

NPTEL

NPTEL ONLINE COURSE

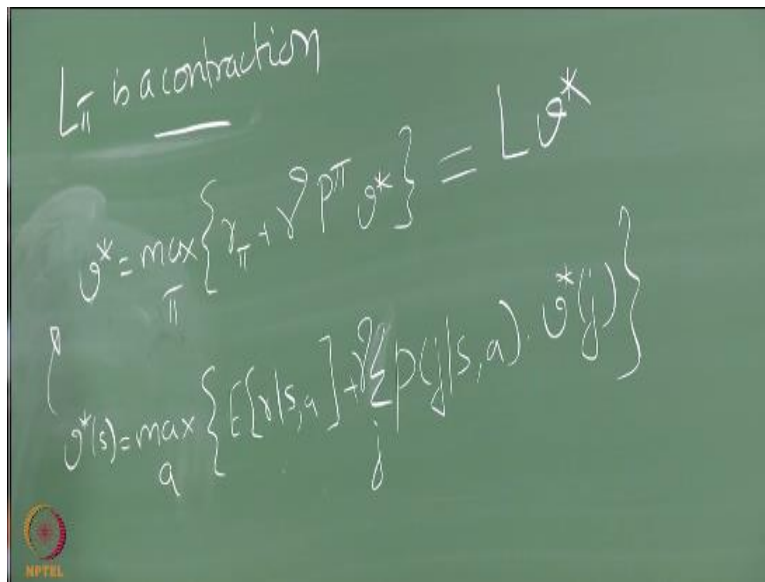
REINFORCEMENT LEARNING

L_{π} Convergence

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

So these guys are going to follow today so last week we left off by showing that.

(Refer Slide Time: 00:28)



The image shows a chalkboard with handwritten mathematical expressions. At the top, it says " L_{π} is a contraction". Below that, there are two equations:
$$v^* = \max_{\pi} \{ r_{\pi} + \gamma P^{\pi} v^* \} = L v^*$$
 and
$$v^*(s) = \max_a \left\{ E[r|s, a] + \gamma \sum_j P(j|s, a) \cdot v^*(j) \right\}$$

L_{π} is a contraction and it is others may leave if they want to the guys who would not actually revised last week's lecture if you are so interested or so disinterested you could leave right so just teach for the six people could read okay I do not care right so there is one question last class to show that this actually works we need to show that we say V is a Banach space okay how hard is it to show that V is a Banach space what is V what is $v \in V$ what is the space V they want to show as a Banach space of all valid functions okay.

Well value functions is little tricky because value functions means that there is some policy or something that generates that right need that need not necessarily be true it is very hard to quantify it such a space right so V is a space of all functions that satisfy something for us what is that very simple what is the most basic condition we put bounded right so component wise the function is bounded see so think of an element here right and then I take $V(S)$ which is one component of the vector V like that component will be always bounded right.

Okay and think of what is the norm that we are considering here in this space so essentially what capital V is it is a space of all bounded what functions right the space of all bounded functions right so because we have this function operating on a discrete space right so we can just think of this as points in space right but it is actually is a function space is actually space of functions right so V is a space of bounded functions right and what is the norm that we consider max now right the max now right and what is it that we need to show if it is a barracks phase it converts right but what we know here all Cauchy sequences what does the Cauchy sequence a as you keep growing difference becomes smaller and smaller and smaller so what does that mean it converges it should be in the same space right.

So what can you tell me about any such space with the max norm on it so it is a complete it is a full space right I am not excluding anything the only problem I have is that my functions have to be component ways bounded right the functions have to be component wise bounded right and if you look at the right so essentially the idea is that since I am assuming that my functions are getting epsilon close to each other my vectors are getting epsilon close to each other and given the fact that their component wise they are all bounded right.

So the limit also has to be a bounded function and as long as it is a bounded function it will exist in this space so the easier thing to show is RM is bounded under the equilibrium norms right under the max norm so RN is bounded under any now okay so the tricky part is this is not RN it is a subset of RN right why is it a subset of RN v and they say all components have to be bounded right so RL does not have that restriction I could have some components that can grow unbounded right.

So the volatility part here is so that okay since I am talking about bounded vectors okay the point is the limit of this bounded vector should also be bounded is actually not always the case right I mean I can actually set up space set up situations where I'm considering bound identities but the limit of the bound identities happens to be at set up a sequence where I am considering only bound identities and but the limit of the boundary identities happens to be something that is unbounded right but in this case I am not setting up any kind of a generic sequence like this right so I am actually setting up a sequence of bounded function so the sequence eventually has to be bounded.

