NPTEL

NPTEL ONLINE CERTIFICATION COURSE

REINFORCEMENT LEARNING

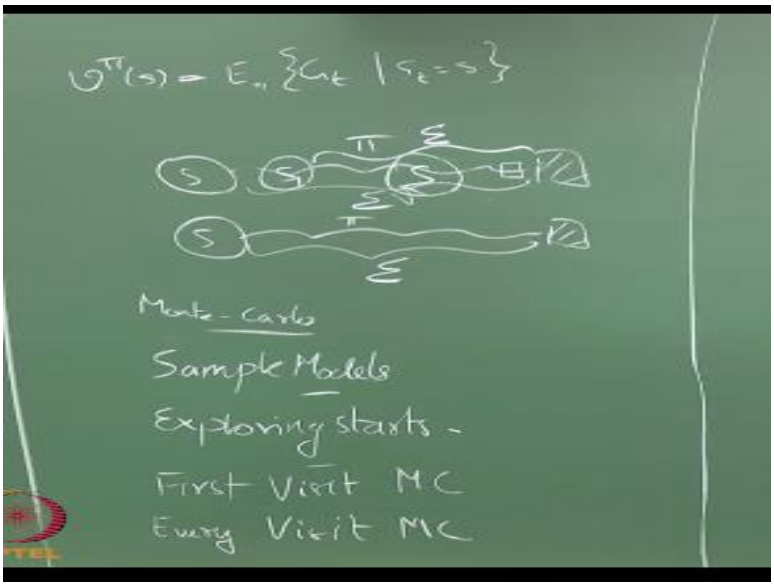
Monte Carlo

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

So let us move on to a situation where I do not have access to the MDP I am interested in solving a problem but I do not have access to the MTP and this is the most cases right remember I motivated you about cycling pretty with this whole thing about cycling and what is your PS' Sa in cycling all if you know what the states are what the actions are you might not be able to enumerate it but you can kind of an intuitive feeling about what the states and actions are.

But you cannot even begin to think of writing down your PS' given S, A right so maybe you can some rough idea of what the expected reward is right so if you fall down your bleeding and things like that so negative reward and those kinds of things but the transition for dynamic sare hard to write but you still want to solve this problems right.

(Refer Slide Time: 01:18)



So I want to look at the whole definition of $V^{\Pi}(S)$ All over again, so what is $V^{\Pi}(S)$ right? Such that is where we started off right and then we wrote GTS GT+ λ may have an RT + λ GT + 1 and so on so forth which are there simplifying but if I had not told you anything else about this whole MDP business right and if I told you that this is a random variable GT is a random variable right and $V^{\Pi}$is the expectation of the random variable and I asked you to give me a way of estimating this expectation.

How will you do that, some random variable I want to find the expectation of the random variable sample and a lot of times and take the average right that is the most straightforward way of solving this right so what we mean by saying sample GT a lot of times for other policy Π so starts in some state S follow policy Π until you come to some end and then what you do you add up all the in some of this some of the reward or some of the discount is the word whatever you do.

Whether you use a λ or not in fact his here I am not even telling you what the reward is neither the λ is there or not is itself a good thing so sum up everything that will give me one sample right then what I do is shattered S again that gives me the second sample so like this I draw many samples by executing Π multiple times and then I take the average that gives me the value in sC test.

He said all I can do with this trajectory so from S I have gone to S1 right so can we use this summation as a sample for V(S1) why is that fine policy same and we had also assuming things are Markov so it does not matter whether it came from S2, S1 or for s' to S1 let him at all it matters is I started in this one and from there on what I did right so since I am assuming things are all markup so I can do this.

So once what essentially what I should do is you should generate a trajectory first look at all the states that occur in the trajectory and then after that take the sum of the rewards that I get and you set as a sample so what I do is a fix a policy Π and I am NOT going to write down the pseudo code for this please look it up in the book I am going to fix a policy Π generate multiple trajectories, okay.

And use that to evaluate the policy so I could get a P' right so we call these kinds of techniques as Monte Carlo methods right because we are in some sense using some form of a simulation to get these samples right so I have this model and they run trajectories on that I get these samples and use that to update the value function so these are called Monte Carlo methods right and so this is a Monte Carlo method for policy evaluation right sided policy evaluation.

So what are the things that you have to note here first thing is I do not really need the model I do not really need the model I do not really need the you are all I need is a ways sampling from that system right I need to be able to run trajectories according to Π okay. I really need to be able to run trajectories coordinates all I need is a cycle and again it is getting on the cycle and start pedaling around and fall over fall down get hurt and things like this all I need is a cycle.

