STOCHASTIC APPROXIMATION: THEORY AND APPLICATIONS

Dr. Gugan Thope

Department of Computer Science and Engineering

Indian Institute of Science, Bangalore

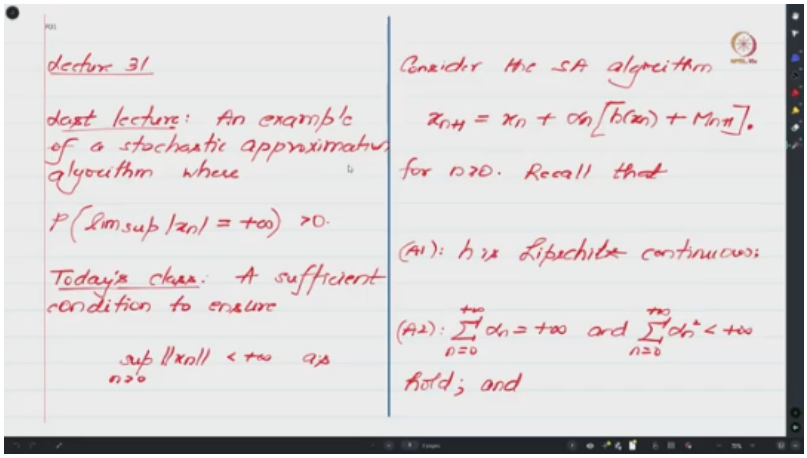
Week 8

Lecture 31

Almost Sure Boundedness of Iterates: Theorem and Example

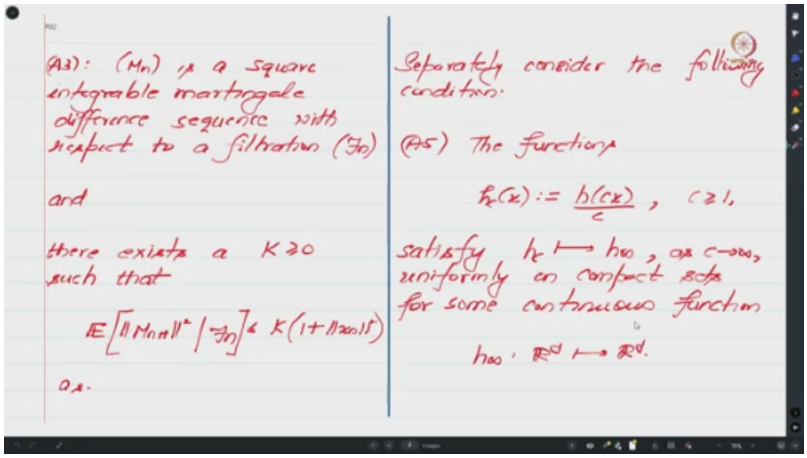
Hello and Namaste, everyone. Welcome to Lecture 31 of this NPTEL course on Stochastic Approximation. So, let us do a quick recap of what we have been doing this week. So, if you recall, in Weeks 5 and 6, we looked at almost sure convergence of stochastic approximation algorithms, and there was an assumption A4, which required that the iterates of your stochastic approximation algorithm be almost surely bounded. So, this week, we are focusing on that assumption.

And in the first class—that was the lecture before this—we looked at an example where, you know, the iterates have a positive probability of racing off to infinity, right? And hence, almost surely, the iterates are not bounded, which implies that for that example, the assumption A4 does not hold, right? And what we will do in today's class is come up with a sufficient condition under which the almost sure boundedness of your iterates is guaranteed to hold. So, with that, let us begin a formal discussion. So, in the last class, we looked at an example of a stochastic approximation algorithm which had this property: that is, with positive probability, the lim sup of the absolute value of Xn was plus infinity.



Now, in today's class, what we will do is look at a sufficient condition that ensures that almost surely, the iterates are bounded. So, toward describing the sufficient condition, let us set the stage. That is, let us suppose that we have a stochastic approximation algorithm of the form: Xn+1 equals Xn plus alpha_n times h of Xn plus Mn+1. And let us quickly recall the assumptions that we had imposed during our discussions from Week 5 and Week 6. We required that this function h be Lipschitz continuous, and this is what we denoted as Assumption A1.

