STOCHASTIC APPROXIMATION: THEORY AND APPLICATIONS

Dr. Gugan Thope

Department of Computer Science and Engineering

Indian Institute of Science, Bangalore

Week 11

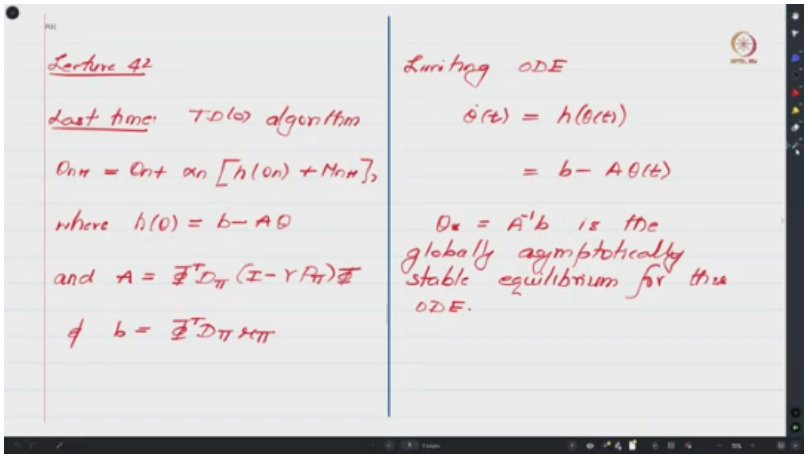
Lecture 42

Almost Sure Convergence Analysis of Temporal Difference Learning via the ODE Method

Hello and Namaste everyone. Welcome to lecture 42 of this NPTEL course on Stochastic Approximation. In this week and also in the next couple of lectures, we are looking at applications of Stochastic Approximation to reinforcement learning. In particular, we are trying to understand the asymptotic behavior of the so-called TD0 algorithm, which is used for estimating the value of a policy π in the function approximation setting. In the previous class, we looked at the limiting ODE associated with the TD0 algorithm and showed that, at least for the limiting ODE, there exists a globally asymptotically stable equilibrium, which we denoted as θ^*.

Then, our goal was to understand how good $\varphi\theta^*$ is. In particular, we wanted to compare the distance of $\varphi\theta^*$ to $V\pi$ with the distance of $\varphi\theta^{*\prime}$ to $V\pi$, where $\varphi\theta^{*\prime}$ is the projection of $V\pi$ onto the column space of φ. By this definition, one can immediately see that $\varphi\theta^{*\prime}$, or the projection of $V\pi$ onto φ, would have the smallest distance to $V\pi$, right? However, the TD0 algorithm—at least we have not yet guessed—would, I mean, at least we have guessed that the TD0 algorithm would not go to $\varphi\theta^{*\prime}$; instead, it will go to $\varphi\theta^*$.

So, right now, we are in the process of understanding how good $\varphi\theta^*$ is compared to $\varphi\theta^{*\prime}$. In the previous lecture, we derived some intermediate results. We will use those intermediate results to obtain the final bound that quantifies the distance of $\varphi\theta^*$ to $V\pi$ vis-à-vis the distance between $\varphi\theta^{*\prime}$ and $V\pi$. Let us begin the formal analysis. So, recall the TD0 algorithm has the update rule as given here, which is that $\theta_{n+1} = \theta_n + \alpha_n(H(\theta_n) + m_{n+1})$, and $H(\theta) = B - A\theta$, where the matrix A has the formula given here, and B has the expression given here.



So, for this algorithm, the limiting ODE can be guessed to be theta dot of t equals h theta of t, which, because of the nature of h of theta, will turn out to be b minus a theta of t. And from this, and also the fact that your matrix A is positive definite, we managed to conclude that theta star, equals A inverse B, is a globally asymptotically stable equilibrium with respect to the ODE that we have over here.

$$\theta_{n+1} = \theta_n + \alpha_n\left[h\left(O_n\right) + M_{n+1}\right]$$

$$h(\theta) = b - A\theta$$

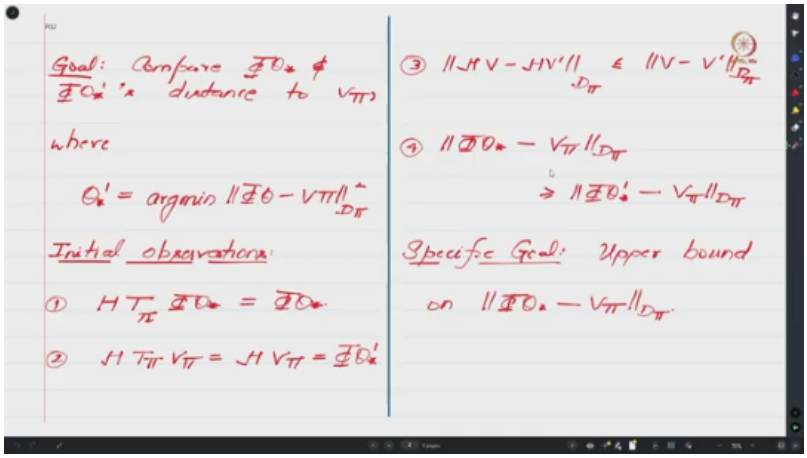
$$A = \Phi^T D_\Pi\left(I - YP_\Pi\right)\Phi$$

