STOCHASTIC APPROXIMATION: THEORY AND APPLICATIONS

Dr. Gugan Thope

Department of Computer Science and Engineering

Indian Institute of Science, Bangalore

Week 12

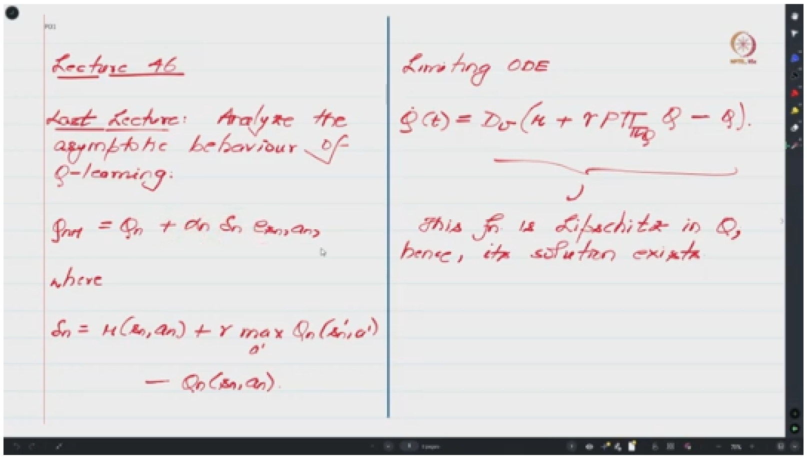
Lecture 46

Concluding the Asymptotic Analysis of Q-Learning

Hello and Namaste everyone. Welcome to lecture 46 of this NPTEL course on stochastic approximation. This week, we have been trying to see how we can use stochastic approximation principles to understand the behavior of the Q-learning algorithm. In particular, we were interested in seeing if the Q-learning algorithm can be used to find the optimal policy. In the previous class, we discussed the Q-learning algorithm's update rule, and we showed that the iterates are almost surely bounded.

Furthermore, we showed that its limiting ODE solution trajectory exists and is unique. We also introduced the notion of the vector comparison lemma and the switched system theory, which tells us that for an arbitrary switched system, there is a sufficient condition that guarantees all its solution trajectories—well, I should not say almost surely because the limiting ODE is deterministic—what is the sufficient condition under which the origin is globally asymptotically stable. We discussed some conditions that ensure that, and now we will see how such a condition and the vector comparison lemma can be used to analyze the behavior of the solution trajectories of the limiting ODE associated with the Q-learning algorithm itself. So, with this background in mind, let us begin our formal discussion. Our goal is to examine the asymptotic behavior of the Q-learning algorithm.

So, our goal is to look at the asymptotic behavior of the Q learning algorithm. So, recall that the Q learning algorithm has the update rule that is given over here that is Qn plus 1 equals Qn times 1. This operation picks, for every state s, the action that maximizes your Q(s,a) value minus Q. This was the limiting ODE associated with the algorithm, and in the last few minutes of the previous class, we showed that this function is Lipschitz continuous in Q, so its solution exists and is unique. For every state S the action that maximizes your QSA value minus Q. So this was the limiting OD that is associated with this algorithm right and in the last few minutes of the previous class. we showed that this function over here is Lipschitz continuous in Q and hence its solution exists and is unique.



Lecture 46

Last Lecture: Analyze the asymptotic behaviour of Q-learning:

$Q_{n+1} = Q_n + \alpha_n \delta_n e_{s_n,a_n}$

where

$\delta_n = r(s_n,a_n) + \gamma \max_{a'} Q_n(s_n',a') - Q_n(s_n,a_n).$

Limiting ODE

$\dot{Q}(t) = D_\nu(\underbrace{r + \gamma P\Pi_{\pi_Q} Q - Q}).$

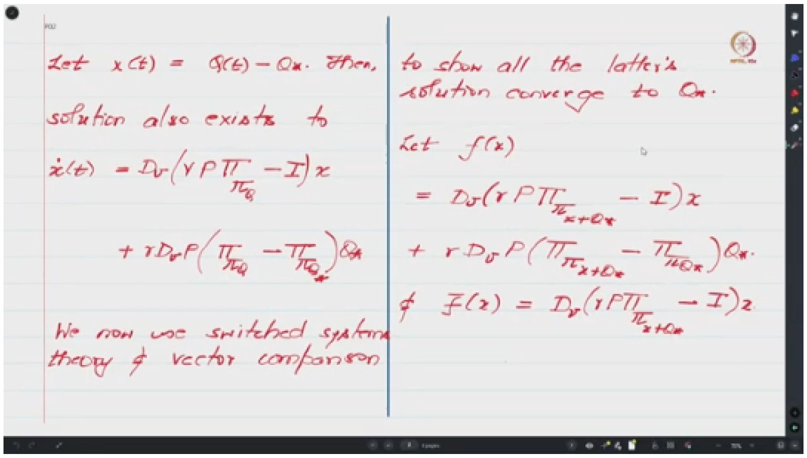
This fn is Lipschitz in Q, hence, its solution exists

$$Q_{n+1} = Q_n + \alpha_n \delta_n e s_n, a_n$$

$$\delta_n = \nu\left(s_n, a_n\right) + rQ_n\left(s_n', a'\right) \ - Q_n\left(s_n,\ a_n\right)$$

$$\dot{Q}(t) = D_\nu\left(\nu - \gamma P\Pi_{\pi Q}Q - Q\right)$$

So, our goal today is to ask if the solution exists, can we somehow guarantee that the solution will actually converge to Q star. Is this okay? So, then what we did was, you know, we want to make use of the switch systems theory and, you know, for the switch system, we have to somehow make the origin the equilibrium point and hence we looked at this, you know, shifted, you know, trajectory. In particular, we looked at x of t equals q of t minus q star and then we showed that x dot of t satisfies this equation where this q is basically x of t plus q star. Right and we wanted to see I mean first of all you know because the solution existence and uniqueness of solution for the Q dot T equals you know this ODE is guaranteed one can conclude that you know this ODE also is well post in that for any initial point



Let $x(t) = Q(t) - Q_*$. Then, solution also exists to

$$\dot{x}(t) = D_\nu(\gamma P \Pi_{\pi_Q} - I)x + \gamma D_\nu P(\Pi_{\pi_Q} - \Pi_{\pi_{Q_*}})Q_*$$

We now use switched system theory & vector comparison to show all the latter's solution converge to Q_*.

