NPTEL

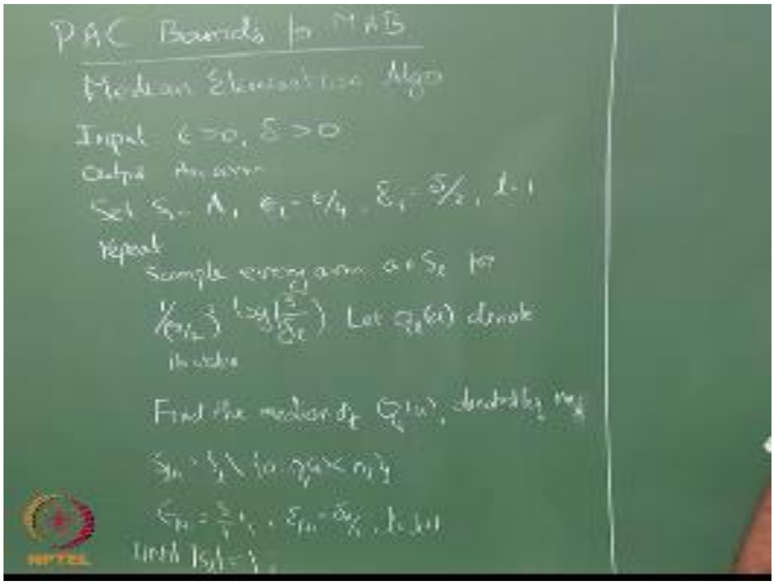
NPTEL ONLINE COURSE

REINFORCEMENT LEARNING

Media Elimination

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

(Refer Slide Time: 00:16)



```
PAC Bounds for MAB
Median Elimination Algo
Input $\epsilon > 0$, $\delta > 0$
Output An arm
Set $S_1 = A$, $\epsilon_1 = \epsilon/4$, $\delta_1 = \delta/2$, $l = 1$
Repeat
    Sample every arm $a \in S_l$ for
    $\frac{1}{(\epsilon_l/2)^2}\log\left(\frac{3}{\delta_l}\right)$. Let $\hat{q}_l(a)$ denote
    its value
    Find the median of $\hat{q}_l(a)$, denoted by $m_l$
    $S_{l+1} = S_l \setminus \{a : \hat{q}_l(a) < m_l\}$
    $\epsilon_{l+1} = \frac{3}{4}\epsilon_l$, $\delta_{l+1} = \delta_l/2$, $l = l+1$
Until $|S_l| = 1$
```

Okay, good took a while to write so essentially what we do is we start off with k arms right and then in the first round okay so median elimination is going to proceed in rounds okay. In the first round we pull each arm some number of times right determined by this magic quantity. So can people read this magic quantity it is 1 by ∑L/2 the whole square better log 3 by ΔL okay.

So I pull each arm that many number of times okay and then I estimate their empirical value right QLA I will estimate it where L indicates that this is the Q function I am forming at the LF stage okay. Then what I do is I take these QLA arrange them in ascending, descending whatever order I find the median of all these arms. And then eliminate all those arms whose values is below the median right I am eliminating. So this is the set different so I take my current set SL that these are all the arms that are under consideration.

So S1 will be all the arms and then I am eliminating all those arms such that so my QLA is less than ml where ml is the median at the Lth level right. And then I change my constants by magic amounts again. So my ∑L + 1 becomes three-fourths of ∑L, ∑L itself started as ∑/4 okay.

Now it becomes 3/4 of ∑/4 and my ΔL+1 becomes ΔL/2 which already started as Δ/2 so it becomes Δ/4, and L=L+1 and I keep going until my SL becomes 1, the number of arms left becomes one right. So how many rounds will I go block k rounds because every time I am eliminating the half the arms right, because I am looking at the median and I am eliminating everything below the median.

So I am guaranteed eliminate half the arms in every round, so I will end in lock K rounds okay. But the trick to provide proving any kind of sample complexity here is to show that okay, this quantity right when I sum it over this log K rounds is some bounded quantity right and that gives me the sample complexity round.

I have already told you what the sample complexity is going to be, what was it k by ∑^{2}log 1/Δ some constant by Δ so as far as order notations we can ignore the constant so log 1/Δ. So that is what this is going to be the sample complexity, so I will have to show you that first. Second in fact first I will have to show you that this is actually a pack algorithm right, that it actually gives you a ∑ optimal armed with a high probability right.