$$b = \Phi^T D_\Pi r_\Pi$$

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

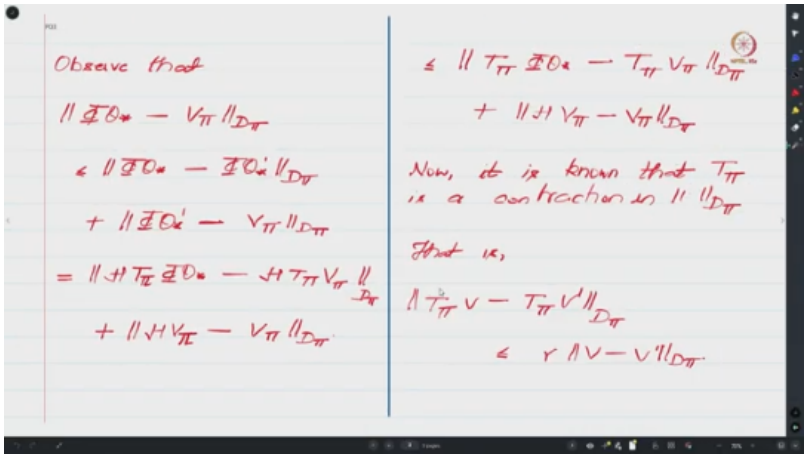
$$= b - A\theta(t)$$

This was something that we proved in the next class—and sorry, not in the next class but the previous class—and our goal in the previous class was to somehow compare the distance of phi theta star to that of phi theta star prime, and theta star prime is the theta value which minimizes the distance to V pi under this norm that is induced by this matrix D pi. And we had made some initial observations in the previous class.



In particular, we had shown that this phi theta star is special in that it is the fixed point of this projected Bellman operator, that is, the operator H T pi. And furthermore, one can see that H T pi V pi—one can show that V pi is actually a fixed point of the operator T pi, and hence T pi V pi equals V pi. And from there, one can see that H T pi V pi is actually H V pi, and H V pi is defined to be phi theta star prime. This is the way we had defined your theta star prime. And in the last lecture, we had also seen that if you take any arbitrary vectors v and v prime, look at their projections, and look at the distance between the projected quantities, then this distance is upper-bounded by the distance between the unprojected quantities, which is v and v prime.

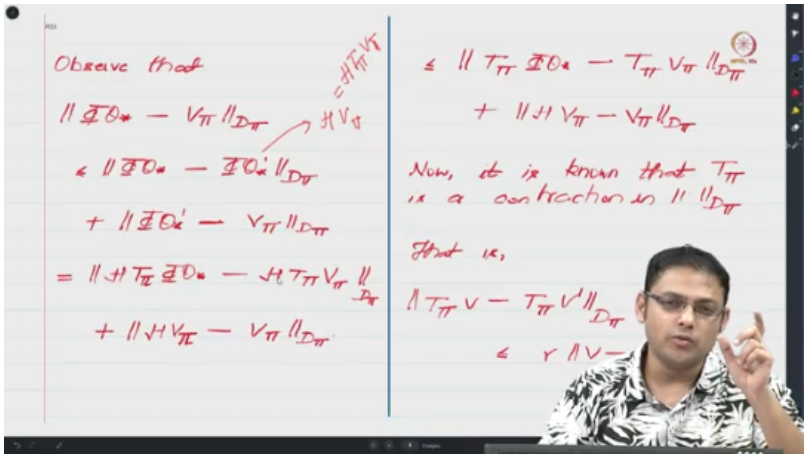
And separately, we had also seen in the previous class that phi theta star minus V pi—this distance will be lower-bounded by the distance between phi theta star prime and V pi, and this inequality holds because theta star prime is defined to be the theta that minimizes this quantity, and hence, by this very definition, one can see that if you substitute any other vector other than theta star prime, the distance of phi theta star to V pi would be larger than that of phi theta star prime to V pi. So, our goal now is to obtain an upper bound on phi theta star—phi theta star's distance to V pi, right? So, let us begin that computation. So, I hope you agree that, by the triangle inequality, phi theta star pi



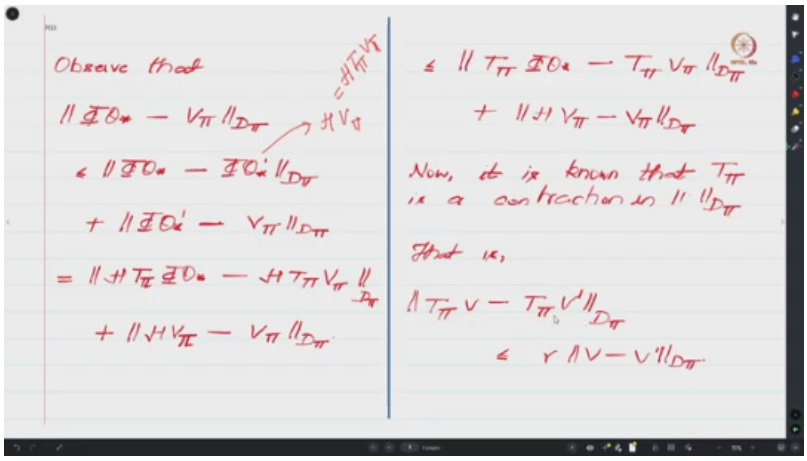
The distance to V pi can be upper bounded by the distance between phi theta star and phi theta star prime and phi theta star prime's distance to V pi. So, this just follows from the triangle inequality, right? And now we know that phi theta star is the fixed point of the projected Bellman operator, and hence wherever you have phi theta star, you can replace it by H T pi phi theta star. And we also know that phi theta star prime is actually H T pi. So, this quantity is actually H V pi.

