NPTEL

NPTEL ONLINE COURSE

REINFORCEMENT LEARNING

Bellman Optimality Equation

Prof.Balaraman Ravindran

Department of Computer Science and Engineering

Indian Institute of Technology Madras

great so now I want to introduce a notion of something called the optimal value function so any questions about the bellman equations all crystal clear right so if i give you an MDP and ask you to form the valve in equation solve it and give me the value function Vл can you do that? I will give you a л also can you solve it and give me a Vл, may be? kind of computers ,oh and we're at some point I'm going to give you like MDP’s with millions of states and ask you to compute value function server, don't try to do that the offensive in paper.

So maybe we should give them the head management program, you think we should? Did he do the head management when he did the course? I don't think so right so this is that's a problem that we formulated when I was once a TF or Andy right, so he wanted people to appreciate how hard it is for how hard is to solve dynamic programming how hard is to solve MDP’s, large MDP and this was a problem that was inspired by a real life problem some sheep or farmers like dairy farmers in New Zealand actually solved, okay.

So they had lots of cows and they have to decide when to milk cows and you know, when meant to use cause for something else I got an Indian now becoming a political statement, right but yeah so I think sheep, okay. So when you when you milk the cow and when do you retire the cows you know so, they have to figure out so let us assume that cows have a retirement home and, so that's the process that's a problem so then the cows could be young you know they could be in breed able they could be old, when they were like something like 14 or 17different states that they had the cows be in right, and then they have rewards okay if you can milk the cow in a particular state then you get so much milk, you know if you milk a cow in a later stage the milk comes down blah blah blah right, so whole bunch of things and there's a small probability that old what have been classified as old cows also have calf, you know me so basically they become milk able and so all kinds a very complex situation.

And somebody actually wrote the paper out of this, they wrote a research paper out of solving this problem, how do you cast it as a as an MDP and then come up with a way of solving it because when in that case when you actually cast it as an MDP ended up with like a few million variables, few million states on variables a few million states and how do you go about solving the problem it was not an order all clear. So what we did was we came up with a very simplified version of it, so we're there instead of like 17 states there are 3 states in which the cows could be right and the transition dynamics were much much simplified right.

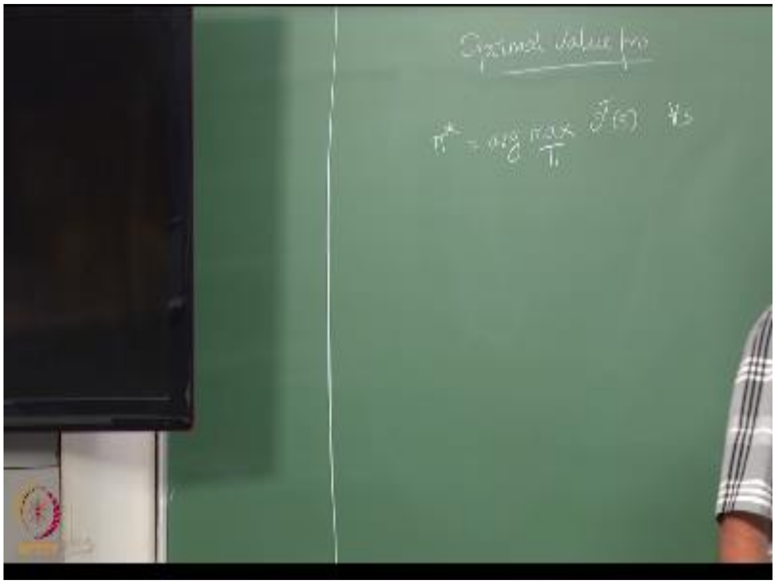
So it could from young it could become breed able and it could pick up old and then from old it could become breed able again so it doesn't become n and so, so they're like a very few transition and so on so forth but it still turns out that you have to for computing the exact transition probabilities right you have to come up with a combinatorial summation right, it's actually it turns out to be incredibly hard to get the right transition probabilities just the P alone and everything else is simple, we give you what the state is we give you what the action is right, just coming up with the P and the are the exact closed expression for P and R itself turns out to be very very hard problem, right.

And it's not it's not so there are I don't know about 300 states or something or 158 states, it is a small problem right but computing that transition probability even when it is just 158 states computing the transition probability turns out to require a combinatorial some rights little tricky. So many people actually fail in getting the right transition probability they would write the correct code for solving the problem right but then they would not know how to compute the transition probability, just to get you'd appreciate the difficulty of coming up with that p and therefore, you will appreciate why reinforcement learning algorithms are useful because they can work without the p.

