NPTEL

NPTEL ONLINE COURSE

REINFORCEMENT LEARNING

Policy Gradient with
Function Approximation

Prof. Balaraman Ravindran

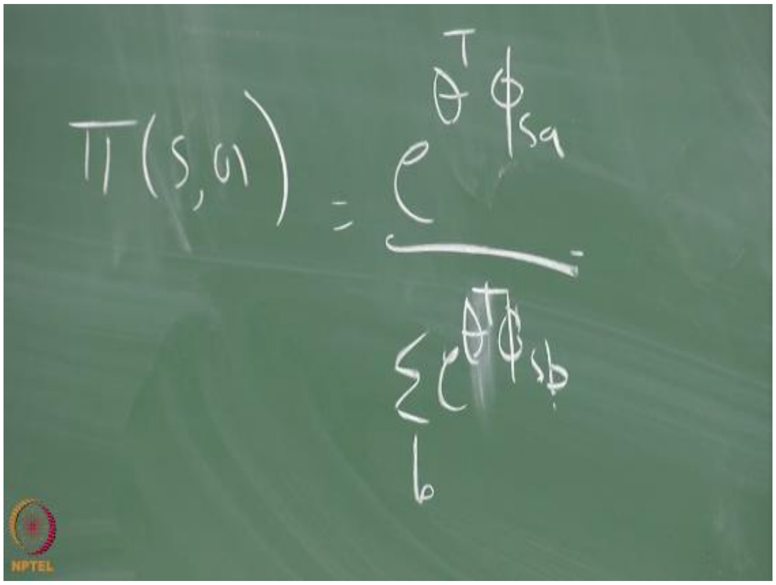
Department of Computer Science and Engineering

Indian Institute of Technology Madras

So the more interesting things that they showed here was a that like this actually is good I mean obviously since this is essentially estimating the gradient right if you follow this gradient estimates you will converge the local Optima there is nothing really to show that is this is the exact estimate of the gradient approximation is done right so as long as you have a way of doing this is all great but then we know that there is no way that you get q π directly right and that is as always some kind of an approximation associate with π right.

So now what happens is if I cannot estimate q π right and if I am going approximate q π let us say I have another function approximate right so I have π that is parameters by θ let us say I have Q that is parameters by some other set of things let us call them w and I have some parameters W that represent the Q function some parameters θ that represents the C policy right now when can a guarantee some kind of convergence right so here is an interesting result.

(Refer Slide Time: 01:55)



$$\pi(s,a) = \frac{e^{\theta^T \phi_{sa}}}{\sum_b e^{\theta^T \phi_{sb}}}$$

Let us assume that fw(s,$_a$) comma a sum Q hat right so some Q hat of so it is an approximation right I am going to denote that way fw and I am going to look at the error right so let us say look at the squared error between the true Q function which I do not really know but let us assume I am just looking at it for showing some steady state results right so I will say that has a Q function right and this is the prediction I am making am looking at the squared error right so if I am going to make changes to my weights right.

So as to minimize the squared errors so what will be my changes be proportional to the gradient of this function is what will be the gradient of this function again so basically I am looking at the τ by τw of the error right so τ by τ w error will be I can lose the two that into τ f w right τ w right and I have to go in the opposite direction of the gradient range that is a change so x minus will go away right this is clear so this will be the change I will make to a weight correct okay so how often will I leg will am a likely to make this change as often as I π (s,a) the same right as often as π (s,a).

