NPTEL
NPTEL ONLINE COURESE

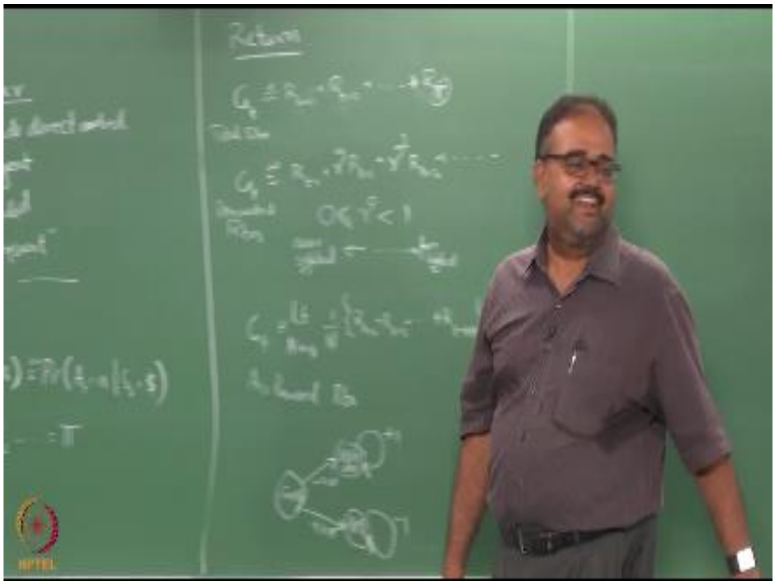
REINFORCEMENT LEARNING

Returns, Value-functions and MDPs

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

Right, so I keep saying long term I said this multiple times in the earlier lectures also, but what really is long term right.

(Refer Slide Time: 00:23)



So we will try to quantify that now right, by introducing the notion of something called return right, so this is a common notation that I will use throughout where T indicates the timestamp of the end of the episode T means the end of the episode rate essentially I have reached S+ right, so this could this is the most natural way of thinking about long term return correct, is the most natural way of thinking about long-term return and if there, if that T is actually finite right, so this is actually is not a bad idea and I will sum up all the word I'm going to get in the feature and I will try to optimize that, okay.

So again you can see that if I try to optimize each one of these individually what will happen I might get a very high RT+1 and RT+2 but it might put me in such a bad ST+2 from then on even if I behave optimally individually for each of this these numbers might be small, right. So what I really want to do is behave in such a way that even if some of these numbers are small right overall in the future the somewhere I am going to get should be the maximum, right.

So overall sum should be maximum, so that is essentially the goal that I have here, okay so this is fine it works fine that is one circumstance where this would not work when T is I mean there is no end right, so I just keep doing this for forever if I keep solving a problem forever then it becomes a problem and not necessarily forever if I am also doing something over very, very, very long horizon if T is very large not necessarily infinite but T is very large also I get into some issues, okay because this will become quantity with a large variance, okay. So one other thing I should point out this is something which people miss and get confused sometimes while reading on their own so T here is a random variable, right.

So every time I run the system it might stop at a different time I think of playing chess right, that is a finite endpoint chess will end, but every time I play the game I do not stop at the same number of moves, right in something like tic-tac-toe right. Even though there is an upper limit on the number of moves you can make right, it is not necessary that you will always end up playing if I moves that you might end you might win in four moves you might win in well yeah, if you win in three moves well, okay.

My opponent is really bad but so it could win in three moves you could win in four moves or you could win in five moves, right or you cannot win at all win so you could still end up draw in the game so like that so T is a random variable so when I put T here does not mean it always runs for some T steps okay, so every time I run it will be a different value for T so remember that, right and consequently G_t is also a random variable because all of these are random variables I am talking about some sort random variables, okay.

And a random number of those random variables are being summed up okay, so G_t is also a random variable that denotes the return okay, great. So another notion of return that people typically use is right, in fact I can even do that, okay. So because I can think about it purely from a mathematical perspective so if my son is going to run till infinity and the γ is less than 1 and remember that each individually each r is bounded okay, so this sum will converge to something finite so I do not have to worry about what I had in this case this sum will not explode okay, so that is fine, right.

