 at some point Si will have P2 right, at that point this expression will surely be true because I know from the Min and Max argument that we have earlier that will be true.

So I will certainly be counting this right I will be over counting I will be counting other things also but I will certainly be counting this, so essentially so far we have not lost anything we are essentially still be maintaining at as Upper bond okay, so we started of with this event right now we have done a lot of simplifications and kept loosening the bound to this was a strict equality and there was nothing here, right.

This was strict equality no approximation here, so from there at every step I am essentially counting a larger and larger event, now the thing is I have come to a point where I have a simple enough expression right that I can try to bound directly now, okay. So you can start thinking know this kind of should lead you to the shun of opting bond I can use the bound shun of opting bound to try and bound this expression.

This is simple enough expression, okay. So this is true okay this is true when can this happen? So my claim is this can be true if one of several conditions hold right, so this can be true if, okay this can be true if I am either I am grossly under estimating my optimal actions value, right. So this is my optimal actions estimate now this is the true value right so the true value minus the confidence bound.

Are other way around think about it, right. This Qs(A*) plus the confidence bound is less than the true value that means I am grossly under estimating the true value that is one thing right. So that could be one condition under which this could be true or it could be another condition your grossly over estimating the $I^{th}$ value you are getting it, so either I have to make a big mistake for my Q*, A* or I have to make a big mistake the other way for my I right you see that, right.

So I have to either come very down from a true Q* value I mean a Q*, A* or I have to be much above my normal value otherwise this is not going to happen, right. So second condition is, okay. Or when can it happen without me grossly over estimating or under estimating it? When the two values are close right if the two values the A* value and the I value are close to each other then I really do not have to gross and under estimate these things.

So it is possible that I still could make errors, right. So essentially the other condition is right so I have my twice my take the confidents interval right so I have a confidents interval add here I will add from that side, right. So I will add a confidents interval right so I am taking twice the confidents interval from my lower value right, from the lower value I take twice the confidents interval.

In the topper value there is optimal actions value lies within that right then I can make a mistake without making a gross error also, right I simply I am saying that these two are close together, right. And close is a relative term close is defined in terms of the current uncertainty I have, right. Because this is going to change every time I take an action this is going to become smaller and smaller, right.

So at the beginning of the learning this can be fairly large so I can still make a mistake right even if the means are slightly far apart, right. So this is the 3 conditions one of this at least will have to be satisfied for these event to become true, second one says when second one is easy man, so the first one says okay I have my Q(A*) and I have my confidents bound this is the first one okay and the true Q*, A* is there, okay.

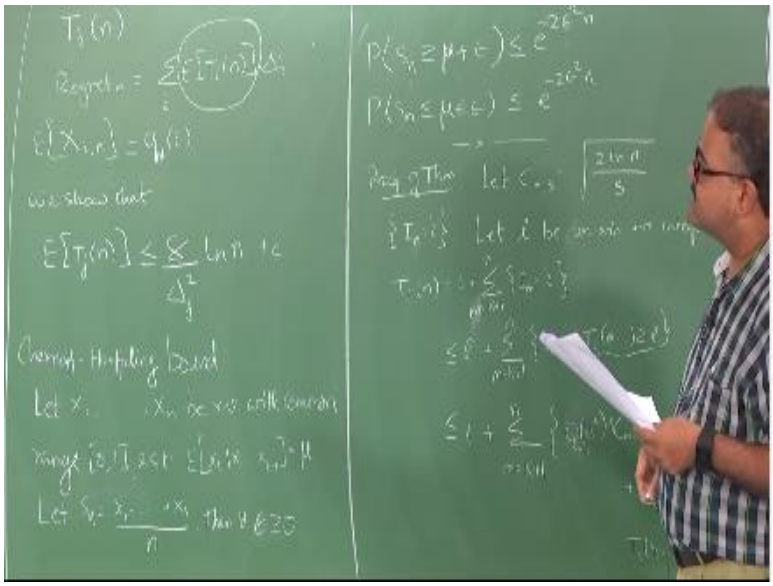
That means I have grossly under estimated my Q*A* this is one condition, so the second condition says this is Q(I) right and this is Q*(I) so grossly overestimated by Q*(I) this is essentially what the second condition says so this is QI > Q*I + the confidents one so this is the confidents bound so Q*I + this confidents bound will be only till here, right. But QI is greater than that.