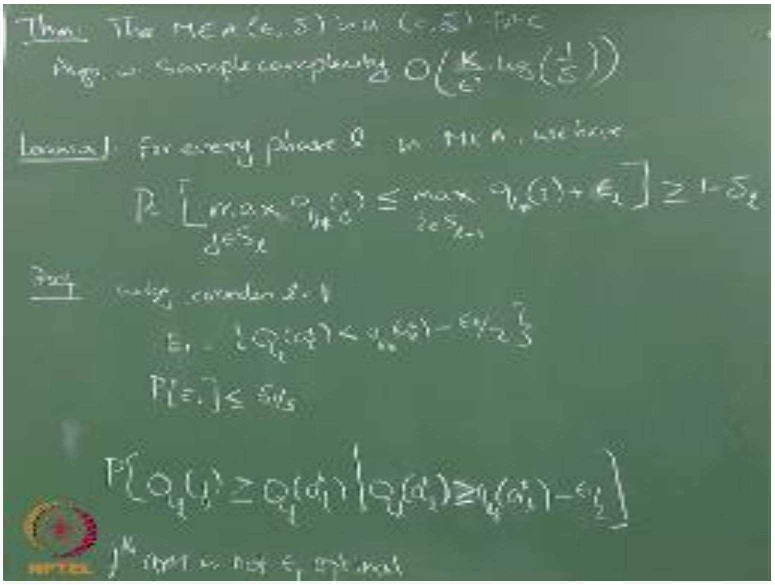
It becomes a little tricky here, because I am eliminating half the arms every times. So now I will have to think about what is the probability that I have eliminated all the ∑ optimal arms right. So I have to, in some round along the way before I reach the final round okay I should not eliminate all the ∑ optimal arms right.

So what we will do is we will see the trick, because at the end of the thing is that one arm left is some ∑ optimal arm and safe and that is along the way if I eliminate all the ∑ optimal arms then

I am in trouble. So what I will do now is to show that at every round the probability of me doing that is very small.

So small that when you add up the probability of me eliminating across all the rounds it is still below Δ still below Δ so that is essentially what I am going to try and show did that make sense. So the every round the probability of me eliminating all the ∑ optimal arms is small, so small that when I add it up across all the rounds it is still less than Δ okay. So that is essentially what I am going to what I mean to show.

(Refer Slide Time: 09:27)



Thm: The ME A(ϵ, δ) is an (ϵ, δ)-PAC
Alg. w. sample complexity $O\left(\frac{K}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right)$

Lemma 1: For every phase ℓ in ME A, we have

$$P\left[\max_{j \in S_\ell} q_*(j) \le \max_{i \in S_{\ell+1}} q_*(i) + \epsilon_\ell\right] \ge 1 - \delta_\ell$$

Proof: wlog, consider $\ell = 1$

$$E_1 = \{Q_1(a_1^*) < q_*(a_1^*) - \epsilon_1/2\}$$

$$P[E_1] \le \delta_1/3$$

$$P\left[Q_1(j) \ge Q_1(a_1^*) \mid Q_1(a_1^*) \ge q_*(a_1^*) - \epsilon_1/2\right]$$

j that is not ϵ_1 optimal

Okay so first theorem that I will state it, but I will come back and prove this later is the median elimination algorithm with takes ∑Δ, this is ∑Δ pack okay. So we will prove this in two parts the first lambda will help us establish that okay, and then we will have another result later that will help us establish the sample complex.

Sample complex is actually trivial and if you think about it, it is essentially summing this quantity up right starting from ∑/4 and Δ/2 and successively changing the value of ∑ and Δ okay. It is just a bunch of algebra race is just move numbers around and simplify things you will get it conceptually there is nothing deep about looking at the sample complexity. So I said yeah, so I essentially have to sum up this quantity right for every arm.

So this will be multiplied by the size of the arms which will be, which will start with k and then it will become k/2, k/4 successively becomes smaller and smaller and likewise the ∑ and my Δ also becomes smaller and smaller. So the idea is just to submit over L this quantity and then replace this by a function of ∑ and L. I think we can replace ∑L right by a function of ∑.

So what is ∑ 1 is ∑/4 what is ∑2 3/4∑/2∑/4 ∑3 is $(3/4)^2$ so ∑L is essentially (¾)L-1x ∑/4. So like that so you can write this as a similar expression for Δ and we can essentially simplify right. So that part is fairly straightforward so if you have time I will just give you an intuition to them otherwise you can work it out easily.

