NPTEL

NPTEL ONLINE CERTIFICATION COURSE

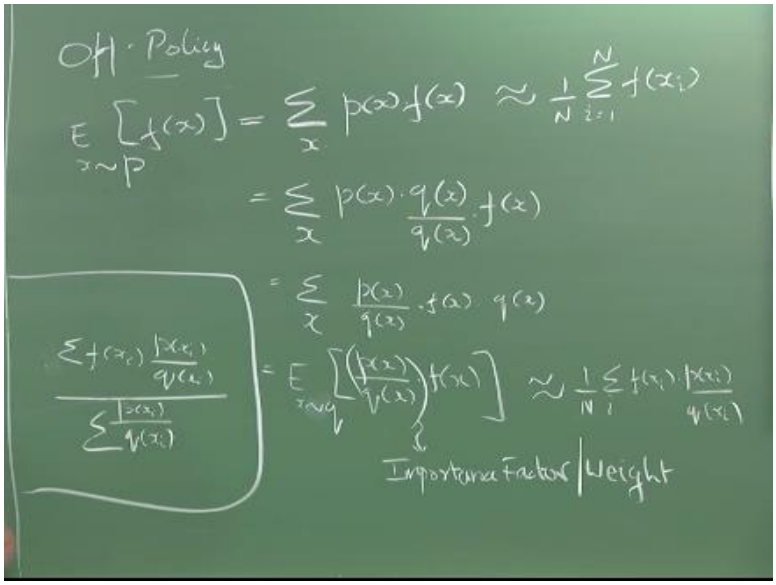
REINFORCEMENT LEARNING

Control in Monte Carlo

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

Okay now what do we do what is next exactly getting there right so we have now told you how to evaluate a policy.

(Refer Slide Time: 00:41)



Now the trick comes in or how to how to improve the policy how to solve the MDP basically I mean so this is called evaluation so this is called control sometimes right how do I learn to control the MDP so this is kind of terminology inherited from I mean so legacy terminology so this is inherited from control theory so now I want to know control so I have evaluation and I have control right so I have evaluation i have done that.

So how do I how do I do control very simple right so I evaluate pixel policy by evaluate I do generalize policy iteration so not the extreme form with that I just basically the answer is positive fraction so I fix the policy Π evaluated using this kinds of trajectories I do not run it until convergence right so I just evaluated on few trajectories right what do I mean the evaluation on a few trajectories here I start from SIM status.

Just keep running the trajectory multiple times and then take the average and you say the C value for status right I have no guarantees that this value that I have computed is a true value for status under the policy Π look in some approximation of that but I am happy I will stop there right oh yeah this is one thing about generalized policy trajectory I did mention right so you do not have to value we do not have to find the true value of a state also.
As long as you can find an approximate value for the state right you can keep going and do here then release policy information that is essentially what we will do here so we take a state we take some samples from the states right and then find the value and then you can go and try to be greedy with respect to that value function okay so any problems that you see here how will you be greedy with respect to value function that way what is the problem being greedy that way I need that Π need the expected RSA.

I just told you we doing Monte carol methods because we do not have any of those or we do not need any of those and I can work with a sample model right but now you are telling you just use that expression to be greedy I cannot how do I get it on that you already know how to do it good throw a chocolate you of all people know the answer do not use the v function use the Q function is the Q value function.

I do remember I told you that and you want to recover a policy from the queue all you need to be you just look at the values of those actions at the state I do not need to look ahead, right so when I do control I use the Q function, now things get really interesting so how will they just do Monte Carlo evaluation for the Q function, exactly so basically starts with an s fix in A right and then follow Π thereafter.

Then some of the rewards the summation and that gives me one sample for sa right let us straight forward enough I then I do this again I had another than other physically no physically no physically in it so the crucial thing is I start from a sign it take a okay, suppose I reach S again what do I do okay forget it let us not even go there okay there is even more fundamental question than that okay.

So I did S I did A I get s one here okay I did some A one that is how action that suggested by policy Π right so this is equal to right so Π of s1 is a one so I do a one right now can I consider this as a sample taken from s1A1 you would think so it turns out that if you do this you get into all kinds of problems mainly because you draw a disproportionately larger number of samples from actions corresponding to Π.
As opposed to actions that are not corresponding to Π, so it tends to terms of it sets up some conditions in which converges cannot be guaranteed right so you will get somewhere answers so you only have to know I want to evaluate all sa paths right only then I can do a policy improvement right there only evaluate the actions corresponding to Π then I cannot do policy improvement right so turns out that I will be disproportionately evaluating actions corresponding to Π right and therefore we never do this right.

This whole sample and I use it only for updating the value of q SI at all the sample value is only for q SI and never use it for qs any one so if I want update the value of Q s 1 a 1 I need to start from s 1 a 2 a 1 and then follow up I that after another thing you are drawing disproportionately large number of samples about one thing right it turns out it sets up some kind of instability in the convergence it is a little tricky.