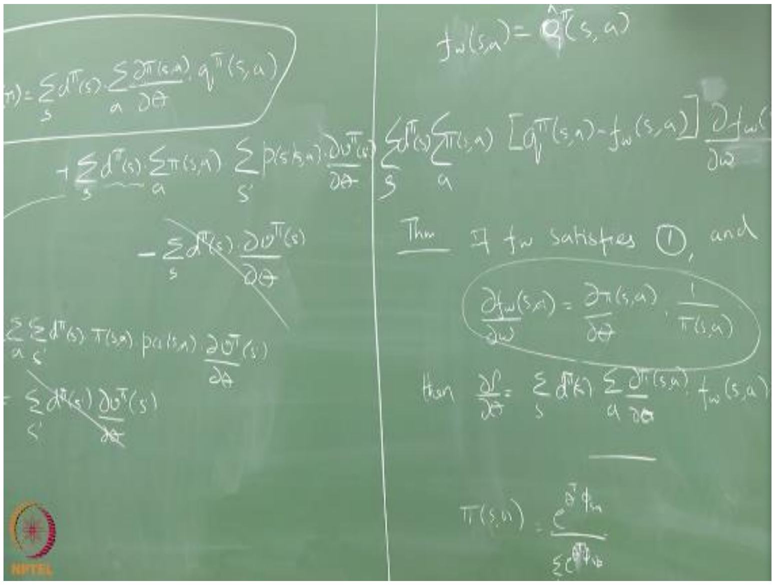
So how often will π (s,a) well that is one part of it right and then I have to be nice but if I have been running π for a long time what is the probability that I will be in nice right and so what will be the total changes in making to the weight function right this will be the total change I make into the weight function right so if the total changes I make to the weight function goes to 0 then converged right so let us assume that what has happened let us assume that my value function has converged okay.

So now we have a theorem that says that this looks remarkably similar to the theorem that we had earlier right there in the absence of any approximation on the value function right we wrote this saying that τρ / τ by θs this quantity right so when I am doing function approximation and if the function approximation has converged to some local point then so that the weight changes of 0 and if my for the parameterization I have chosen for fw is consistent this condition is called the consistency condition is consistent with respect to the parameterization chosen for π okay in the way they define consistence is like this that τ f/ τ w = τπ/ τθ x 1πsa that is basically τln π/ τθ right if the ist he value function here if the gradient of the value function is equal to the gradient of ln π right.

Then I get a very similar condition but with my approximate function put it right so essentially what it tells you is in rough way what it really tells you is that you really only need to get your in some sense your relative ordering correct my fw can be very wrong so what does this convergence condition tell you here okay let us do a special case then it becomes a little clearer let us take our favorites of max right so θ^Tvsa /Σ/ veθ^T ψsp right so this is like the soft max kind of a value function okay.

So what ψsp is ψs,a like you wrote their work I just made it shorter for me to write it some field vector that represents the state action per sa well yeah so I know is little suddenly a switch back to being more mathematical initial but I hope people are coping right that is fine good now what does this condition mean.

(Refer Slide Time: 10:32)



I really need my so what should that be full of you know how to take the derivative of soft max come on so if you put this back in here so guys what does this mean so this essentially you will get τ/ τθ x 1/ πsa so 1/πsa will get cancelled right so you will essentially get τπsa τθ right so what it essentially tells you is in the direction in which the policy is varying right so the error is 0 that is what the they equal to 0 part is so looking at the direction in which the policy parameterization is pushing you know τπ/ τθ you are taking that and you are projecting the error q π- fw on that diver direction right on that.

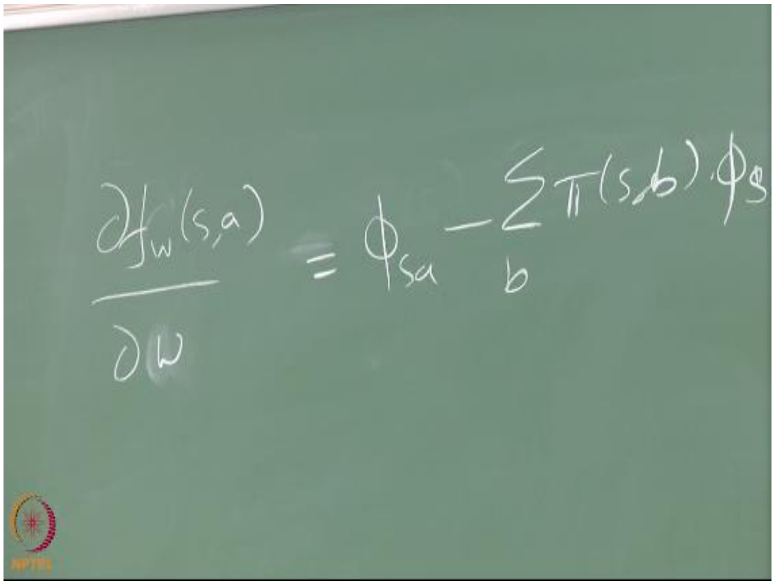
What you are getting summed over all s is 0 essentially it shows that my value function approximation is sufficient to represent any variations I would have in this policy that if I increase the value of action in little bit more like the probability of picking an action little bit more right so the error that I have in approximating the value function in that direction is essentially zero so that so that orthogonality is whatever is what you are looking for from the ovarian function approximation.