Separately, we required that the step sizes satisfy the following two properties. On the one hand, we required that the step sizes not be absolutely summable. On the other hand, we required that the step sizes be square summable. So, one can ensure this assumption by an appropriate choice of step size, and we also require that this Mn be square integrable—that is, its second moment should be finite—and that Mn should be a square integrable martingale difference sequence with respect to some filtration Fn. Furthermore, we require the existence of a constant k greater than or equal to 0 such that the conditional expectation of the square of this norm is upper-bounded by k times 1 plus the norm xn squared.



Now, in our discussion during weeks 5 and 6, we required a separate assumption, which we denoted by A4, under which we presumed that the iterates of your stochastic approximation algorithm are almost surely bounded. In today's discussion, we are going to drop that assumption. Instead, we are going to presume this assumption, which we denote as A5, following similar notations from chapter 3 of this stochastic approximation textbook by Professor Vivek Borkar. So, what does this condition say? So, firstly, it defines this scaled function hc of x. So, what is this scaled function?

So, hc of x is defined to be h of cx divided by c for all c greater than or equal to 1. So, let c be a scalar and define hc of x as the value of h evaluated at c times x, divided So, this is like scaling the input, and whatever the output is, you divide it by c, right? So, if H is from RD to RD, then X will be a D-dimensional input. C is a scalar.

So, C times X is again a d-dimensional input. So, C times X will be like a scaling of the vector X, right? And you evaluate H at that scaled input and whatever is the output that which will be a d-dimensional vector, you divide each coordinate of that d-dimensional vector by C. And notice that we presume that C is bigger than equal to 1 and hence, you know, the division by 0 and all those concerns, right? are not valid here, right? So, this condition requires that

this you know set of functions right has the property that as your C goes to infinity this sequence of functions actually you know converges to another function which we denote it as H subscript infinity and we require that this convergence hold uniformly on compact

sets I will define what it means. And we require that this H infinity be a continuous function which again takes as input something in Rd and spits out something in Rd.

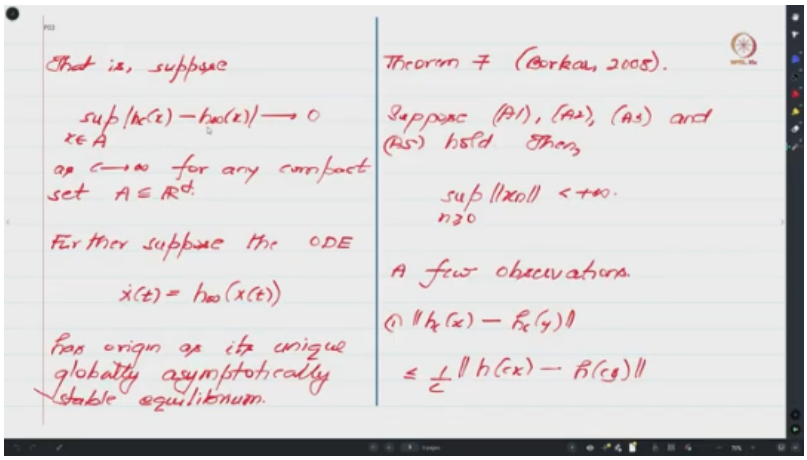
$$E\left[\left|\left|M_{n+1}\right|\right|^2|_{F_n}\right] \leq K\left(1 + \left|\left|x_n\right|\right|^2\right)$$

$$h_c(x) := \frac{h(cx)}{c},\ c \geq 1$$

$$h_c \to h_\infty$$

$$h_\infty : R^d \mapsto R^d$$

So, what does this uniform convergence over compact sets mean? Not only do we require that hc of x go to h infinity of x for all x, we also require that if you look at this distance between hc of x and h infinity of x and we take the supremum over x belonging to some compact set A. We require that this supremum also go to 0.