No, no is that eternal condition that can hold right when both of these do not hold also I could still be making a mistake right I could still this could still be one if the true means or not far apart with respect to this error bar I have. See basically what I am saying is, a true Q* could be somewhere here right but I still can make a mistake because it is too close. So probably draw all this pictures and put it somewhere permanently.

Because every time I seem to be re-creating this pictures, right. So I have first thing I said okay the Q * is somewhere there right I make a mistake right so and then I have this guy right is possible for me to make an error, right and then the other case was okay this the true Q was somewhere here this possible for me to make a mistake right this third case say no none of this really need to happen, right the true Q* could be well within the boards right, but given the current uncertainty I have in my Q(I) right, so the true means are not so far apart right this is uncertainty cannot over well made right the true means are actually within the uncertainty. Therefore, I could still make a mistake so the first condition is not true the second condition is not true but I still made the mistake, okay.

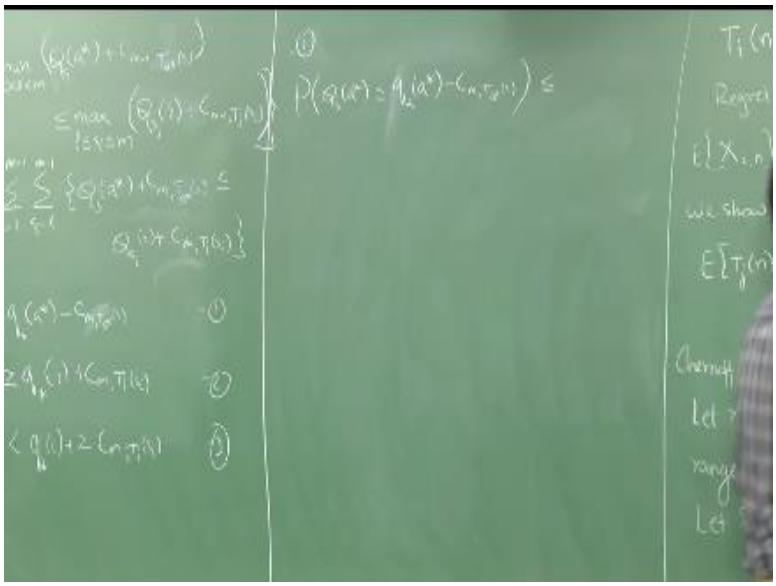
So people have all of this expression everyone has in their notebooks people who are writing notes, people were not writing notes you have this in your memory, because I am going to erase it but I am going to write it again, okay so I want to all of you to remember this.

(Refer Slide Time: 37:35)



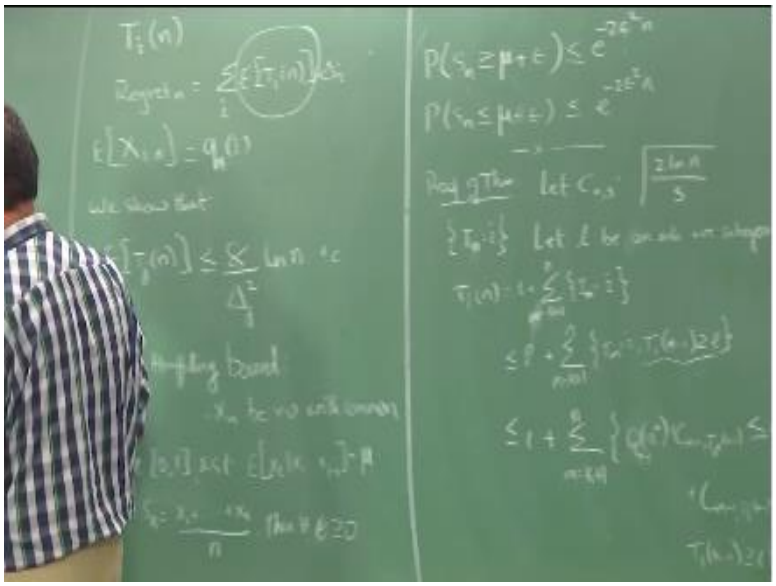
Okay, so let us look at 1, so I want to look at the probability that a* what did I do so I want to bound that probability, right so can I do that remember I told you $S_n$ is Q, I told you μ is Q* right, so we have they already right, if you remember when I wrote down the Hoeffding machine of bound, I told you their comparison right, I told you that $S_n$ is Q μ is Q* so you have those there, right. So essentially then a ε becomes CM right and by bound is $e^{-2\varepsilon 2n}$ right, and what is Cn so it is $2\ln n/s^2$.