So what is optimal here what am I talking about as being optimal so, whenever talked about that so far right so what is optimality in the context of MDP’s or in the context of RL so, we already talked about optimality in the context of RL what is optimality in the context of RL? I want the л says that for no other л can I get a better return than expected return than this point right, so I want a л right so, I want a л that is the best overall л what should be the quantity here? But that is not defined we know that right, so the expectation over return this does not make sense, so what should i do I should qualify it over a state right, I should talk about where i start the return and then i can take expectations over that or whatever.

But even better I can do this point wise so what do I mean by that you can say I need something like this, л ,thanks yeah this is on the flight translation problems edition one has RT capital R T is written okay, and finishing to capital GT is written anyway so this is this is what I am looking for right and we know what that quantity inside there is right, so point wise i am trying to find max over л Vл(s) for all is right, so this is I want a state I mean I want a policy that achieves the maximum value for every state right. So i am going to call this yeah i am going to call this л * right.

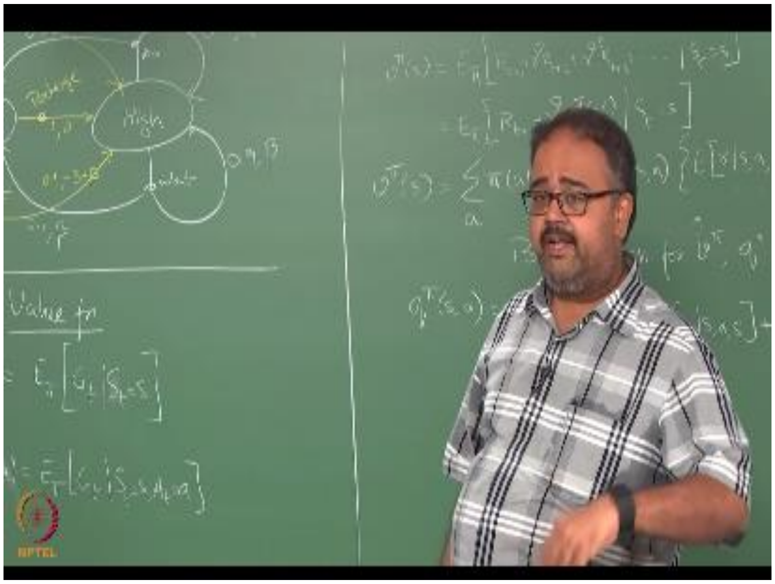
(Refer Slide Time: 07:47)



If I yeah I can do organized so one thing one caveat to note here is that arc max might not return a unique л, that could be multiple policies for which this maximum is achieved in which case any one of them can be called the optimal policy. So there is no the optimal policy right but that is a optimal policy that could be multiple optimal policies, just remember that it went very simple example I can give you think of like a grid well so, this is something we will revisit often right.

let us say that every state has four actions I can go up down left or right wait then every action is going to get a reward of- one right, and in the data ballistic world so if I go up I will move up by one square and if you go right I will go right by one square and so on so forth, like in the indicator directions or you could think of a stochastic world where you're trying to go up with probability 0.9 I will go up the probability 0. 1,i'll just stay where i am and you can think of this as some wheels slipping when you have a bought that is trying to move forward and with some probability that we wheel slip and then you stay in the same place and you probably have to try moving again and again until you succeed right.

(Refer Slide Time: 09:07)