So in some sense, this is like a stronger notion of convergence of functions, right? So of course, if you have convergence which has this stronger property, it also implies pointwise convergence. However, you know, just because you have pointwise convergence, it does not imply that you have uniform convergence over compact sets, right? However, in the textbook it is also discussed that you know because of this Lipschitz continuity and other things in this particular case you know this uniform convergence over compact sets can also be verified using point wise convergence but for

that I request you to you know look at the textbook. At this point we will presume that this stronger notion actually holds for your hc of x functions

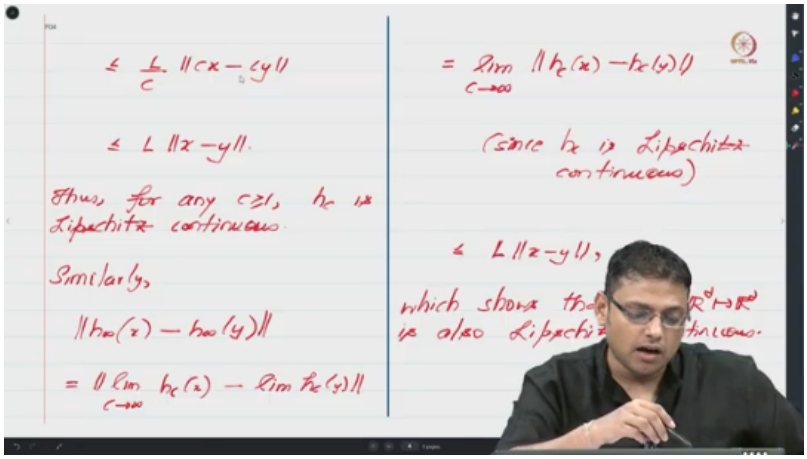
And this condition separately requires that whatever is your H infinity function, if you look at the ODE whose dynamics is governed by H infinity, this ODE has origin as its unique function. globally asymptotically stable equilibrium, right? So, if this, you know, ODE has origin as its globally asymptotically stable equilibrium, then what this theorem 7 says is that if these assumptions A1, A2, A3 and A5 hold, so notice that here the assumption A4 is missing, right? Instead, we have replaced this assumption with A5. So, this result says that suppose a1, a2, a3 and a5 hold, then almost surely your iterates are bounded.

Is this okay? Now, I would like to add a caveat here. This assumption A5 in general stochastic approximation algorithms will not hold. So, there will be some special situations where this A5 assumption can hold. In those situations, one can guarantee that the iterates are almost surely bounded.

But in my experience, I have observed that, you know, for many algorithms, these conditions do not hold and one has to, you know, separately figure out a way to show almost sure boundedness of your iterates, right? But whenever this assumption A5 holds, then of course, this theorem tells us that, you know, the iterates will be almost surely bounded. So, proving this result will be beyond the scope of our course and hence we will not try to prove this thing. Instead, we will try to verify when the assumption A5 can hold. So, we will verify that and before we verify, let us make a few observations.

So, recall that assumption A1 said that this function h is Lipschitz continuous. So, keeping that in mind, let us ask ourselves if this scaled function hc will also be Lipschitz continuous or not. Towards that, if you look at the difference between hc of x and hc of y, from the definition we know that hc of x is h of cx. And hc of y is h of cy, and this is the difference between the two. You take the norm between them and let us divide it by c. This expression, I hope you agree, is equal to—in fact, it is not less than or equal to, but equal to—from the definition, 1 over c times the distance of h of cx from h of cy, right? Now, we know that this function h is Lipschitz continuous, and hence whatever

difference I had written can be upper bounded by L times CX minus CY. From here, we can take C out as common, and that will cancel off with C because the norm of a scalar multiple of a vector equals the absolute value of that scalar times the norm of the original vector.