Let $f(x)$

$$= D_\nu(\gamma P \Pi_{\pi_{x+Q_*}} - I)x + \gamma D_\nu P(\Pi_{\pi_{x+Q_*}} - \Pi_{\pi_{Q_*}})Q_*.$$

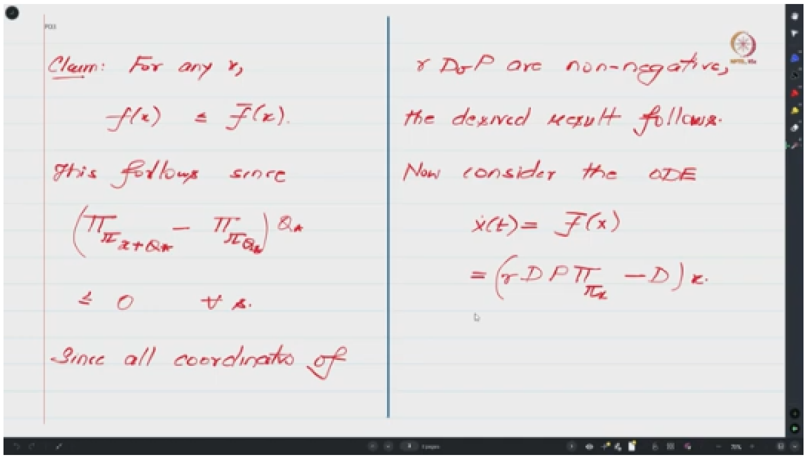
& $\bar{f}(x) = D_\nu(\gamma P \Pi_{\pi_{x+Q_*}} - I)x$

one can guarantee the existence and uniqueness of solution and now what we will do is we will make use of this vector comparison principle and switch system theory to say something about the limiting behavior of the solution trajectories of this ODE. And as I said this OD is slightly challenging and we want to somehow make use of this vector comparison lemma to come up with a simpler OD that is easy to analyze. So, towards that what we are going to do is we are going to define f to be this function. So, f of x is basically whatever you have here gamma p and wherever there is this little pi key I am going to replace it with x plus q star. And this identity is as it is, right?

And similarly, this expression is now replaced by gamma d nu times P capital pi subscript pi of X plus Q star minus pi of pi Q star what we have over here times Q star. Right? And now what we are going to do is this is f of x. Now similarly define f bar of x where you drop this latter term and make use of only this expression. So basically your f bar of x is d nu times gamma p. capital pi subscript pi x plus q star which is what you have over here minus this expression right now one can ask why are we looking at this expression you will soon see the idea is that this expression is simpler to work with and somehow we will you know use this switch system theory to understand the dynamical system which is governed by this f bar driving function okay let us see how we can do that

So, the first claim is that for any x, your f of x is upper bounded by f bar x, right? So, first keep in mind that your f and f bar take in as input a vector and spit out a vector. So, when I say this inequality holds, what I mean is that, coordinate-wise, the output of f of x and the output of f bar x, you know, coordinate-wise, f of x is less than or equal to f bar x.

That is what I wanted to say. And why does this follow? It follows because in the definition of f of x, you have this additional term.



Claim: For any x,

$f(x) \leq \bar{f}(x)$.

This follows since

$\left(\Pi_{\pi_{x+Q_*}} - \Pi_{\pi_{Q_*}}\right) Q_*$

$\leq 0 \quad \forall x$.

Since all coordinates of

γ D & P are non-negative,

the desired result follows.

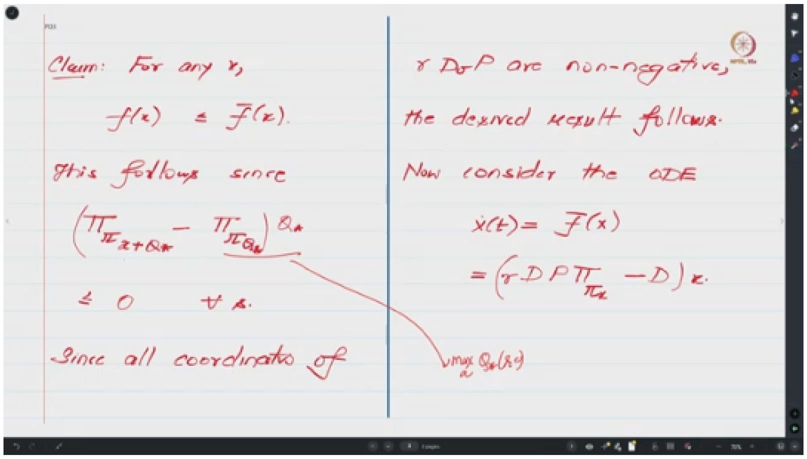
Now consider the ODE

$\dot{x}(t) = \bar{f}(x)$

$= \left(\gamma D P \Pi_{\pi_x} - D\right) x$.

