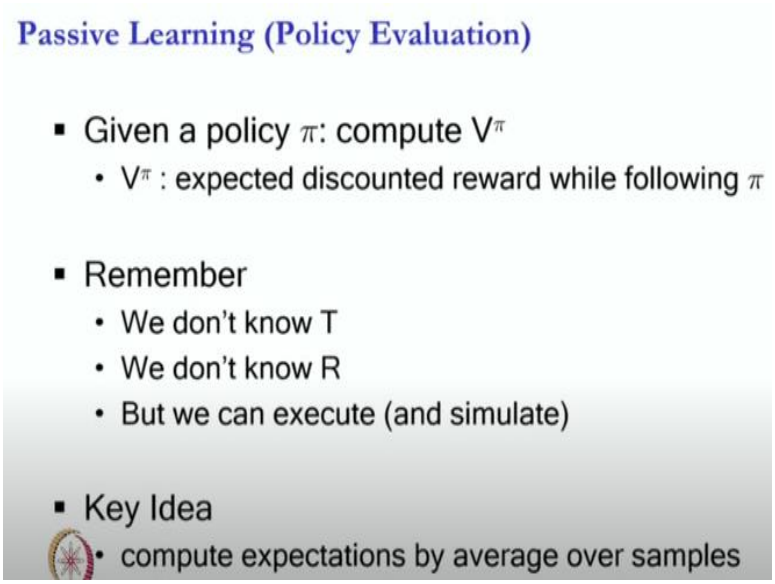
Artificial Intelligence
Prof. Mausam
Department of Computer Science and Engineering
Indian Institute of Technology-Delhi

Lecture - 77
Reinforcement Learning: Model-based Learning for Policy Evaluation (Passive Learning)

Okay, so with this background let us get started on our first problem.

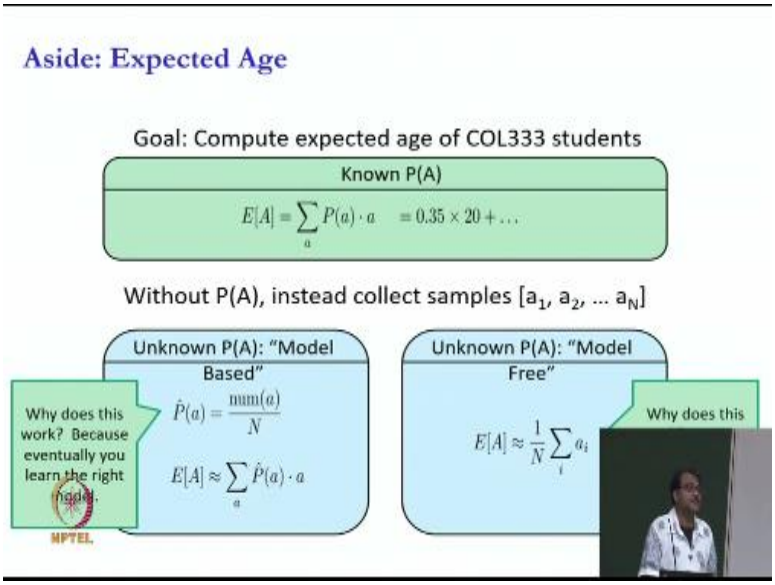
(Refer Slide Time: 00:27)



And the problem will be obviously passive learning because we will build some passive to active we will do policy evaluation pretty much like we did in the MDP lectures. So we are given a policy pi we want to compute V pi, the same thing. And V pi is defined as the expected discounted reward while following pi starting in a given state. And of course, remember, we do not know transition function, we do not know reward function, but we can execute the policy and simulate.

Okay, and let us assume we are in a weak simulator world. So we want to solve a general enough. And the key idea that you will learn today is to compute expectation by taking average over samples. That is it. If you know this idea, well, and I think you know but you have not maybe explicitly seen it in this form, or maybe you have, then everything is a just an application of that carefully and intelligently. So what is this idea of computing expectations by taking average?