I do not really need you to explain to me what this PS' Sa right so or if you want to think of it in the herd management case all you really need is a infinite supply of course okay you start breeding them you see what happens to them and then you know that you can solve it right so some kind of gruesome but you could you could do that is that is essentially what we are talking about here okay.

But you could also get by without the real system right I do not give you an infinite supply of cows or something but I give you a simulation model so what I mean by a simulation model let us give you an example so how many of you have written programs to play games anybody like play card games okay something if you know about the game blackjack is an example given in the book.

How many of you have heard of the name game blackjack, really heard, heard only I am not going to ask you to play here today okay, so it is like how many of you saw the movie what 21 yeah everyone says then you have a blackjack now they are more confidents of putting your hands up when they said how many of you seen the movie21 do not worry I am not going to play the movie here either.

So it is a card game where you deal a hand of cards to people right and then people can start turning over their cards and then they have to make sure they get as close as two 21 as possible right so the state space would be what, what cards you see right what cause has your opponent has right, so you know the first two cards then you know your cards and then you can keep going if you want or stop right.

Do you see your opponent cards you do not have to stop before you see your opponent's cards yeah you know to stop before I see up front all right so you do not see opponent card so you see whatever cards you have in the keep going and what are their actions yeah was it flush no I am not going to it I am talking about 21blackjack so essentially it is hit me or stay right again it is state means you do not do this I hit me it give me a more card right.

So that is all the two actions either continue or stop right so that is basically it is a very simple problem right but no if you want to think of it as okay what is the probability that the next card that flipped will be an ace what is the probability the next car they flip will be a king right then it becomes a little tricky and it is right so it is you have to do this on a specific game right and then what is the probability that the opponent will win given that these are the cars I have turned over.

Now we have to come into your comment Oriole summation all right what are many different combinations that the operand can flip over and that will actually be controlled by what cards you have already flipped over all right so this becomes a little tricky on the other hand suppose I asked you to write a computer program that will play 21 with me or blackjack with me it is a fairly straightforward thing right.

You just need to keep track of which cards you have dealt and then randomly sample from the other cards and then give it I do not I can write the program that will play blackjack with me much easily write much more easily than actually writing a program that will compute that PS' S, A blackjack right, so such models are called simulation models or sample models okay this is already generate samples.

But they are not a full specification of the transition probabilities right, so I can run these Monte Carlo methods with a sample model I do not have to use the sample model to compute my transition probabilities I can just leave the sample one I can run Monte Carlo method to the sample mall so that is one of the nice things about Monte Carlo methods right, one thing which I want you to note here is the samples I draw like the trajectories are run here.

Not only pick the states at which I will be doing the evaluation they also give me the value with which I will be doing the evaluation right this expectation rate so whatever the expectation computation that is also determined by the sample later right so here when we draw when we did trajectories you are only using them for picking the states but the updation was using this kind of a recursive evaluation right using the model.

But here the trajectory is I draw not only give me the state but they also give me the rewards with which I am updating the value of the state okay good, so this is called Monte Carlo policy evaluation is there any other question that you want to ask me here and the Π could be stochastic why it should be I am evaluating a Π and I am only interested in evaluating a Π why it should be if Π is deterministic you need to do one small additional thing what is that.

Very simple I am talking about very practical things here yeah I have a fixed Π okay and look for the back of a policy evaluation I have a fixed Π there is no greedy no nothing business okay no exploration business if I explore the policy will change right I am giving you a fixed Π if I am going to explore then that if the Π is probabilistic then already and doing some kind of explains to Casting I am doing some kind of exploration.

If the Π is deterministic if you know if we explore on the Π then it is not no longer Π right I will come to that case later but this is a good point now so what should I do, I am interested in finding v Π (S) right if I am only interested in the small subset then a sufficient for me to start my trajectory is only from that small subset but if I am interested in finding the value for all these TX in the MDP then I allowed to make sure that I atleast start from each and every state.

Because it is a deterministic policy right I have no guarantee that the states should be visiting only a specific trajectory that I will be following repeatedly right there is no guarantee that the states that I do not start from will ever be visited for me to get samples from them so I have to start from many different states so that I can get samples from those states okay, so infact even if it is a stochastic policy you will have to do this exploring starts.

Because there is no guarantee that you will hit every state even with the stochastic policy, right unless the policy happens to be something that log you to explore the entire state spaces like a random policy if you have like a completely random policy every state I pick an action with probability half or whatever 1/n, so I have n actions in each state like each action with probability 1 by n.