So, let us focus on this expression, which is what I have written over here. And let us try to understand what is happening. So, if I multiply pi subscript pi q star with q star, so this will be, you know, the s-th coordinate will basically be the max of q star s-a, with the max taken over a. This is what will be the s-th coordinate. Now, similarly, what will be this times this?

Now notice that here the greedy action is taken with respect to this value. In other words, if I look at the s-th coordinate of this expression, it will be you know, Q star of S, right? And instead of now being, you know, the action that maximizes Q star, it will be some A, let us call it A bar of S, where A bar of S is basically the argmax of Let me just see if I can move this to the left so that I have some space.



Claim: For any x,

$f(x) \leq \bar{f}(x)$.

This follows since

$\left(\Pi_{\pi_{x+Q_*}} - \Pi_{\pi_{Q_*}}\right) Q_*$

$\leq 0 \quad \forall x$.

Since all coordinates of

γ D & P are non-negative,

the desired result follows.

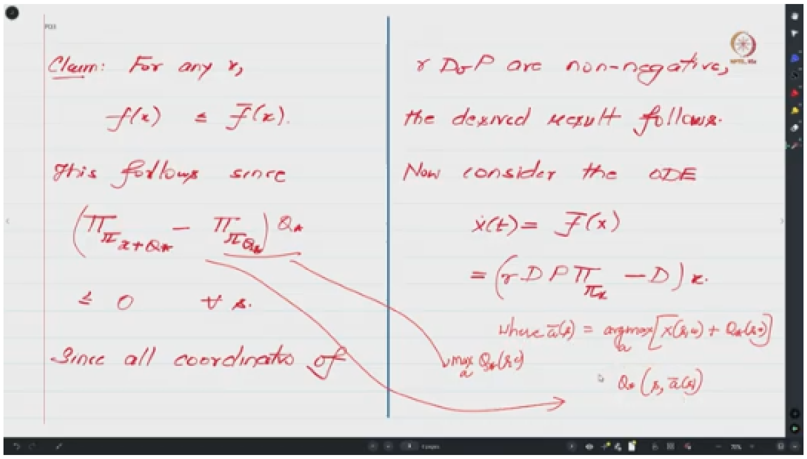
Now consider the ODE

$\dot{x}(t) = \bar{f}(x)$

$= \left(\gamma D P \Pi_{\pi_x} - D\right) x$.

$\max_a Q_*(s,a)$

So, where your A bar of S is basically the argmax over A of X of SA plus Q star of SA. So, you basically look at this vector for different values of S A, pick the action which maximizes this and use this action over here to get Q star S A bar S. That is what the definition of this matrix times Q star implies. In particular, this quantity over here dictates which action we will be selecting. And since this quantity leads to the max of Q star SA and whatever we have here will be less than the max. So, one can easily see that this expression minus this expression will be less than equal to 0 from which we can conclude that F of X is upper bounded by F bar X. And since the coordinates of this expression, so notice that you are multiplying this quantity with gamma denu P.



Claim: For any x,

$f(x) \le \bar{f}(x)$.

This follows since

$\left(\Pi_{\pi_{x+Q_*}} - \underline{\Pi_{\pi_{Q_*}}}\right) Q_*$

$\le 0 \quad \forall x$.

Since all coordinates of

γ D & P are non-negative, the desired result follows.

Now consider the ODE

$\dot{x}(t) = \bar{f}(x)$

$= \left(\gamma D P \Pi_{\pi_x} - D\right) x$.

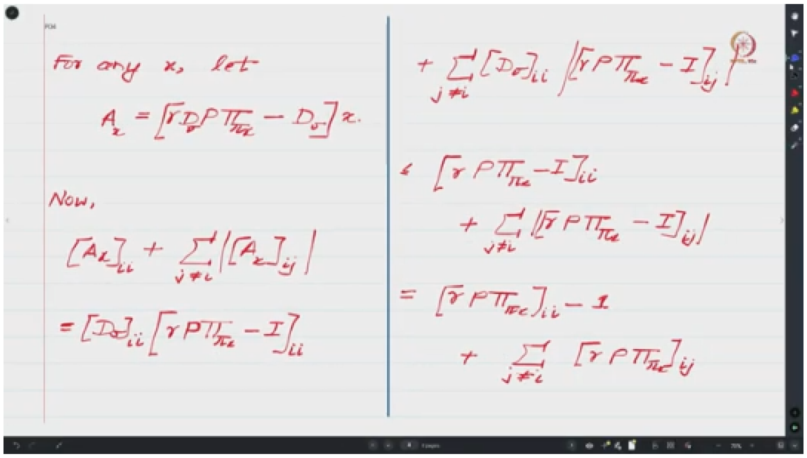
where $\bar{a}(s) = \arg\max_a \left[x(s,a) + Q_*(s,a)\right]$

$\max_a Q_*(s,a)$

$Q_*(s, \bar{a}(s))$

right, and all coordinates of this matrix are non-negative, one can show that, you know, indeed this conclusion holds. Is this okay? So, let me again repeat what we have done. We first show that this expression, right, is a vector whose all entries are non-negative, right, and since I am multiplying this, sorry, I should be careful. I want to say that the entries of this vector are non-positive.