(Refer Slide Time: 01:27)



So let us say my goal is to compute the average age of all of you, okay. And while you think that the average age would be you know 20, but it may not be. Because there are some masters student. There is a couple of PhD students. There is some advanced students possibly you know who did not do it last year. There are some students who are from a different university outside the country.

So their ages may be very different, right? So let us say this is our goal, right? And so we can always use the formula for expectation. Now suppose I give you the probability of each age. This is akin to known model, I know the model. Like I know the transition function and reward function. So keep thinking at the back of your mind. So suppose I give you the probability of each age.

So I tell you the probability of the age being 20 is 0.6 and probability of age being 21 is 0.2 and probability of age being 22 is 0.1 and so on so forth. Then can you compute the expectation? You will simply use the formula for expectation of a discrete random variable, and you will just do a sum over a P of a times a. Where a is your age, P of a is given to us. It is the probability of each age.

And so you can say 0.35 into 20 plus point whatever into 20 in our world 0.6 into 20 plus 0.2 into 20 okay. Now this is the easy part. Suppose I do not tell you the probability of the age. What would you do? Come on. The question is clear, right? So I am not giving you probability of the age. So what will you do? Ask people. So you

ask somebody, they say 20. You ask another person, they say 21. You ask somebody else, they say 20.

You ask somebody else they say 19. You get this data. So you have now generated samples for the age a 1, a 2 up to a n. What do you do next? Well, you can do two things. Can somebody see that we can do two things? Okay, so Poorva what is the obvious next thing we can do? Exactly. So Poorva says the obvious next thing possibly to do is to compute probability of this age wise sampling, right?

How would you compute probability? You will take samples of various ages. And you would take how many times the age came out to be that number divided by the total number of samples, and that gives you the estimated probability. So this you would do where you would compute P hat of a. Let us say hat means that this is my estimated version.

And you compute P hat of a is number of times a came divided by the total number. And now that you have the probability distribution, you can compute an estimated expectation by simply taking the formula and putting in P hat instead of P. And in the limit of infinite samples, this gives you the right expectation in the case in the limit of all of your ages being sampled, this gives you the right number.

This is what I am calling the model based computation of expectation. And this works, because eventually you learn the right model, right. So in the limit of lot of samples, you finally learn the right model. Or as more and more samples come your P hat becomes closer and closer to P, and eventually your expectation becomes closer and closer to the origin expected. However, you can alternatively do a second thing.

What is the second thing you can do? You can forget estimating probability away and simply take the average of all the ages. Now why does this give us the right expectation? This gives us the right expectation because the ages are not uniformly distributed in this computation. And age, which has higher probability will come many more times the same fraction more times compared to a different age, right.

So if I had probability a 1 and probability a 2, a 1 was 3 times more likely as a 2, then I will have 3 times more samples of a 1 and you know less samples for a 2. So when I do the sum in the numerator, when I take the average a 1 would be, you know added 3 times more than a 2 and this would be the correct expectation. So this works because samples appear with the right distributions. And this is your key point.

If you understand the slide you will understand a lot of what is happening in this particular topic. Whenever you see that I have to compute expectation and I have to do it by sampling, you should quickly recognize that there are two ways to do this. One is to estimate probabilities and then take the expectation or other is to not estimate the probabilities, get samples with the distribution of the probabilities and then take the average. Any questions on this? Yes.

"Professor - student conversation starts" Sir, if you in the unknown model paste the if you substitute P a hat and find then you are getting the same formula as the models there. So in one just instance of data collection how can you say that these two methods are different. **"Professor - student conversation ends".** So Jay says why are these two methods different? And there is a short answer to this.