And in fact for a long time the convergence was of open question for updating on s1e1 also right in fact If we look at look at the older version of the textbook they actually say in the textbook it is still open question because the text book was written in 97 and it was proven only in 2002 that it will not converge if you do that and they had to actually work on for two decades to get to the answer is a little tricky to give you the intuition.

But the intuition is that so you are getting too many samples of one thing therefore the rate of convergence for this that SA is much faster than s1a1 will be much faster than the other kinds of things and therefore that you get up to get all kinds of problems, so 11 probably convergent phase 2 only to use the first state for state in first action even longer than what yeah I mean I can estimate the value.

But what will they do with it right I mean I can assume it only Π right I cannot do policy improvement on that having said all of this it is only for convergence issues right it turns out most cases in practice if you update for s1e1 also it works okay in practice it works but you cannot guarantee convenience so guaranteeing convergence you should do it only on the initial state.
So now once I say that you have to do it only on the initial state you exploring starts becomes even more important and because you have to start at every state every action you have to do that often right so that you can get a good estimate for it and so you cannot be greedy at a particular state if unless you have taken at least a good fraction of the actions from the state right and try to estimate the thing so those becomes important.

So one way of avoiding exploring starts right again it is not clear whether this kind of approaches is going to be convergent is to use something called a epsilon soft policy right, so whenever I say do a policy improvement I do not give you the greedy policy I give you a policy that is mostly greedy right so most of the probability will go to one action and some of the probability will get distributed to all the other actions okay.

So when I start from anywhere and earn a trajectory there is a probability of me sampling all the actions from all the states is not just the Π action corresponding to Π but I can sample all the actions from all the states so these are called epsilon soft policies, right so epsilon soft it means that every action has at least a probability of epsilon being taken from every state right so uniformly random policy is an epsilon soft policy right.

An epsilon greedy execution of a policy is also epsilon soft policy and where the epsilon is a epsilon by n soft policy epsilon greedy policy is an epsilon by n soft policy right a purely random

policy would be some for some epsilon it is a 1/n soft policy rate if I have n actions it is 1 by n soft policy state action tap they did can you see how much yes says this is a very simple straightforward thing right.

It depends on the number of samples you draw right so there are rates of convergence that you can establish based on the number of samples inequality if you knew the transition probabilities and other things you could apply concentration inequalities there yeah good I do not recall the actual results but you could apply concentration abilities yeah sadly same innocence more like a textbook exercise application of concentration inequalities.

But little complicated because the distribution you are sampling from is this horrendous the distribution that rolls in your oh in this case it is me fine right it is just a Markov chain I fixed a Π right I have fixed up I so there is no question of picking in actions here right so it just becomes a Markov chain and essentially I can drink this valley computational approach and it is very easy to show keen convergence and rates of convergence all of the things you can go here it become tricky.

Because they keep changing the policy from one step to the other so it becomes a little tricky to show convenient right so epsilon soft policy so now the question becomes that we showed you policy iteration converges right but that is assuming that you are greedy with respect to the previous policies value function now I am saying do not be greedy be epsilon soft with respect to the previous value.

And will this converge and if so to what okay and we are running out of time so it will converge and it converges to an optimal epsilon soft policy where the more highest probability will be given to the optimal action provider do you have a very specific way of generating the epsilon soft policy from the previous iteration right so it is essentially you take the greedy policy okay make it epsilon greedy or whatever epsilon by n greedy or something n epsilon greedy.

And then keep doing this so you will converge to something that is optimal in the sense the highest probability will be on the optimal action okay there is a very simple proof given in the

textbook I encourage you to look at it just so you know and themes any questions from this, there is one other topic that I have to cover under Monte Carlo methods which is do with something called important sampling.

This goes back to one of the questions raised earlier about doing epsilon greedy exploration and so on so forth so if I am using an epsilon soft policy right whenever for non exploring starts right I am not really getting what I want right I am actually solving for something else right and if you want to evaluate a deterministic policy right we saw that you needed through exploring starts right if you want to evaluate a deterministic policy certainly have to do exploring stats.
And I cannot do my usual exploration of the entire state space and then estimate the value of the policy Π so why because this will no longer be the policy Π right so ideally what I want is a situation where I can behave however I want in the state space but should give you the value for any policy that you ask me is it possible can a d couple the policy that I am evaluating from the policy that I am using for behaving right.

So such kinds of methods are called of policy methods so whatever you have talked about so far are called on policy methods why because the samples were generated by the policy which I am trying to evaluate whether it is here or here right in both cases when the samples are being generated it is with respect to the policy that I am trying to evaluate right in our policy methods the sample will be generated by a different policy.

But I will try to evaluate a different policy technique we have 10 more minutes okay let me let me try and do the simple thing right so essentially so I what I am looking at this I am I want to compute the expectation of some function f (X) where X is sample from some distribution P this is the rotation for this sum of f of X is there and X is some random variable right and X is sample from some distribution P.

So this is the expectation that I want to compute let us keep it simple so that is assume it is a discrete distribution so essentially a sum over X right so this is the expectation now suppose I do not have access to p okay but I have some way of sampling X this is exactly one of the

distribution we have whatever we have the setup that we are here right so then what will I do so I use P and asked me to give you ask you to give me an X.

