STOCHASTIC APPROXIMATION: THEORY AND APPLICATIONS

Dr. Gugan Thope

Department of Computer Science and Engineering

Indian Institute of Science, Bangalore

Week 12

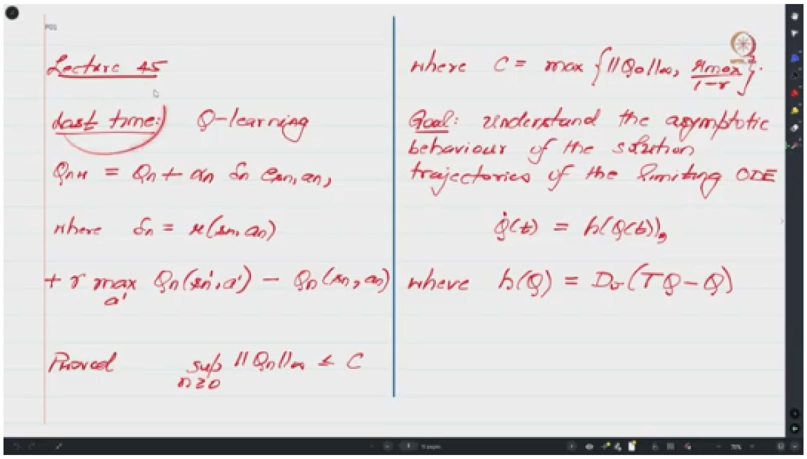
Lecture 45

Asymptotic Behaviour of the Q-Learning Limit ODE — A Switching Systems Perspective

Hello and Namaste everyone. Welcome to lecture 45 of this NPTEL course on Stochastic Approximation. During this week, we are trying to see if we can use stochastic approximation ideas for analyzing an algorithm that can be used for finding the optimal policy. So, in the last two lectures, we came up with an algorithm—I mean, we came up with an algorithm which we refer to as the Q-learning algorithm. And then we showed that the iterates of this Q-learning algorithm are almost surely bounded, which tells us that if the limiting ODE is well-behaved, right, then the asymptotic behavior of the Q-learning iterates will be dictated by that of the solution trajectories of this limiting ODE.

So, in this class and the next, we will try to analyze the asymptotic behavior of the solution trajectories of this limiting ODE and see how we can use that to say something about the limiting behavior of the Q-learning iterates itself. And, if you saw in the previous class, when we tried showing that the iterates are bounded, we did not make use of the scaling ODE or scaled ODE. Instead, we used some very basic principles to establish that the iterates are almost surely bounded. In this class and the next class, you will see that again, to understand the asymptotic behavior of the solution trajectories of this limiting ODE associated with the Q-learning algorithm, we would not make use of Lyapunov functions and things like that. Instead, we will use a very new idea called the vector comparison lemma, and you will see how that can be used to discuss the asymptotic behavior of the limiting ODE associated with the Q-learning algorithm.

Let us now discuss things in a formal way. So, as I said last time, you know, we have been focusing on the Q-learning algorithm whose update rule is given over here, right? Here, delta n has the formula R of S n A n plus gamma times max of this expression minus Q n S n A_n. And as I told you, this delta n is referred to as the temporal difference error because this quantity, in some sense, is an estimate of Q star S_n A_n, and this is another estimate of Q star S_n A_n, and we are basically comparing the difference between the two. And in the previous class, we showed that the iterates are almost surely bounded, where C is defined to be this.



Lecture 45

last time: Q-learning

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

where $\delta_n = r(s_n, a_n)$

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

Proved $\sup_{n \geq 0} \|Q_n\|_\infty \leq C$

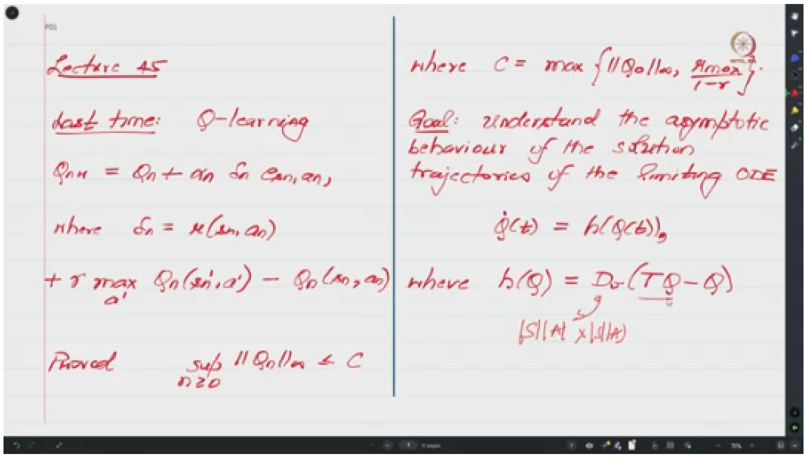
where $C = \max\left\{\|Q_0\|_\infty, \frac{R_{max}}{1-\gamma}\right\}$.

Goal: understand the asymptotic behaviour of the solution trajectories of the limiting ODE

$\dot{Q}(t) = h(Q(t))$,

where $h(Q) = D_\mu(TQ - Q)$

And in this class and the next, we will try to understand the asymptotic behavior of the solution trajectories of the limiting ODE that is associated with this algorithm, which can be described as Q dot of T equals H of Q of T, where H of Q is given in the following way. So, this expression we had established in the previous class. So, H of Q is equal to D mu times TQ minus Q. And this is a SAC. cross SA matrix right and T is your Bellman optimality operator it takes as input a SA dimensional vector and spits out another SA dimensional vector that is what you have over here and this is another SA dimensional vector and if you recall TQ of SA equals RSA plus gamma times sum over S prime P of S prime given SA times max of Q of S prime A prime where the max is taken over A prime.



Lecture 45

Last time: Q-learning

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

where $\delta_n = r(s_n, a_n)$

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

Proved $\sup_{n \geq 0} \|Q_n\|_\infty \leq C$

where $C = \max\left\{\|Q_0\|_\infty, \frac{r_{max}}{1-\gamma}\right\}$.

Goal: understand the asymptotic behaviour of the solution trajectories of the limiting ODE

$\dot{Q}(t) = h(Q(t))$,

where $h(Q) = \underset{|S||A| \times |S||A|}{D_\mu}(\underline{TQ} - Q)$

$$Q_{n+1} = Q_n + \alpha_n \delta_n \left(s_n, a_n\right)$$

$$\delta_n = r\left(s_n, a_n\right)$$

$$+ \ rQ_n\left(s_n^{'}, a^{'}\right) - Q_n\left(s_n, a_n\right)$$

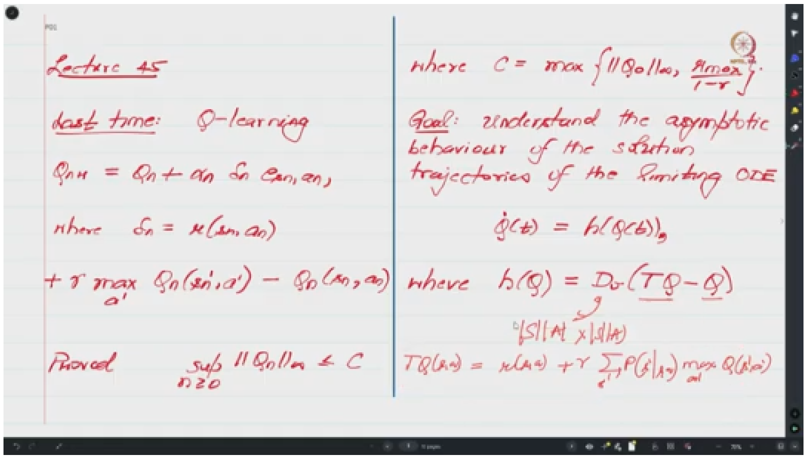
$$\left|\left|Q_n\right|\right|_\infty \leq C$$

$$C = \left\{\left|\left|Q_0\right|\right|_\infty, \frac{r_{max}}{1-r}\right\}$$

$$\dot{Q}(t) = h(Q(t))$$

$$h(Q) = D_\mu(TQ - Q)$$

So, this is the expression for TQ that we had discussed from the previous class and this T over here is referred to as the Bellman optimality operator. And to analyze the behavior of the solution trajectories of this limiting ODE, we will make use of this paper called Unified Switching Systems Perspective and ODE Analysis of Q Learning Algorithms by Dongwan Li and Niao He. Right and you can look up this paper in case you miss any details during the class basically I am going to summarize the discussion that is there from this paper right. So, first let me talk about what is known as the switch systems right and we will make use of the switch system idea to understand the behavior of the Q learning algorithm. And why are we making use of this switch system also I will soon explain but the high level idea is that if you remember you know in the T operator there is a max expression and the max will depend on the Q value.



