NPTEL

NPTEL ONLINE COURSE

REINFORCEMENT LEARNING

Proof of Therom I

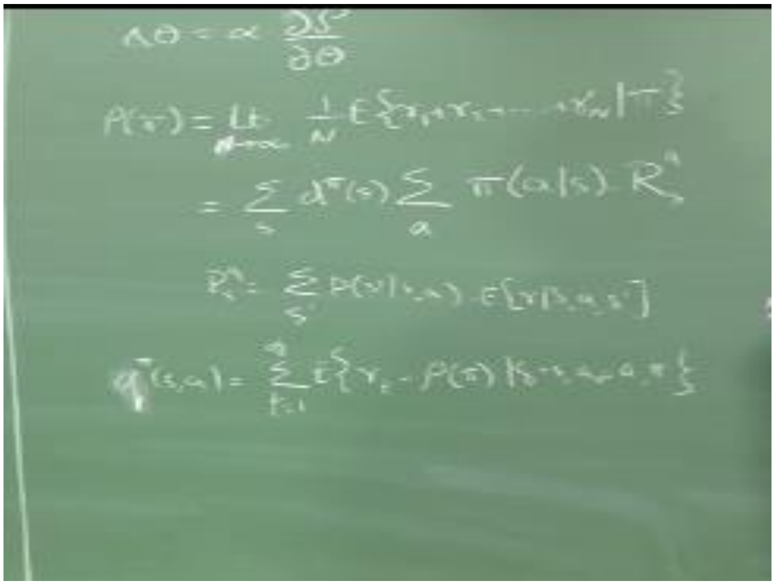
Prof. Balaraman Ravindran

Department of Computer Science and Engineering

Indian Institute of Technology Madras

So if you remember the way we defined our Δθ.

(Refer Slide Time: 00:17)



$$\Delta\theta = \alpha \frac{\partial \rho}{\partial \theta}$$

$$\rho(\pi) = \lim_{N \to \infty} \frac{1}{N} E\{r_1 + r_2 + \cdots + r_N | \pi\}$$

$$= \sum_s d^\pi(s) \sum_a \pi(a|s) R_s^a$$

$$R_s^a = \sum_{s'} P(s'|s,a) E[r|s,a,s']$$

$$Q^\pi(s,a) = \sum_{t=1}^{\infty} E\{r_t - \rho(\pi) | s_0 = s, a_0 = a, \pi\}$$

Was going to be some α times its some ∂ ρ/ ∂ T right or ∂ η by we defined it as ∂ η / ∂ η θ where η was some performance measure right so one popular performance measure that we could have is essentially the average reward which is yes, so limit n tending to infinity of the sum of 1 R1 to Rn and we did this last time I put the end within the expectation right now I have taken the n outside expectation otherwise it is exactly the same expression I wrote in the last class right for this is the average of all right.

So one way of thinking about the average reward is that it is given by the following where RSA is the reward you get for taking action a in state S right, so what have I done here I have marginalized over s prime I just summed over all s' that you could land up just to make it easy when you could you can always write an extra Σ here over s' and keep carrying it around so RSA is essentially using your old notation.

But this is what our essay is if you want you can write the Σ inside that and you can keep carrying it around I just want to simplify this I put in a new notation for the Σ okay so RSA is that and I am also going to define without giving you any reasons for it a new Q function Qπ SA for the average reward case okay, which tells me if I start by taking action A right if I start by taking action A so how much more than the average reward will I expect to get right.

It should not really be a difference if you think about it if I do not if I remove this difference and take the summation in right the ΣRT t equal 1 to infinity Σ ρ π T equals one to infinity should be looks like it should be the same right so such look silly counter in today I am just pointing it out to you already right but then this is called the this is called the ρ is typically called the gain of a policy right and this Q function is called the bias of a policy okay I think I got it right I think that is the gain and this is the bias right and this can actually be used to compare different actions right.

So if I start off by taking this action so there will be some initial transient terms that might not get cancelled out so that will tell me that I am slightly better than the other thing so this is why I said we need to do the whole couple of days of average reward RL so that you can understand what the difference between the gain and the bias and other things but right now I am defining this without any further explanation.

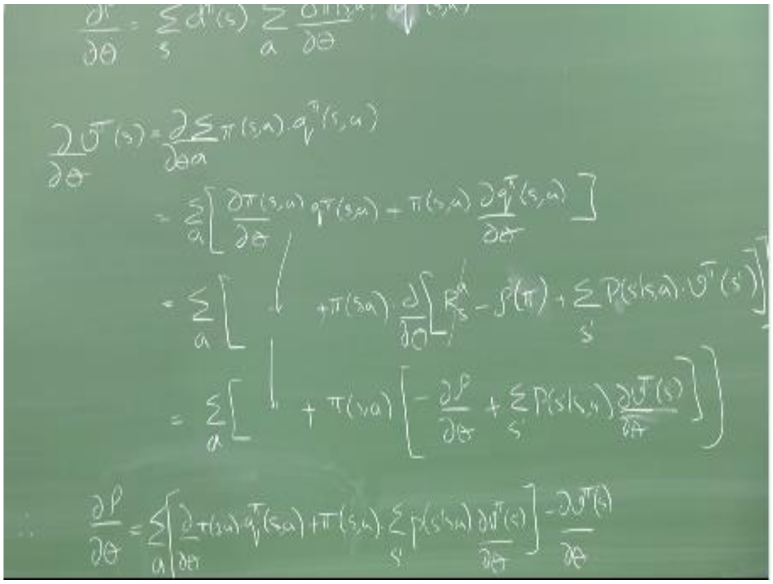
So let us take this as the Q function definition for average reward are essentially tells you how much starting off with action a let us say a1is better than starting off with action a2 to this more of a comparative value function that you should be using why not quite it goes it if this action is very good and the rewards are always much higher than the ρ so the value will be a very high value is the summation on that there are how will hold that because of how does that happen.