So it is little tricky to show that but the intuition is that it is going to be a set of sequence of bounded functions and I am assuming that the difference between them is becoming smaller and smaller and sway infinitely smaller right as it becomes larger and larger and therefore you cannot have an unbounded value that is infinitely close to a bounded value so therefore we assume that it is finally the limit exists in V so I am not doing a formal proof for it because it involves introducing more concepts but the intuition is there.

So next we have to move on to the optimality equation right you know the bellman optimality equation order dated is all these foundations so we talked about this right marginalizing over s' identities this notation in the last line right so this is essentially the expected value of the reward given the state and the action right I am marginalizing over a spread so basically summing over the probability of seeing all possible expense and this we spoke about right and yeah so this is this is the optimality equation right.

So what we are doing to go from here to here right is writing this in a vector notation right so translating all of this into vector notation is if remember I defined $R \pi$ as the expected value will get following π right for the next immediate step right so where so the only difference here is here I assume that you use a here at but here I am assuming you are using a π okay and so this is the probability of that of the next state following policy π right here again I am assuming a here I am assuming fight and here I am saying v^*_j and here I am saying v^* the summation is gone.

$\forall \pi$ no so I will tell you why right so I am doing max over a right so for each state I get one max or A so I will get 1 action as the max action let us say let us assume that there is one action as the max action for each yes right so what essentially this is give me it gives me a mapping from state to action so what is that mapping it is a policy so that is a policy π we are talking about here right this is essentially a component wise maximization of this expression so for every state I will write out this expression for every state I will find the a that correspondingly maximizes that expression.

Let us just like it I there right so every state I will solve this thing I will find out what that is I can set that as the max by right so this max here is a cut for saying it is a component wise max right for every S I learn this max independently I could afford to do I am not making any assumptions about the structure of the policy π right so for every state I can set an arbitrary action as the action for π right.

So this is my well-being optimality equation return in vector form right makes it is so now I am going to define this as an operator.

(Refer Slide Time: 10:06)

$$LQ \triangleq \max_{\pi} \{ r_{\pi} + \gamma P^{\pi} Q \}$$

(Claim: L is a contraction.)

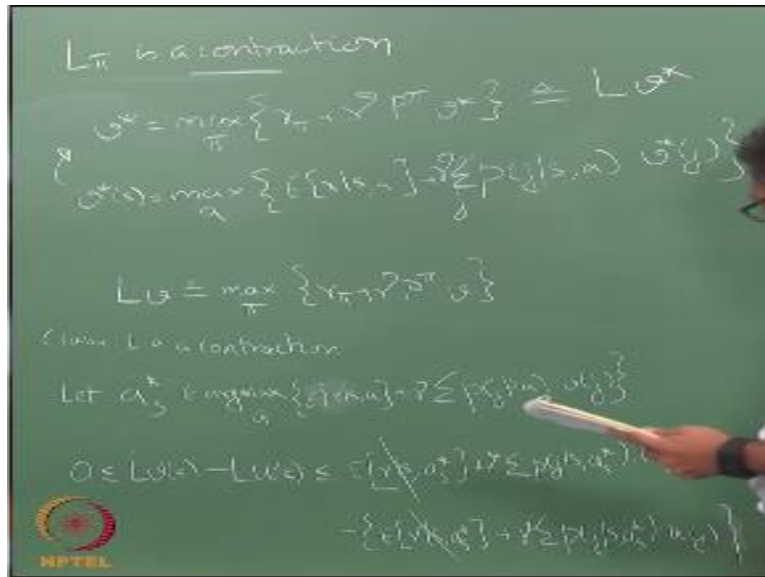
$$\text{Let } a_s^* = \arg \max_a \{ r(s,a) + \gamma \sum p_j(p,a) v_j \}$$

So I am going to define L as that operator that takes in V^* unless this computation okay and gives you back another point in the space let L is an operator so L is essentially so LV and what we are claiming is that V^* is the fixed point of n right so last class we looked at $L\pi$ right which was the iterative process for value an equation for $V \pi$ now this is a bellman equation for V^* so I have written it like this right and then taken this whole operation and converted that into a single matrix operation right and so now the function essentially becomes LV is this.

Say L is a not a linear function of V because of the max it is L is not a linear function of wave just for convenience I am writing it as LV it looks like it is a matrix multiplication it is not okay and that is basically the claim here is that what is the claim L is a contraction so if I say L is a contraction what do you get you get lots of things for free right so what do you do what do you get if L is a contraction in kind of hand waving lee argued that V is a Barak space L is a contraction so what do you get a that is a unique fixed point V^* right.