And, you know, from the theory of reinforcement learning, we know that V pi is T pi V pi. In other words, V pi is the fixed point of your Bellman operator T pi. Hence, this quantity equals H T pi V pi. So, this is exactly what we have written over here. H T pi V pi in place of phi theta star prime, and wherever you have this phi theta star prime in the next expression, we will actually replace it with H V pi.

You will soon see why I am doing it in this fashion. And here one can see that you have H times some vector minus H times another vector. And if you recall, we had shown that the distance between the projected quantities is always upper bounded by the distance between the unprojected quantities. So by using that inequality, one can see that this expression is upper bounded by this expression over here. Is this okay? And this expression, we write it as it is. Now, you know from the theory of reinforcement learning, it is also known that your T pi operator is actually a contraction.

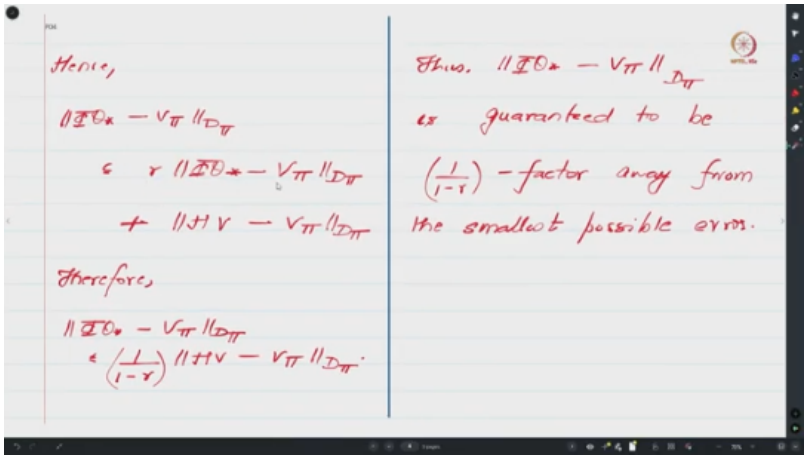


So contraction in this norm. So by contraction in this norm what it means is that if you take any two vectors V and V prime and look at their outputs under T pi. That is T pi V and T pi V prime and look at the distance between the outputs under this metric induced by this matrix D pi. Then the contraction over here means that this quantity is upper bounded by gamma times the distance between V and V prime. So, because of the presence of gamma which is the discount factor that is present in your you know MDP definition right and if you recall when I was discussing about this value of gamma I had told you that this value of gamma will be strictly less than 1.



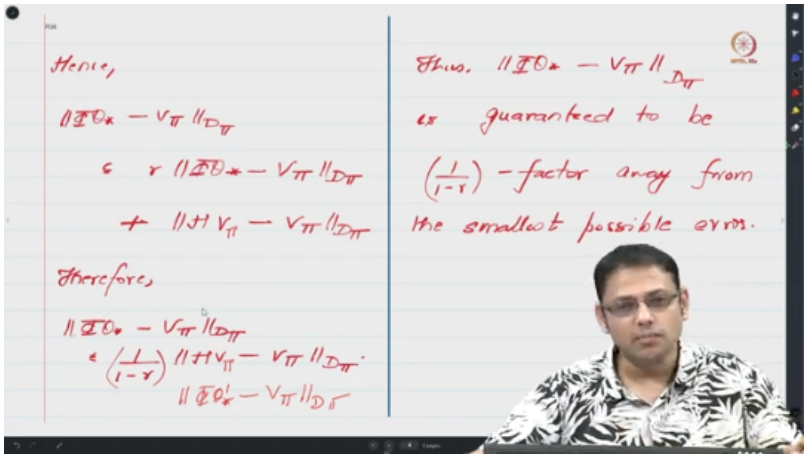
So, because it is strictly less than 1, what this inequality tells us is that the distance of the outputs of V and V prime under this T pi operator will be strictly less than the distance between the inputs which is V and V prime. So, this expression is true from the theory of reinforcement learning. Now, we can apply this inequality and show that this quantity

over here is upper bounded by gamma times the distance between phi theta star and V pi. right? So, this is what we end up with.