That essentially means I leveled up covering the entire state space provided no provided the entire state space is reachable so I can I can have a grid word like this that is a wall right so everything here is I can go everything here they can go if I start from here and go even if I am acting randomly there is no way I will go to this part of the world, right if I start from here even if I am acting completely randomly is no way I will go to this part of the world right.

So they have not reachable from one another was it how do you find do you need to is a question so if you need to there are ways of doing it for smaller problems for larger and larger problems they become not ready tractable right so you can use techniques from dynamical systems you can use techniques from physics you can use techniques from graph theory depending on what assistance is set up that you are interested in solving right.

So where were we so sample models then exploring starts so is one other question that I am expecting people to ask me no one is happening you know I start from s I run a trajectory so use the trajectory value for updating the value of s, okay so what you hear sub KS again in the physically this is two trajectory starting from s or one trajectory sorry promise why is it one when we started from S.

I said this is next state after case 1 then we can consider that as a trajectory starting from s 1 right and we are talking about probability of transition is that even if I am using a deterministic policy as a probability of transition right so I may come to s and then I may skip and go out of s and extreme as possible I mean here is a simple example so I have S only one action from S I do that so twenty percent of the times I am going to come to S and then I am going to stop right.

So S stop SSS stop so that is also going to be perfection it is possible, which is correct one trajectory or two trajectories? One trajectory how many of you say one trajectory, no my question was how many of you say one trajectory either put your hand up or down put your hand up so many said one trajectory yeah only one person is putting his hand up okay, how many of you say two trajectories.

What about the rest a tragic releases zero you have to have an opinion I am going to start from 0 again so I am going to keep going until everybody put their hand up 0 okay turns out both are right one and two right, so when you do one it is actually call first we set Monte Carlo when you say to what is called close, it can be better than that right, so if you use the only the first occurrence of the state is called first visit protocol know if you use every occurrence of the state is called every visit want to call.

So when you say you use the first occurrence what does it mean so I take the return occurring after that let the entire thing you set as one sample for updating v Π of it when you say every visit that means I take the return after every occurrence oh right so the first return will be this much right now the second return will be using will be this much okay, so what is the real problem here is that when I do every visit Monte Carlo.

I have just taken one sample trajectory right but I am using the sample trajectory twice atleast part of the sample trajectory I am using it twice okay why is that a problem yeah so is that a problem so when I approximate an expectation with samples I am assuming that the samples were drawn according to the distribution on which I am taking the expectation right, so the distribution yield at me this one sample once right that one trajectory delayed it once.

But I am using it twice essentially I am changing the distribution right so that is a pain problem so I am not using the same distribution that I am taking the expectation with respect to but it turns out that all this adds is a little bit more biased estimation people remember bias-variance strain of it is not very important to know if you do not remember bias-variance trade or never heard of bias variance.

So it induces a little additional bias into estimator but I turns out that in the limit right if you do a lot a lot of samples in this particular case it turns out both converge to the same value function so it is right, see I roll the dice and take a sample right I mean so essentially I am for every step I take I roll the dice first pick an action according to my Π then I pick a next state according to PS'SA and then I pick a reward according to the expected reward so on so forth, I so I do a lot of simulation.

So a specific trajectory is now appeared so S, s1 s2 s3 s4 and then for each one of these I get r1 r2 r3 r4 and I would up into actions a1 a2 a3 a4 right but then how many times this kind of a thing occur should be determined by rolling the dice again and again rolling my coins again whatever right tossing the coins again and again but I did not do that right I just generated this once sample once and I kept saying okay.

So this part is one sample now this part is another sample this part is another sample so for one roll of the die and this taking multiple subsections of it and then repeatedly using it right, so if I had started from S here and rerun the experiment and I might not that this might not be the one that I produced that I will actually produce something else again alright so essentially that is the pet sitter keeper that is why I am saying you are changing the distribution to some extent.

Was it should not hurt the same saying it will eventually converge to the same thing so if at all there was a thought about why you should not do every visit that was the reason or it affects its effects in some sense but it is it just how do I put it let us say that there is something here that happened right that is very rare very that it occurred once right it will probably occur once in a million times if you are actually random independently it occurred once on this trajectory .

But you are going to count it twice once for this is one saw that is right but if you had drawn independent samples it will happen only once every million through the fleece now it is happening twice everybody into the cliffs which is a significant change, in something you double the probability of it is missing so that is that ids why it is it could be a problem using a very visit NC could be a problem right.

So there any questions on this so far the thing you have covered all the things I wanted to do so Monte Carlo equal to sample models exploding starts first was it revisit.

IIT Madras Production

Funded by

Department of Higher Education

Ministry of Human Resource Development

Government of India

www.nptel.ac.in