And if you do it iterative it converges to be stuff okay so it starts from some arbitrary V or repeatedly apply a L it converge to Easter so we got both this thing for free so let us look at showing this right so when they say a star office it means that the best action to do at least one of the best actions to do in status.

(Refer Slide Time: 13:13)



Remember we talked about the world they could go in any combination so you could reach the top right necessarily so it could go either right or you could go up any combination of right and up will work to get you to the final state right so either both right and up where the best actions it could be more than one best action both of this do not preclude multiple things being best okay and this is just finding out the value corresponding to the best action right I am not so we argued that earlier also there could be multiple best actions but there will be only one value corresponding to those best actions.

Regardless of whether we go up where you go right you will take the same number of steps to reach the goal right so the value will be the same but you could have multiple policies so I'm not procuring this any one of those next a you plug in here right let us likewise any one of the best π you plug in here and re compute we stop so I am not saying you have to find one π or anything right so this here I am picking one such a that is why I did not say a is equal to a^* is equal to $R \max$ I am saying a^* belongs to our max okay.

So what we really have to establish is that later on which will be useful for you in your doing the assignment is that you do not really have to consider probability stochastic policies when you are

doing dealing with MDPs at least you do not have to consider sarcastic policies for finite MDPs there will always be a deterministic optimal policy so it is enough if you only search for one deterministic policy you do not have to search for stochastic policies that will have to show right what I will do is I will not do it in class so I will give you a small right up to read so from my notes I will make up right up and put it online so you can read it from so this just to convince you that is a little bit of algebra not too much right.

So just to convince you that deterministic policies are sufficient okay but right now so far I have not assumed anything like that and this assuming that is one action that is the best right and let us assume that I have two functions V and U okay and I have L operating on V and L operating on U remember from our discussion in the last class that LV is a function LV is a function $LV(s)$ is the s^{th} of that component of that function or LV acting on s whatever is the output okay so the function is LV L is not operating on VS I think a lot of you had a confusion last class L is not operating on $V(S)$ L is something that takes a function as an argument and outputs a function right so LV is the function right you take V you apply L on it and then the resulting function you apply on S right.

So likewise U take you apply L on it and the resulting function you apply on S okay so I am first making the assumption that LV of s is greater than $LU(S)$ okay so what do I need to show come closer than VS and US that is basic to order remember we did this last class also we did the same thing so we assumed one way we showed it and then we flip the other way showed it and then we showed that for every state s the absolute value in contracted that for the norm will also contract exactly.

The same thing that we did last class that is why I asked people to revise from the last class and come so that I can just go through this very easily right and this clarifying this because last class after the class a lot of people are asking me questions about this $LVLU$ business so that is clear right so LV is a function which is operating on S okay right so a^* is the one that gives you the max right so that gives you the max here so I remove the max and wrote it in terms of a a^* right so this is essentially a $L(V)$ okay.

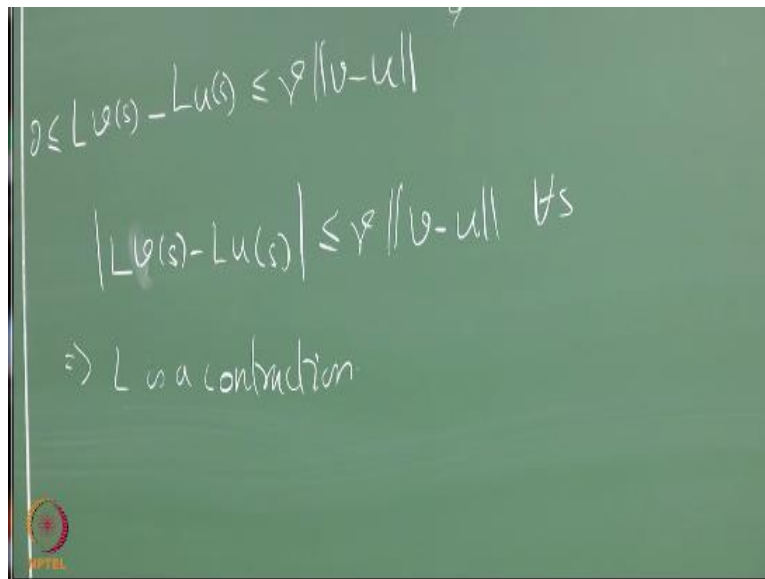
So I put a less than or equal to here so I am going to do something on the other term which $L(U)$ okay so if I am going to expand $L(U)$ I might as $L =$ equal to right so I am going to do something tricky with the other term so what is it tricky things I recommended so sick eating at it I got rid of the max here also but instead of using the max corresponding to you U_j I am using the max corresponding to V right so a^* is what is a max when V is the function here right a^* is the max and V is the function here but I am plugging in a^* itself here right.