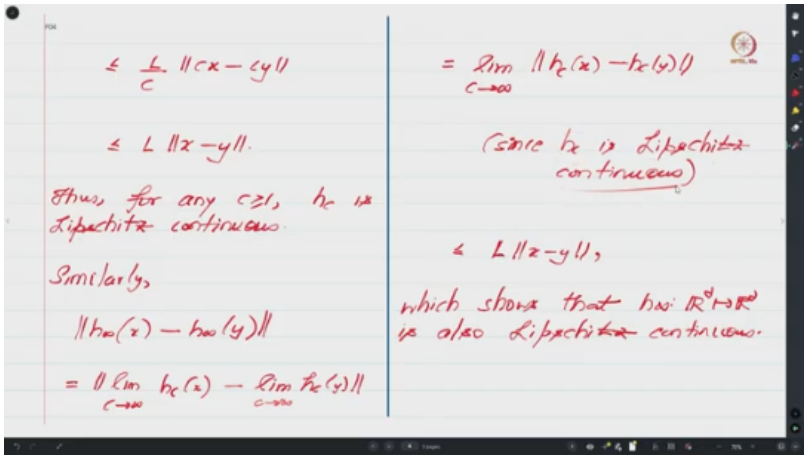


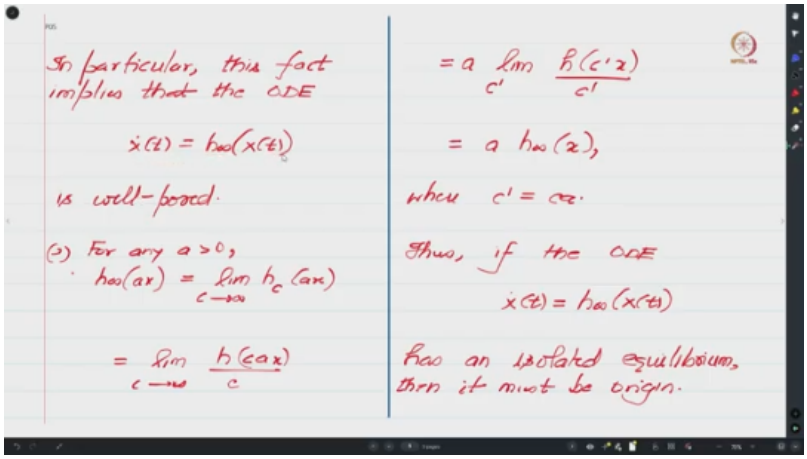
So, if I invoke that property over here, one can see that these C's will cancel off, and we will be left with L times the norm of XY. So, from this, one can conclude that for any c greater than or equal to 1, your hc function is also Lipschitz continuous. In fact, it is Lipschitz continuous with the same Lipschitz constant that governs the Lipschitz continuity of your original h function. So, that is one observation that one can easily make. Now, let us see if we can make use of this to say something about h infinity.

Now, recall that h infinity of x is the limit of hc of x as c goes to infinity. Hence, if you look at h infinity of x minus h infinity of y, then this distance is equal to the distance between the limit of hc's at x and the limit of hc evaluated at y. So, this is limit c tending to infinity. And because this norm is a continuous function, one can actually pull the limit outside this norm expression and say that the norm of the limits is equal to the limit of the norm. So this holds because of the continuous nature of the norm expression, and we know that this

The quantity is upper bounded by L times x minus y because we just showed that your hc function is Lipschitz continuous. Hence, this expression is upper bounded by x minus y, and this expression has no c. And since this expression is upper bounded by this for every C, one can conclude that in the limit as well, this expression should be upper bounded by

L times X minus Y, the norm of X minus Y, which shows that this function H infinity, which is defined from RD to RD, is also Lipschitz continuous. And this immediately implies that if you look at this limiting ODE, which was considered in assumption A5, then this limiting ODE is well-posed. Meaning, for any initial condition, first of all, there is a solution, and that solution trajectory is unique.





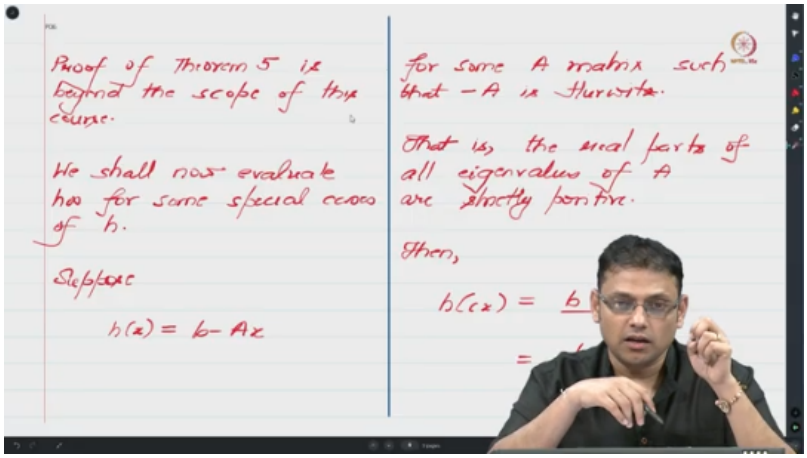
And so this is something that one needs to ensure before one can talk about asymptotically stable equilibrium and so on and so forth. And the second observation that we can make is for any a bigger than or equal to 0, that is, for any positive scalar, if you look at h infinity of ax, right, one can see that this equals, from the definition of h infinity, the limit as c goes to infinity of hc of ax. Right. And this expression, one can write it as H of C AX by C. Right. And since your limit C is going to infinity and A is greater than 0, what one can do is one can multiply and divide by A here.