Lecture 45

Last time: Q-learning

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

where $\delta_n = r(s_n, a_n)$

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

Proved $\sup_{n \geq 0} \|Q_n\|_\infty \leq C$

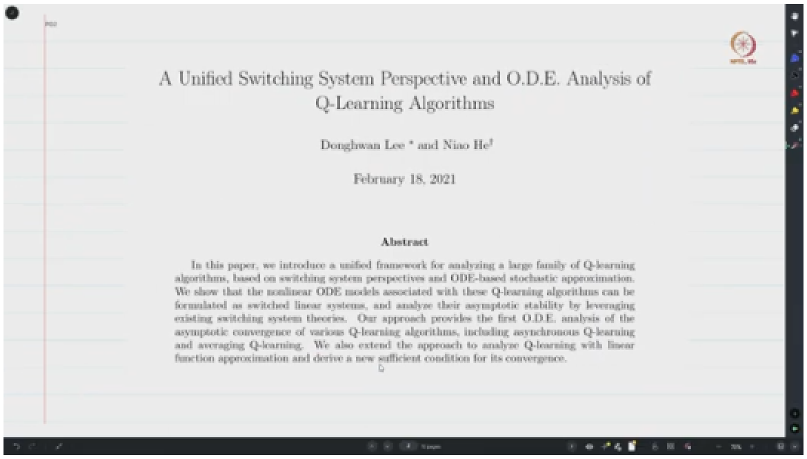
where $C = \max\left\{\|Q_0\|_\infty, \frac{R_{max}}{1-\gamma}\right\}$.

Goal: understand the asymptotic behaviour of the solution trajectories of the limiting ODE

$$\dot{Q}(t) = h(Q(t)),$$

where $h(Q) = \underset{|S||A| \times |S||A|}{D_\sigma}(\underline{T Q} - \underline{Q})$

$TQ(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q(s',a')$



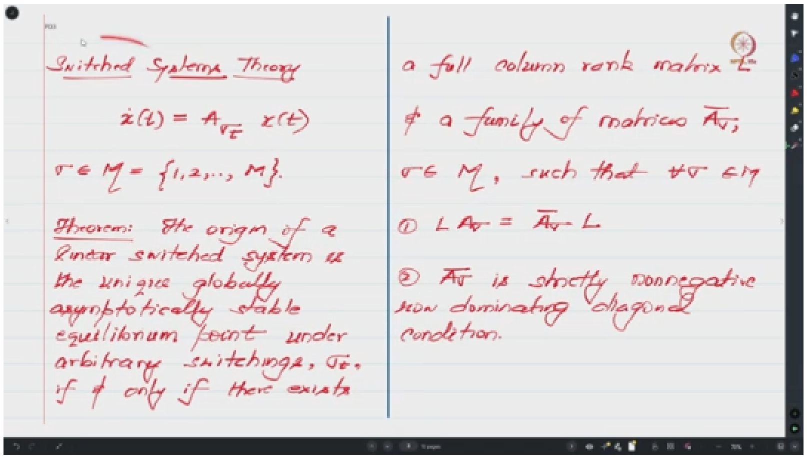
A Unified Switching System Perspective and O.D.E. Analysis of Q-Learning Algorithms

Donghwan Lee * and Niao He†

February 18, 2021

Abstract

In this paper, we introduce a unified framework for analyzing a large family of Q-learning algorithms, based on switching system perspectives and ODE-based stochastic approximation. We show that the nonlinear ODE models associated with these Q-learning algorithms can be formulated as switched linear systems, and analyze their asymptotic stability by leveraging existing switching system theories. Our approach provides the first O.D.E. analysis of the asymptotic convergence of various Q-learning algorithms, including asynchronous Q-learning and averaging Q-learning. We also extend the approach to analyze Q-learning with linear function approximation and derive a new sufficient condition for its convergence.



Switched Systems Theory

$$\dot{x}(t) = A_{\sigma_t} x(t)$$

$\sigma \in \mathcal{M} = \{1, 2, \dots, M\}$.

Theorem: The origin of a linear switched system is the unique globally asymptotically stable equilibrium point under arbitrary switchings, σ_t, if & only if there exists a full column rank matrix L & a family of matrices $\bar{A}_\sigma$, $\sigma \in \mathcal{M}$, such that $\forall \sigma \in \mathcal{M}$

① $L A_\sigma = \bar{A}_\sigma L$

② $\bar{A}_\sigma$ is strictly nonnegative row dominating diagonal condition.

And, you know, as the Q value changes, right, like, for example, if your Q was two dimensional, and if you are below the X equals Y diagonal, right, then the, you know, max will be the first coordinate. And if you are above the diagonal, then the max will be the second coordinate, right. So, you can see that the dynamics actually switches, right. when the iterates lie below the diagonal or above the diagonal. And in general, when your iterates lie in a SA dimensional space, right?

Similarly, the dynamics will change depending on where the Q values lie, right? Because of the presence of this max operation. And that is the reason we have to sort of work with this switched system theory. So, here is an example of a general switched system. So, the general switched system is you have x dot of t equals a subscript sigma t times x of t. So, if the sigma t was not there, then this would have just been a times x of t, in which case we would have the vanilla linear ODE system.

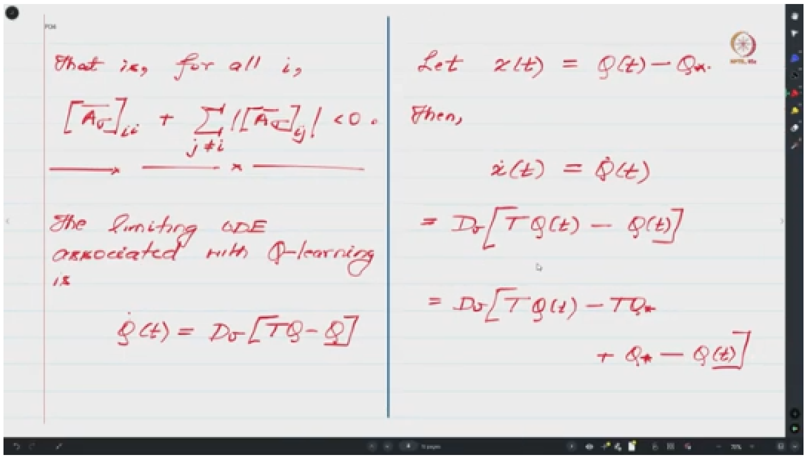
However, here what is happening is the a is indexed by sigma t, and sigma in particular depends on t. So, you know, as your t changes, right, the driving matrix also changes, and because of this, such a dynamical system is referred to as a switched system, right? And sometimes this x can depend on t, sometimes it can depend on t via x of t, and sometimes it can depend on both t and x of t, and so on and so forth. Now, for such systems, of course, the origin is an equilibrium point, but one would like to know: can we somehow identify some conditions on these switching matrices to say something about, you know, when the—I mean, to check whether the origin is a globally asymptotically stable equilibrium for this ODE or not. So, let us give some formal notation. So, we will presume that A is a function of some switching signal sigma t, and sigma t can take any value in this calligraphic M set, and for simplicity, we will presume that this calligraphic M set is made up of M nodes.

So, 1, 2, capital M. So, sigma of t sometimes can be 1, sigma of t sometimes can be 2, and you know, sometimes it can be any number between 1 to capital M. So, in that way, we have presumed that you know sigma of t can take any of these values, and A sigma t accordingly will be one of the M many matrices. So, you can presume that we have M many choices here, and which choice you use is based on this switching signal sigma of t.

So then there is this theorem, and you can find the references for this theorem from that Li-He paper. So, this theorem says that the origin

Of a linear switch system of the form that is given over here, there is a unique globally asymptotically stable equilibrium. So, because you have a switch system, it is not clear if the origin will always be a globally asymptotically stable equilibrium or not, and this theorem says that it indeed is under arbitrary switching if and only if There exists some full column rank matrix L and a family of matrices A bar sigma such that for every sigma in this calligraphic set M, you have the following two conditions. On the one hand, you have that L times A sigma. So, A sigma is what you have over here where sigma T is sigma.

So, L times A sigma equals A bar sigma times L. So, A bar sigma is over here. Furthermore, you require that this A bar sigma be strictly non-negative row dominating—I mean, this A bar sigma should have a strictly non-negative row dominating diagonal condition. So, I should say, instead of saying this, I should say this has a strictly non-negative row dominating diagonal condition. What does that mean? It means that if you look at the ith diagonal entry of this matrix, right?