So what can you tell me about this it will be less than the actual max when L operates on U see that so when the L is operating on this right so essentially there is a max over here and there will be U here right but instead of taking the max action I am taking some arbitrary action which is a^* and they plugging it in there for this will be less than the max right there for this difference will be greater than this because this in this term this is also the max corresponding to V and this will be the max corresponding to U right so I am taking the difference but now what I have done is have replaced this term with something that smaller therefore this will be greater this difference will be greater okay.

Is it clear V is not optimal but a^* is optimal for V when I am doing that maxing maximizing thing here right so a^* is the solution for this maximizing this optimization this is an optimization problem right so I am doing our max over a so as a^* is one of the solutions for solving this problem when the function here is V it may or may not be the solution when the function is U so if it is the solution then this will be inequality if it is not a solution then this will be lesser than okay great.

So now the advantage of doing this we can start canceling out some terms and try to make this simpler so what will you end up with is.

(Refer Slide Time: 21:30)

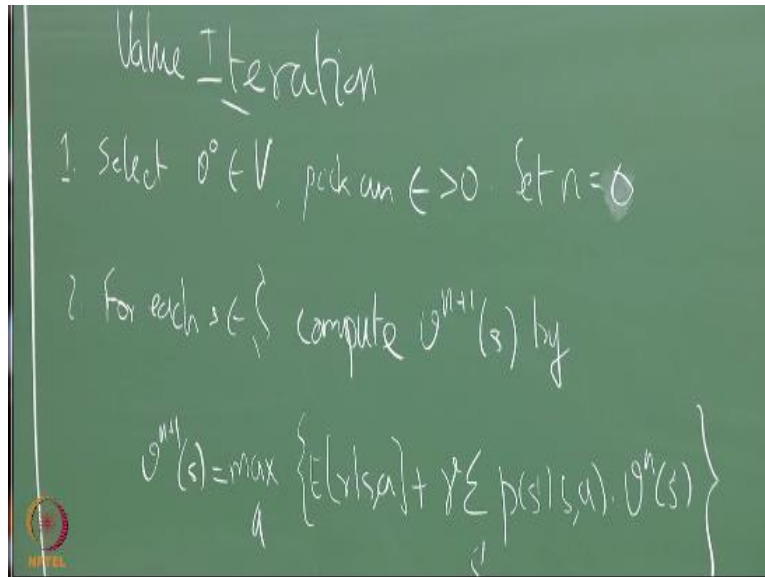


The image shows a green chalkboard with handwritten mathematical expressions. The first line is $0 \leq L(v(s)) - L(u(s)) \leq \gamma \|v - u\|$. The second line is $|L(v(s)) - L(u(s))| \leq \gamma \|v - u\| \quad \forall s$. The third line is $\Rightarrow L$ is a contraction. In the bottom left corner, there is a small circular logo with a red and yellow design.

So this term is equal to there is a same trick that we did head right so this is component-wise difference it will be bounded by the max difference which is the norm right and once I write it as the norm it becomes independent of J so I can take that out and they have this summation and the summation is 1 right so what do I have here so I have so I can do the same thing I can flip things around right I can start with 0 less than or equal to $LU - LV$ right and then pick the action that maximizes the right-hand side for U and then do the whole thing right I will get exactly the same thing.

So showing the other way round also right so I finally get that okay so now this gives us a way of solving the solving in MTP.

(Refer Slide Time: 23:36)



How do I do that start with an arbitrary value function just keep applying L repeatedly I will end up with an optimal value function so once I have an optimal value function how will I recover an optimal policy I can do this essentially what I did here run this on V^* right this will give me an optimal policy then I can pick any action in the arbitrary action from that from the R max and take that as may optimal action or I could just define some kind of a probability distribution over that take the dress off Matt great so we will just write that down so this is a very popular algorithm called value iteration because you start with the value and then you iterate over it until you come to the final convergence value right.

So this is yeah so for people who are wondering about their programming assignment herd management assignment so value iteration is one of the things I am asking you to do right so you just start with an arbitrary value function later then repeatedly implement L and run it until you come to a fixed point the only interesting thing now here is what is the what is what how many more times do you have to say arbitrary what is the more interesting point is we have talked about one end I said it is not the right answer.

When do I stop and wait not the number but when do I stop it cannot be a simple number I just cannot say run a thousand times okay so I should have some other convergent test that tells me when it has topped right essentially I keep doing it until successive iterates or within a few are same you are a great event but same will not happen same will take forever very close let us listen ah but then you are not computing the policy every time right the policy itself a painful process you have to be greedy with respect to the value function.