But there is even something more interesting about this discounted reward formulation, in fact originally when people propose a discounted or what formulation it was not really from the bounded ness point of view there is more from a from a psychology point of view immediate gratification which is delay the gratification, right so if you let us look at what happens if γ is 0. All thing will go right, it is just RT+1 this goes back to just optimizing for the immediate payoff, right so this gives you essentially you are going to behave in a way so that the next time step what is the maximum reward that you are going to get so you are going to behave only in that way, right.

So if you think about what γ is doing so it is making suppose I get a 10 here which is getting a 10 here so 10 here is more valuable than 10 here right, of course 10 here is more valuable than 10 here, right. So let us go back to our robot grasping problem we are talking about having a -1 all the time and then giving you 100 when you actually grasp so instead of that I can say I am going to use the discounted rewards my γ say 0.8 so what does it mean, so the sooner graspit the larger the value I am going to see, correct.

So the overall reward I am starting from here I should put myself in a trajectory so that a graspit picks quickly. Otherwise I will see some $\gamma^{10,000x100}$ so that is very, very small value so the shorter eyes I know I can get rid of the -1 right. So I do not really have to do the -1 but then what becomes tricky here we have to figure out what is the right value of gamma right, so that becomes tricky.

So this is, so they use this kind of γ to control for whether you are a near sighted agent will be worried about immediate gratification or whether you are okay delaying the gratification little bit so that you get a higher payoff further down the line. The only 1, 1payoff right, so just pull what we do not cumulative or so you are talking about regrets you are talking about regrets right, we are talking about the total reward case, right.

So yeah, having a discounter really does not make sense, right so anyway you are trying to optimize regret right, so in some case your γ is as small as possible right when you say you are optimizing regret that means you are trying to keep your γ as small as possible so essentially you want all your winnings to come in one pool right and one pull I want to be optimal, so that is your optimal regret case, right.

So when you are trying to do achieve something like this setting a γ does not make sense right, but even if you can come up with a different objective function maybe there is something that you can think of or the entire learning process, okay. So one thing which I should point out just let me finish one thing I should point out is that classically the RL people right, look at bandit problems as immediate RL right, I mean the learning another the whole interaction is just that one arm pull, okay that is how we think of it while the, in the long-term case is the full RL problem.

But if you look at the people who work in the Bandit theory community right, they think of the interaction as the entire learning process, right that they do not think of it as one arm pull they think of it as the whole process itself is the learning process so for them it is like an episode, right so there is a very different mindset so the other people do not think of it that way and the bandit people do not think of discount factors and other things because that is something which the RL just came up with for looking at long-term payoffs so there is a mismatch of vocabulary there so maybe there is something to find that in the middle ground but I am not sure.

Return is a function of the time step, yeah here let us put it this way so I should optimize my policy in such a way that wherever I am starting from that and going forward I should get maximum possible return. So even if I am the last step I do not want to go from the last step but if I am see, one thing you should remember is that if I know where I am going to end if that capital T is given to me before hand then it becomes a non-stationary problem.

Because I know okay, I have five more steps to go what should I do I take one more set I have only four more steps to go what should I do so immediately the problem changes but if I know the T then it is a non-stationary sitting, right so we will not will not look at cases where the T is known, okay we will only look at cases where T is known to be finite but not deterministic, right it is not known if you are so it is not like I can say that I will optimize only at the last step or anything but I do not know what the last step is, right.

I might recognize it when you actually reach it but I do not know beforehand that okay, I am going to have only so many substance step to go, I mean this is the normal seeing that we assume but you can immediately see it is not realistic you know I mean when you are playing chess you can always say I am going to win in three more moves or well I am going to lose in three moves it is depending on right, so how well the game is going. So in some cases you do have this kind of a little bit of a forecasting ability as to what the T is, but typical in the basic setup that we will talk about will assume T is unknown.

The people understand why $\gamma=1$ closer to 1 is far sighted, right and closer to 0 is near sighted okay. In the second equation R and T bios to the action which is given as fast as ever say I have five action possible and I picked an action which is giving you everybody in the first and there is another property where I could pick another action first and then this action later which is higher reward in the longer run, so then in this equation we are actually giving more reward to the actually which is done.