He had a long question, but this is a short answer to this. Short answer is that in the second version, even though I can estimate P hat a I am not estimating P hat a, that is it. That is the only difference. Eventually, these are mathematically consistent things. So if you try to estimate P hat with the same distribution, you are going to get the right thing. If you take the model base you are going to get the right thing.

The beauty here is, and this is, I mean, this does not look apparent in the slide, but it will look apparent when we are doing one step at a time and doing online estimation and so on so forth in RL. But basically, the beauty is that if our goal is expectation, in the first case, we first estimate P hat and then estimate expectation. In the second case, we can estimate P hat, but we do not estimate P hat and just take the average.

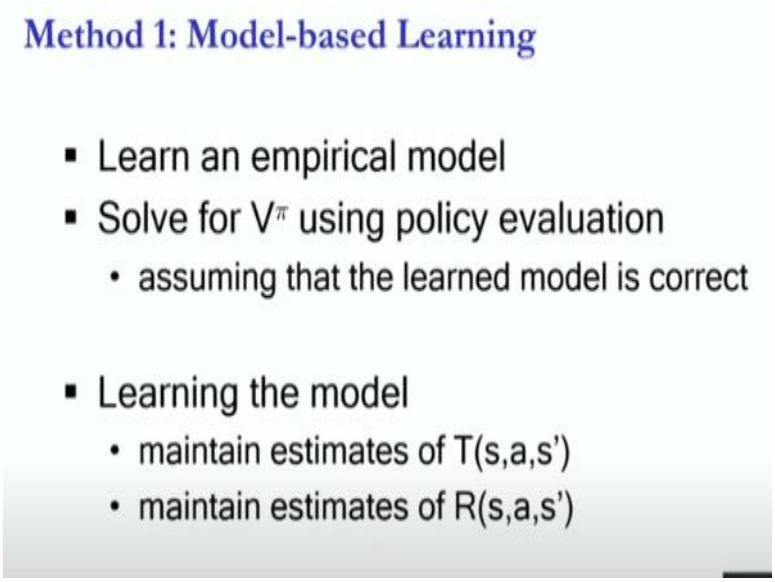
That is the only difference. And now we will even think about how many parameters are we estimating? Let us think about how many parameters are we estimating? In the unknown model case, how many parameters are we estimating? Let us say your ages

can be anywhere from 18 to 60. How many parameters are we estimating? We are estimating 43 parameters in the first version right?

In the second case, how much are we how many parameters are we estimating? Do you see the difference? Now you were asking what is the difference? The difference is that in one case I am estimating too many more parameters. And if I believe that estimating them correctly requires me to take more and more samples then I might get a better of convergence in the second case with fewer number of samples right.

So there is a sample complexity issue here as well. Okay. So this is our goal and this is what we are going to do. So we will first talk about model based learning. Model based learning is going to be simple. Then we will talk about two methods for model free learning. We will see how far we can go today.

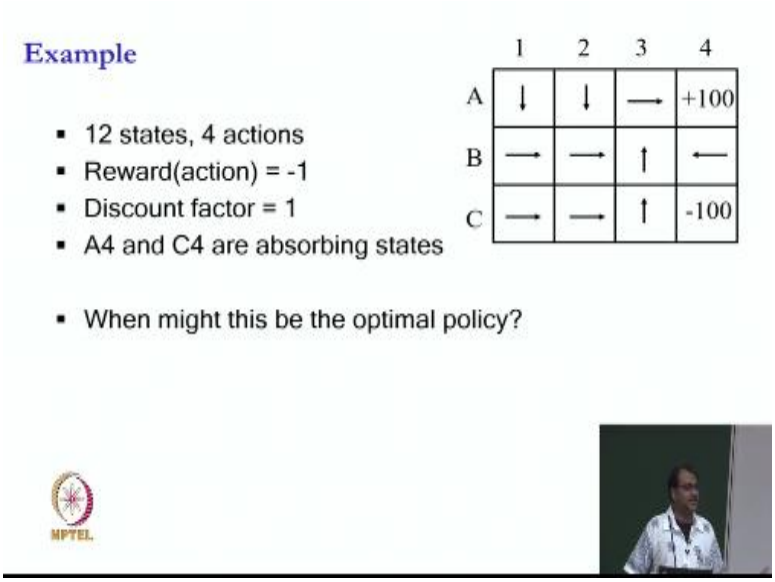
(Refer Slide Time: 09:35)