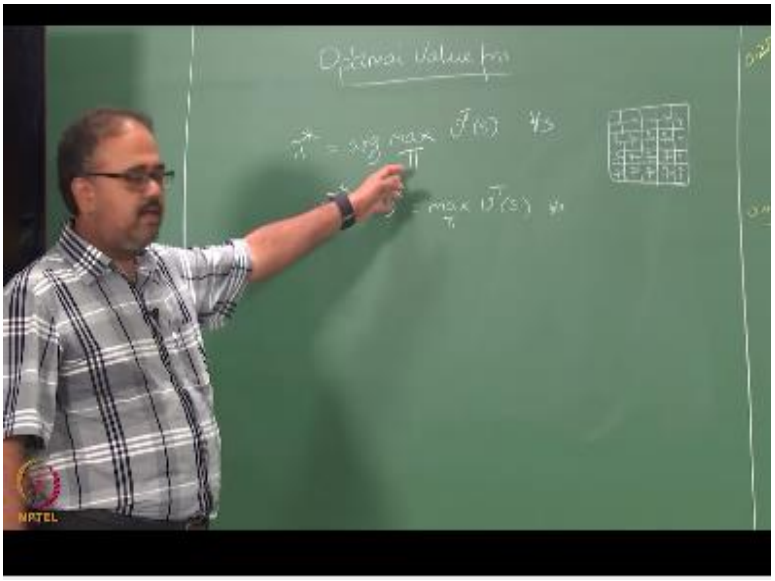


So you could think of it that way right that is a stochastic world and every time you make an action regardless of whether you move or not, you get a -1 right, so if we try to go up in the top row, what will happen? You stay there you stay there whether the action succeeds or not will stay there and you will get a lot of- 1 okay. So and this is an episodic task right, the whole thing will end when you reach the state mark G right so what will be behaving optimally here we get to G as quickly as possible right, so the shortest path to G alright so, that is be the optimal way of behaving in this world right.

So how many optimal ways are there of behaving in this world? A policy is a mapping from States to actions it doesn't matter very structural but this is the mark of earnest right so we are assuming everything is Marco is really shouldn't depend on very start from right so, so regardless of where I start from for example when I am here the optimal action is to go this way wait I do not care whether they started here and came here or whether i am starting here or not right regardless of where I'm starting the action is like that likewise, okay things become interesting from here okay.

The number of optimal policies we have so, any one of these states i can choose any one of those actions right so, basically i have about 12 states and I can choose choose any of the two actions in those 12 states right so, I have a huge number of optimal policies right you can convince yourself these are optimal because they will take me to the expected number of steps if it is deterministic well, I can tell you exactly how many steps it will take me but if it is a stochastic world depending on what is the probability of failure of the action right I'll have some probability of reaching the goal right, so great.

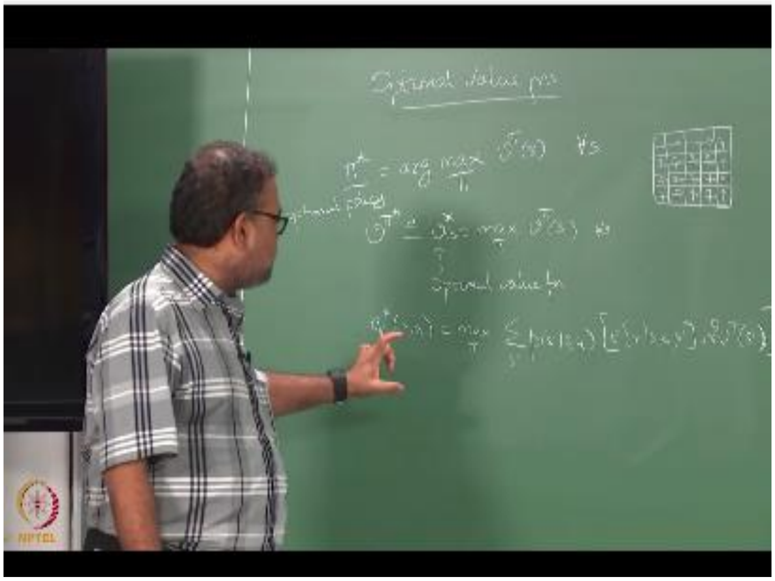
(Refer Slide Time: 12:57)



So there are too many optimal policies it says something common across these entire policies okay and you can you rephrase it? okay not in terms of steps terms of something else that we know net reward the expected reward is going to be the same so for any state regardless of what is the optimal policy I choose to execute the value of the state will be the same, correct is it clear? Regardless of the policy I choose to execute the value of the state is going to be the same correct so i can write this and I got rid of the л* here, I am going to call this V*and V* is unique given an MDP V* will be unique right.

Л* need not be unique, but regardless of how many different Л* you have right V* is going to be the same whichever Л* you consider V* will be the same so regardless of VЛ* regardless of the Л* you choose I am going to call the optimal value function as V* okay, so this is called the so this is an optimal policy its optimal value function is I said ,there's only one thing which you guys should be asking me now how do I know that the same Л will reach the max across all s, so if you have finite MDP’s, it is actually straightforward to show that okay if you have continuous MDP’s right then it becomes a little trickier, so we will discuss this when I am actually starting the proofs in the next class, why is it that there exists a optimal policy at all right, so we will discuss this in the next class when I'm talking about proves right.