Right. So one can write A here. Divide by A, recall that A is strictly positive, and hence one can do that right, and let's call this quantity as C prime right, so this we replace it by C prime, and this expression also we replace by C prime, and since C goes to infinity and A is positive, C times A, which is C prime now, will also go to infinity. Hence, this expression, one can write it as equal to A times the limit as C prime goes to infinity of H of C prime X by C prime, right?

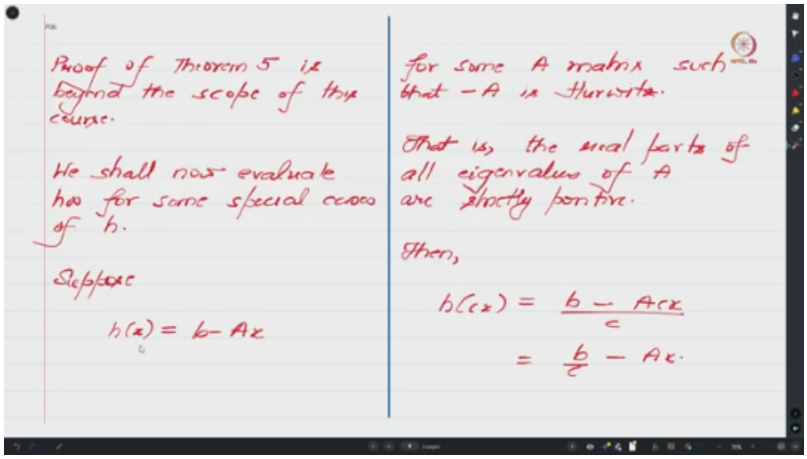
And if you look at this expression and, instead of thinking about C, if you look at it from the C prime perspective, well, since C prime goes to infinity, this expression is precisely H infinity of X. And whatever your 'a' was, you know, bring it down here, and now we can conclude that h infinity of ax is a times h infinity of x, which means that If x is an equilibrium point for this ODE—by that, I mean h infinity of x is 0—then for every a greater than or equal to 0, h infinity of ax is also 0. So, what this implies is that, You know, if you have an equilibrium point which is not the origin—that is, if the equilibrium point is not 0—then every scalar multiple of this equilibrium point will also be an equilibrium point.

In other words, if I can somehow say that, you know, this ODE x dot of t equals h infinity of x of t, has an isolated equilibrium point, then that isolated equilibrium point must be the origin. Is this okay? So, one actually makes use of these different facts to prove the theorem that I just stated, which is that if assumptions A1, A2, A3, and A5 hold, then the iterates are almost surely bounded. So, as I said, you know, verifying this theorem—or proving this theorem—is beyond the scope of this course, right?

And what we will instead do now is to check situations where this assumption A5 can be verified, right? So, towards that, let us consider the case where we have a stochastic approximation algorithm where the driving function has a form like this. And if you have been exposed to reinforcement learning, you can look at some algorithm called TD learning—temporal difference learning—for policy evaluation. And one can check that for that algorithm, the driving function indeed has a form of this kind. And let us presume that H of X is B minus AX for some matrix A such that the negative of this matrix is Hurwitz.



That is, the real parts of all eigenvalues of A are strictly positive. When we say minus A is Hurwitz, we require that all real parts of the eigenvalues of A be strictly positive. So, for such a driving function, let us look at the HC functions and, in particular, the H-infinity functions. So, if you look at H of CX, let us first understand what it is—I think there is a typo here. So, this is HC of X, this is H of CX divided by C, right?