Because this is π anyway and behaving according to π am just starting off with a after that I am behaving according to π right so the long-term behavior for our three should that should be tending towards ρ π right I mean sometimes it is how much above ρ π sometimes it should be much below ρ π otherwise ρ π cannot be the average of all right you after the first action I am always behaving according to π right so it cannot be the average of all it is only the first action that is different right so if the first action gives me a slight advantage then I am going to say that it is a better action than starting off with something else okay.

It is essentially used as a comparative measure between actions does it make sense great so likewise we could define ρ for the discount rate or what case also right I am few remember if I am going to use discounted rewards what was the assumption I made for the policy Grade in case what is a something maybe let us not take that is what you need guess not and take the value of s not as my value for the policy right.

So like this I can define the same thing I can look at the expected discounted reward starting from s not keep that as ρ and I can define my EULA the usual way starting from s not doing action a following policy π director okay so that is my Q function and I need $\partial \rho / \partial \Delta$ right so let us see what $\partial \rho / \partial \Delta$ should be.

(Refer Slide Time: 07:03)



$$\frac{\partial\rho}{\partial\theta} = \sum_s d^\pi(s) \sum_a \frac{\partial\pi(s,a)}{\partial\theta} q^\pi(s,a)$$

$$\frac{\partial V^\pi(s)}{\partial\theta} = \frac{\partial}{\partial\theta}\sum_a \pi(s,a)\cdot q^\pi(s,a)$$

$$= \sum_a \left[\frac{\partial\pi(s,a)}{\partial\theta} q^\pi(s,a) + \pi(s,a) \frac{\partial q^\pi(s,a)}{\partial\theta} \right]$$

$$= \sum_a \left[\quad + \pi(s,a) \frac{\partial}{\partial\theta}\left[R_s^a - \rho(\pi) + \sum_{s'} P(s'|s,a)\cdot V^\pi(s') \right] \right]$$

$$= \sum_a \left[\quad + \pi(s,a) \left[-\frac{\partial\rho}{\partial\theta} + \sum_{s'} P(s'|s,a) \frac{\partial V^\pi(s')}{\partial\theta} \right] \right]$$

$$\therefore \frac{\partial\rho}{\partial\theta} = \sum_a \left[\frac{\partial}{\partial\theta}\pi(s,a) q^\pi(s,a) + \pi(s,a)\cdot\sum_{s'} P(s'|s,a) \frac{\partial V^\pi(s')}{\partial\theta} \right] - \frac{\partial V^\pi(s)}{\partial\theta}$$

This was a pretty amazing result so if you think about it see π depends on θ right, all of you know that is why I have the ρ π / ρ θ but deep I also depends on π, so that should have a that should also figure in the different derivative right q π also depends on π that should also be a figure in the derivative so if you just think about applying the applying the derivative directly to this expression and essentially allowed to do it through the product rule right so there are three terms here and I should differentiate everything but then it turns out that that is not really the case right and the derivative of ρ/θ is you just basically have to take the derivative of the policy with respect to theta evaluated at a right.

And basically you are done this is amazing result right, so there are variations of this results already in the literature but then the this guy Peter Marbach came up with a proof for this right it turns into several pages maybe 15 pages or 20 pages or something of his thesis the proof of this theorem runs into several pages in his in his thesis right and then a little later without really knowing about bar box result before they were set out to prove these things.

So rich certain nothing just seeing David McAllister and a bunch of people wrote this paper and that is the proof okay so mod box proof runs for several pages because he was working from a different set of initial principles and so rich came up with this proof where he starts off with an expression that is completely un related to what they are trying to prove right, and then somewhere along the line suddenly like a magician he derives his relation okay so I will show you this proof just because it is so elegant and so simple and so if you know what the expression for V π is do you?