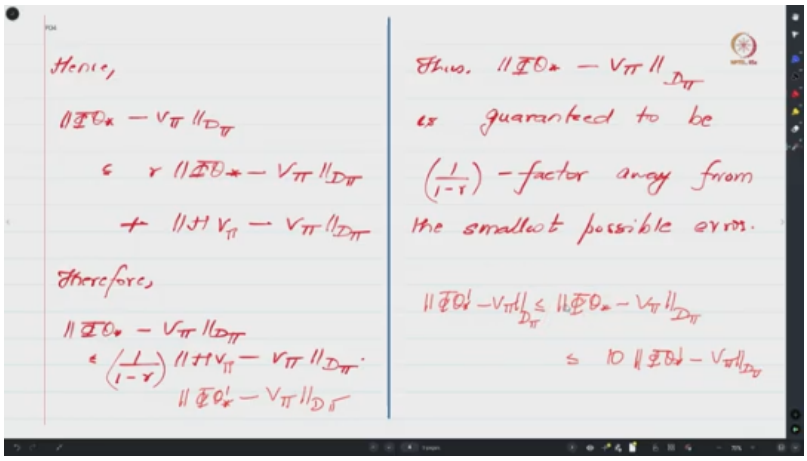
Now, the nice thing about this quantity is that this quantity is exactly what we have over on the left hand side. Hence, we can take this quantity to the left hand side. So, you will have a 1 minus gamma and if you take the 1 minus gamma on the right hand side, we will eventually end up with an inequality of the following form which is that phi theta star's distance to V pi is less than 1 over 1 minus gamma times distance of HV to HV pi to V pi. So, I should say this is V pi here and this also is V pi.

Let me just make sure if I have V pi everywhere. Yes. So, I mean, this is also equivalent to your So, HV is precisely phi theta star prime minus V pi d pi. So, what this means is that the distance of phi theta star to V pi in this norm is at most 1 over 1 minus gamma factor away from the smallest possible error that one can incur by trying to approximate V pi within the column space of phi. So, this is the inequality that we will end up with, and this was the inequality that we were chasing, and we have finally managed to show.

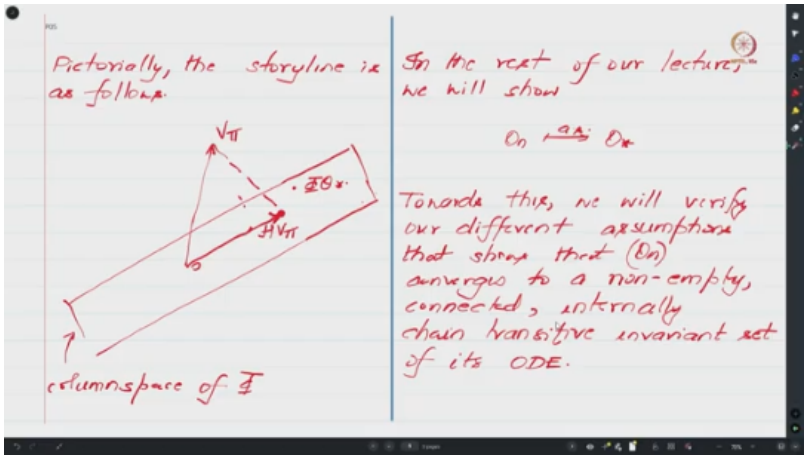


So, what this means is that if your gamma is, you know, let us say 0.9, right, then 1 minus gamma will be 0.1, and 1 over 1 minus gamma will be 10. So, what this means is that your phi theta star prime minus V pi, right. will be less than or equal to the distance of phi theta star minus V pi, and all of this should be under this norm. This will be less than 1 over 1 minus gamma, where you know, if you substitute gamma equals 0.9, you will end up with 10 times phi theta star minus prime minus V pi d pi, right?

So, this distance is the smallest distance that we can get, and theta star is a potential place to which your PD0 algorithm will converge. I say potential because we have not yet established that, right? But what this result is trying to tell us is that even if we manage to find phi theta star, the distance, or how bad this quantity will be, will be at most 1 over 1 minus gamma away from the best error that we can obtain. So, this is what we have managed to show.

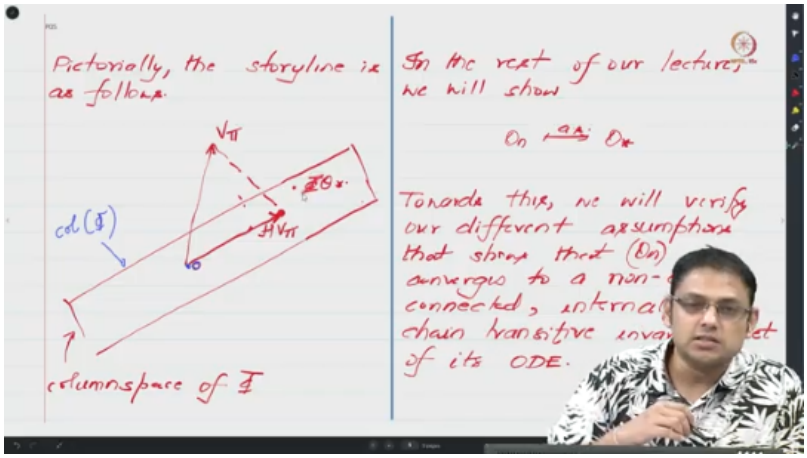


And pictorially, the story that we have managed to show show so far can be seen as follows, right? So, here is the column space of phi. So, this is your column space of phi, right? And here is the origin sitting.