So, this expression H of CX equals—wherever you have X, you replace it with C, right? And hence, H of CX would be B minus ACX, right? And you have to divide this whole expression by C, right? And if you rearrange things—because your C is greater than or equal to 1—this expression will lead to something like this. The C and this C will cancel out, and this part will lead to minus Ax. You can see that there is no C influencing this term here, so this is very important in that all the linear terms— the C influence will be canceled out. So, the X will be replaced by CX, and this whole expression will be divided

by C. The C's will cancel out, and hence, whatever this expression is, we will get it back as it was in the definition of H itself.

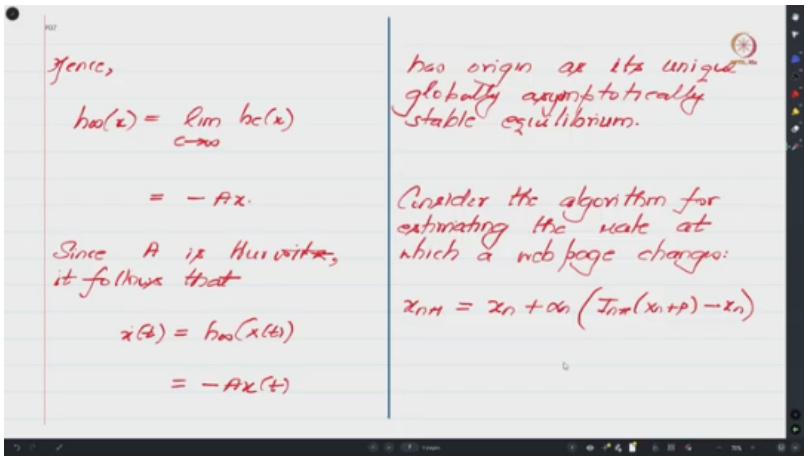
So, all linear terms will remain as they are in the definition of HC of X.

$$h(x) = b - Ax$$

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

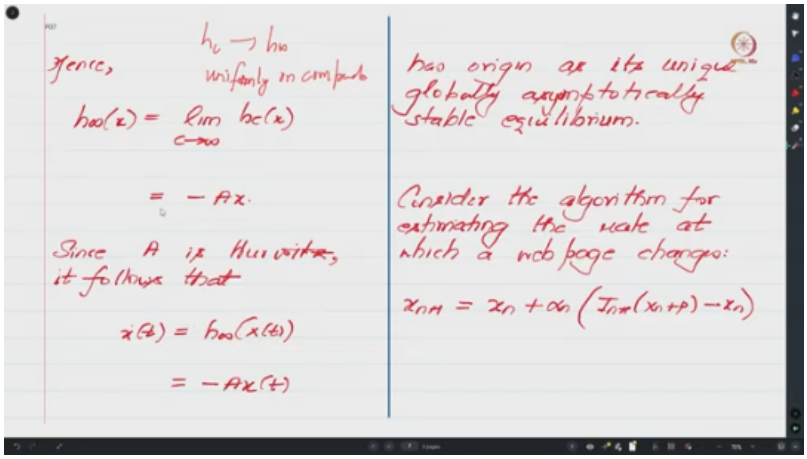
$$= \frac{b}{c} - Ax$$

Now, if you take C going to infinity—since B is some finite vector, right?—this term will be eliminated, and from that, one can guess that your H-infinity of X should basically be minus AX, right? And hence, in fact, one can show for this choice of HC and H-infinity that HC of X—or HC—converges to H-infinity uniformly, right? on compacts. One can actually show this very easily.