Well, depends on many things right, depends on the value of γ and depends on the magnitude of the other rewards that you are getting right, so let us say γ is 0.9999999. When the bios is slow. When the bios is slow depending on how far sighted you want to be you have to pick your γ, right so they are unsatisfactory you have another free parameter really free parameter that you have no way how to tune but yeah it is there and it is a well talk about another reward formulation shortly.

Which gets results of the γ but makes it very hard to optimize anyway so any other questions on this. Yeah, sure, yeah but then that means I am fall, I am failing after more time steps, right so I am okay maybe that is positive or negative it does not matter as far as this return definitions are concerned, you are essentially interested in maximizing the return therefore if you are going to take an action that gives you a -100 reward the farther out in the future you take it is better, right.

So succeed for as long as you can and then die you know right, so yeah it does not matter so the rewards can be positive, negative. So reinforcement learning is really the community is little regret right, so we talk about positive rewards and negative rewards then if you look at the psychology literature they talk about rewards and punishments right, for some reason the RL community and if you look at the, you know other places they call it payoffs right or they call it cost negative rewards are called costs but the RL community somehow things of negative things also as rewarding so we have positive rewards and you have negative rewards but in the larger scheme of things you know negative things also improve your experience you know the enrich your life in some way so you can think of negative rewards also anyway.

So good so this is sign then we will move on to one more definition how to give this right, so what is that do just gives me the average right, so the more the reward accumulate the larger the average is going to be right so for a finite number of n this makes sense right, but what if I have an infinite number of rewards in the future. Under mild regularity assumptions and bounded-ness of the individual Rs this will be bounded, okay.

What is not clear is that it will have actually have a limit, might not right in general it might not have a limit so you have to impose additional constraints on the kind of problems that you are solving not all problems have this defined right or when you can make some small variations to this to make sure that all problems have this limit defined, okay. So there are slight different the limit can be we can play around with the limit.

For example one case where limit will not exceed exists is this if the sequence is periodic, right if the sequence is periodic then the limit or not exist right, because it will just keep oscillating all the time, right. So in this case we use something called the let me know how to handle limits of periodic sequences use something called the cesaro’s limit which essentially looks at the average over periods and let us the number of periods run to infinity, right so in the reason the limit does not exist is because you cut the sequence in the middle of a period where so it is going to be periodic so you only take a period as a unit and take the reward over the period, right and then you keep adding that over subsequent periods and take the average, right and limit as the number of such periods you consider it goes to infinity is called the cesaro's limit.

Then in such cases limit might exist because there are bunch of other condition you have to put on the actual structure of the problem itself for it ensure that this limit exists but for a very broad class of problems you can talk about this limits as well. So the nice thing about this kind of the return definition is that it does not have a pesky γ parameter right so it makes it nice from a computational point of view not to have the gamma parameter.

But it makes it little weaken from with the one of the original motivations of RL point of view because it does not have the γ parameter, okay. So γ parameter allows you to have you know some sense personality for your agent that is it the near sighted agent or a far sighted agent and so on so forth that so especially people in neuroscience when they are looking at modeling reinforcement learning humans right, so they actually have they can think of estimating what is your γ by running experiments with you right.

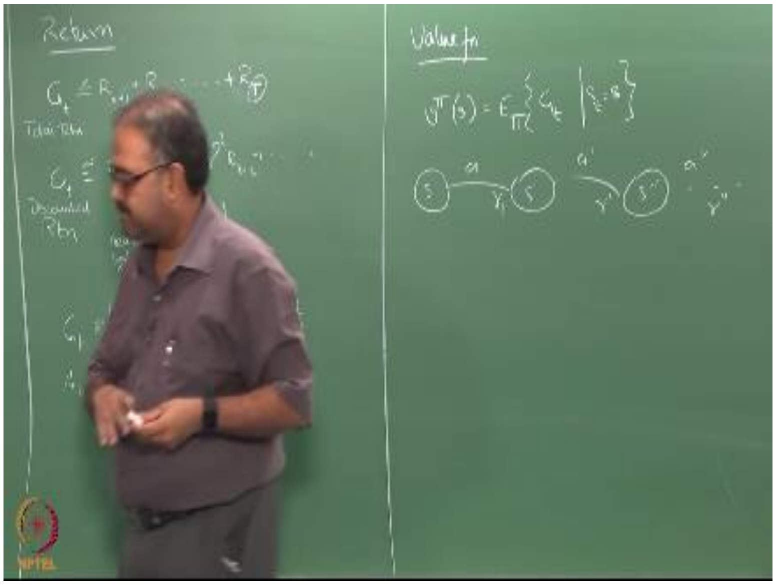
Because it is kind of you know it is a personality issue thing right, so some people will tend to be in fact there are studies that try to correlate the levels of serotonin in your head to γ, right so more the serotonin the larger the γ value you have in your decision making and so on so forth I mean it is like this is really weird how we seem to have all these parameters in our head tied up with the different kinds of neurotransmitters, right anyway.