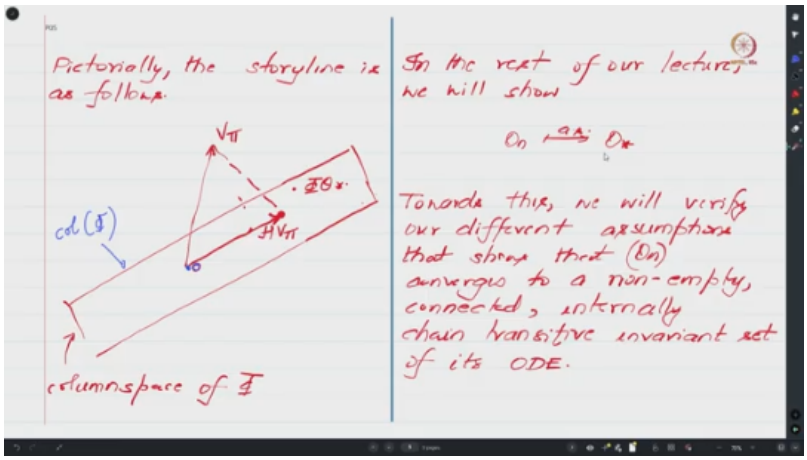


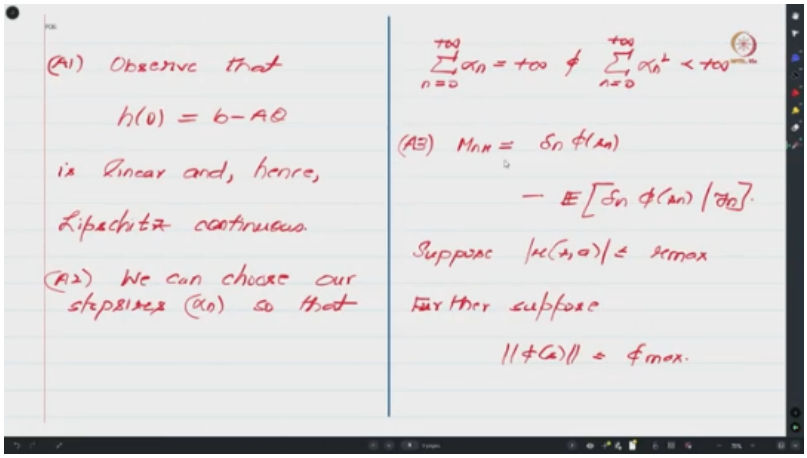
So, this is your origin, right? And somewhere in this column space of phi, you will have the projection of h v pi. So, v pi is the vector whose approximation you want to find in the column space of V and H V pi is that best approximation to V pi in the column space of V. Ideally, we would have liked to go over here, but the SGD algorithm that we tried working with couldn't be implemented because we don't know V pi, right.

Hence, we, you know, did some approximation to that algorithm, and we ended up with the TD0 algorithm. And the TD0 algorithm, we now, you know, conjecture that it will converge to phi theta star. And what we have shown is that this distance of phi theta star to V pi will be at most 1 over 1 minus gamma factor of the distance between V pi and H pi. So, this is the pictorial interpretation of what we have shown so far. And in the rest of this lecture, what we are going to show is that your theta n, that is the iterates of your stochastic approximation TD0 algorithm, would almost surely converge to theta star. And in order to show this result, we are going to make use of the different assumptions that we have studied in the first few weeks of this course.



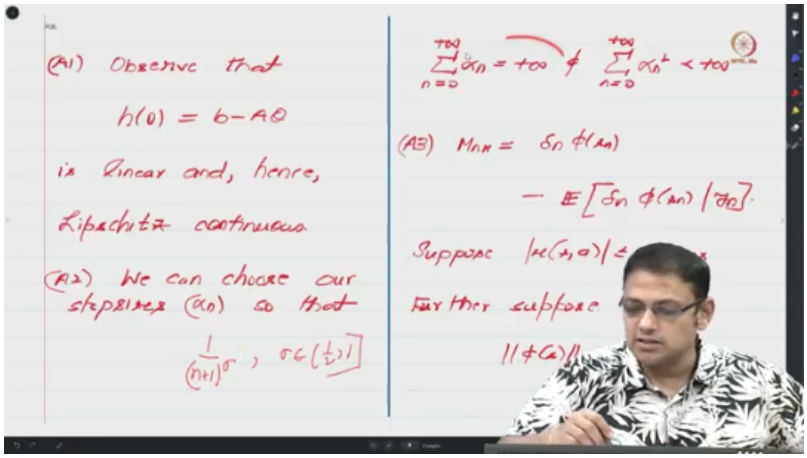
And if you recall, if we verify certain assumptions, we would be able to conclude that these stochastic approximation iterates converge to a non-empty, connected, internally chain transitive compact set of its limiting ODE. So, we are going to verify this assumption and then, you know, use this conclusion to show that theta n indeed converges to theta star almost surely. So, let us verify the different assumptions, and I hope you are able to remember, you know, the symbols for the different assumptions. The first assumption was A1, which required us to show that the driving function was Lipschitz continuous. In the TD0 case with linear function approximation, H theta turns out to be B minus A theta, which means that your H function is linear in nature.





And because it is linear in nature, it is trivially Lipschitz continuous. One can, you know, show that it is very, very easy to prove. So, hence, this assumption A1 holds true in our context. Now, the next thing that we need to verify is that your step size sequences satisfy some nice properties, which is that they should not be summable but should be square summable, right? And, you know, to ensure this condition, we will basically pick a step size choice that satisfies those conditions. And as I have told you, you know, particular examples of such step sizes include 1 over n plus 1 to the power sigma, where sigma is some number, let us say between half and one, and one is close.