So our goal today first goal is to given a policy pi estimate V pi using model based learning, right. Model based learning means learn a model, learn a model empirically. What does that mean? That you maintain estimates of transition function, you maintain estimates for the reward function. And then, after you have given a certain estimates, you just use the equations for V pi, which are the system of linear equations that we did in the last week.

And then you plug in T hat and R hat, which are the estimated versions. And then that is it.

(Refer Slide Time: 10:14)



And let us take a very simple example. This is the example that we will work with for the first few slides. So let us say I have this grid world, I have 12 states and 4 possible actions up, down, left, right. My reward of every action is -1, except when I reach A4 or C4. When I reach A4 I get a reward of 100, I stop. When I get C4 I get a reward of -100, I stop. So these are absorbing states. Discount factor is one okay.

Somebody gave me this policy pi. This policy pi says, these are the actions you should do in any of these in all of these states. By the way, any intuition on when might this policy be the optimal policy? But this is, is this suggesting the Manhattan distance version. It is saying go down in A1. And why do I want to go down in A1 if I finally want to reach A4 with the 100 reward?

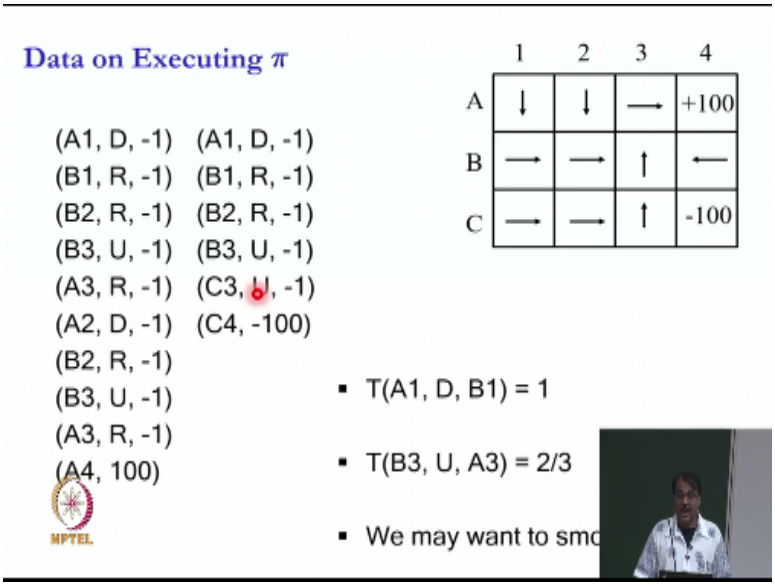
"Professor - student conversation starts" Yes, Parth. It might happen that the A2 to A3 very low probability. **"Professor - student conversation ends".** Yeah, so it is possible that going from A2 to A3 is low. When might that happen? Let us say when there is a wind blowing on the opposite direction. But there is a very strong wind blowing. So when I try to go from A1 to A2 taking the right action, maybe I you know go elsewhere, it is possible.

So therefore, maybe the alternative policy is to go through the B row because the B row there is no wind so I can be more careful about what is going to happen, right. So you never know. See we are assuming that by taking the right action we go to the next

grid. But the transition function is not known to us. Anything could be happening here. It could be that there is a you know big, big you know monster sitting in one of the grids.

Then he is going to eat you. We know nothing. We are not even given this +100, -100 and -1. We are not even given that. We are only given these actions in the each grid. And our goal is to figure out how good is this policy? Right. So all our intuitions about transition in the world are the intuitions that we may have, but the model does not have that. And how we are going to do this is that we will be given some data.