(Refer Slide Time: 39:43)



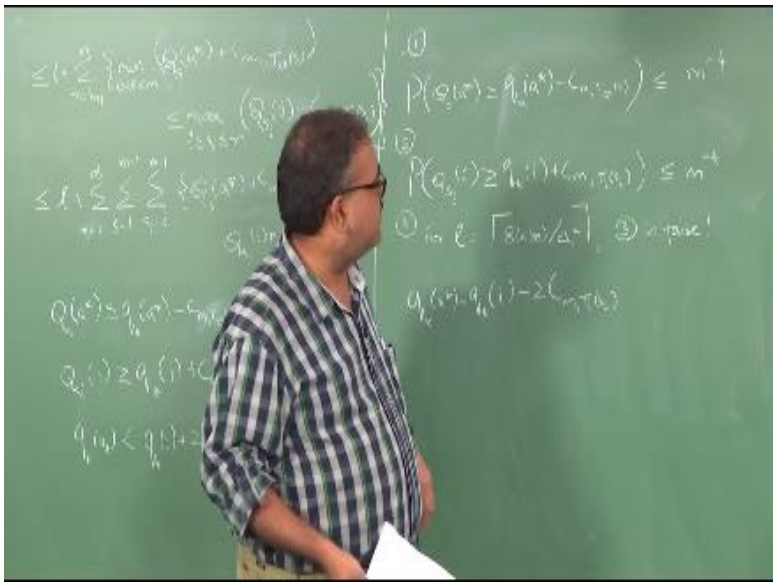
Okay, then simplified and tell what is the answer? So what is your n there, now what is the n that you have yo plug in to the expression, S this is where the thing became a little messy because the paper is S as the, when they introduce an addition notation, right but I did not use that I try to use the t* ta* notation itself, so it biomes ta* of S that is that thing that you have to plug in, okay so simplified and tell me what is the answer.

(Refer Slide Time: 40:46)



So I plug in ta* of S here M here right, and then I have to square it basically that is my ε that will get squared, right so my ta* of S and my ta* of S get cancelled, right I get ln m come on, after all that we have done so far you could do this simplification easily enough just substitute this into that and then cancel things out and tell me what is happening.

(Refer Slide Time: 41:28)