In other words, they are less than equal to 0. And since I am multiplying it with a matrix made up of non-negative entries, every coordinate of this vector will be less than equal to 0. One can easily see that. And from this fact, one can see that f of x is less than f bar of x. this okay so now what we are going to do is we are going to you know look at the OB which is driven by F bar of X and let's see what we can say right so so you have this you know X dot of T equals F bar of X



For any x, let

$$A_x = [\gamma D_\sigma P \Pi_{\pi_x} - D_\sigma] x.$$

Now,

$$[A_x]_{ii} + \sum_{j \neq i} |[A_x]_{ij}|$$

$$= [D_\sigma]_{ii} [\gamma P \Pi_{\pi_x} - I]_{ii}$$

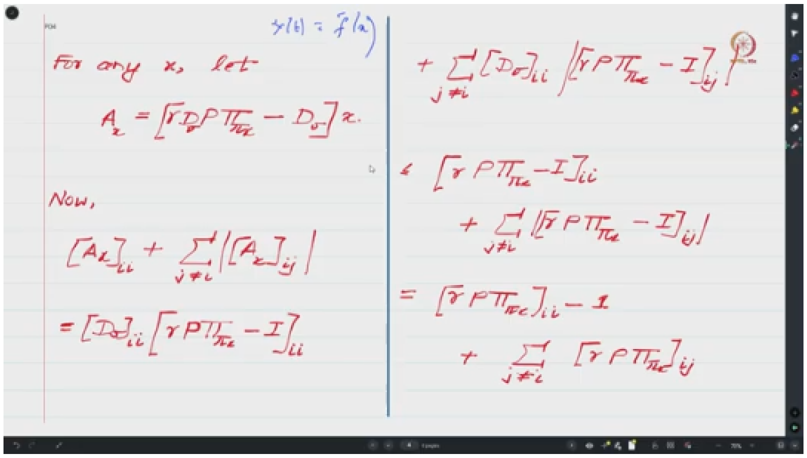
$$+ \sum_{j \neq i} [D_\sigma]_{ii} |[\gamma P \Pi_{\pi_x} - I]_{ij}|$$

$$\leq [\gamma P \Pi_{\pi_x} - I]_{ii} + \sum_{j \neq i} |[\gamma P \Pi_{\pi_x} - I]_{ij}|$$

$$= [\gamma P \Pi_{\pi_x}]_{ii} - 1 + \sum_{j \neq i} [\gamma P \Pi_{\pi_x}]_{ij}$$

And in f bar of x, you do not have the second term. So notice that there is nothing like this. So now for every x, you will think of your sigma of x basically being dictated by where x is. So if your current x is x or x of t is x, then the driving matrix will be a subscript x which is given over here. And if you remember the switched systems result that I had spoken about in the previous lecture, in order to show that the origin of the switched systems ODE is globally asymptotically stable, we need to show that there exists a matrix L and some complementary matrices A bar sigma.

In this case, it will be A bar x. Such that LAx equals A bar x L and this A bar x needs to have this negative diagonal dominating condition. So here, what we will show is that this Ax matrix itself has this diagonal dominating condition, and hence we can work with L which is equal to the identity matrix. In other words, we will show that this A subscript x, whatever your x is, okay, which is basically, you know, the submatrix is decided by this quantity over here, right? This matrix has a negative diagonal dominating condition and hence,



$\dot{x}(t) = \bar{f}(x)$

For any x, let

$$A_x = [\gamma D_\sigma P \Pi_{\pi_x} - D_\sigma] x.$$

Now,

$$[A_x]_{ii} + \sum_{j \neq i} |[A_x]_{ij}|$$

$$= [D_\sigma]_{ii} [\gamma P \Pi_{\pi_x} - I]_{ii}$$

$$+ \sum_{j \neq i} [D_\sigma]_{ii} |[\gamma P \Pi_{\pi_x} - I]_{ij}|$$

$$\leq [\gamma P \Pi_{\pi_x} - I]_{ii} + \sum_{j \neq i} |[\gamma P \Pi_{\pi_x} - I]_{ij}|$$

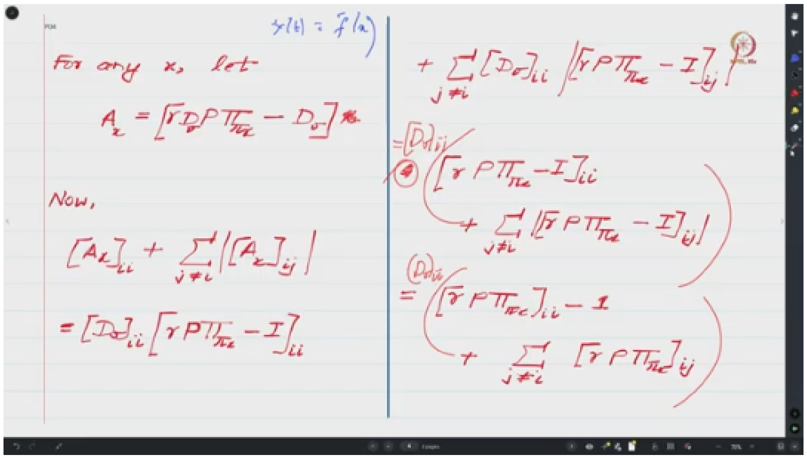
$$= [\gamma P \Pi_{\pi_x}]_{ii} - 1 + \sum_{j \neq i} [\gamma P \Pi_{\pi_x}]_{ij}$$

So, I should actually be careful. This is without x. So, the matrix is actually this, and then you multiply x with this matrix. So, we want to show that this matrix over here has a negative diagonal dominating condition, and hence one can use the switched system ODE theory to conclude that the origin is a globally asymptotically stable equilibrium for the ODE x dot of t equals f upper bar x. So, to show that this A subscript x has a negative diagonal dominating condition, what we will do is we will pick the ith diagonal entry of this and then we will look at the sum of absolute values of the off-diagonal entries, right? So, the ith diagonal entry of this matrix is basically the ith diagonal entry of this, and the ith diagonal entry is basically the ith diagonal entry of DV, right?