And you will say three then you will say for you will say three 4839 let us keep telling me numbers like this now what I will do is I will plug in 3 for 8 9 whatever I plug that into f of X I will add up all the values of f of x divided by n it will give you a number right so I how I approximate this is essentially very checks I was sampled from people this is how i do this right.

But now let us say that I know what I know something about the probability of certain outcomes okay but you can only tell me things according to Q let us say so you are going to give me numbers according to Q so Q is a different coin that you are tossing right so instead of selling me 338 or something that now we are going to tell me 88 97 3874 or something that I so three and four occurred a lot of times in the first sequence of numbers I gave you.

But three or four occur very few times now but eight and nine occur more right so now what I do is I am saying that I do not have p but I have some other distribution how do I do this so I have some other function Q alright so I am going to do this is it finds people are okay with that so I am going to rewrite this as right, so I have taken the expectation off of X with respect to P and I written it as the expectation of some quantity with respect to the distribution q.

So what is this quantity it is f of X but waited by the ratio of the probabilities right if are equated with ratio of the probability now I want to approximate this so what will I do I will do this so for every sample you give me I will compute f of X right and also the x this ratio of P by Q and then we will add up everything and take the average that will give me the average of f of X if the samples are drawn according to p right.

So I said fine so everybody is with me write this equivalence you are all agreeing so this quantity is called the importance factor it tells you how important is the specific X in determining the expectation of f of X with respect to P when I am drawing samples according to Q so it gives me

the relative difference in their probability so if P is very large and Q is very small then the importance will be very large for that X.

So if P of X is very small and pure fix is very large then the importance will be small I do not care right because I am only interested in P of X if there are comparable probabilities then I will just use it as it is right P of XY q of x will be 1 I will be using as it is so it is fine right so if Q if X happens to be more frequent under Q then I want to give it a lower weight when they compute the expectation of X happens to be less frequent under Q then under p I want to give it more weight when I am computing the expectation so this way it is called importance waiting.
And importance weight I am sorry important factor or importance weight right and this whole way of sampling to determine the expectations is called any guesses important sampling yeah I am assuming p and q are no it is just that you are not able to sample from P notable to draw samples from p to draw samples from q can you think of a situation where that might happen when you want to draw a samples from a distribution I told you that I have to find thee this CD F.

So for serious one or the CDF is hard to compute right4p so the Q CDF might be easier to compute earth might be easier to actually simulate samples from Q then from p right so in such cases I might want to do this important sample right or in the cases where we are talking about I will have a deterministic policy or I have a policy that is hard for me to execute I have some random behavior policy right.

So I use the random behavior policy and generate trajectories some my f of X will become my gt-right the random sample I draw for the return I take from the trajectory right and my P and Q will be policies according to which I am behaving but so is there any condition that we have to be careful about here QX will become zero right away if Q X becomes 0 then it is going to become undefined.

So one DC wait assume is that the Q X is not 0 for any X but that is too strong a condition what we need to assume is QX is nonzero for any X for which px is non zero so 40 x 0 again kind of fed gen say that okay does not matter but if px is nonzero then Q X has to be non 0 and 1 DC

way to ensure that QX is nonzero for any PX is epsilon soft okay yeah so whatever yes of soft matte is another way of generating in epsilon policy here right.

So however you generate it which could be epsilon DD epsilon of whatever the saying they should ensure that every action has atleast some small epsilon probability of being picked so that way for whatever be the policy that you want to estimate the value for this will never become zero whenever you want right so that is a good thing so this is all fine right so one small aside it turns out that this can suffer from instabilities if you are trying to do this numerically we are trying to do this why is that.
Because for cases where px is non zero but QX is close to zero the importance weights can become very large it rare events under QX right which are not right under PX the importance can become very large disproportionately large and so you can cause severe oscillations in your estimations that if you add 1 1 1sample like this so what people typically do is instead of doing this right instead of dividing by 1 by n.

They do so the ratio PA x qi becomes very large it will happen in the numerator and in the denominator so it kind of the ratio will be kept reasonable it would not oscillator too much okay so this is called the normalized important sampling this is please do not read anything in this place okay, so this is also an estimate of this but this turns out to have slightly higher bias than that when that will have a higher variance.

So for every sample that you draw it can fluctuate tremendously right there which will not fluctuate as much for every sample and in the limit both converge to the same answer so we are all happy right but so this is weight over this normalized important sampling is considered better in practice because it is a lot more stable okay, so this let us it with the Polish sampling so I am going to stop today.

So keep the important sampling in your mind somewhere right when you come to class on Wednesday I will tell you how to use important sampling to do of policy learn okay we will work out exactly what is that p and what is the cube that we have to use for important sampling okay that will take me about 15 minutes and after that we will move to chapter 6 that will be the

end of chapter five so we are done most of certified today so there will be small bit of chapter five left in the next class and after that we will move chapter six.

IIT Madras Production

Funded by

Department of Higher Education

Ministry of Human Resource Development

Government of India

www.nptel.ac.in