That is, for all i,

$$[\bar{A}_\sigma]_{ii} + \sum_{j \neq i} |[\bar{A}_\sigma]_{ij}| < 0.$$

The limiting ODE associated with Q-learning is

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

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

Then,

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

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

$$= D_\sigma[TQ(t) - TQ_* + Q_* - Q(t)]$$

So, A bar sigma is a matrix. So, you look at the ith diagonal entry of it, right? And then you look at all the off-diagonal entries in the ith row, right? And you take their absolute values and then sum them up. So, let me again repeat.

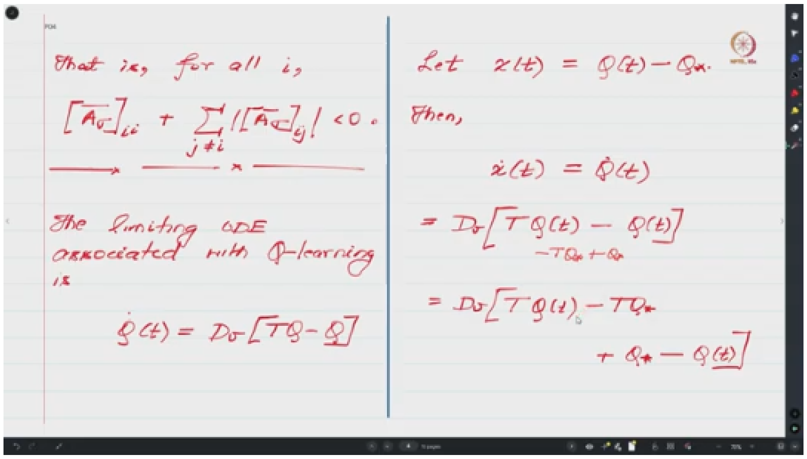
Here you look at the ith diagonal entry and notice that you do not take the absolute value here and here you take the absolute values of the off diagonal entries and put the absolute value over here and you want to add them up. And you want that this sum be strictly negative. And this should be true for every i. So, which means the diagonal entry should be negative. On top of that, it should be sufficiently negative so that even if you take all the off diagonal entries individually, take their absolute values and add them up, their sum when it is added to the ith diagonal entry should not make it 0 or positive, right?

So that is the condition. So in other words, the diagonal entry is negative and it is dominating, right? That is what this condition over here means. And just to summarize this result quickly, it says that look you may have these arbitrary switchings but if every switching matrix has some complementary matrix which we denote here by A bar sigma and if this A bar sigma is negative diagonal dominating, right then for the switching system the origin will be a globally asymptotically stable equilibrium which means that even though you are switching right every solution trajectory of the switch system will always converge to the origin.

So now we are going to make use of this result to say something about the limiting behavior of the solution trajectories I mean of the solution trajectories of the limiting OD that is associated with your Q learning algorithm. So, recall that the limiting OD that is associated with your Q learning algorithm has the form that is given over here which is Q dot of T equals D nu times TQ minus Q. So, recall that this nu is some behavior policy. This is the policy with which you interact with the environment, right? And we will presume that this behavior policy is such that this D nu matrix has strictly positive entries. Is this okay?

So now what we are going to do is, you know, we would like that, you know, the solution trajectories of this ODE go to Q star, right? But in order to apply the previous result, okay, your equilibrium point should be shifted to the origin. And in order to do that, what we will do is we will define this new signal X of t, which is basically X of t minus X star. So, in case we want to show Q of t—sorry, I should say X of t is Q of t minus Q star—and since we want to show Q of t goes to Q star, it suffices to show that X of t goes to the origin. And in that sense, you know, for the dynamical system that is associated with X t, the hope is that we can somehow make use of this switched system theory that we just saw. And if you denote X of t in this fashion, if you take the derivative of X of t—because this Q star is a constant—one can see that X dot of t will indeed be Q dot of t. And recall that Q dot of t is exactly this expression, and this expression we can substitute it back over here. And in order to do that, what we will do is we will add and subtract

Q star and make use of the fact that TQ star is Q star, right? Or Q star, which we are interested in finding. If you remember, I said that it satisfies this Bellman optimality equation, which is basically saying that TQ star is equal to Q star. So, instead of subtracting Q star and adding Q star, what I have done is I have subtracted TQ star and added Q star, which is one and the same because TQ star equals Q star. So if I add and subtract Q star—or equivalently, minus TQ star and Q star—so I can write it as minus TQ star here and Q star over here, and this expression is precisely minus X of T, which is what we have defined over here. And this expression—notice that T is not a linear operator—so I cannot, you know, write this as T times Q of T minus Q star because T is not a linear operator. In fact, it is non-linear because of the presence of the max expression, and in fact, this is what makes



That is, for all i,

$$[A_\sigma]_{ii} + \sum_{j \neq i} |[A_\sigma]_{ij}| < 0.$$

The limiting ODE associated with Q-learning is

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

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

Then,

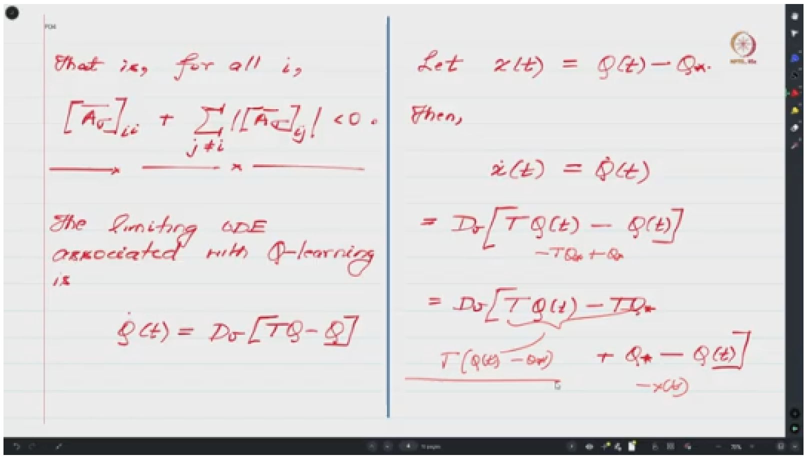
$$\dot{x}(t) = \dot{Q}(t)$$

$$= D_\sigma[TQ(t) \underset{-TQ_* + Q_*}{-} Q(t)]$$

$$= D_\sigma[TQ(t) - TQ_* + Q_* - Q(t)]$$

the analysis of your limiting ODE associated with your Q-learning algorithm a bit challenging, right. So, this is what we have over here—I mean, we cannot write it in this fashion. So, now let us see how we can express this TQT minus TQ star. So what we will do is we will expand TQT using the formula for T. So TQT what it will do is it will take R plus gamma times P Q bar T, where—recall—whenever I write Q bar, I mean that your

Q bar T is basically a vector in SA dimensional—or maybe I will write this without the T over here so that it can be defined in general. So, this is an S-dimensional vector and Q bar of S.



That is, for all i,

$$[A_\sigma]_{ii} + \sum_{j \neq i} |[A_\sigma]_{ij}| < 0.$$

The limiting ODE associated with Q-learning is

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

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

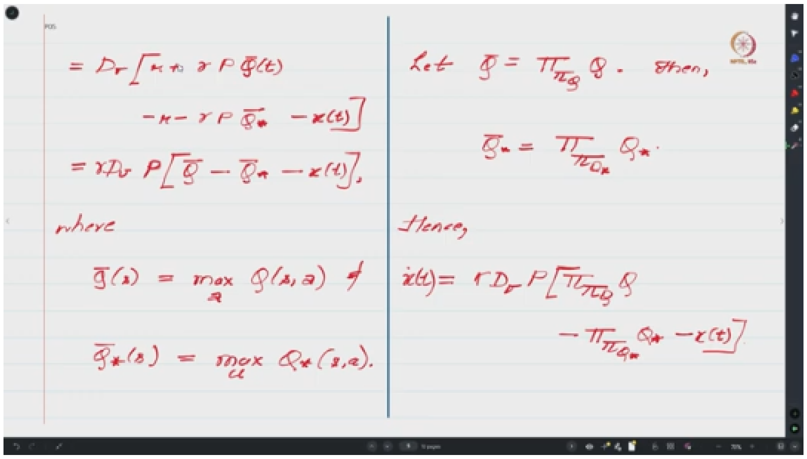
Then,

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

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

$$-TQ_* + Q_*$$

$$= D_\sigma[\underbrace{TQ(t) - TQ_*}_{T(Q(t) - Q_*)} + \underbrace{Q_* - Q(t)}_{-x(t)}]$$