And since this is a diagonal matrix, one can show that the ith diagonal entry of this matrix is basically the ith diagonal entry of DV times the ith diagonal entry of this matrix over here. Similarly, the ijth off-diagonal entry of this matrix is basically dvii times the ijth entry of this matrix over here. Since the ijth diagonal entry of I is 1, And the ijth off-diagonal entry of I is 0, right? One can drop this I from this expression and write it in the following way.

And since the ijth diagonal entry of this identity matrix is 1, one can see that this expression can be written in the following way, right? And since, you know, all these—just one minute. Okay, so what we will do is—I have actually written—one needs to be careful here. Okay, so I have written this upper bound over here, but I am just checking why this is guaranteed to be positive here. So, let me just make sure.

So, I have to be careful here. So, what we will do is we will take this ith diagonal entry outside with an equality here and I will write this as DVII and multiply it with this expression. So, I will have the ith diagonal entry of II and multiply it with this expression. And one can then see—so forget the inequality—I am working with equality. So, I am basically taking this expression outside in common and then looking at what we have over here.



$y(t) = f(x)$

For any x, let

$A_x = [\gamma D_\sigma P\Pi_{\pi_x} - D_\sigma]$

Now,

$[A_x]_{ii} + \sum_{j\neq i} |[A_x]_{ij}|$

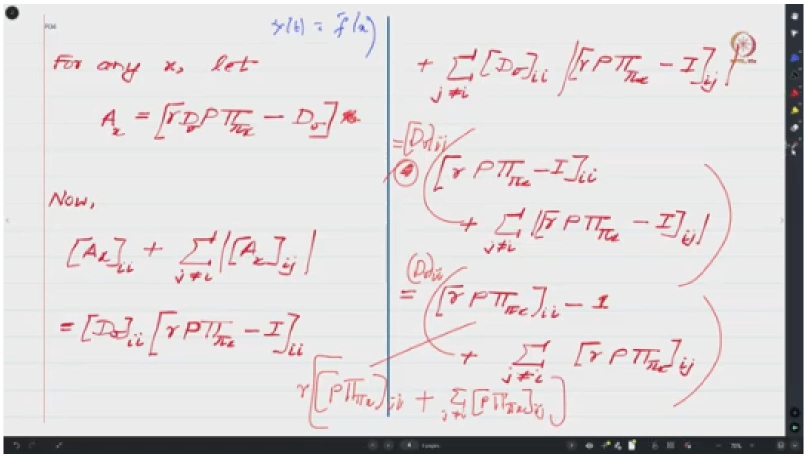
$= [D_\sigma]_{ii} [\gamma P\Pi_{\pi_x} - I]_{ii}$

$+ \sum_{j\neq i} [D_\sigma]_{ii} |[\gamma P\Pi_{\pi_x} - I]_{ij}|$

$= [D_\sigma]_{ii} \left([\gamma P\Pi_{\pi_x} - I]_{ii} + \sum_{j\neq i} |[\gamma P\Pi_{\pi_x} - I]_{ij}| \right)$

$= (D_\sigma)_{ii} \left([\gamma P\Pi_{\pi_x}]_{ii} - 1 + \sum_{j\neq i} [\gamma P\Pi_{\pi_x}]_{ij} \right)$

So, now let us see what is happening here. Here you have gamma, here also you have gamma. And then one can see that this expression is basically gamma times p pi pi x the i-th diagonal entry of this right plus the absolute value of this thing but because these expressions are all non-negative one can drop the absolute value and one can hence get the expression without the absolute value and this will be j not equals i the i-j-th entry of this quantity over here so one can see that you know this matrix your p pi pi x actually is a row stochastic matrix which means that its entries add up to 1 and hence one can see that this expression which doesn't have any absolute values actually adds up to 1 right and because you are multiplying it with gamma you will end up with gamma right and hence you will end up with gamma minus 1.



$\dot{x}(t) = \bar{f}(x)$

For any x, let

$$A_x = [\gamma D_\sigma P \Pi_{\pi_x} - D_\sigma]$$

Now,

$$[A_x]_{ii} + \sum_{j \neq i} |[A_x]_{ij}|$$

$$= [D_\sigma]_{ii} [\gamma P \Pi_{\pi_x} - I]_{ii}$$

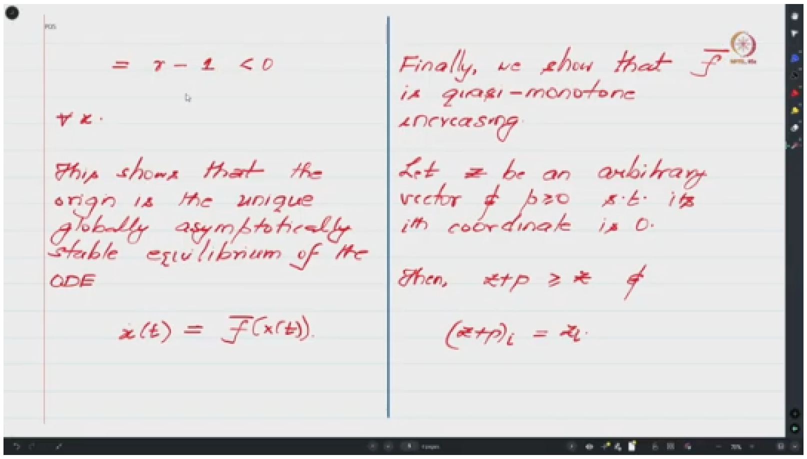
$$+ \sum_{j \neq i} [D_\sigma]_{ii} |[\gamma P \Pi_{\pi_x} - I]_{ij}|$$

$$= [D_\sigma]_{ii} \left([\gamma P \Pi_{\pi_x} - I]_{ii} + \sum_{j \neq i} |[\gamma P \Pi_{\pi_x} - I]_{ij}| \right)$$

$$= [D_\sigma]_{ii} \left([\gamma P \Pi_{\pi_x}]_{ii} - 1 + \sum_{j \neq i} [\gamma P \Pi_{\pi_x}]_{ij} \right)$$

$$\gamma \left[[P \Pi_{\pi_x}]_{ii} + \sum_{j \neq i} [P \Pi_{\pi_x}]_{ij} \right]$$