Now the first part is one that requires some work right. So this is what we will do for every phase L in the median elimination algorithm we have the probability. So this requires a little bit of thought, so what I am saying so what is this, this is the two best arm at level L right. So at level 1 I would expect this to be A* right.

So and the true best arm I am not talking about the Q right I am talking about Q* here, so I am talking about the true best arm right. So in this case I would expect the true this this should be Q* A* right in level 1, but in level two itself it could be something different because I might have lost A* in the first round itself right.

What I am saying is okay this is the best I could have done in the Lth round right potentially right, and what is this, this is the best I could have done in the L+1th round right. So the bad cases when this is less than that, and that means I have eliminated some high-ranking arms well all I have left or low ranking arms right.

So the bad cases when this quantity is less than that quantity right but how much it should be less right, that is a question we are asking here. So if I add an ∑ to this so essentially I have my, and let us call this J* right that is a maximum in the Lth level. So I have my Q* J* here and I have my Q* I* here which is the K that gives me the max here right.

If I add an ∑ to this if I add an ∑ to this and this is within that that means this is ∑ optimal when you can consider against that right, do you see that if an add an ∑ to this lower value and the higher value is actually less than that then it is an acceptable situation, because the lower value is within ∑ of the higher value correct, so this is a good outcome.

So I am saying the probability of the good outcome is at least 1-Δ that means the probability is high okay, do that makes sense the probability of the good outcome is at least 1-Δ okay. Now the thing to notice here is this is not truly ∑ this ∑L right and this is not Δ this is ΔL, so why is that the case because every round will lose ∑L right at most ∑L right.

So at the end of the day I should not have lost more than ∑ right. So what I am going to show you is that every round I will lose ∑L and my ∑L is chosen such that even if you lose the ∑L for every L overall loss will be ∑ right. So that because we know that when the S, in S1 the best number is A* right.

And if you overall loss is less than ∑ then at the end of this I will have at least 1∑ optimal on left okay, so that is basically what we are trying to show here. So this relation makes sense why we are trying to show this okay. And likewise this 1-ΔL is there because at the end of the day if all the events hold true also my probability must be at least 1-Δ right.

Therefore, we use ΔL and that is the reason your ∑1s and Δ1s are already smaller than ∑ and Δ, because if I start off with ∑ here and Δ here I am gone okay. So third one w log means ray without loss of generality I am going to consider L=1 right. So whatever results I show here will will hold for the subsequent L right.

So we start off by looking at the first, the event E1 which is first thing I am looking at it I am grossly under estimating the true value right. I mean this should now start looking familiar to you, because in everything we have been doing the same thing we start off by say okay, first thing that has to go wrong is I grossly underestimate the true optimal actions value.

And then we will think about okay what happens if I overestimate the sub optimal actions value but I start off by say okay grossly underestimate this right. So what is the probability of this, now we already have our magic number here, so substitute everything so you have the expression here right and we have the magic number there.

That this K will get cancelled out, again we have a four problem there so well the square should be outside I am just writing it down from the algorithm here, so their results are wrong right. So it should be $\sum L^2/2$ $\sum(L/2)^2$ this is a one-sided bound. So the outside 2 will not be there, right so everything will simplify and you will get Δ1/Δ3 right.

This is the first E1 suppose it does not hold okay, suppose E1 does not know that it essentially I have not made a bad mistake under estimating the best arm right. And since it is maybe I should have done it slightly differently I have said without loss of generality consider L=1, but then I put an A* here which actually violates the without loss of generality part okay.

So the A* here is actually the best arm left at level L that is how we should think of it, it is not the overall based arm, whatever be the level L it is a best arm left at level L and for L=1 it happens to be A* the true A* right for L=3 and 4 this will essentially become the best arm at that level.

So I will just put a L there so that you know what I mean okay. This is the best arm in the Lth level that is left among all the arms that is left true best arm in the Lth level okay. So next I am going to look at the probability that as well as translation skill this one, I should put the 2 in the numerator what do you need for it to cancel out.

∑1/2 so ∑1/4 no yeah, so this is ∑/2 so substitute that like an $\sum^2/4$ Δ2 will become ½ so that is $1/\sum^2/2$ so I will get that is all fine okay good. So I am going to look at this problem right, so sorry not this, so E1 does not hold that means that this is kept to right. So even if E1 where I have underestimated it so I am saying E1 does not hold that means that there is no longer an underestimate.