$$= D_\sigma[r + \gamma P\bar{Q}(t) - r - \gamma P\bar{Q}_* - x(t)]$$

$$= \gamma D_\sigma P[\bar{Q} - \bar{Q}_* - x(t)],$$

where

$$\bar{Q}(s) = \max_a Q(s,a) \ \&$$

$$\bar{Q}_*(s) = \max_a Q_*(s,a).$$

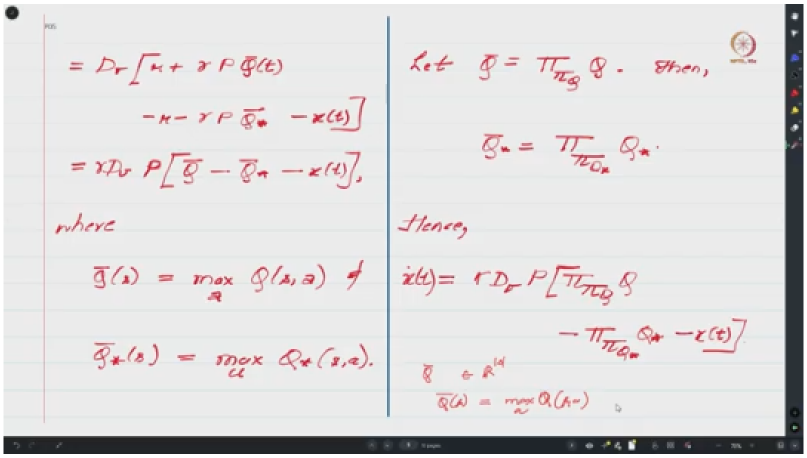
Let $\bar{Q} = \Pi_{\pi_Q} Q$. Then,

$$\bar{Q}_* = \Pi_{\pi_{Q_*}} Q_*.$$

Hence,

$$\dot{x}(t) = \gamma D_\sigma P[\Pi_{\pi_Q} Q - \Pi_{\pi_{Q_*}} Q_* - x(t)]$$

is basically max of Q of SA where the max is taken over A. So, I will sort of use this as shorthand. So, I can express TQT in the following way R plus gamma P Q bar T and TQ star I will express as R plus gamma P Q bar star and this R and this R will cancel off and hence By taking this gamma P in common, I can write it as gamma times d nu P times Q bar minus Q star bar and this X of P, I will write it as it is. And as I have defined here, Q bar of S is basically max of QSA and Q bar star of S is defined in this way. So, notice that here I have a round bracket T ideally I should put round bracket T here as well but you know sometimes I will not write it so that the notation becomes slightly easy.



$$= D_r[\kappa + \gamma P \bar{Q}(t) - \kappa - \gamma P \bar{Q}_* - x(t)]$$

$$= \gamma D_r P[\bar{Q} - \bar{Q}_* - x(t)],$$

where

$$\bar{Q}(s) = \max_a Q(s,a) \; \&$$

$$\bar{Q}_*(s) = \max_a Q_*(s,a).$$

Let $\bar{Q} = \Pi_{\pi_Q} Q$. Then,

$$\bar{Q}_* = \Pi_{\pi_{Q_*}} Q_*.$$

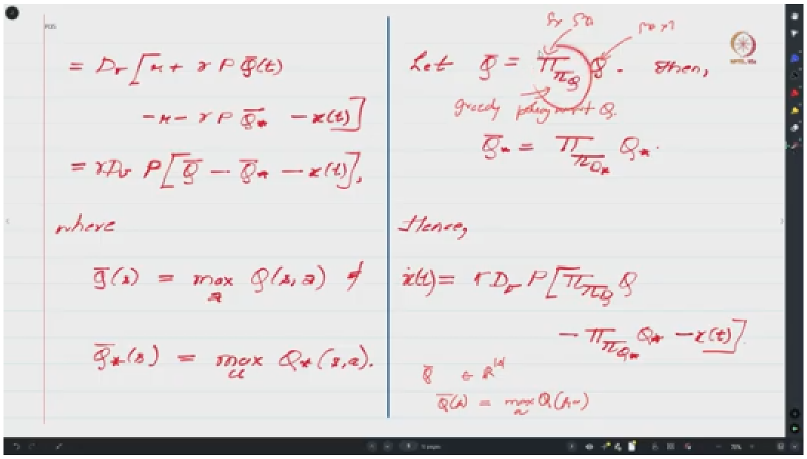
Hence,

$$\dot{x}(t) = \gamma D_r P[\Pi_{\pi_Q} Q - \Pi_{\pi_{Q_*}} Q_* - x(t)]$$

$Q \in \mathbb{R}^{|d|}$

$\bar{Q}(s) = \max_a Q(s,a)$

So, of course this is a constant but this quantity indeed depends on T. So, now what we will do is instead of writing these bars what we will do is we will define an operator called superscript sorry capital pi subscript little pi Q right so the little pi Q over here basically denotes the greedy policy with respect to Q okay and this here will be a S cross S A dimensional matrix right and recall that this is basically S A cross one vector. So, the idea is that after you take this matrix and multiply it with Q you end up with the S dimensional vector and the output should be such that the S th coordinate of this output you know satisfies this relation over here. Is this okay?



$$= D_r[\kappa + \gamma P \bar{Q}(t) - \kappa - \gamma P \bar{Q}_* - x(t)]$$

$$= \gamma D_r P[\bar{Q} - \bar{Q}_* - x(t)],$$

where

$$\bar{Q}(s) = \max_a Q(s,a) \; \&$$

$$\bar{Q}_*(s) = \max_a Q_*(s,a).$$

$S \times SA$ $SA \times 1$

Let $\bar{Q} = \Pi_{\pi_Q} Q$. Then,

greedy policy w.r.t Q.

$$\bar{Q}_* = \Pi_{\pi_{Q_*}} Q_*.$$

Hence,

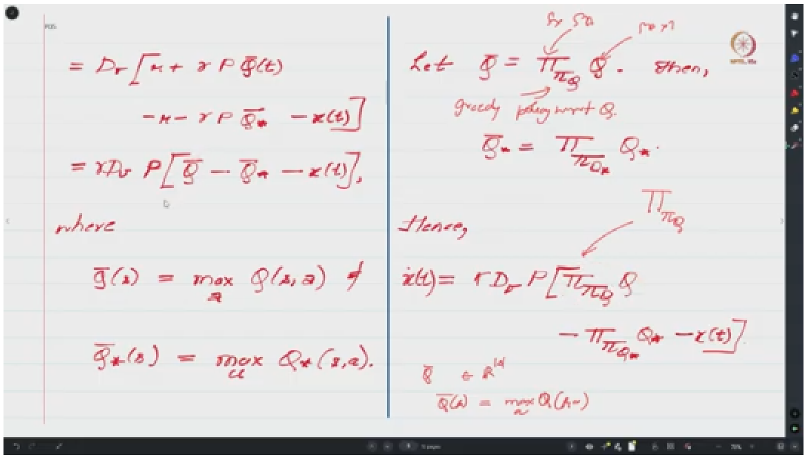
$$\dot{x}(t) = \gamma D_r P[\Pi_{\pi_Q} Q - \Pi_{\pi_{Q_*}} Q_* - x(t)]$$

$Q \in \mathbb{R}^{|d|}$

$\bar{Q}(s) = \max_a Q(s,a)$

And one can define a 0, 1 matrix suitably so that, you know, this definition holds true, right? So, I can define capital Pi, subscript Pi Q star to denote, you know, picking that action which maximizes your Q star values, right? So, in this way, we can define it favorably. as follows and hence one can show that x dot of p is gamma d nu p which I have copied over here and instead of writing q bar I am now going to write it as capital pi subscript little pi q I mean I think I should write this carefully so this is capital pi subscript little pi Q, right?