Is this okay? And since this quantity is strictly less than 0 because we have presumed that the discount factor is strictly less than 1. Hence, this quantity is strictly less than 0. And hence, if we multiply this with a non-negative quantity, that inequality continues to hold true. In particular, if we presume that the Markov chain induced by your behavior policy mu is



$$= \gamma - 1 < 0$$

$\forall x$.

This shows that the origin is the unique globally asymptotically stable equilibrium of the ODE

$$\dot{x}(t) = \bar{f}(x(t)).$$

Finally, we show that $\bar{f}$ is quasi-monotone increasing.

Let z be an arbitrary vector & $p \geq 0$ s.t. its ith coordinate is 0.

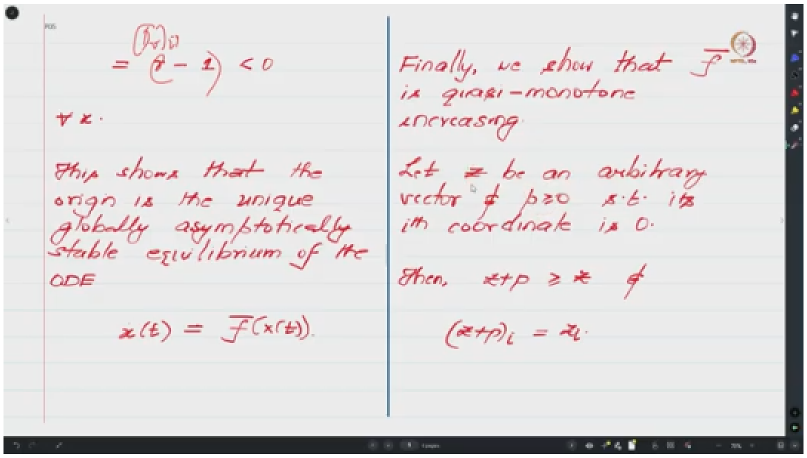
Then, $z + p \geq z$ &

$$(z+p)_i = z_i.$$

is ergodic then this entries will be strictly positive which using this one can show that if I multiply this with DVII right this inequality will continue to hold true right and since this inequality holds true one can show that this AX matrix for any X is has this negative diagonal dominating condition and because it has this negative diagonal dominating condition one can show that the origin is a globally asymptotically stable equilibrium for this ODE. So one can see that by working with this F upper bar you know checking whether the origin is a globally asymptotically stable equilibrium or not becomes very easy. We just have to verify some algebraic expression. Furthermore, what we will now do is we will show that this F upper bar is actually quasi-monotone increasing whose definition I had described in the previous class.

So, if you have forgotten, you may either look up the definition from the previous class video or you can go to that paper by Dongwon Lee and Niao He to check the definition of when do we say a function is quasi-monotone increasing. right so before we show that let me just tell you why we are showing it right so here is the f of x update rule so we have this od x dot of t equals f of x of t and then there is another od where the driving function is f bar right so what we have shown is that all solution trajectories of you know the limiting OD where the driving function is F bar actually goes to the origin. Now we would like to see if we can say something similar for the solution trajectories of the OD where the driving function is F. In order to be able to show that We want to make use of the vector comparison principle and for that we need that F be less than F bar which we have already shown.

Furthermore, we require that this F bar be quasi-monotone increasing. And if we show that then one can eventually claim that the solution trajectories of the ODE where the driving function is f is upper bounded by the solution trajectories of the ODE which is governed by the driving function f bar. And we know that all solution trajectories of f bar go to the origin using which we can say something about the solution trajectories of the ODE which is governed by f. So, let us try to show that our goal is to show F bar is quasi monotone increasing and the way we will prove this is we will consider an arbitrary input vector let us say Z and another arbitrary vector P whose all coordinates are greater than equal to 0. So because all coordinates are greater than equal to 0, Z plus P will be greater than equal to 0 on all coordinates.



$(D_x)_{ii}$

$= (\gamma - 1) < 0$

$\forall x.$

This shows that the origin is the unique globally asymptotically stable equilibrium of the ODE

$\dot{x}(t) = \bar{f}(x(t)).$

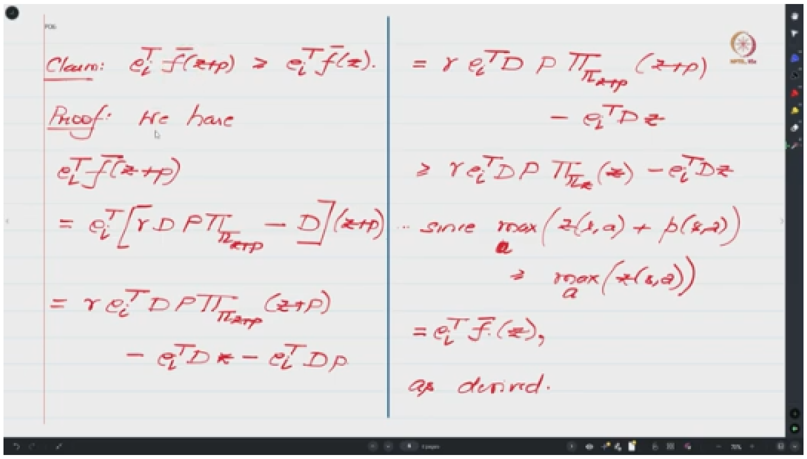
Finally, we show that $\bar{f}$ is quasi-monotone increasing.