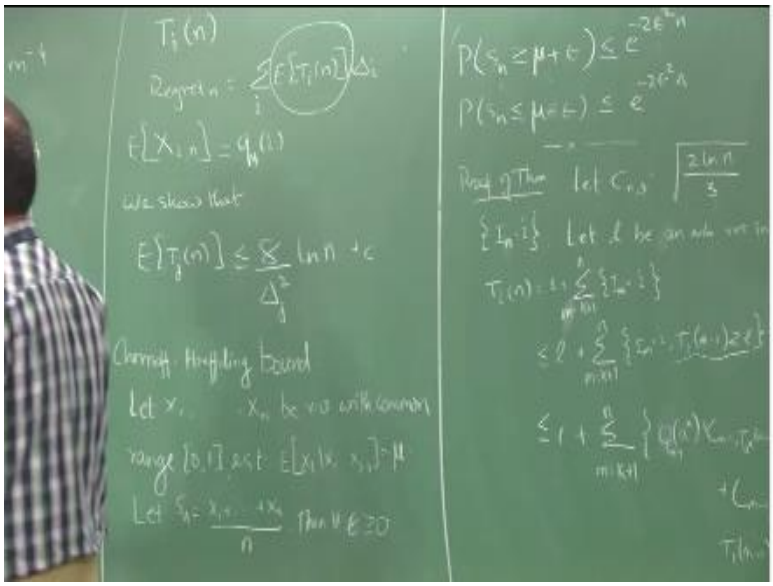


$1/m^4$ right, so essentially this will be $S_n$ is equal to okay, so what about the second term okay, what about the third term okay, here is where we are going to use our mysterious L right, what did L what did I introduce L for I said that hey, I have done so many number of trails already, okay I am only looking at these events after that many number of times I have taken our MI, right.

So I am going to make the following statement if I taken sufficient number of trials initially okay, if I have taken sufficient number of pulls of form i initially this event will never happen that is what I am claiming, right this rather easy to see right, so what do we have, we have q*a*-a*i-2Cm, okay right. SO essentially I am saying that my TiSi is at least this much, right so that is what this means right okay, say L is equal to this much that means the number of times I have taken are my is at least this much.

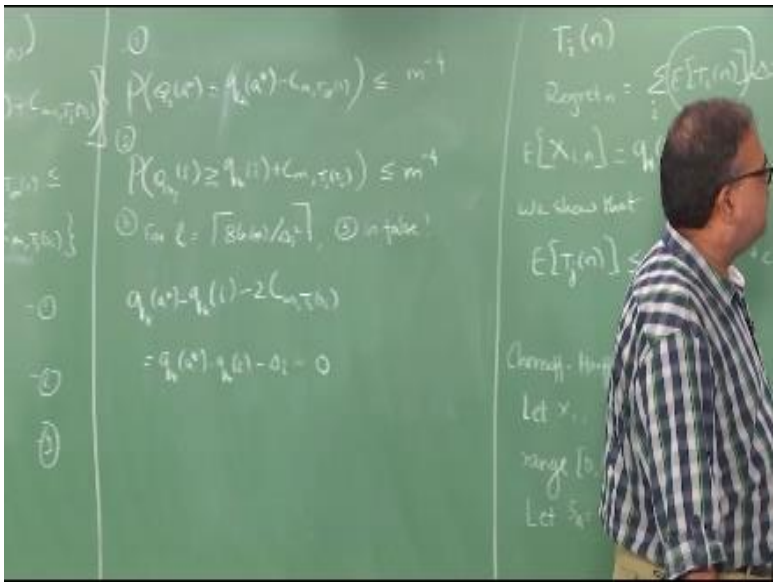
If you remember my Si $\sum$ starts from L right, I consider Si only from L so if say L is this so that means my TiSi is at least this much, correct go and plug this into my Cn expression.

(Refer Slide Time: 45:24)




So if my S is this much at least this much so what do I get, I get 2/8 ln m I mean $\Delta a^2$ will go up so ln m, ln m will get cancel, right so 2/8 will become 1/4 that is 1/2, 1/2 x$\Delta$i right, so 1/2 and 1/2 2 will get cancelled I will essentially get $\Delta$i, right so this expression will become.

(Refer Slide Time: 45:58)



Okay, and what is this equal to, so for any value of l greater than this, this quantity will be greater than 0 that means this is inequality will not fall, correct you people see there you want me to do that again, yes okay, so I am saying L is going to be at least this much so then if you remember my TiSi that the Si∑ is starting from L I am assuming that at least L times I have taken this already, right.

So I am assuming you have taken this L times already so what I will do is L will plug the, this value for the number of times I have taken this already right this value, so I will plug in 8 ln m this is Cm, S right so this case look at the expression m, this so this will be ln m right I am plugging in here instead of S I will plug in 8 times ln $m/\Delta a^2$ right, ln m, ln m will get cancelled so 2/8 will become 1/4 and divide by $\Delta a^2$ will go up right, so I will basically get $\sqrt{\Delta a^2}/4$ which is $\Delta i/2$, right.