So this is called the average reward, this is called average reward return so any case is what this is called discounted return that sorry I cannot hear, cumulative or total return on right so people if you are going to call that cumulating then you have to call this cumulative discount, okay so there is a very nice problem that people use for motivating why discounted return is bad, okay. So let me let me try and do that here, okay so I am implying no theological leanings here okay this is just a fun example, so that is earth and it is heaven and that is hell so the path to heaven has a -100 reward and the path to hell has a +1000 reward right but once you reach heaven you get a +1 infinitely once you reach hell you get a -1 infinitely I said I am not getting into theology here okay, this is just a fun example of course you can always think about other things here, know you what for now anyway are finished here right, this it is a transient state we do not care.

But then if I said γ is 0.1 or something, right you would prefer going to hell right if γ is like 0.9 I said is γ has to be pretty high for you to go to heaven right, so it is the same problem so now γ becomes part of the problem definition right, depending on what value I set for γ my optimal policy itself will change right, so there is not a desirable state of affairs I mean if you are thinking about saying okay here is a problem solve it and it turns out that if I use average reward scenario now which is better going to heaven is better always does not on depend on anything what does it depend on.

I mean if infinite number of +1s right that will help any my negative things I get initially I am getting a finite value for negative here I am going to get a minus infinity there and a plus infinity here so that trumps anything that happens here, so average reward will tell me that this is the best thing to do. But if I am going to use discounted returns so I will go up or down depending on what my γ is right. So now you can immediately see that discounted rate or is a more realistic return because where everybody is going to have it right anyway that is the fun part.

(Refer Slide Time: 24:09)



They come to the really one of the clerks of a reinforcement learning here, right in the full RL and a lot of the algorithm that we look at and the notion of a value function. But we already looked at the value function right, but the value function we looked at was very simplistic right so what did we do that we said okay, average of all the words I obtained so far right that is the value function then we have another value which he looked at which is the expected to expected reward that I should get if I take an action, right so we denote that by q* and then we noted by q the average that I maintaining for this right.

But here they are something more complex, what do we have here we have a policy π which determines not just the current reward but also the sequence that I am going to see hence forth right, so when I am going to talk about a value function here I will talk about something called V^{π} right, so V^{π} is the value function associated with a policy π, right. So when you say $V^{\pi}(s)$ so this is essentially the expectation of, so what should I be taking the expectation of rewards or returns what do you think it is I should be taking the expectation of the returns, right.

Because the return is what I am looking to optimize this is what I should be predicting so in the bandit case I was only looking at the reward so I was just making out is the expectation of the reward and here I should make it out as the expectation of return so given anything. Anything else, Oh quite I need to condition it on π right, because I already have a π here so why do I need to condition on π because with G_t is going to depend on π. Regardless of what formulation I am using the subsequent rewards I am going to get depend on the actions I take, right so I nearly need to know what my π is right.