Let z be an arbitrary vector & $p \geq 0$ s.t. its ith coordinate is 0.

Then, $z + p \geq z$ &

$(z+p)_i = z_i.$

And furthermore, let us presume that the ith coordinate of P is 0 so that the ith coordinate of Z plus P equals the ith coordinate of Z. This is what I have written here. In order to show quasi-monotone increasing, what we now need to show is that if I give Z plus P as input to F bar, and look at the ith output or ith coordinate of the output, that should be greater than the output when z is given as input, right? So, let us formally state that, right? In order to show f bar is, you know, quasi-monotone increasing, it suffices to show that ei transpose f bar z plus p is greater than ei transpose f bar z. So, let us verify this.



Claim: $e_i^T \bar{f}(z+p) \geq e_i^T \bar{f}(z).$

Proof: We have

$e_i^T \bar{f}(z+p)$

$= e_i^T [\gamma D P \Pi_{\pi_{z+p}} - D](z+p)$

$= \gamma e_i^T D P \Pi_{\pi_{z+p}} (z+p) - e_i^T D z - e_i^T D p$

$= \gamma e_i^T D P \Pi_{\pi_{z+p}} (z+p) - e_i^T D z$

$\geq \gamma e_i^T D P \Pi_{\pi_z} (z) - e_i^T D z$

$\because$ since $\max_a (z(s,a) + p(s,a)) \geq \max_a (z(s,a))$

$= e_i^T \bar{f}(z),$

as desired.

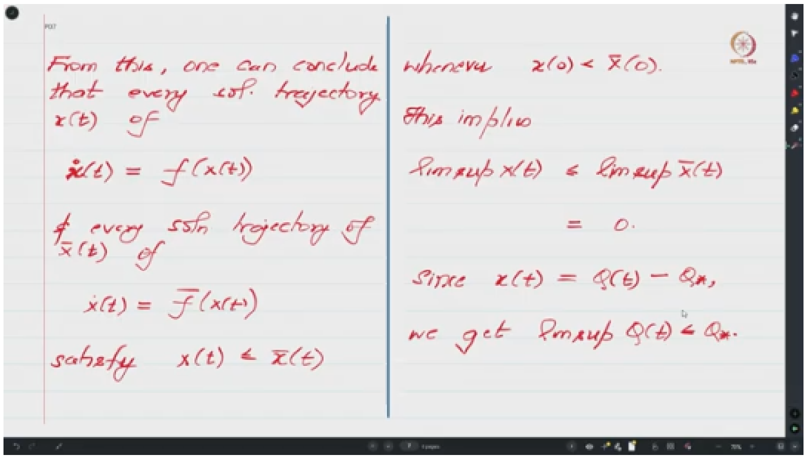
And notice that I only talk about I because this is what the quasi-monotone increasing condition is. Only on those coordinates where Z plus P and Z are equal, I need to show that EI transpose F bar Z plus P is greater than or equal to EI transpose F bar Z. So let us check what is the i-th coordinate of this expression over here. The i-th coordinate is basically EI transpose. So EI over here is the standard basis vector whose i-th coordinate is 1.

So EI transpose, when I multiply it with this, it will give me the i-th coordinate. So EI transpose F bar Z plus P, and now recalling what is the definition of F bar. It is basically gamma D mu. P times pi, subscript pi, and since the input is Z plus P, it will be pi Z plus P minus D times Z plus P, or this D again should be D nu, right? So, this EI transpose, I can take it inside, and this will become EI transpose D nu P, right?

Pi subscript pi Z plus P times Z plus P, I am bringing it here, minus EI transpose D nu times Z minus EI transpose D P. Is this okay? Now, all coordinates of P are non-negative, right? And in particular, the i-th coordinate of P is 0, and D is a diagonal matrix, right? So, since the i-th coordinate of P is 0, one can see that EI transpose DP will actually be 0.

Hence, this expression can be dropped and one can then see that, you know, this whole thing equals what I have written over here. right and now one can see that because your Z plus P is non-negative right I mean I should claim that because your P is you know sorry P is non-negative one can see that max of ZSA plus PSA the max over this quantity is actually bigger than the max of ZSA so I inadvertently said that the Z is also non-negative that is not true It is you know your P is only non-negative but one can see that because your P is non-negative if I take the max over ZSA plus PSA over A this will be bigger than this. Now what is this expression? This is exactly what you will get when you multiply Z plus P with this quantity and the right hand side is what you will get when you multiply Z with this expression.

So from this inequality one can then see that you know this expression will be greater than equal to this expression and hence this whole thing is greater than this expression minus this expression and what we have over here is precisely the definition of F bar Z and hence one can conclude that your EI transpose F bar of Z plus P is greater than equal to EI transpose F bar Z. In other words, by making use of this relation, one can indeed show that your F bar has this quasi-monotone increasing nature as desired. And because F bar also has this quasi-monotone nature, one can now conclude that every solution trajectory of this ODE and every solution trajectory of this ODE satisfy this relation whenever X0 is upper bounded by X bar 0.



From this, one can conclude that every sol. trajectory $x(t)$ of

$$\dot{x}(t) = f(x(t))$$

& every ssln trajectory of $\bar{x}(t)$ of

$$\dot{x}(t) = \bar{f}(x(t))$$

satisfy $x(t) \le \bar{x}(t)$

whenever $x(0) < \bar{x}(0)$.

