NPTEL

NPTEL ONLINE COURSE

REINFORCEMENT LEARNING

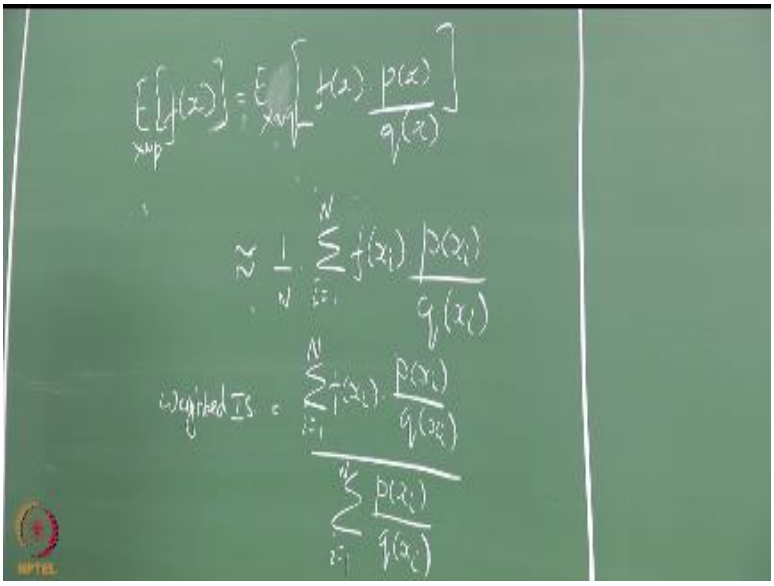
Off Policy MC

Prof. Balaraman Ravindran

Department of Computer Science and Engineering

Indian Institute of Technology Madras

(Refer Slide Time: 00:15)



So we started looking at off policy Monte Carlo methods in the last class, right and let us continue from there, so people remember important sampling, yes I said something like expected value of what is the, right so expected value of f(x) where is x is sampled according to p can be given by expected value of f(x).p(x)/q(x) where f is sample from q, I mean x is sample from q,

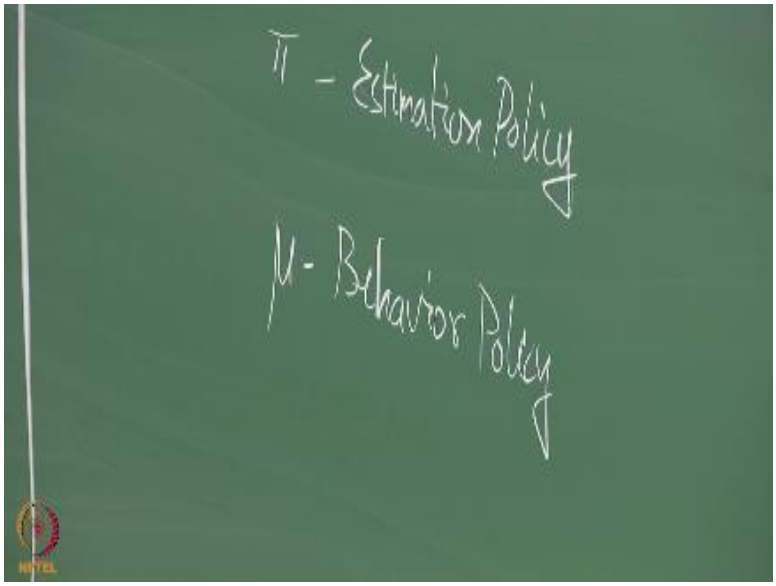
okay so this is the basic area been an important sampling and we saw some expression for it, so which was essentially so notice that there is no approximation here, right.

So this is approximation we do 1/n right, we also wrote another sampler which was okay, so this is the I think this is called the weighted important sampling estimated right, this is a weighted important sampling or normalize right, so different ways in which people calling, okay. so how are we going to use it in Monte Carlo learning so what are the samples we are talking here, what is our f(xi) what is xi what is f(xi) right, and what is p(xi) and q(xi) what are these quantities in a Monte Carlo setting.