(Refer Slide Time: 12:45)



Let us say we are given this data or how we can collect this data given policy, we can just take those actions from different starting states and generate data, right? We do not it is not a controlled problem, it is not deciding the action, we know the action, because the policy is given to us. So we can just quickly generate data. So let us say this is the two trajectories that we generated.

We started in A1 and kept taking the actions that were given to us up, down, left, right. And then in one case, we went to A1, B1; B2, B3, which is here and then we tried to go up, but we went up A3 when we tried to go right, we went to A2 for some reason. Then we went down then we went to B2 then we went B3 and then we went up again and then we went to this 100. So this is one of the trajectories that we are given.

A different trajectory you know when we tried to go B3 up, we go to C3. And then we tried to go to C3 up, we go to C4, right? So this is a different trajectory that is given to us. We do not know the probability distribution. Now given this, can I estimate the probability of let us say transition function A1 D B1? What is the probability that in action instead A1 when I take the down action I go to B1. One.

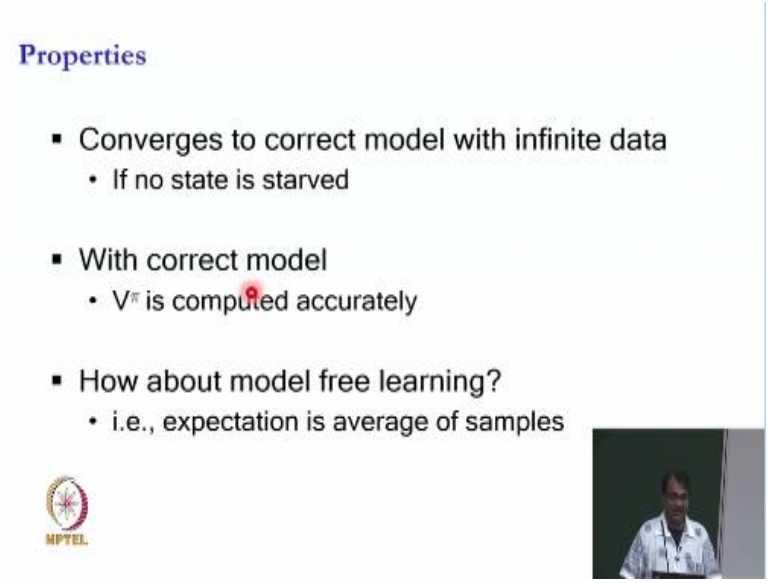
What is the probability that instead B3, I take the up action and go to A3? So instead, B2 I B3 I take the up action and I go to A3, what is the probability? Well, B3 to A3 happens once B3 to A3 happens twice, but B3 to C3 happens the third time, so that probability would be 2 by 3. And of course, we do not want ones and zeros. So we may want to smooth these things. We can also check rewards.

What is the reward that when from A1 to down action we go to B1 and get reward of -1. What is the reward we get? We get -1. And if different times you get different rewards, then we can take an expectation of the rewards as well, or average of the rewards. See, rewards need not be deterministic, it is possible that somebody is tossing a coin and giving us a reward like in the slot machine.

I do the slot and I know sometimes I win a lot of money, but most often times they take my money. So that is also reward distribution. So somebody sampling our coin and giving me a reward that is also possible. And the model takes care of it by just saying that I will compute the average reward. So now I have been able to compute the transition function.

Once I have been able to compute the transition function, I can just put it in the equation that I had and compute V pi.

(Refer Slide Time: 15:30)



This is called model based learning. It converges to the correct model with infinite data, if no state is starved. With the correct model, V pi is computed accurately. And now the question is how do we do model free learning, right? By using expectation as the average of samples. So this is model based learning because I am estimating the model and then taking expectation by using the formula for expectation.