That is what I have over here times Q, right? This is basically picking the greedy action so that you end up with Q bar over here. And similarly, this Q bar star can be written as capital pi subscript pi Q star times Q star and whatever this minus Xt is, I have written it as it is, okay? So, in this way, one can see that we can, you know, get back this X dot of t can be represented in this fashion. And I am doing all this so that eventually instead of Q, I can replace it with X of t, right?



$$= D_\nu \Big[r + \gamma P \bar{Q}(t) - r - \gamma P \bar{Q}_* - x(t) \Big]$$

$$= \gamma D_\nu P \Big[\bar{Q} - \bar{Q}_* - x(t) \Big],$$

where

$$\bar{Q}(s) = \max_a Q(s,a) \ \forall$$

$$\bar{Q}_*(s) = \max_a Q_*(s,a).$$

Let $\bar{Q} = \Pi_{\pi_Q} Q$. Then,

greedy policy wrt Q.

$$\bar{Q}_* = \Pi_{\pi_{Q_*}} Q_*.$$

Hence,

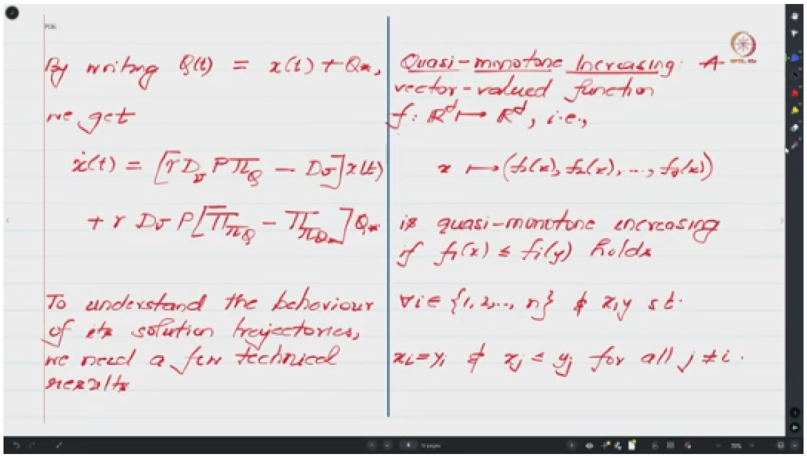
Π_{π_Q}

$$\dot{x}(t) = \gamma D_\nu P \Big[\Pi_{\pi_Q} Q - \Pi_{\pi_{Q_*}} Q_* - x(t) \Big]$$

$Q \in \mathbb{R}^{|S|}$

$\bar{Q}(s) = \max_a Q(s,a)$

And in order to do that, what I will do is I will, you know, take this, subtract Q star and add Q star, right? So, I can do it in this way and this quantity over here, right? I will denote it by X, right? So, if I write it in this fashion, one can see that your X dot of t, right? X dot of t will equal gamma d nu p and you can see that this p capital pi little pi q that is what we have over here.



By writing $Q(t) = x(t) + Q_*$, we get

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

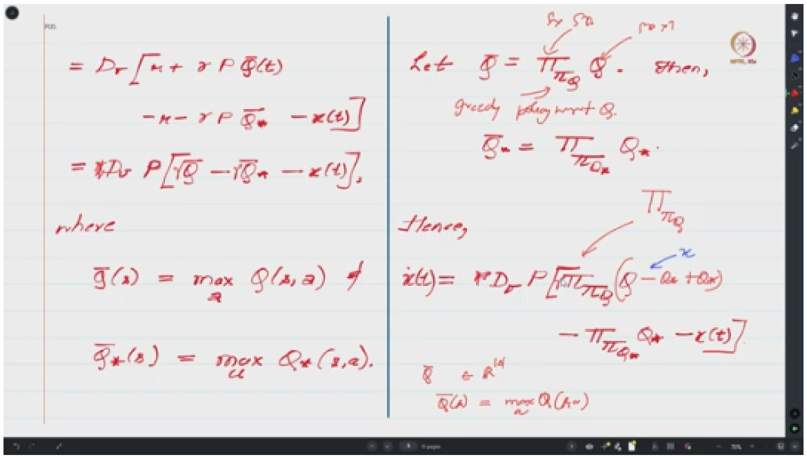
To understand the behaviour of the solution trajectories, we need a few technical results

Quasi-monotone Increasing: A vector-valued function $f : \mathbb{R}^d \mapsto \mathbb{R}^d$, i.e.,

$$x \mapsto (f_1(x), f_2(x), \dots, f_d(x))$$

is quasi-monotone increasing if $f_i(x) \le f_i(y)$ holds $\forall i \in \{1, 2, \dots, n\}$ & x, y s.t. $x_i = y_i$ & $x_j \le y_j$ for all $j \ne i$.

I think I should write this as capital pi pi q and this expression comes from this quantity over here. I think there is a mistake here this gamma should actually be inside. I think I have made a mistake this gamma should be over here and this should be without gamma. So, this also should be with gamma and similarly, this gamma should not be there outside rather it should be multiplying this expression over here. Sorry about this.



$$= D_\nu[r + \gamma P \bar{Q}(t) - r - \gamma P \bar{Q}_* - x(t)]$$

$$= \cancel{\gamma} D_\nu P[\gamma \bar{Q} - \gamma \bar{Q}_* - x(t)],$$

where

$\bar{Q}(s) = \max_a Q(s,a)$ &

$\bar{Q}_*(s) = \max_a Q_*(s,a)$.

Let $\bar{Q} = \Pi_{\pi_Q} Q$. Then,

for SA

[SA×1]

greedy policy w.r.t Q.

$\bar{Q}_* = \Pi_{\pi_{Q_*}} Q_*$.

Hence,

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

Π_{π_Q}

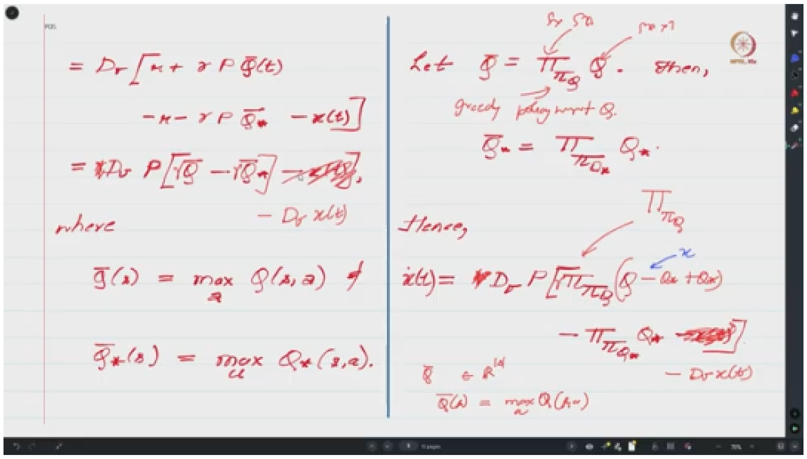
x

$Q \in \mathbb{R}^{|s|}$

$\bar{Q}(s) = \max_a Q(s,a)$

So, one can, you know, check this and then one can conclude that your x dot of t will have the form that is given over here. So, this quantity comes because you will take this d nu. Okay, I think I have made a mistake here. So, this square bracket should also—I mean, maybe I will just write this part clearly, okay. So, this should be minus x. d nu x of t. Here also, I have made this mistake.

So, this should actually be this minus d nu x of t. So, the x of t—let me just see here—is. d nu times minus x of t is what you should have. I inadvertently took this p and said that, you know, the p multiplies this x of t as well. So, that was a mistake, right? So, this d nu x of t is there.



$$= D_\nu \left[r + \gamma P \bar{Q}(t) - r - \gamma P \bar{Q}_* - x(t) \right]$$

$$= D_\nu P \left[\gamma \bar{Q} - \gamma \bar{Q}_* \right] - D_\nu x(t)$$

where

$$\bar{Q}(s) = \max_a Q(s,a) \;\&$$

$$\bar{Q}_*(s) = \max_a Q_*(s,a).$$

Let $\bar{Q} = \Pi_{\pi_Q} Q$. Then,

$$\bar{Q}_* = \Pi_{\pi_{Q_*}} Q_*.$$

Hence,

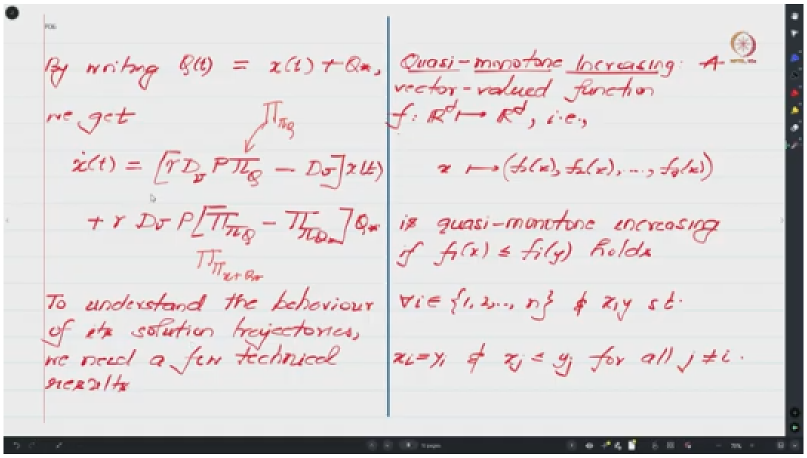
$$\dot{x}(t) = D_\nu P \left[\Pi_{\pi_Q} (\overset{x}{Q - Q_*} + Q_*) - \Pi_{\pi_{Q_*}} Q_* \right] - D_\nu x(t)$$

$Q \in \mathbb{R}^{|A|}$

$\bar{Q}(s) = \max_a Q(s,a)$

And if you take d nu outside, you would see this, you know, expression that I have written over here. So, this will multiply x of t, and the remaining expression indeed will have gamma d nu p multiplied. Pi subscript pi Q minus pi subscript pi Q star times Q star. So, one can see that we will end up with some expression like this, and, you know, for, you know, replacing this last Q, one can also see that this can be written as pi, you know, X plus Q star. Recall that X of t is Q of t minus Q star, and hence Q of t will be X of t plus Q star.