So you should if your policy parameterization is very weak right then consequently your value function approximation also can be pretty weak the sense that you do not really have to make too many fine distinctions because there are only a few directions in which you can change your policy I bet your value function with policy parameterization is very rich and you can represent all possible policies then it when your value function approximation also will have to be equally in the ones so that is essentially what the condition means right so people see that right if you take this expression which is τf / τw substitute that here and you will get a $\tau\pi/ \tau\theta$ and $1/\pi sa$ $1/ \pi sa$ will get cancelled with this π sa so all will be left out with this error term that q π- fw into gradient of πsa,/ $\tau\pi$ $\tau\theta$ and that is a gradient of πsa so in the direction of the gradient right this error should be 0.

So when you project this error in the direction of the gradient you should get a 0 so that is essentially what you are or the expected value of the error in the direction of the gradient expectation taken with respect to $d\pi s$ is should be 0 so that is that is the consistent condition right and yeah there are a lot of can theories about what will satisfy this and one thing which I will show now I will give you one example of what we satisfy this if somebody will actually give me the derivative well I have I there but somebody will give me the derivative that will be good okay.

Let this Ø it sends me nobody did the reinforce homework forgot all about it this is exactly what I have to do in the reinforce somewhere like find the upgrade of updates for the soft max.

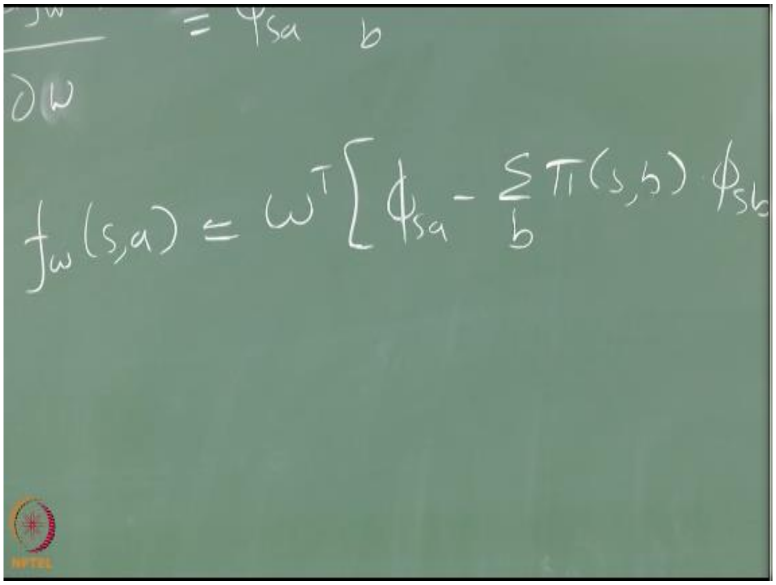
(Refer Slide Time: 14:35)



$$\frac{\partial f_w(s,a)}{\partial w} = \phi_{sa} - \sum_{b} \pi(s,b) \cdot \phi_{s}$$

So that is that so what is one easy way of achieving this.