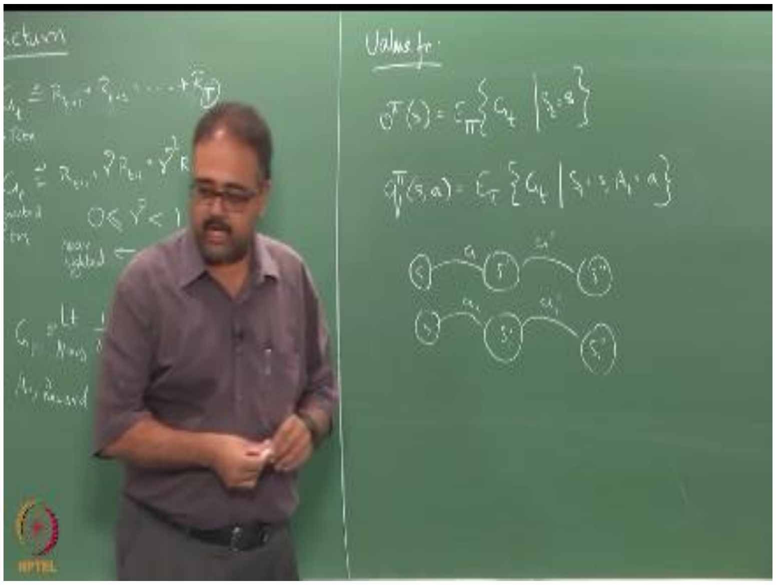
So I need to condition it on π so we will denote it like this, right when I write the π here that means that is a policy I am conditioning off, okay the expectation with respect to the distribution generated by following π okay, of G_t given that at time Ti started with state s okay, does it make sense everyone on board, right so I am starting from state s I am following pi am generating one trajectory right like this right so I start from s okay, go to some s′ go to s″ so on so every time I take an action so I will take a, a′, a″ and so on and I will get some r, right this is one trajectory.

So like that I will start from s, I will do another trajectory then I will start from s I will do another trajectory, I will start from s I will do another trajectory then I will be getting all these rewards like I have r,r′, r″ like that I will get some r1′, r1″ and so on so for each trajectory I will get one set sequence of rewards right, so I will add up all of these or I will take a discounted at sum of all of this right, so I will do that right and I will take your average that is that essentially whatever that average will converge to is my expected value for this is this is essentially what I mean.

When i say G_t says state t=s that means I will start with s okay, and compute this rewards okay, right. So is it clear why we need to condition on the π as well write the future is going to depend on π yeah, yeah they are talking about the expectation here I just explaining to you what the expectation is so every time I generate the trajectory through π right, two things could happen A I could pick different actions because π is stochastic B the world could change in a different way because the world is stochastic, so the same s I could take the same A I might end up in a different desperate every time right, so there are multiple ways in which the stochastic can operate so one π itself is stochastic therefore in the same s I need not take the same action again.

Say first time I can take A1 next time I can take A2 and so on so forth and that will completely change the trajectory that will come later or I will take the same A right, I might end up in a different exponent because of the transition probability so every time I do this trajectory I will get some different project development every time I generate a trajectory I will get a different sequence right, and the expectation is over all set sequences you know what is the total reward I will get so that is essentially what the value functions, okay is it clear great.

(Refer Slide Time: 30:14)

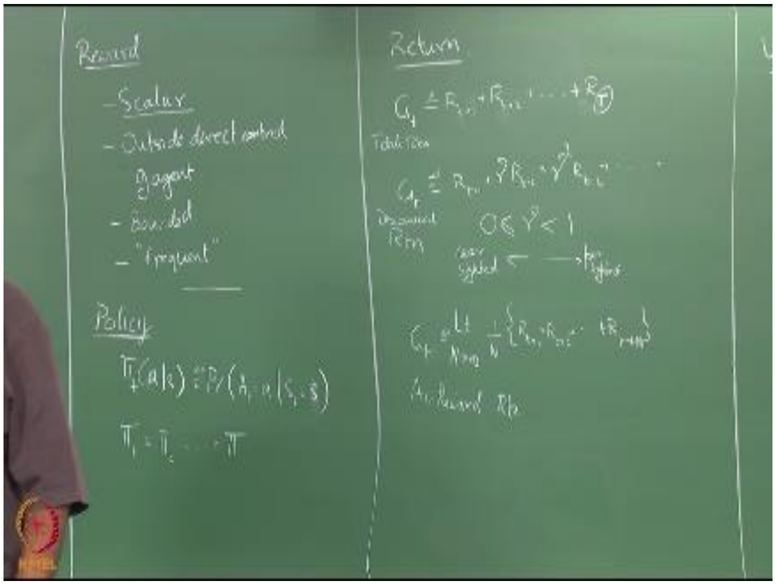

Value fn:
$v^\pi(s) = E_\pi\{G_t \mid S_t = s\}$
$q^\pi(s,a) = E_\pi\{G_t \mid S_t = s, A_t = a\}$
NPTEL