So this is my estimate is within ∑/2 right, but then I am looking at the possibility that yeah, so with my some J^{th} arm which is not ∑1 optimal right. So that is a condition no, this shows I know that current estimate in the electron I am going to decide about eliminating the electron right in the current round I am not underestimated my A*L okay.

So which I am within the ∑ of that, within ∑1/2 of the 2A* value right, but then I take some other arm J which is a bad arm, bad in the sense J is not ∑L close to A* it is a bad arm right. And I figure out what is the probability that the Q estimate for that arm will be greater than the Q estimate for the A* right.

So if this happens right, then there is a chance that A* will get eliminated only a chance right, it has to be in the lower half not just get bitten by one J but it has to be in the lower half of the arms right. So I need at least k/2 or whatever the size of SK SL/2 bad arms to beat me correct, for me to eliminate this arm.

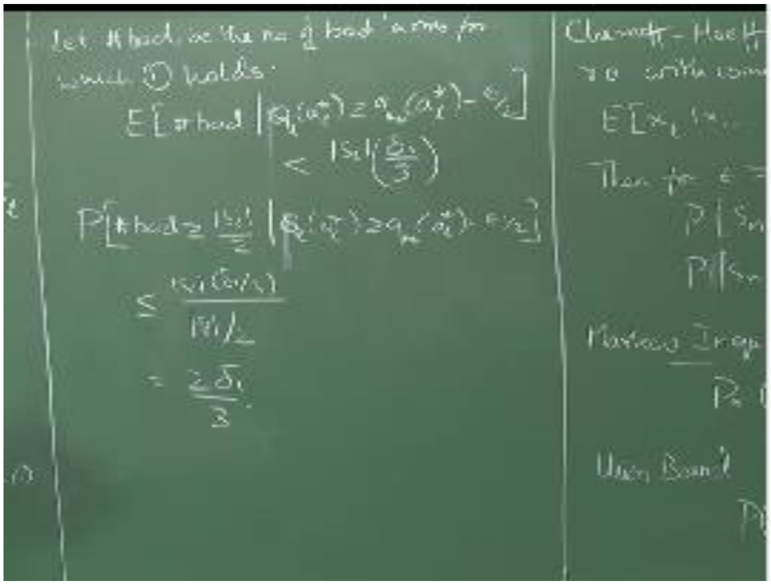
And I need at least half of the arms to beat me not just one arm right, but before we go to that part, so what is the probability of this happening. Well so this conditioning tells us that QLA* is not too bad right. So what does that mean QLJ has to be more than ∑/2 so we remember that figure we do right.

So either this has to be under estimate was if I would erase the figure, but either the A* has to be underestimated by more than ∑/2 or the A' has to be over estimated by more than ∑/2 right, here we are conditioning it on the E1 that A* is not underestimated by more than ∑/2. Therefore, J has to be over estimated by more than ∑/2.

Since this event we have conditioned it as false, so that even has to be true okay. So what do we get in that case same thing here, so kind of skipping a step here but people are all fine with that right, I am skipping, rewriting this bond right. So essentially you will have to now replace this with over estimating by ∑/2 right. In fact I should probably just make you write that missing step as a homework.

So that I am sure that people are all comfortable with the notation nothing more the notation can get really complex, otherwise the concepts are very simple, that is I am taking so much time talking over it and if I could just write the proof and leave it at that because it is fairly easy, these things are fairly easy. So this is for one arm right and like this I have k arms right, so the probability good.

(Refer Slide Time: 29:55)



Let #bad be the no of bad arms for which (1) holds.

$$E\left[\#bad \mid \hat{Q}_l(a_l^*) \ge q_*(a_l^*) - \epsilon_l/2\right] < |S_l|\left(\frac{\delta_l}{3}\right)$$

$$P\left[\#bad \ge \frac{|S_l|}{2} \mid \hat{Q}_l(a_l^*) \ge q_*(a_l^*) - \epsilon_l/2\right] \le \frac{|S_l|(\delta_l/3)}{|S_l|/2} = \frac{2\delta_l}{3}$$

So let now use the same symbols here, I do not agree with let # bad be the number of bad arms okay, # bad be the number of bad arms such that this this condition holds later # bad be the number of bad arms such that, that condition holds. So these are the number of bad arms that are beating the best arm.