So, we can pick one as well. So, if you pick a step size like this, these conditions can be ensured, and because of that reason, your assumption A2 can also be trivially ensured. Now, with regards to assumption A3, we want to first show that your MN sequence is a square integrable martingale difference sequence. Furthermore, it satisfies certain linear growth conditions.



So, those conditions we will now verify. So, recall that Mn plus 1 is basically delta n times phi of Sn minus these conditional expectation of delta n times phi Sn with respect to the information that we have until time n. And what we will do is we will presume that the rewards are upper bounded by some maximum quantity that is R max. Furthermore, we are presuming that the length of your feature vectors, they are also upper bounded by phi max. So, let us make such assumptions like this and let us presume that R max and phi max are less than infinity.

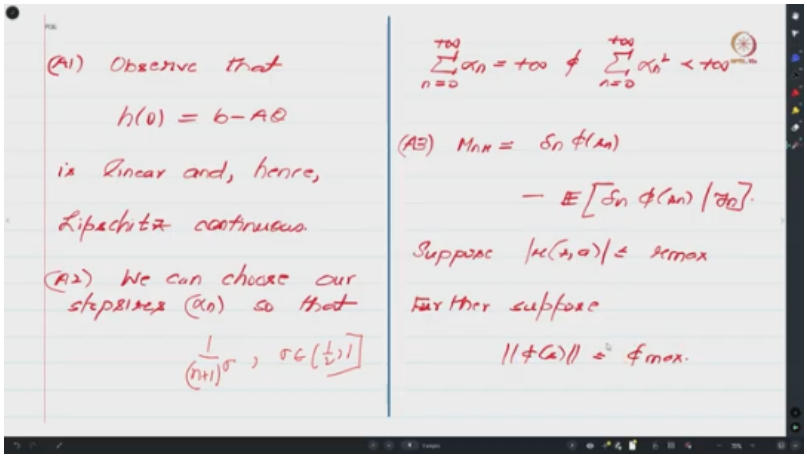
$$h(\theta) = b - A\theta$$

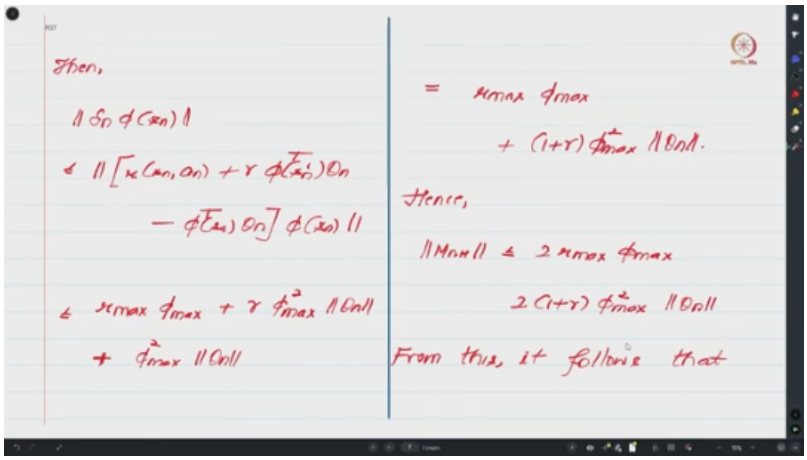
$$\sum_{n=0}^{+\infty} \alpha_n = +\infty$$

$$\sum_{n=0}^{+\infty} \alpha_n^2 < +\infty$$

$$M_{n+1} = \delta_n \phi(s_n) - E[\delta_n \phi(s_n) | F_n]$$

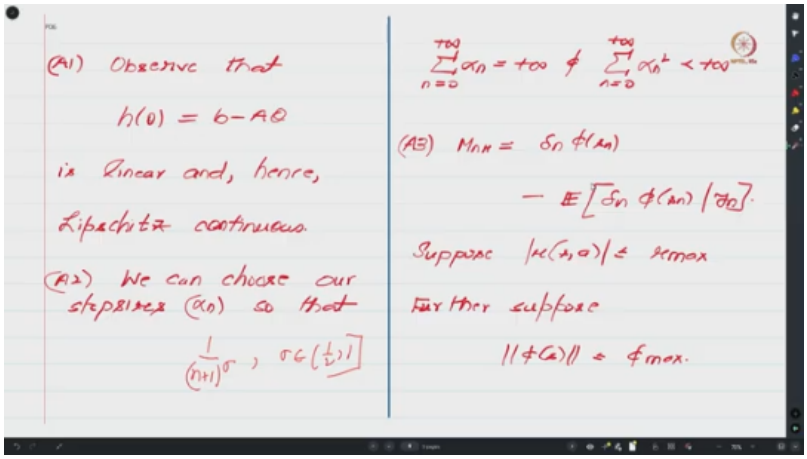
So, in which case one can show that if you take the norm of delta n times phi of Sn.