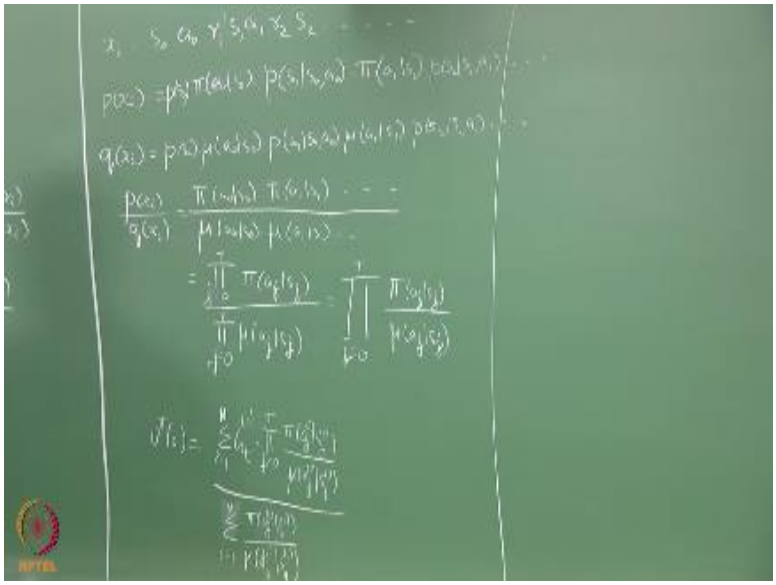
Start with the x, exactly it is the trajectory where you saw doubtful about it if you say trajectory yeah, okay good. Well, I am happy nobody said it depends because it does not okay, so in this case xi is the trajectory and so you are drawing samples you are drawing trajectories so what f(xi), the return right f(xi) is a return on the trajectory, okay. Now what is p(xi), little bit more, little bit more you have to related xi, right probability of the trajectory given that I am following policy π, right.

So p(xi) is a probability of the trajectory xi given that I am following the policy π according to which, for which I am trying to do the evaluation, okay. So what will be q(xi) probability of the trajectory under the policy that I am actually following, okay so for to sinking up with the text book case we will π is the policy that you are interested in and μ is the policy that you are following, okay sometimes it is called the behavior policy.

(Refer Slide Time: 05:17)



(Refer Slide Time: 05:26)



So let us say xi is some trajectory okay, so x is some trajectory which starts of with some state s_0 right, what will I do after that right, so that will be a trajectory correct. So there is some, so what will be the probability what will be the p(xi) so in fact I can leave out the rs because they will be captured in the f part, right so I can actually ignore I mean the r are there, right but when I am computing the probability of xi I can ignore the r parts, because they will be captured in the f, okay.

So I must put them back here, yeah so what is the probability of xi what it will look like, $\pi(a_0)$ it will looks like a q, a_0 given s_0(p(s1)) given s_0,a_0 into anything else, right so I need a some probability that I will start with s_0 okay, that is the USL thing, is it fine, okay. So what will be q(xi) right, so can you compute p(xi) trick question. I need the ps right, I do not know the p I mean the whole idea behind going to Monte Carlo methods is we do not want the ps right, likewise exactly but if you think about it where we need p is only as a ratio with q.

So I never need to compute p and q separately all I need to do is compute the ratio so if I take the ratio what happens it reduces to, right if I can write this as a product of the ratios that is fine, so I think that might been easier way of doing it, so I mean historically I mean π is used for both the product as well as the policy so please this ambiguity on context, okay it is clear. So now I know how to find the ratio p(xi)/q(xi) so would this kind be then, so if you think about it little tricky I should not have used i here I apologies, really apologies.

Okay, we have i on the left hand side I should not have use i as the running variable, sorry okay, so what do we have here, so what is this correspond to what does it weighted important sample here correspond to the expected value of f(x) right, so what is that, what is that value function right, so the value function of some state right, $v^{\pi}(s)$ right some s, right. So essentially we have now $v^{\pi}(s)$ is equal to, so let us put a super script in brackets i to denote that this quantities come from the i^{th} trajectory, right.

So this i=1 to n means I am running n different trajectories so I put this super script to denote that they are coming from the ith trajectory and the return is computed starting from state s, okay so likewise I mean I need to have a thing here, so is that fine, so this is how I will do Monte Carlo policy evaluation, okay in off policy fashion, okay. So why do I want to into off policy learning right, so this is allow me to explore lot of lot more states right, then if I am following policy π so I can get larger samples of states values estimated and so on and so forth.

And it can also be that maybe view and so easier exploration policy than what I have with π I am trying to sampling from so that could be variety of reasons well I want to do this and I encourage all of you to read the text book, right and yeah I miss something in the denominator I need a big

π thanks, right I just indicative here all know what I missed there, so just filled in, okay. So there were few tricks that they do in their book they take this and convert this into an incremental method where you can keep I mean you do not have to wait till if you finish running the trajectory to compute this importance weight.

You can keep updating it as we go along right and then you can keep updating this estimate also as you go along in the trajectory you do not have to remember everything right, so they give you all kinds of small arithmetic tricks to make it incremental right, so again I encourage all of you to look it up right, if you are going to ever implement a Monte Carlo method that is the way you will do it not like this, right so I am really, really trying to get you guys to read the book, right.

So somebody other than Sekhar I mean try to guilt him into reading the book yeah, anyway great, so this is all fine for evaluation so what do you do for control is there anything else you have to be careful about for control. What said, yeah so how will it change, how will this change for first visit and every visit, will it change, we need action values so what is the problem with using action values there is any specific problem if receiver using action values, okay.

So I am going to integrate actually go read the book, okay I am not going to tell you just go read the book and figure out how would you do off policy Monte Carlo control, okay.

IIT Madras Production

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

www.nptel.ac.in