NPTEL

NPTEL ONLINE COURSE

REINFORCEMENT LEARNING

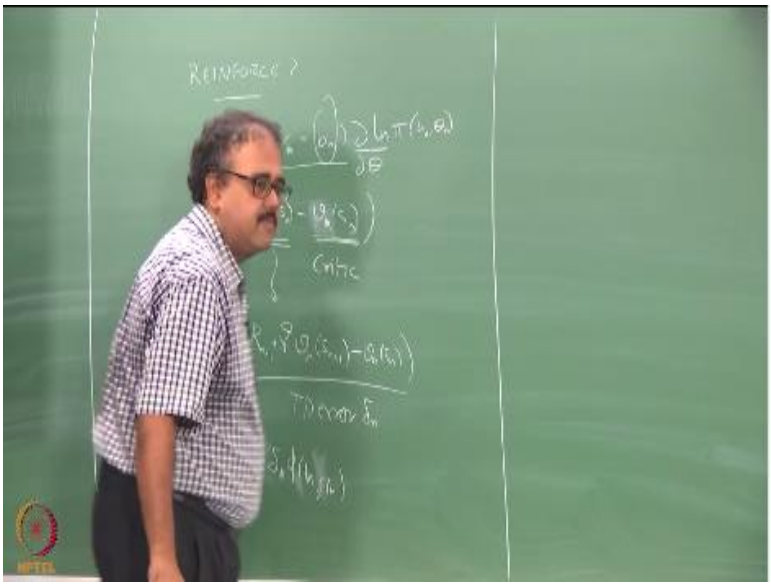
Actor-Critic and REINFORCE

Prof. Balaraman Ravindran

Department of Computer Science and Engineering

Indian Institute of Technology Madras

(Refer Slide Time: 00:15)



So people remember reinforce memory so they were what we did with reinforce anyone can give me the update rule for reinforce. All of you remember reinforce right, so give me the update rule for reinforce the change in Θ will be I use Ti right, and r_n-a δ a_n yeah $a_n\Theta_n$ so this is the update equation right, correct so what was this guy reinforcements baseline and we looked at one example of using this base length did we talked about an example for what this baseline will be. We did talk about it being the average of all the previous rewards right, so and what was the thing I said so b_n is kind of the average of all the previous rewards and if r_n this above b_n then it is a good action.

If r_n is below b_n then it is a bad action that we said that b_n is a baseline right, took it as average of everything that went on before it right. Suppose I want to extend this idea to the full RL problem suppose I want to extend this idea of a reinforcement baseline to the full RL problem so how will I do it, make it as a function of the state right, instead of having the b_n right, I will probably make it as $b_n(s)$ right, and what will r_n become return, so it has to become the return starting from s, so whatever and so something like this if you will start getting right.

It should it be the return can you be tell me something else okay, we will come to that will come to that okay, so this is essentially what I am going to look at right, so what is this guy exactly can you define it before for you say its value function what is it what would it be in words generalizing from b_n what was b_n the average of all the rewards that you would get from that place, so what should $b_n(s_n)$ be the average of all the returns you should get starting from s_n what is that, that is the definition of value function, right.

So if I extend this reinforcement baseline idea to this I essentially end up getting a value function right, but then nothing else has changed I am still solving for the what policy parameters and I am still solving for the policy parameters but in the evaluation I am using a value function and I am still solving for the policy parameters but in the evaluation I am using a value function.

So this kind of generalization of the reinforcement comparison idea to the full RL problem suddenly gives rise to both a policy as well as the value function, right. So this is my critic my actor so the Θ gives my actor and the value function gives my critic, so this kind of a generalization of the basic idea of policy parameterization gives rise to actor critic, right. So that is one way of thinking about what actor critic is correct.

So I can replace this G_n at what right, what is that the TD error right, so this term somehow magically became a TD error, right. So if you think about what is happening with your critic updates I mean sorry the actor updates it becomes some alpha times TD error times this gradient right, and depending on how we choose the policy parameterization this gradient can also vanish. What would be the policy parameterization so that this gradient term goes away softmax will make the gradient involving.

Some kind of a lookup table kind of a representation for them it is not a softmax kind of thing it said I can think of a lookup table I think if it is one if it is that action is taken in that state it is 0 for everything else, right so I can look remember when we talked about value function approximation we said we could have indicator variables like parameterization you can have an indicator variable like parameterization for policies as well right.

In which case you can just write like a TD update rule, right you will just look like a value function of tater will essentially ΔT will start looking like we will start looking α_n Δ_n into whatever is the ϕs_n whatever is your state encoding at that point, right so if you some kind of linear lookup table like parameterization for it will be either 1 or 0 depending on what is happening there right so this is essentially what you will end up with, right. So this is the updation for the policy parameters.

What will be the updation for the value parameters I have to learn v here, right you already looked at V so what will be the updation for the value parameters the TD update rule so what is the difference between this role in the TD update rule you still have not found out what is wrong with this rule of these four bases well I do not know why it is five figures in the policy update was it what is the five fuel in the policy of it I need some kind of parameterization I am using for my policy, right.

Let us say I say this will be like 1 or 0 depending on what parameterization it could be something else also, but if I am using a lookup table like parameterization this will be 1 or 0 there is something missing here nothing big missing here, so we will do look up table like parameterization I will leave the actions also there right, so if so for this state action pair if it is one that means that is action i am going to do in the state if it is zero that is not that answer is going to be that, right.

So essentially what will happen in this case is that I will do the same update except that in the actor I will do the update corresponding to that particular action if I took that action this time I will change the parameters otherwise I would not and for the state case regardless of that action I take I will update the value for every action every time I visit the state right, you understand the

difference so the actor and the critic will always be changed by the same amount which is α_n and Δ_n right.

Both of them will get updated by α_n and Δ_n this the only difference is the critic gets updated the same parameters get updated for the critic regardless of what action you take as long as you visit the state you will update the value function for the state right, this common sense in this is telling you just falls out naturally by whatever the framework that we are setting up okay, and for the actor parameterization right, the actor corresponding to the action that you took will get updated by α_n and Δ_n, right.

So you have to choose the parameterizations appropriately so that you get that but that is the basic idea, so this is the basic actor critic method that is actually presented in the book the edition of the book as it stands now, okay they do not actually do the function approximation part actor critic is presented in the book before the before the function approximation, right so no is actually presented after the function approximation in the second edition of the book but the text that was written was copy pasted from the first edition they have not adapted it yet, and in the first edition it appeared before the function approximation chapter. So the second edition also they do not make reference to function approximation yet, right so as which keeps adding more material to the second edition that will also change this is one peril of using a book that is being written, right.

As we speak so chapters are fluid they may change next week but right now they are the version that we downloaded at the beginning of the semester as this inconsistency it has the same text from the first edition but in a different part of the textbook okay, this is roughly people got what we are doing here right. The second motivation for actor critic is as follows or I already told you the second motivation for actor critic the first motivation is what we are generalizing from reinforcement comparison like using a reinforcement baseline.

We start using a reinforcement baseline per state then you kind of get into an actor critic kind of a setup, right. So the other motivation for doing actor critic is to go back to policy gradient again and say that a policy gradient has a lot of variance so we would like to reduce the variance right,

and therefore we instead of using returns as evaluation for a policy I am going to somehow try to bring in the value function, okay so that is essentially what we are going to see next how do you bring in the value function in that setting there will be slightly different motivation for this right.

IIT Madras Production

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

www.nptel.ac.in