So you have this expectation be mod L/2 yeah, not really but I agree with the mod SL part but not SL/2 I have a probability for that happening right. And how many such experiments I actually mod SL-1 right potentially everything other than the best arm could be a bad arm with maximum thing is SL-1 right.

And so it is a cell -1Δ1/3, so that is the probability of one arm being bad and one arm actually for which this first relation poles for one bad arm and there are SL-1 I am just making it SLl to make our life easier. So SL such bad arms so the expected number of bad arms will be SLxΔ1/3. Now I want to look at the probability that # bad this is why we are going to use the Markov inequality right.

So I have already written the expectation here, so I am asking you the probability of X greater than equal to A is the expected value of X divided by A right, so this is less than or equal to the expected value which is divided by A okay. So what have you shown there is that if E1 does not hold if I am good in my estimate for the best arm okay, the probability that half the I mean, so many bad arms will beat me is bounded by 2Δ1/3 right.

But I am also considering E1 as a bad event that E1 is also bad because my estimate for the best arm is really bad you know. So all bets are off in some sense that because I do not really, I cannot really identify my best term because it is so bad right, so all bets are off so E1 is also bad event that all of this is bad and E1 is also bad.

So what is the total probability of something bad happening the two together, but they are you sticking in if you want to use a fancy term we use the Union bound and then we say the probability of something bad happening is bounded by Δ1 right. So this bad is Δ1/3, that bad is 2Δ1/3 and this is actually the complementary event of E1 in fact I can just add the probabilities, I do not really need the Union bound.

So the total probability of something of bad happening is Δ1 right. So that is essentially what we are started off by saying here. So the probability of good is 1-Δ now we have shown that the probability of bad is ΔL, so the probability of good is 1-Δ okay. So one of the nice things about the median elimination algorithm is that it was one of the earliest papers to introduce this kind of a round based algorithms to the bounded literature right.

Now a lot of later improvements on both regret based case and people trying to play around with the the pack guarantees all switch to some kind of round based ideas, you know so you do a certain number of pulls and then you eliminate a certain number of arms right, and then go back and pull the remaining arm some more and then eliminate some more arms.

So this kind of a round based elimination methods have become very popular in the Pandit literature and the meeting our nation is one of the earliest such round based elimination approaches then after this is became very popular and a lot of papers started being written using this kind of a round based idea right.

So we have so far what we have shown is that it is have you shown that it is back what is it that is fine I mean that I have already spoken about this lambda is taken as proved right, this lambda is taken as prove, so you have to still use the Union bound for every round that I am eliminating right, so what we have to show is well I have all these Δs right, so we have to keep adding them up to make sure that eventually you are less than overall you are less than Δ right.

So every round the probability of failure is bounded by $\Delta 1$ right so for the K rounds the probability of failure is I mean whatever the log K rounds the probability of failure is bounded by summing it up. So $\Delta 2$ $\Delta/2$, $\Delta/4$, $\Delta/8$ and so on so forth until you reach 1 right, so that is the summation and not until you reach log K rounds.

So it will get $\Delta/2^{logk}$ okay, and the summation of log will go to Δ it is essentially Δx1/2 +one-fourth +one-eighth right if there is some runs to infinity you will get one, but the sum is not running to infinity it stops at some finite time, so it will be less than 1, so the summation will be less than Δ proved right.

So some probability of something bad happening will be less than Δ. So likewise you can show these same things for the $\sum$ right, so at every level you drop utmost by $\sum$L right. So the total you would have dropped this $\sum$/4+ so it is essentially one-fourth plus what is it three-fourth into one or write $\sum$/4x1+1 fourth and so on so forth right.

3/4L-1 yeah yeah that was also upper bounded by 1 if you run it to infinite note some and suppose wanted by 1 if I run the summation to infinity but you are stopping at some finite amount so it will be less than ∑ okay. So both of that are satisfied that is easy enough, the harder, the trickier part is to show the sample complexity right.

It was not hard, it is just algebra can you let you guys do it or you want me to do the sample complexity part of the proof as well no is it depends yeah nobody say depends depends on what that is a paper keys are in the paper yeah it is there in the paper you can look at it. And so I will stop it here for the pack the pack complexity will come back to all of this regulator and pack and everything when we look at the full RL problem at some point. Now right now for the bandits we will stop here.

IIT Madras Production

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

www.nptel.ac.in