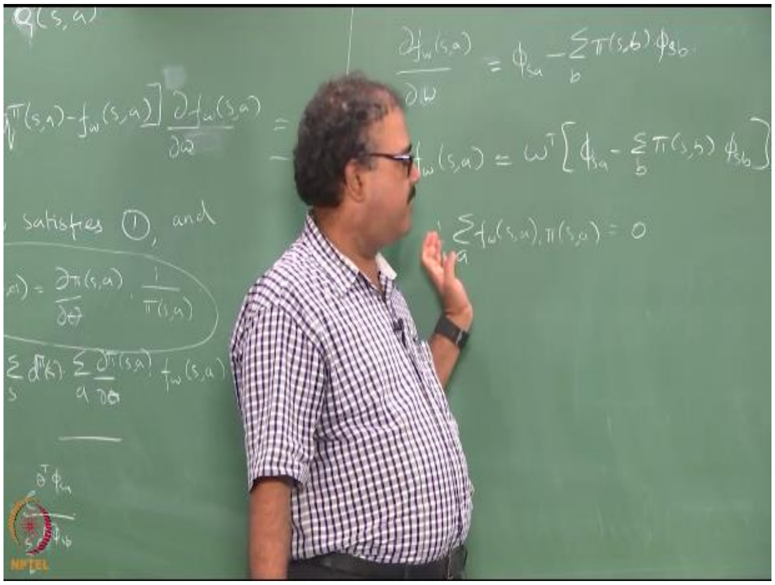
(Refer Slide Time: 14:45)


$= \phi_{sa} \quad b$
∂w
$f_w(s,a) = \omega^T [\phi_{sa} - \sum_b \pi(s,b) \phi_{sb}$

What is the easiest to have achieving this now it is making linear in those and if I want the gradient of f with respect to W to that quantity and make it linear in that quantity right so W^T that now take the gradient it will be there very simple way of doing it in fact since it is hypothesized that this is the only way of doing it okay so in general it is going to be hard to do anything else other than coming up with a linear combination of the what are Ø the representation of using for the policies it Ø is the representation amusing for the policy and so the representation I am using for the value is essentially some combination of Ø right some linear combination of the Ø.

So Ø sa - that so for this is Ø so people agree with this is clear it okay so now let us look at another thing so for a given state for a given state what will be the mean of the feature vector.

(Refer Slide Time: 16:28)



So this is the summation over the mean this mean of the feature vector what will this be so something else I need to do okay so I need to know how often I will be taking action A yeah zero right so it should not will be π a to sa and that will go away because that is when there is only b terms I will take it out so π sa will become one there so it will be summation over a a π sa Ø sa - Σ over b / π sb Ø sb which will be 0 right this 1/ a will not matter at all right so this will be 0 so essentially what is happening here is I have taken the same parameterization the same features that I use for the policy made it zero mean in some sense with respect to the π sa probabilities because that is a probability of me taking inaction right.

So made it zero mean for each state and that is a parameterization I am using for the value function okay so that is essentially the relationship that you would need between the policy and the value parameterizations okay so what does this guarantee it guarantee is that I can just plug in the approximate value function for the actual value function in that expression and I can still estimate me gradient so where is that expression right so instead of the Q function I plugged in fw so fw is a sufficient enough approximation because in the direction I am going to make any changes right that is what we are interested right because this is the direction in which I am going to make changes in the direction I am going to make changes the error between Q and f is 0.

That is essentially what we got here right if we substitute this consistency condition here you will get that in the direction of the changes which is this guy the expected error will be 0 so that is essentially what you are getting here it is a very simple thing in hinds right but it was a very powerful result when they initially showed it because it showed that you could actually use value function approximation right under the consistency conditions and you are guaranteed convergence right.

So that is a very powerful result okay this was the first search results for any kind of reinforcement learning algorithm where they showed convergence with arbitrary parameterization only thing they require is that parameterization you should b or then it is not arbitrary right so well you could say an arbitrary linear parameterization no arbitrary differentiable parameterization this is assumed that you have the gradient that existing.

So they had they need to assume that so the thing is cyclists surmises that this consistency condition can be satisfied only if you are going to do something linear like for completely nonlinear for a parameterization it might not be satisfied but we do not know for sure it could very well be that there are actor criticality architecture where this is satisfied and then you get convergence but that is a very powerful result right so good any questions on this will stop.

IIT Madras Production

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

www.nptel.ac.in