So I will come and plug in $\Delta i/2$ here and 2 and 2 will get cancelled I will essentially end up getting $\Delta i$ here, so that is what I have written here. So and by definition of $\Delta i$ this expression is 0, right and for values larger I mean if I take the ai more times than this will essentially become

greater than 0 and therefore that inequality that I wanted will never hold, if I have taken L at least if I have taken action i at least L number times where L is given by that expression.

The third condition will never hold, okay so the probability of the third condition holding is 0 if L is greater than this, right. So now what happens so the probability of this whole thing happening is upper bounded by just two times this probability right, just add these events up so we are still not done. Where are we, so are bounding the expected value of the thing right, so we have expected value of, so and I am going to say you to write this out here okay.

So this is lesser than or equal $8\ln m/\Delta a^2$ because that is L plus the $\sum$ terms that we had earlier right, and I am substituting a specific value for L here okay, then what will be the internal term here. Remember, we have this random variable which could take values of 1 or 0 so what is the expected value of the random variable is just a probability of it being 1, right so there are only two of those things right.

So the probability of this event plus the probability of this event, right that is basically right so I am not writing down this because this expression are going to take a long things so probability of this event plus the probability of that event so in which our case is just m-4, $m^{-4}$ that is essentially, right so it is clear. So you see how that came up about, right so here this was the random variable right. Now once I took the expectation so I can replace that random variable with the probability of it being 1 that will be the expectation of that random variable, okay is it fine so replace that with the probability of it being 1 is upper bounded m-4, m-4 and that is basically here, okay great.

So this $\sum$ okay, I am going to declare it to converge to $1+\pi^2/3$ okay, when you can little involve thing but it is pure algebra okay I mean you can work it out, okay or I mean this is the standard relation you can look it a person, right so that is basically what we have. So this is kind of see that this is going to because you going to be diminishing some so it is going to converge to a limit, right and so what is really interesting is this part, right in some constant whatever I said that it converges to it is not really of interest you know not really of great interest what is really of interest is this $8 \ln m/\Delta a$, right.

Because this is the one where you are going to get mistake a lot because it has the ln m term right, and it also has your $\Delta i^2$, right but this is the only expected value of Tin but you have to multiplied by $\Delta i$ to get the total integrate and you have to some more all i, right so that is essentially the thing that we have finally that expression that I erased here that is what you will end up with that is why I said keep that in your mind because that is the final expression you are going to end up with, right.

This is for each arm this is for one suboptimal arm, right this is a total number of times you applied it right that is all we have derived now, so the regretted due to that suboptimal arm will be $\Delta i$ times this number so that is why the $\Delta i^2$ goes to $\Delta i$ in the denominator in the original expression if you flip back and you look at the original expression we had a $\Delta i$ here and we also had a $1+\pi^2/3.\Delta i$. but this is for one arm, so you have to do this for all the suboptimal arms basically that is why your $\sum$ is overall i which is not optimal, okay great.

So this is you have done with UCB and they have shown that UCB is regretted is bounded and what is the nice thing we shown here is that the regret grows as ln m right, and there is a result from the 80s which tells us that you cannot have a regret optimizing algorithm that does better than ln m, right what you can try to improve on is the 8, right this is also impossible to get of that the $\Delta i$ is also impossible to get regret of you cannot say that I will, I mean if the arms are close in revote right, it becomes hard to eliminate the bad arms, right. If they ate far away in revote then you can quickly figure out which are the bad arms but the arms are close in revote and it becomes harder to figure out so those things are always there.

So those are essential components with sop basically people flied around with that date, okay and yeah so that is essentially lot of work that happens in regulate automate so let to stop here because we have reached 5'o clock I do not know if how many of noticed it so if you have not I will take that as I kept you successfully in gross, but if people were just biting their time to leave when I am apologies where keeping here for so long, but I needed to finish this right, I could not have stopped half way through chart in the next class with $3\sum$ or something, right.

So in the next class right, so we will look at the last optimistic criterion was talking about which was pack optimality, okay.

**IIT Madras Production**

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in