So every time if every iteration you are computing the same policy it becomes a problem we are going to compute the policy in every iteration then it might be a problem because you have to do an additional computation maybe not I mean it is a call that you have to make but if you want to give certain guarantees about what will hold when things converge then you better have a bound on the value function okay.

So we will come to another algorithm called policy iteration where we can work with the space of policies okay but right now I am just going to do this with value functions right yeah so select some V_0 belonging to be pick an epsilon for our stopping criterion and start your iteration another just a little component-wise have written your $V_{N+1} = V_{N+1} = L V_N$ is basically what I am asking you to do here right.

(Refer Slide Time: 28:16)

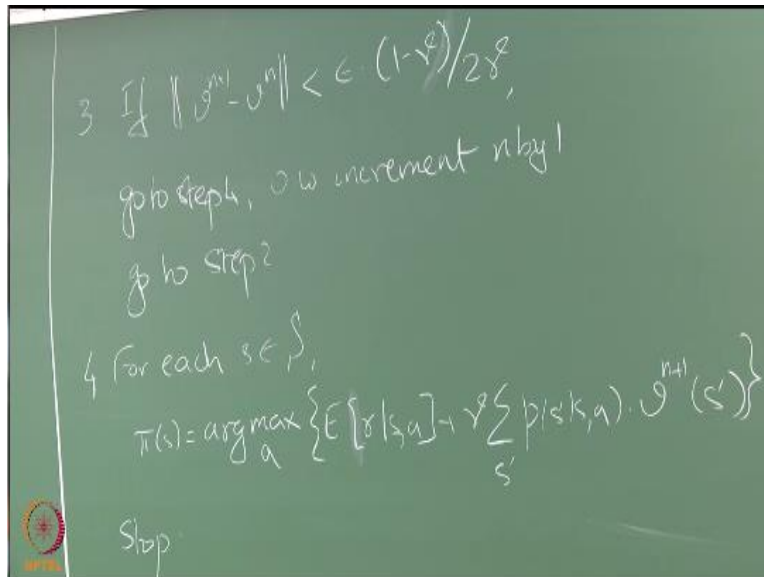
The chalkboard contains the following mathematical content:

- $$L \text{ is a contraction}$$
- $$= \max_{\pi} \left\{ r_{\pi} + \gamma \sum_{s \in \mathcal{S}} P^{\pi}(s, s') \right\} \triangleq L v^*$$
- $$= \max_{\pi} \left\{ \sum_{s \in \mathcal{S}} \left[\gamma \sum_{s' \in \mathcal{S}} P^{\pi}(s, s') v^*(s') \right] \right\}$$
- $$L v \triangleq \max_{\pi} \left\{ r_{\pi} + \gamma \sum_{s \in \mathcal{S}} P^{\pi}(s, s') v \right\}$$
- $$L \text{ is a contraction}$$
- $$\left\| \sum_{s \in \mathcal{S}} \left[\gamma \sum_{s' \in \mathcal{S}} P^{\pi}(s, s') (v(s') - v) \right] \right\|$$
- $$\|L v - L v^*\| \leq \gamma \|v - v^*\|$$
- $$\|L v - L v^*\| \leq \gamma \|v - v^*\| \forall v$$
- $$\Rightarrow L \text{ is a contraction}$$
- Value Iteration
- 1. Select $v^0 \in V$, pick $\epsilon > 0$, let $n = 0$
- 2. compute $v^{n+1}(s)$ by

$$= \max_{\pi} \left\{ r_{\pi}(s) + \gamma \sum_{s' \in \mathcal{S}} P^{\pi}(s, s') v^n(s') \right\}$$

So this is nothing okay so the interesting part is this.

(Refer Slide Time: 28:35)



Is fine all of this is fine the only tricky thing is why do you have this weird expression here why not just epsilon why do you have epsilon into $1 - \gamma / 2 \gamma$ so I am doing value iteration what I am I computing what I am interested in compliment what do I expect output at the end of V^* or Π star or whatever it I'm expecting to output V^* right but my stopping condition is V_N and V_{N+1} right but I would like to give you guarantees on how far away I am from V^* at the end of the day not necessarily how close V_N and V_{N+1} are right so we will take this close class condition and convert it into a guarantee on how close you are to be S^* and then when we do that conversion we will find that you will be what do you think you will be guess guys I mean you should be able to guess by now what do you think will be epsilon close to V^* okay.

IIT Madras Production

Funded by

Department of Higher Education

Ministry of Human Resource Development

Government of India

www.nptel.ac.in

Copyrights Reserved