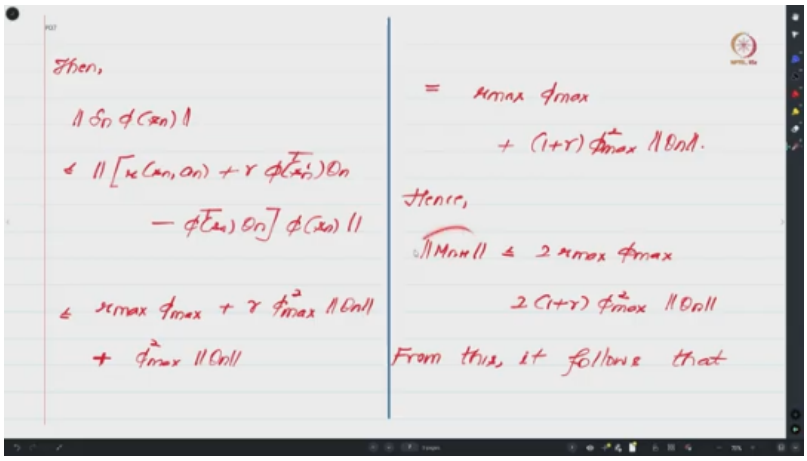




So, recall that delta n is this expression within the square bracket. So, if you you know take the norms one can see that this quantity is upper bounded by R max in particular the absolute value is upper bounded by R max and if you multiply this thing with phi SN this norm is upper bounded by phi max and hence you would end up with R max times phi max. And similarly this quantity and this quantity when you multiply with each other and you know make use of let us say something like the Cauchy-Schwarz inequality then one can see that this norm is upper bounded by 1 phi max this norm is upper bounded by another phi max and hence you end up with a phi max square times the norm of this vector theta n. And finally you know this expression similarly one can show is upper bounded by phi max square times norm of theta n. So if you put all of them together one can see that this norm is upper bounded by r max phi max plus 1 plus gamma times phi max square of norm theta n. And hence, one can conclude that.

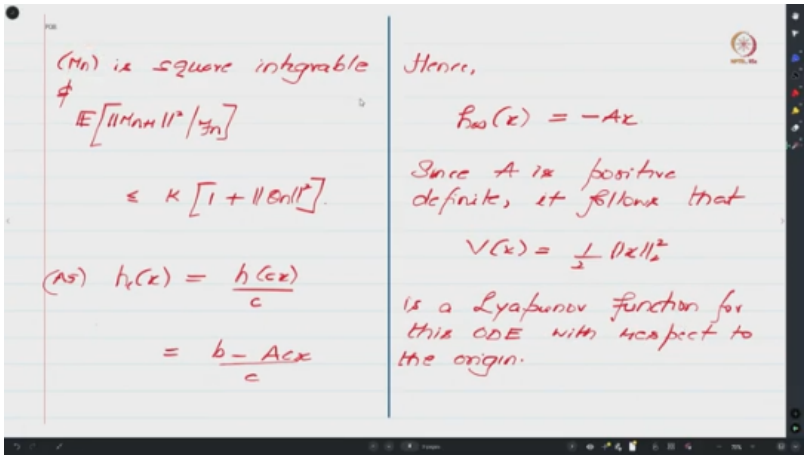
So, if you recall Mn plus 1 is this expression plus this expression. So, if you take the norm of this, it will be the norm of this plus the norm of this thing and one can show that the norm of a conditional expectation is upper bounded by the expectation of the norm. or the conditional expectation of the norm and from that one can show that you know your norm Mn plus 1 is basically upper bounded by twice this quantity which is that 2 times R max times phi max plus. So, here there should be a plus 2 times 1 plus gamma phi max square norm of theta n. So, one can trivially show this expression. And from this, it follows that your Mn is actually square integrable.





So, you are, you know, at any given time instance, you are basically adding terms with finite norms, right? And hence, at any given time instance n, your norm theta n is going to be bounded. And from that fact, one can show that Mn plus 1 is actually bounded. Of course, the bound can keep changing with n. But for every n, we have some bound. And

from that, one can see that because your Mn plus 1 is bounded, its expectation will also be bounded, and hence one can conclude that each Mn is square integrable.



Furthermore, you know, it is based on that inequality that we saw on the previous slide; one can also immediately see that if you take the square of this expression and take the conditional expectation, then one can indeed find the K such that this conditional expectation is upper bounded by K times 1 plus norm theta n square. So, this tells us that your martingale noise sequence indeed satisfies the linear growth condition in terms of norm theta n square. And so, this sort of verifies the first three assumptions. Now, what we need to verify is that the iterates are almost surely bounded.

And towards that, if you remember, we had looked at a sufficient condition which involves looking at the scaled ODE and showing that the origin is the globally asymptotically stable equilibrium with respect to the scaled ODE. So, towards that, we had defined this hc function, which was h of cx by c, where c is any scalar bigger than or equal to 1. And if you substitute, you know, this CX in the definition of H, right, one can see that we would end up with B minus ACX, and this C will cancel off with this C. So, one can conclude that this HC of X will basically be B over C minus AX. And if you take C to infinity, one can now see that H infinity of X will indeed be minus AX.