So, this Q can also be written in this fashion, and this Q can also be written in this fashion. So, that eventually you will not have any Q anywhere. Instead, the whole update, you know, the driving function of this limiting ODE will be a function. So, of course, I would again like to highlight that this x is actually a function of t. However, you know, I am suppressing this t so that the notation becomes slightly manageable. So, this is the, you know, ODE that we have, right, and we would like to understand.



By writing $Q(t) = x(t) + Q_*$, we get

$$\Pi_{\pi_Q}$$

$$\dot{x}(t) = [\gamma D_\mu P \Pi_Q - D_\mu] x(t) + \gamma D_\mu P [\Pi_{\pi_Q} - \Pi_{\pi_{Q_*}}] Q_*$$

$$\Pi_{\pi_{x+Q_*}}$$

To understand the behaviour of the solution trajectories, we need a few technical results

<u>Quasi-monotone Increasing</u>: A vector-valued function $f: \mathbb{R}^d \mapsto \mathbb{R}^d$, i.e.,

$$x \mapsto (f_1(x), f_2(x), \ldots, f_d(x))$$

is quasi-monotone increasing if $f_i(x) \le f_i(y)$ holds $\forall i \in \{1, 2, \ldots, n\}$ & x, y s.t.

$x_i = y_i$ & $x_j \le y_j$ for all $j \ne i$.

The behavior of the solution trajectories of this ODE. So, let me just recall why we have done it in this fashion, right? So, we want to invoke the switched systems theory, and in order to do the switched systems theory, right? We want to somehow come up with a system for which the origin is the equilibrium point, right? And in order to do that, that is what we

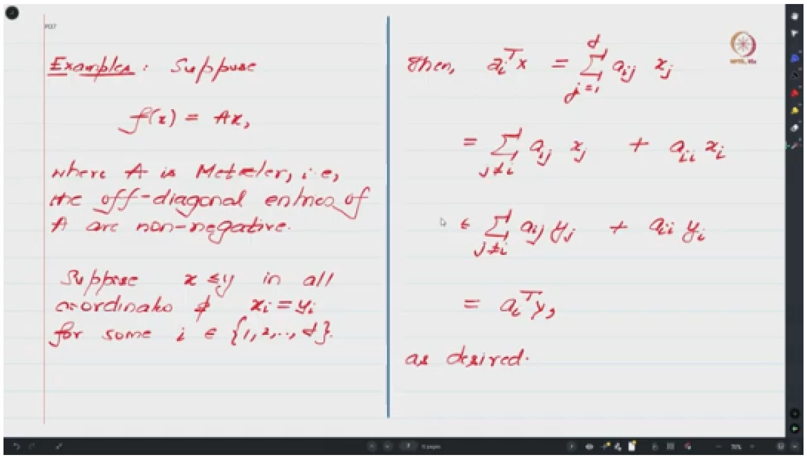
You know, in order to do that, we subtracted this, you know, Q star from Q of t and we have written down an ODE corresponding to Q of t minus Q star. So, this is the ODE that we have ended up with. Right. And we want to now understand the behavior of such an update rule. Right.

Or I shouldn't say update rule, such a limiting ODE. So towards that, what we will do is we will make use of some technical result and the technical result goes as follows. Right. And for that, we need to first define something called as a quasi monotone increasing function. Right.

So let me explain what that is. A vector valued function F is right vector valued means it takes as input some element in rd and spits out an element in rd okay so f is the output of f is vector valued and the input is also vector valued that is what i will refer to as you know a vector valued function whose input is also vector valued So what it means is that you take X which is in D dimensions and split out D real numbers so that the output also lies in RD and let us denote the first output as F1X, the second coordinate of the output as F2 of X and similarly the Dth coordinate of the output as RD. fd of x right and we will say such a function is quasi monotone increasing if it satisfies the property that is listed over here so let us understand what this property says it says that if you give x and y so x and y are elements in rd suppose you give x and y as input to your function f

And suppose x and y have the property that on all coordinates xj is less than yj. So coordinate wise x is below y. Furthermore there is at least one i for which xi equals yi. So in other words on all coordinates x is below y and on some coordinates xi equals yi. Now for all i where xi equals yi we require that fi of x that is the ith coordinate of your output should be less than the ith coordinate of you know the output when the input is y. So the ith coordinate of the output when the input is x should be less than the ith coordinate of the output when the input is y. So, notice that this condition is only needed for those i where xi equals yi, right?

We do not need such a condition for, you know, other coordinates where xj could be strictly less than yj. So, only for the coordinates where xi equals yi, we require that fi of x be less than or equal to fi of y, and hence the word quasi-monotone. Is this okay? So, let us look at an example where this quasi-monotonicity condition holds. Suppose f of x equals a times x, where a is modular, meaning the off-diagonal entries of a are all non-negative, right?



Examples: Suppose

$f(x) = Ax,$

where A is Metzler, i.e, the off-diagonal entries of A are non-negative.

Suppose $x \leq y$ in all coordinates & $x_i = y_i$ for some $i \in \{1, 2, \dots, d\}$.

Then, $a_i^T x = \sum_{j=1}^{d} a_{ij} x_j$

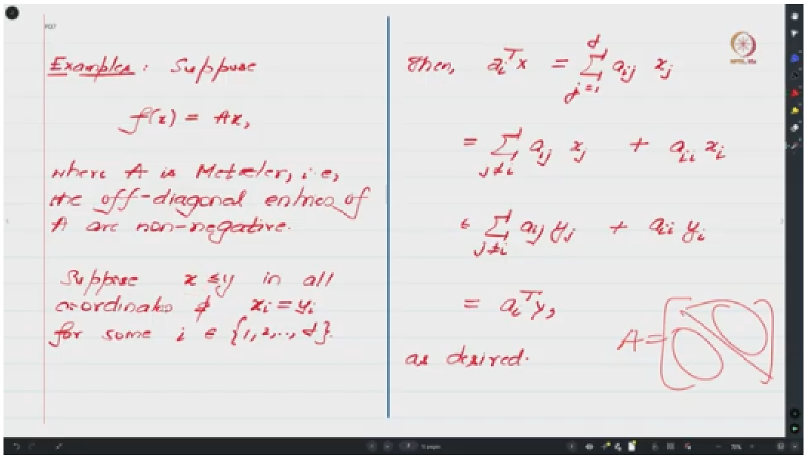
$= \sum_{j \neq i} a_{ij} x_j + a_{ii} x_i$

$\leq \sum_{j \neq i} a_{ij} y_j + a_{ii} y_i$

$= a_i^T y,$

as desired.

So, the diagonal entries can be negative, or you know, zero or positive, but we require that all the off-diagonal entries—which means that if I define, you know, matrix a— So, if this is your diagonal, all the off-diagonal entries should be either zero or positive. You know, the input to this function f is, you know, x and y, and they satisfy the condition x is less than or equal to y in all coordinates, with the exception of the ith coordinate where xi equals yi, right? And let us say i is arbitrary. Then the claim is that if you look at the ith coordinate of the output when the input is x, that will be less than the ith coordinate of the output when the input is y, right?



Examples: Suppose

$f(x) = Ax,$

where A is Metzler, i.e, the off-diagonal entries of A are non-negative.

Suppose $x \leq y$ in all coordinates & $x_i = y_i$ for some $i \in \{1, 2, \dots, d\}$.

Then, $a_i^T x = \sum_{j=1}^{d} a_{ij} x_j$

$= \sum_{j \neq i} a_{ij} x_j + a_{ii} x_i$

$\leq \sum_{j \neq i} a_{ij} y_j + a_{ii} y_i$

$= a_i^T y,$

as desired.

$A =$

So, towards that, let us look at the ith coordinate of f of x, which is basically ai transpose x, where ai transpose is the ith row of a. And one can see that ai transpose x is basically the sum over aij times xj, j equals 1 to d. I can split this sum into two parts. This is basically the sum over the off-diagonal entries aij, and this is the quantity based on the diagonal entry. And we know that the off-diagonal entries are all non-negative. That is what we have been told over here. And xj is less than yj.