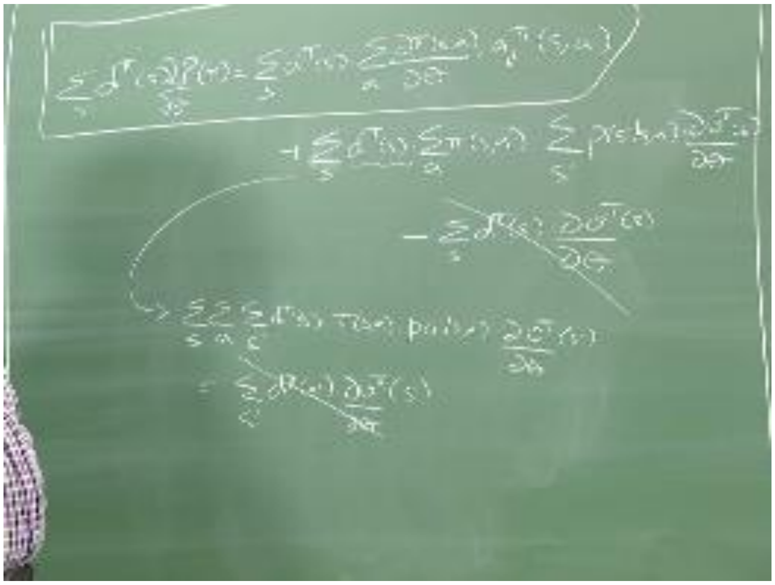
Okay is another expression I am using π of a given s and π s, as the same okay, so I can take the derivative of this all right so what will this be right yes that is rather straight forward okay then what we do okay, so do we need to do anything to this first expression not really that is what we want to be left behind right so we let us leave it as it is they will just leave the first expression as it is and we will play around with the second expression and see if we can make it, so why is it RSA.

So the expected value of $RT - \rho$ the expected value of RT is essentially RSA like why did we write RSA, so this is our essay right so that is expected value I will get for the first step for taking action a right, so expect value RT will be RSA when $-\rho$ π and then what do we have so where is that come let us $t = 1$ to infinity so this is like the first term right so this is the rest of the terms whatever follows later I am writing it in terms of the value function right.

So this will go away because does not depend on θ right so you will get minus okay finally we got our $\partial \rho / \partial \Delta$ so far we did not have that expression at all right what we wanted to drive $\partial \rho / \partial \Delta$ I did not start from this expression I am sorry I did not start from where is ∂ expression this expression I did not take the derivative right I started from some arbitrary equation is $v\pi = Q\pi$ right and then I started making substitutions and then it took the derivatives I finally got my term $\partial \rho / \partial \Delta$.

Now I will have to rewrite it so that this comes to the left hand side so what do I get there for you right so I think that is correct I did nothing I just took the $\partial \rho / \partial \Delta$ other side so if you think about it I multiply this by π SA right but as I can pull it out because it is not dependent on a then sum overall A so that π I say will go to one, right so I will be left with $-\partial \rho / \partial \Delta I$ being $-\partial \rho / \partial \Delta$ here I take ∂V π $\partial \Delta$ that side right when the rest of the terms have written it like this okay is it clear okay.

(Refer Slide Time: 19:32)



So more sleight-of-hand now right so I am going to take expectations with respect to d π s on both sides and I am going to take expectation with respect to deep is on both sides so essentially what that would be Σs Vπ s x ρ(π) = Σ s dπ s x Σa written in correctly right so the first term Σdπs the second term dπs the third term depends okay I do not know this is fine right so order this now comes the amazing trick so we know that dπs is a stationary distribution all right so let me just take this term alone right and write it like this.

So what does this mean ok I am going to start off with the stationary distribution is I mean Dπ that some state s I will take an action according to my policy π then I will make a transition according to p this whole thing what do I expect this distribution to be like because the π is a steady state distribution the condition of steady state distribution is it should be the same right the condition when is when we call when you call Dπ the steady state distribution if I apply π and p on it I should get back the same distribution and only then it is a steady state distribution so that is essentially what is happening it right if you think across all this Σ s' and a essentially I have applied π and P on Dπ at if I evaluate these two summations over this I will essentially end up with Dπ of s' right.

So this is essentially equal to summation over s’ okay, this looks familiar of course I can ignore this because there is the probability distribution it is one right so I can ignore this so basically I have my the theorem $\partial\rho \,/\, \partial \, \Delta = \Sigma s \; d\pi s \text{ x } \Sigma a \; \partial\pi \; sa \; \partial \; \Delta \text{ x } Q\pi \; SA$, amazing right whatever possessed them to start off with the equivalence between V and Q to get this derivation on row right I never ceases to amaze me that the rich managed to find something like this so what is the nice thing is it gives you a good connection between using value functions and so what is the difference between what we are doing now and what is the motivation I gave you earlier in the motivation I gave you earlier you ended up using the v function.

V was your baseline right when we said actor critic so the actor was the policy but the critic was V of s here the critic you are using is QSA right and the amazing thing is it turns out the same form holes regardless of whether you are using the average reward formulation like we did now or whether you are using the discounted reward formulation except that you need to do a difference sleight of hand to get the results but the proof holes whether you are using average reward formulation or whether using discounted reward formulation right.

I encourage you to read the proof for the discounted reward formulation so I am without a TA so you are my TA so just remain the TA to put up this certain McAllister sing and current one so paper online so I increase this is actually a really old paper I think is from 99 or something but still it is amazingly instructive as to how you can think out of the box to get really simple proofs for things okay this is one thing how it is now connected policy gradient approaches to the q value function okay and is still not done.

IIT Madras Production

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

www.nptel.ac.in

Copyrights reserved