(Refer Slide Time: 17:13)



So right now, we will try to do the equivalent of Bellman equations for V* right, I mean likewise i can define an optimal value function for q, what would be the optimal value function for q, yeah another solution is there something that sound right, what is problem here so we need to put that л thing here right, so what I what I want you to appreciate here is that even though I am talking about optimal policies, I still the first action it take easy right the first actually take is A, always here whether I am talking about optimal so, it is only after i take a I behave optimally so that is why this max is after, in fact I can remove the maximum here and I can insert it anything that comes before it is it not influenced by the max right.

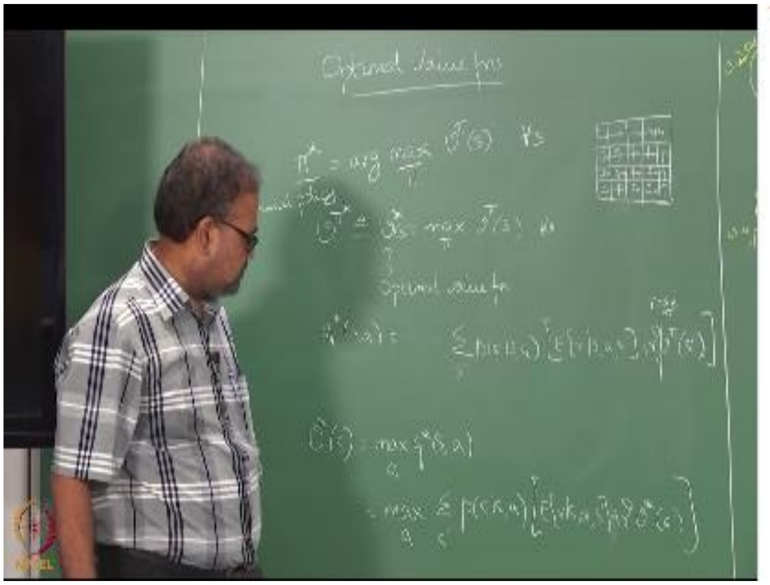
Because I take a fixed action it is something which you have to keep in mind so whenever you talk about Q*S,A right, the A is still fixed it might not be the best A to do right in fact, I can have a Q* a defined for this action here, wait I can have a Q* here defined for this action also for each one of these actions I can have a Q* defined right, this is it so the Q* of the down action will essentially be -1 + V* of going from here to the goal, right in the Q* of this action will be 1 + V* of going here assuming is deterministic, otherwise it will be 0.9 -1+0.9 times the value of going from here and 0.9 times the value of going from if it is sarcastic right.

so I can think of defining Q*for even sub optimal action this is one thing if I want to keep in mind so it's not like Q* is defined only for optimal actions let there, okay great, so now we have these things so let us try to see what we can do so, I want to do v star of yes ray I want to find V*(s), so what would V*(s) equal to? Do you think in terms of Q*? So what do you think V*(s) will be in terms of Q*, so what is Q* tell me? If I take action A and behave optimally thereafter right so, what can V* be, I mean that is essentially behaving optimally from the current state.

Let say I am going to look at okay which one whatever gives me the best reward on taking A and then behaving optimally thereafter right so think about what this expression is? This expression is essentially okay I behave according to A and after that I behave optimally right, so that is essentially this expression right and I am taking a max here which essentially means said okay, which is the action that gives me this? The best of this expression and that I cannot behave more optimally than that, because it is only the first action I have a choice here after that I'm assuming and behaving optimally whatever is optimal behavior right.

So i am not talking about the recovering the optimal behavior yet right, I'm only talking about finding what the optimal value function is right, I am assuming after-action A and behaving optimally, so now the only thing i need to figure out is what is action A and I need to pick that action A says that when you add the reward for taking mediate or what for taking action A and the return for behaving optimally they are after I get the most. So this is essentially very intuitive way of defining what the optimal value function is now I can expand this you can say it's max over A summation s prime and that is what max over л is say max over лVл V*sort of the V*is prime.

(Refer Slide Time: 21:53)