Hence, I can show that this sum is less than or equal to this sum, where I have replaced xj with yj.

$$f(x) = Ax$$

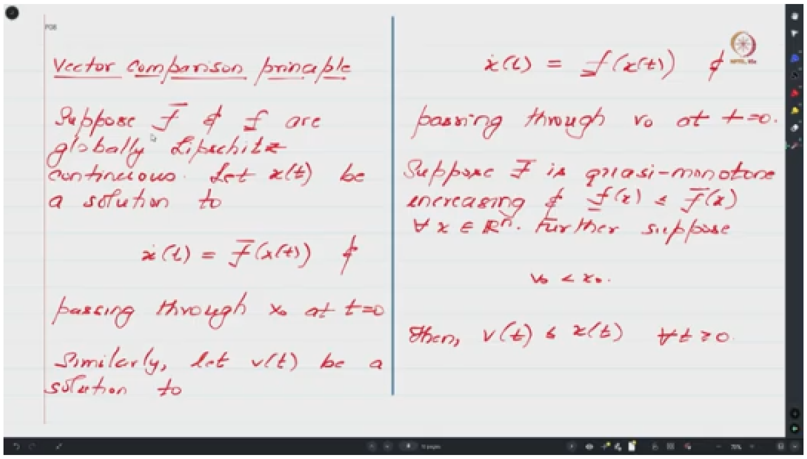
$$a_i^T X = \sum_{j=1}^{d} a_{ij} x_j$$

$$= \sum_{j \neq 1} a_{ij} x_j + a_{ii} x_i$$

$$\leq \sum_{j \neq 1} a_{ij} y_j + a_{ii} y_i$$

$$= a_i^T Y$$

And separately, I also know that for the diagonal entry, xi equals yi, right? So this is something that we have been told over here. Because xi equals yi, right? It does not matter whether aii is negative, zero, or positive, right? And one can then conclude that this sum is less than or equal to this sum. And one can then immediately see that this expression is basically ai transpose y, which is the ith coordinate of this function when the input is y, right? And one can see that for such a function, the quasi-monotonicity condition indeed holds, right. And, we are now going to state another important result, right, which is known as the vector comparison principle, right.



Vector comparison principle

Suppose $\bar{F}$ & $\underline{F}$ are globally Lipschitz continuous. Let $x(t)$ be a solution to

$$\dot{x}(t) = \bar{F}(x(t)) \;\&$$

passing through x_0 at $t=0$

Similarly, let $v(t)$ be a solution to

$$\dot{x}(t) = \underline{F}(x(t)) \;\&$$

passing through v_0 at $t=0$.

Suppose $\bar{F}$ is quasi-monotone increasing & $\underline{F}(x) \leq \bar{F}(x)$ $\forall x \in \mathbb{R}^n$. Further suppose

$$v_0 < x_0.$$

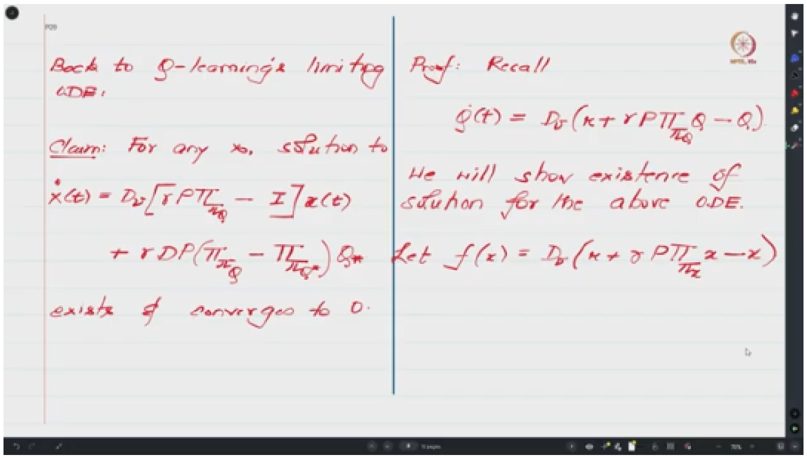
Then, $v(t) \leq x(t)$ $\forall t \geq 0$.

And we will be using this result to study the asymptotic behavior of the limiting ODE that is associated with your Q-learning algorithm, right. So what does the vector comparison principle say? It says that suppose you have two functions going from RD to RD. Let us denote the first function as F upper bar and the second function as F lower bar, and suppose these two functions are globally Lipschitz continuous. Furthermore, let X of P be the solution to this ODE where the driving function is F bar, F upper bar.

And let us say this ODE is passing through X naught. Similarly, let V of t be the solution to this ODE where the driving function is F under bar, right? And let us say at time t equals 0, the solution passes through V0, right? And what this result requires additionally is that this F upper bar, which is over here, be quasi-monotone increasing, right? And F lower bar x be less than F upper bar x for all x in Rn.

So, suppose a bunch of conditions like this hold. And finally, suppose the initial condition of the solution trajectory of this ODE is strictly less than in all coordinates the initial condition of the solution of this ODE; then one can show that So, you know the inequality holds between V of t and X of t for all t greater than or equal to 0. So, let me again repeat what this says: if your F bar and F lower bar are such that F lower bar is less than F upper bar and your F upper bar is quasi-monotone. Then, if you take a solution trajectory of this ODE and a solution trajectory of this ODE starting from V0 and X0 respectively, such that V0 is strictly less than X0.

Then, if this initial condition satisfies this inequality, then V of t and X of t will also satisfy this inequality for all t greater than or equal to 0. Now, what we are going to do is to see how we can use this vector comparison lemma to say something about the limiting behavior of the solution trajectories of this ODE that we had constructed. So, this is the ODE, and as I said, sometimes I will write this as pi Q or sometimes one can think of this expression as pi Q. So, depending on the context, I will sometimes think of it in this fashion or sometimes think of it in this fashion. So, our goal is to show that if you take any solution trajectory of this, first of all, such a solution trajectory exists, and the second thing is we want to show that it indeed converges to 0.



Back to Q-learning's limiting ODE:

Claim: For any x_0, solution to

$\dot{x}(t) = D_\mu[\gamma P\Pi_{\pi_Q} - I]x(t)$

$+ \gamma DP(\Pi_{\pi_Q} - \Pi_{\pi_{Q^*}})Q^*$

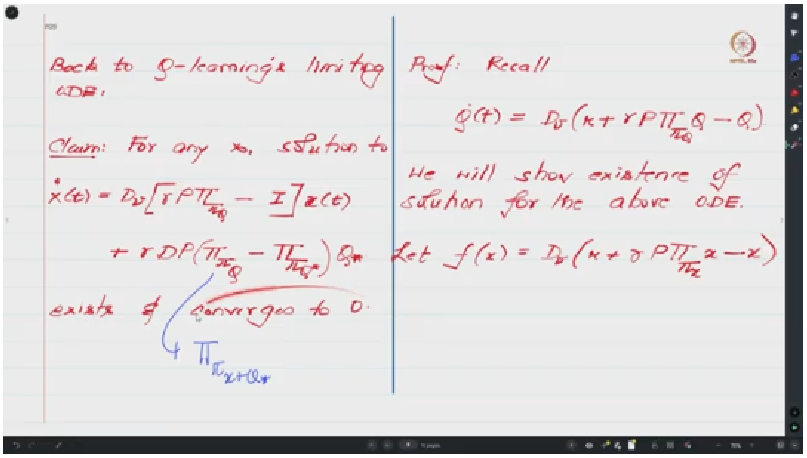
exists & converges to 0.

Proof: Recall

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

We will show existence of solution for the above ODE.

Let $f(x) = D_\mu(r + \gamma P\Pi_{\pi_x} x - x)$



Back to Q-learning's limiting ODE:

Claim: For any x_0, solution to

$$\dot{x}(t) = D_\mu[\gamma P \Pi_{\pi_Q} - I]x(t) + \gamma D P(\Pi_{\pi_Q} - \Pi_{\pi_{Q^*}})Q^*$$

exists & converges to 0.