Now, we have already shown that this matrix A is positive definite. So, from this fact, one can verify that if you define a function V in this fashion—that is, V of X equals half of the norm of X squared—then one can show that. So, this function is actually a limiting ODE with respect to, you know, x dot of t equals H infinity of x of t. So, sorry, I should

have been using theta here. So, let me write it in terms of theta. So, theta dot of t is equal to H infinity of theta of t. So, if you define your V of theta to be half

of the Euclidean norm squared, then one can see that this expression is the Lyapunov function with respect to this ODE with respect to the origin, which means that all solution trajectories of this ODE are guaranteed to converge to the origin. In other words, because of the existence of a Lyapunov function, one can show that the origin is indeed the globally asymptotically stable equilibrium with respect to this limiting ODE, as desired. So, this verifies your assumption A phi as well, and now, if you—so this sort of verifies all the assumptions—and hence one can conclude that indeed the iterates of your TD0 algorithm will converge to some non-empty, connected, compact, internally chain-transitive invariant set of the limiting ODE theta dot of t equals H theta of t.

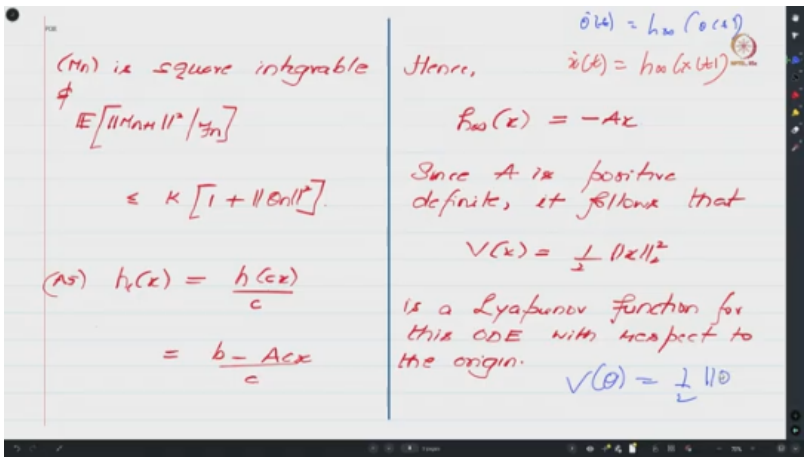
$$h_c(x) = \frac{h(cx)}{c}$$

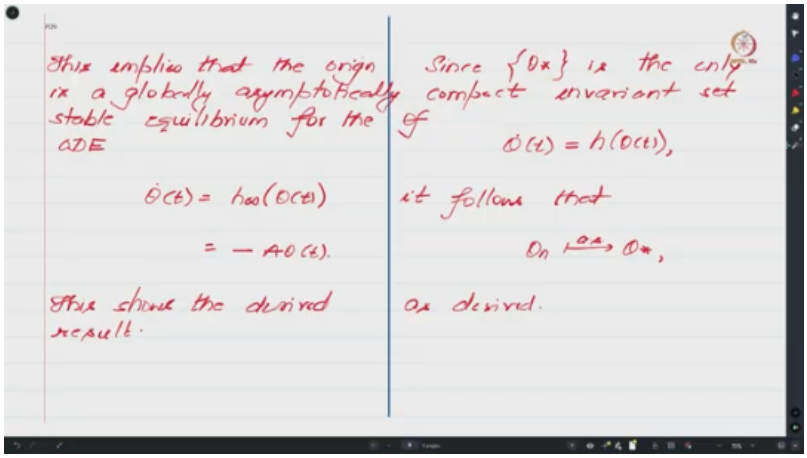
$$= \frac{b - Acx}{C}$$

$$h_\infty(x) = -Ax$$

$$V(x) = \frac{1}{2}||x||_2^2$$

But we have already shown that the only compact invariant set of this limiting ODE is actually the singleton set containing theta star. That is the only compact invariant set. Of course, RD is also an invariant set trivially, but that RD set is not compact.





So, the only other invariant set—because theta star is globally asymptotically stable with respect to this limiting ODE, right—is basically the singleton set containing theta star. So, from this fact and the fact that, because these four assumptions we have verified, so your theta n has to converge to one such invariant set, and there is only one invariant set that satisfies all this criterion, one can conclude that theta n must almost surely converge to theta star. So, let us quickly summarize what we have done so far. We defined the TD0 algorithm, which was the algorithm that we wanted to use to come up with an approximation for V pi in the column space of phi, and what we have shown is that the stochastic algorithm itself converges to theta star, and if it converges to theta star, whatever is the limiting value that will be used as a proxy for V pi.

And separately, we have shown how good is phi theta star compared to phi theta star prime, which is the best approximation to V pi. So, one can see that our original goal was to find a good approximation to V pi. in the column space of phi right and we designed an algorithm to find this approximation and while we could not find the best approximation to V pi in the column space of phi we were able to come up with a reasonable approximation. And the nice thing about this algorithm is that it can take data that is obtained by interacting with the environment. What do I mean by that?

So, when you interact with the environment, you see that the environment is in a certain state, you take some action and then you get to see the next state and During this interaction, you may get some immediate reward, right? So, you can take these information and use this information to get an estimate of how good your strategy is,

right? So in the next week we will focus on generalization of this problem which is known as the control problem. So far we were looking at given a policy how good it is.

In particular we wanted to quantify its quality in terms of the value function. In the next class we will try to use similar principles to design an algorithm that can be used for finding the optimal strategy itself. I hope you will join me for the next week. Until then, goodbye and namaste.