Let us confuse people some more so we had q earlier right so what was q(a)q*(a) was the true think q(a) is the expectation we are expected estimate for taking action A, right but then here we have to worry about taking action A in a specific state because every state it is going to be different, correct every state it will be different so what less will be s will have to do this s still have to do this, so what is a big difference between the v and q sorry, yeah action A is fixed so earlier when you are generating trajectories for Vπ I said you start with s right, then you go to s′ then you go to s″ and so on so forth and then you could I said you could pick any action.

So first time will pick a, a′ so next time you put some a1 so like this right, so you could pick different actions every time we generated the trajectory. So now I'm saying you cannot do that right, so every time I generate the trajectory you have to pick a first right, but subsequently you pick actions according to type the first action is something that is fixed now the first action will be a all the time so what does this tell you it tells you how good is it to perform action a in a given state. So earlier you did not have that information but now you have the information about how good is it to perform actually in a given state, okay is it clear going back and thinking about the problem itself right. So there is a problem that we are trying to solve right, so how do you define the problem how do you characterize the problem right, so I am in a state s I take an action I go to a next state and I also get a reward right.

So for me to characterize the problem I need to tell you how this transitions happen and how the reward generation happens right, so I need to really look at this following question what is a probability of given is sorry.

(Refer Slide Time: 33:53)


Reward
- Scalar
- Outside direct control of agent
- Bounded
- "Frequent"
Policy
$\pi_t(a|s) \doteq Pr(A_t = a | S_t = s)$
$\pi_1 = \pi_2 \ldots = \pi$
Return
$G_t \doteq R_{t+1} + R_{t+2} + \ldots + R_T$
Total Return
$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$
Discounted Return
$0 \leq \gamma < 1$
short sighted
far sighted
$G_t = \lim_{N \to \infty} \frac{1}{N} \{R_{t+1} + R_{t+2} + \ldots + R_{t+N}\}$
Av. Reward Rtn

RT+1 given this is sufficient I mean think about it right, riding a bicycle so you probably need to know what is the velocity momentum that you had before and it is not enough to know how you were on the road and how you are tilting you might want to know something more about the history of how you got there, right. So you really need that quantity so you need to know what is

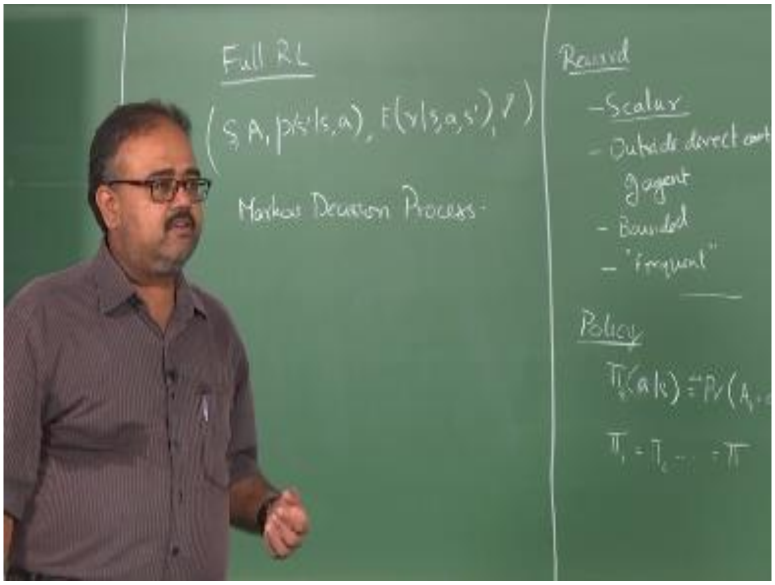
the probability of s T+1 RT+1 given everything that went before that, okay. Now this base is a little complex to model because the number of parameters you will need for modeling something like this becomes pretty huge, right.
So what we typically end up doing is make a very strong assumption that this guy is equal to so what you truly need is the first quantity right, because it becomes very hard to estimate very hard to define you make the assumption that that quantity is really equal to this right so what is this assumption called typically Markov assumption it is essentially a first order Markov assumption where the history does not matter right only the current state and current action matters that will determine what the next state and extraction will be right.