$\Pi_{\pi_{x+Q^*}}$

Proof: Recall

$$\dot{Q}(t) = D_\mu(R + \gamma P \Pi_{\pi_Q} Q - Q).$$

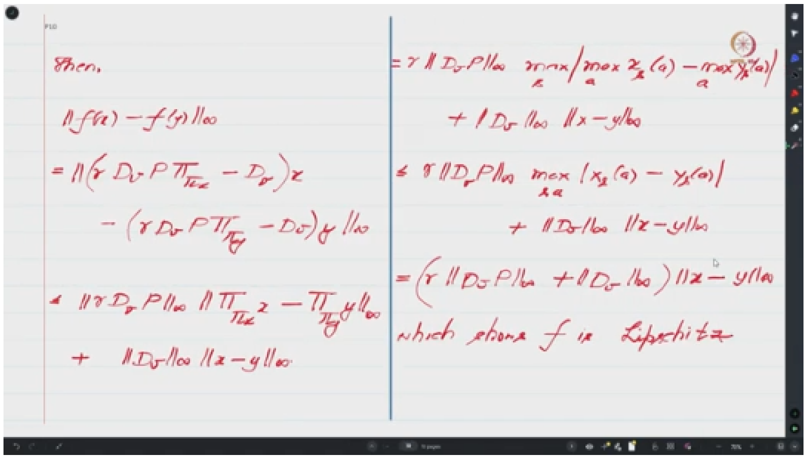
We will show existence of solution for the above ODE.

Let $f(x) = D_\mu(R + \gamma P \Pi_{\pi_x} x - x)$

Now, this ODE is a bit challenging to analyze. In particular, the switch system theory that we had studied is a bit difficult to directly invoke in the context of this ODE. So, what we will do is we will try to approximate this ODE using some other ODEs and apply the switch system ODE to this approximate switch system principle right to these approximate ODEs and, you know, say something about the solution trajectories of this approximate ODE and use that to say something about the solution trajectories of this ODE, okay. So, towards that, the first question that we would like to ask is, you know, does the solution trajectory exist for this ODE?

In other words, if I give you an initial condition, can one conclude that this ODE has a solution guaranteed, right, and it is unique? So, towards that, what we will do is we will go back to this ODE that we had. So, recall that we had obtained this ODE from this ODE by defining X of t to be Q of t minus Q star. So, let us go to this ODE, and you know, we will first show that this ODE, you know, has existence of solution and uniqueness of solution guaranteed, right? And the way we will do it is we will show that this driving function that we have over here, right, is actually Lipschitz continuous.

If it is Lipschitz continuous, then it is known that, you know, this ODE is well-posed, which means that for any initial condition, the solution of this ODE exists and is unique. So, towards that, let us define this function f to be this function where, wherever you have q, you replace it by x. Right? And in order to show this function is Lipschitz continuous, what we are going to do is we are going to look at the difference between f of x and f of y. Right? And in particular, we are going to look at the L-infinity distance between f of x and f of y. So, one can, you know, using the definition of f,



Then,

$\|f(x) - f(y)\|_\infty$

$= \|(\gamma D_\nu P \Pi_{\pi_x} - D_\nu)x - (\gamma D_\nu P \Pi_{\pi_y} - D_\nu)y\|_\infty$

$\le \|\gamma D_\nu P\|_\infty \|\Pi_{\pi_x} x - \Pi_{\pi_y} y\|_\infty + \|D_\nu\|_\infty \|x - y\|_\infty$

$= \gamma \|D_\nu P\|_\infty \max_s |\max_a x_s(a) - \max_a y_s(a)| + \|D_\nu\|_\infty \|x - y\|_\infty$

$\le \gamma \|D_\nu P\|_\infty \max_{s,a} |x_s(a) - y_s(a)| + \|D_\nu\|_\infty \|x - y\|_\infty$

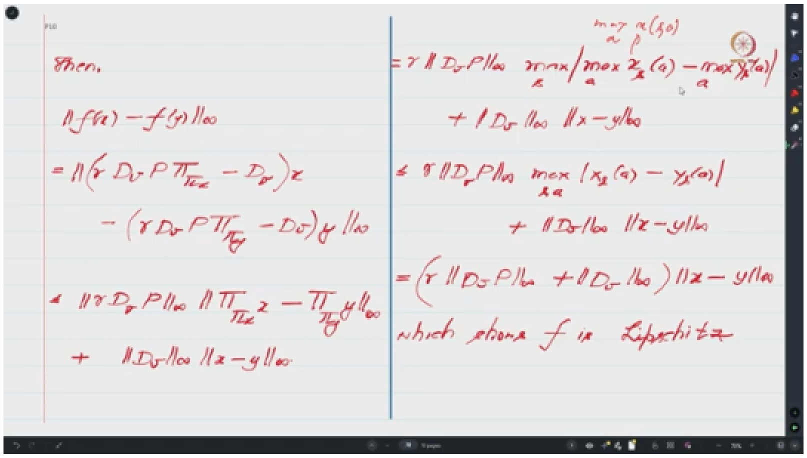
$= (\gamma \|D_\nu P\|_\infty + \|D_\nu\|_\infty)\|x - y\|_\infty$

which shows f is Lipschitz

one can see that f of x is basically this quantity, right? And f of y is basically this quantity, right? And one can, you know, sort of collect similar terms. So, we are going to collect this expression and this expression. And this expression is gamma d nu p common.

So, let us write it together and take the L-infinity norm. And what we will be left with is capital Pi little pi x times x. Minus capital Pi little pi y times y—that is what we have written—and this expression minus this expression will result in d nu infinity times x minus y infinity. So this expression is nice because this is what we would need in order to show the function is Lipschitz continuous. What is not clear here is whether this difference will also decay as x and y become closer to each other.

So, let us focus on this quantity and let us see what we can say, right? So, as I told you, this is an S A cross 1 vector and this is an S cross 1 vector. And since we are looking at the L-infinity norm, one can see that this expression is precisely max over S because of the L-infinity. And this thing is max over A X of S A, or alternatively this expression is S of S. x of s, a, and you take the max over a. So this is the expression, and similarly this quantity over here is max over a y of s, a. So this is exactly what we have over here.



P10

Then,

$$\|f(x) - f(y)\|_\infty$$

$$= \|(\gamma D_\nu P \Pi_{\pi_x} - D_\nu)x - (\gamma D_\nu P \Pi_{\pi_y} - D_\nu)y\|_\infty$$

$$\le \|\gamma D_\nu P\|_\infty \|\Pi_{\pi_x} x - \Pi_{\pi_y} y\|_\infty + \|D_\nu\|_\infty \|x - y\|_\infty$$

$$\max_{a} \alpha(s,a)$$

$$= \gamma \|D_\nu P\|_\infty \max_{s} \left|\max_{a} x_s(a) - \max_{a} y_s(a)\right| + \|D_\nu\|_\infty \|x - y\|_\infty$$

$$\le \gamma \|D_\nu P\|_\infty \max_{s,a} |x_s(a) - y_s(a)| + \|D_\nu\|_\infty \|x - y\|_\infty$$

$$= (\gamma \|D_\nu P\|_\infty + \|D_\nu\|_\infty)\|x - y\|_\infty$$

which shows f is Lipschitz

You can check the definition of this capital Pi little pi x and capital Pi little pi y to see that this is exactly this. And now, you know, one can—by making use of the max that is present over here—show that the max of the difference between max and max is upper bounded by the max over XSA of the absolute value of XSA minus YSA. So one can show that this quantity actually upper bounds this quantity, and one can then see that this expression is precisely X minus Y. L-infinity norm, and hence one can see that this whole expression is upper bounded by gamma times this quantity, which I have written here, and the L-infinity norm of the d nu matrix times x minus y. x minus y is L-infinity norm. So finally, one can see that the L-infinity norm of fx minus fy is upper bounded by some constant times the L-infinity norm of x minus y, which can be used to finally conclude that this f—

The function is Lipschitz continuous in the L-infinity norm, which is good enough to show that the solution trajectories of this ODE exist and are unique. This ODE is obtained, basically, by taking a solution trajectory of this and subtracting it with Q star. So, that relation can then be used to conclude that the solution trajectories of this ODE also exist and are unique. So, this brings us to the end of this class. So, what we have managed to show is that we discussed this vector comparison principle, then we looked at a sufficient condition for showing the global asymptotic convergence to the origin for a switched dynamical system. And finally, we want to use all these principles to say something about the limiting behavior of the limiting ODE associated with your Q-learning algorithm.

And in order to do that, the first question we asked was: Can we say something about the existence of solutions of this limiting ODE? For that, we showed that this f-function has the Lipschitz continuity property, right? And hence, the solution trajectories of the limiting ODE associated with your Q-learning algorithm exist. In the next class, what we will do is look at how we can use the vector comparison lemma and the switched system theory to say that the solution trajectories of the limiting ODE converge to Q star and use that to conclude that your Q-learning iterates will almost surely converge to Q star. Until then, goodbye, thank you, and namaste.