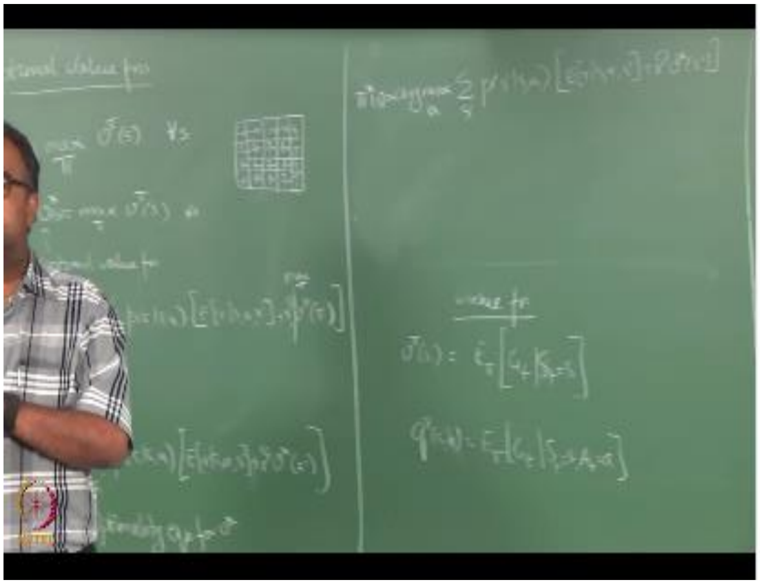
e

This is how we got that so, this is called the bellman optimality equation. For V* likewise you can write a bell but optimality equation for, you start have to be why not wait whether it some policy there is no question of some policy will give you a higher policy even why I given you all the optimal policies tell me which one will do that? Just think about it if that is some policy that will do that then I will basically change the action to that I mean that has to be optimum right if there is going right this does not have the same values going up going up as a higher value than going right, then going up has to be optimal going right cannot be optimal because then if I do the max only one will apply the other one won't appear right.

so one other thing so I have been talking only about deterministic optimal policies right I said there are like there are 12elements and there are two options there did not be so, as long as I have a policy which are saying only non zero probabilities to the optimal actions right that is also an optimal policy so, in this state I can choose with probability 0.5 to go up or probability 0.5 to go right, right so, 0.5 do up, 0.5 go right that is also an optimal policy right it could be any combination as long as I have zero probability for going right or going down right then I am self happy then it is an optimal policy right so, there is an infinite number of optimal policies is an infinite number of twin powers itself, is it clear? Okay.

so nowhere is the next question I want to ask so given a policy so Q* value function how will you recover optimal policy from that because, I give you v star how you recover an optimal policy from that so I have to know figure out what action I have to take in each state right and I have to take that action that achieves the max right so, what I mean by achieves the max so in the case of V* actually take an action figure out where I will end up it right, and take that action that gives me the max over that right so, what I mean by where I land up at so I pick some action right then i sum over s prime sorry, so this is the utility for taking an action right.

(Refer Slide Time: 26:10)


optimal value fns
value fn
$v^\pi(s) = E_\pi[G_t | S_t = s]$
$q^\pi(s,a) = E_\pi[G_t | S_t = s, A_t = a]$

so I have to pick an action that gives me so given a V* this is how I will recover a PI star right yeah depends on how you how you implement the arc max that if the ark max you can write it such a way that you will return any of the maximizes then you will get a deterministic policy or you can return your arm access that it returns all the maximizes then you can think of having some kind of uniform probability distribution over the Maximize it can be uniform or it could be one way or but if, you think of arc max is returning some arbitrarily returning one of the maximizes will get a deterministic policy right this is one way of recovering optimal policies from the optimal value function okay.

so the little trickier I still need to know p and c need to know the expected values and other things even if somebody gives me the value function I still need to know the rest of the parameters of the MDP for me to recover the optimal policy. But suppose somebody gives me the Q function how I will recommend the optimal policy, we think about look at the expression there right in the arc max is exactly this right. so if i give you the cue functions if I give me Q* I do not need all this that what I am doing that is essentially looking at one step right if I do a okay what will happen? Right.

so I am looking at the outcome and then trying to make the decision the best decision based on that outcome right but here the Q* I do not even have to do that one step look at it right I can just be greedy with respect to my current Q*and I will л* makes sense so, if i give you V*and Q* now we can know how to find л* and we also we spoke about how we can solve the bellman equation for V л and Q л, can you solve the bellman equation like that for V*and Q*, so I have a system of n equations in n variables same thing as before can I solve it what's the problem? it's not linear right so it says i have this max here it makes it the non linear system of equations so you'll have to figure out how we are going to solve it right.

So this is what we will look at in the next class the next layer what I will do is first show that well in multiple ways so that Vл is unique and then show that V* is unique right, and then come up with a talk about multiple ways of solving for we started well not only this will b Vл can we'll talk about multiple ways of solving for V*, in fact the rest of the course in some senses figuring out multiple ways of solving for V*and Q* under different assumptions okay, how much information you have how much data you have and so on so forth right, so great so any questions?

IIT Madras Production

Funded by

Department of Higher Education