So we make a further assumption which is where s is the assumption I made stationary, right so we will typically operated it mark of environments right and will typically assume their stationary, okay. So for me to characterize a problem now what should I give you what I need is what is the question yeah, to specify the problem what are the things I should give you okay, I should give you the states over which the problem is defined right.

I need to do that right I need to define the state space over which the problem is defined and then the set of actions that I can take and write the P function right, so I need all of this to define the problem completely anything else I need depends on what I see return I am using I mean also need a γ right. If my return if I am using a discounted return depending on what the value of gamma is the problem changes right that is what we discussed.

So if I am going to use discounted return I should also specify gamma as part of the problem, okay this is clear so I need all of this to define my problem right in some cases we use a slightly simplified definition where instead of having the single joint distribution s/s′ and r right, we also give it as two separate quantities.

(Refer Slide Time: 39:08)



So I will suggest s I mean I have to specify sa then I will specify that so what is that call sometimes called the transition function or transition probabilities okay, and then I will specify the expected value of the return given that a Ti was in s I did action a underwent a s′ right, so essentially I am writing this out splitting this up right, so instead of giving the joint Distribution I am saying okay no, no I will specify the s′ separately and for the R I will condition it on s′ also right.

So I can we can think of some sense I am using the chain rule and splitting that joint probability as p(s′) into P(r) given sa, sa′ but what is the difference here instead of specifying p(r) given sa′ I am giving you only the first movement or zeroth moment in the first moment of the distribution right, so why is this enough because I am going to be optimizing only the expected value of Gt, okay and then Gt is composed of these guys right.

So essentially I will be taking the individual expectations of each one of these right, and that is enough for me, I do not really need to know the distribution we generated these rewards because individually I will be taking the expectations of each one of these rewards and therefore knowing the expectation alone is enough so sometimes you specify the problem like this right, and this is typically how you would I would specify a what is called a Markov decision process.

So it is a system is a decision system that you have to for every state you have to actually keep giving a decision right, and it follows Markov in dynamics right, both the transition and the rewards follow how when satisfy the Markov property right, so such a system is called a Markov decision process right so we have states you have set of decisions right and the evolution of the states follows the Markov process okay so they are called Markov decision process, okay and what we will assume is that the reinforcement learning problems that we are trying to solve can be modeled as MDPs, okay.

We will assume that the reinforcement problems you are trying to solve can be modeled as MDPs so in some sense a value function really you know is useful only if you have the Markov property, why because I am assigning a value to a state regardless of how I got to the state right, so if the Markova property is not satisfied how I got to the state will influence what happens in the future right.

Since I am saying that no I do not care about what happened in the past I am saying I am starting in state s okay, I am going to be following policy π here after what is the expected return that for I am saying history is irrelevant right, when I am defining a value function another way of thinking about it is here the value function is marginalizing over history so it is essentially taking the expectation regardless of what the history was that is one way of looking at it I mean in some cases that is how you make sense out of the value functions.

Because as we go along later we will see that we actually apply RL blind lead to problems are non-Markov right and we will be using the v function in the q function in such cases. So the only way to make sense out of what is happening in such cases is to say that essentially marginalizing the history say we are just doing the taking the expectation over all possible industries as well and may or may not be a sensible quantity.

But if it works you just use it right, but normally you value function makes sense only if you have the Markov property forget about the future right, then the because you are ignoring the past it makes sense only if you have the Markov property right, it makes sense right so that is why I was saying that I should have probably defined value functions after telling about Markova addition process but it is okay, you can still understand it in the general sense as well great. So the next thing we have to talk about as optimal value functions, right when I will start that in the next class.

IIT Madras Production

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

www.nptel.ac.in