This implies

$$\limsup x(t) \le \limsup \bar{x}(t)$$
$$= 0.$$

Since $x(t) = Q(t) - Q_*$,

we get $\limsup Q(t) \le Q_*$.

So, let me elaborate. You take any solution of this ODE and let us call it as X of t and let us take any solution of this and let us call it as X bar of t and suppose X of t and X bar of t satisfy this initial condition relation where the strict inequality needs to hold coordinate wise. then one can show that this inequality carries over to X of t and X bar of t for every t bigger than equal to 0, right? And because X of t is upper bounded by X bar of t for every t bigger than equal to 0, one can conclude that the limb soup of X of t is upper bounded by the limb soup of X bar of t. However, we know that every solution trajectory of this ODE goes to the origin, hence this expression must be 0. from which and using the fact that X of t is Qt minus Q star, one can conclude that limb soup of Q of t is less than equal to Q star.

$$\dot{x}(t) = f(x(t))$$

$$\dot{x}(t) = \bar{f}(x(t))$$

$$x(t) \le \bar{x}(t)$$

$$x(0) \le \bar{x}(0)$$

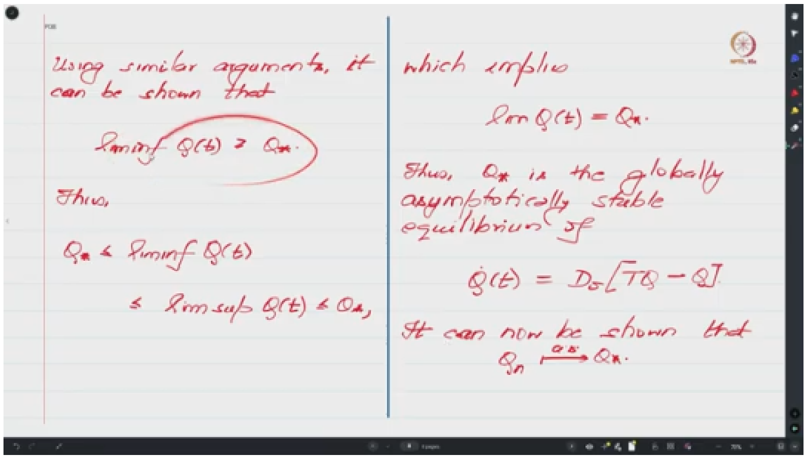
$$x(t) \le \bar{x}(t)$$

$$= 0$$

$$x(t) = Q(t) - Q_*$$

$$Q(t) \le Q_*$$

So, let me just summarize what we have managed to do. On the one hand, we showed that every solution trajectory of this ODE goes to the origin and every solution trajectory of this is upper bounded by a suitable solution trajectory of this, right? And from this, we managed to conclude this and hence, one can, you know, make use of the fact that X of t is Q of t minus Q star where Q of t is the suitable solution trajectory of the limiting OD associated with your Q learning algorithm, one can use this to conclude that lim sup of Q of t is less than equal to Q star wherever the solution trajectory starts from. And using similar arguments, one can also show that lim inf of Q of t is greater than equal to Q star.



Using similar arguments, it can be shown that

$$\liminf Q(t) \ge Q_*.$$

Thus,

$$Q_* \le \liminf Q(t)$$

$$\le \limsup Q(t) \le Q_*,$$

which implies

$$\lim Q(t) = Q_*.$$

Thus, Q_* is the globally asymptotically stable equilibrium of

$$\dot{Q}(t) = D_{\sigma}[TQ - Q]$$

It can now be shown that

$$Q_n \xrightarrow{a.s.} Q_*.$$

You can look at the Dong-Wan Lee and Hay paper to verify this. So what we have now managed to show is that for any solution trajectories of the limiting OD associated with your Q learning algorithm lim inf of Q of t and lim soup of Q of t which have this relation by default right will always be lower bounded by Q star and upper bounded by Q star and then by invoking the sandwich property one can show that for any solution trajectory of the limiting ODE of your Q learning algorithms limiting ODE right. the limit exists and the limit will be Q star right. In other words for any solution trajectory you know the limiting I mean the solution trajectory will have a limit and that limit will be Q star from which we can conclude that Q star is a globally asymptotically stable equilibrium of this ODE.

And now by using your conditions from the stochastic approximation theory that we had studied that is by verifying your assumptions A1, A2, A3 and A4 and some of which I have already proved. Right, one can see that one can show that the limiting behavior that we have shown for the solution trajectories of the limiting OD associated with the Q learning algorithm also carries over to this iterates of your Q learning algorithm and one can then conclude that the limiting behavior of the Q learning algorithm will also be Q star. And in this way, one can show that if you have this Q learning algorithm, which actually makes use of data obtained by interacting with the environment using this behavior policy mu, one can show that the limiting behavior will mirror that of the limiting behavior of your limiting ODE. And hence, one can show that your Q learning algorithm will converge to Q star. And as I told you, once you get Q star, the optimal policy can be easily estimated because the optimal policy is basically greedy with respect to Q star.

So, if you can somehow get a good enough estimate of Q-star such that the ordering of the different state-action values matches those of your Q-star's ordering, then from this estimate itself, one can get an estimate of Q-star. So, of course, the problem with this algorithm is that it runs in a space whose dimension equals the product of the state and action spaces, right? And when the state and action spaces are very large, then running this algorithm becomes problematic. So, in summary, in the no-function-approximation case, the Q-learning algorithm is good enough to find Q-star. However, in practice, the state and action spaces are very large, and we cannot run this algorithm. So, now the question becomes: how can we run it in an approximate way?

This will be the discussion in the forthcoming class. Until then, thank you and namaste.