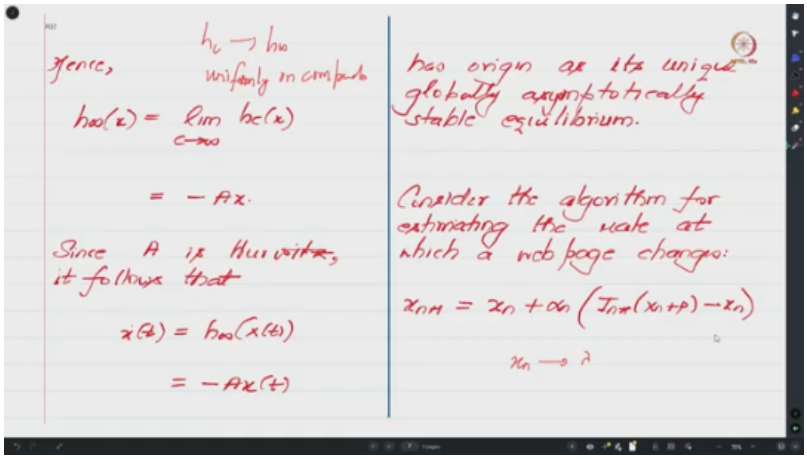


So, now if you look at this ODE, x dot of t equals h infinity of x of t, because h infinity we have guessed it to be minus a x, one can see that the h infinity related ODE will basically be x dot of t equals minus a times x of t. And since your minus a is actually a Hurwitz matrix, one can show that for this ODE, your origin is indeed a unique globally asymptotically stable equilibrium. Now, because we have verified assumption A5, one can actually show that for this particular example, A1, A2, A3 also hold. I mean, if you can come up with a stochastic approximation algorithm where A1, A2, A3 also hold. In addition to that, what we have verified shows that fi also holds. So, one can then show

that for that stochastic approximation algorithm, your iterates will be almost surely bounded, right?

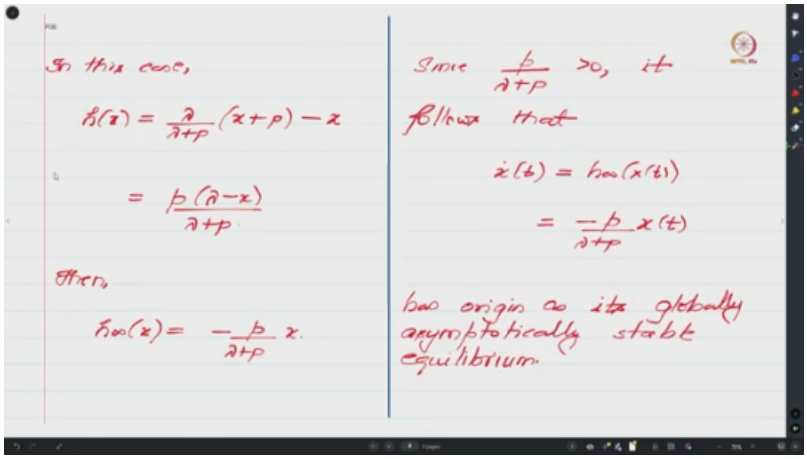


So now, what we will do is we will build upon this idea and actually look at this example that we have been considering multiple times throughout this course. So, in this algorithm, previously we had shown that this algorithm actually converges to lambda, which is the rate at which your web page changes and is unknown. However, in that discussion, we had presumed that the iterates are almost surely bounded. So, what we will do today is, for this algorithm, we will verify this assumption A5 and use that to conclude that the iterates are almost surely bounded, and hence this conclusion actually holds.

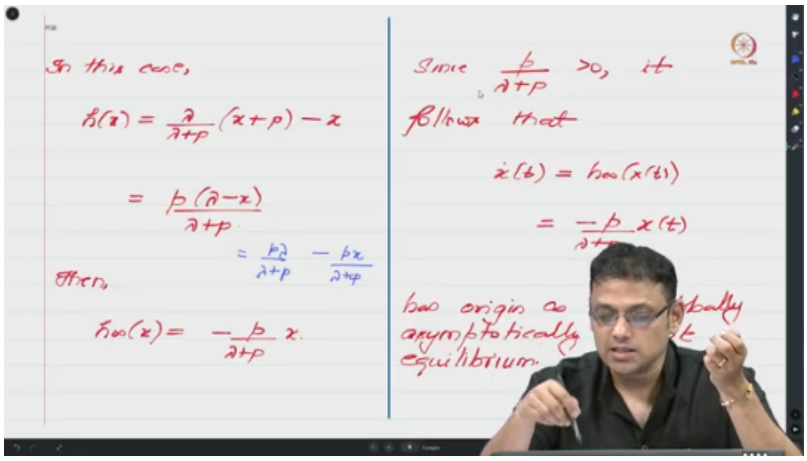


So, one can see that in this case, h of x—I mean, this is something that we have verified previously—has this form, right? And which can be rewritten in this way, right? So here

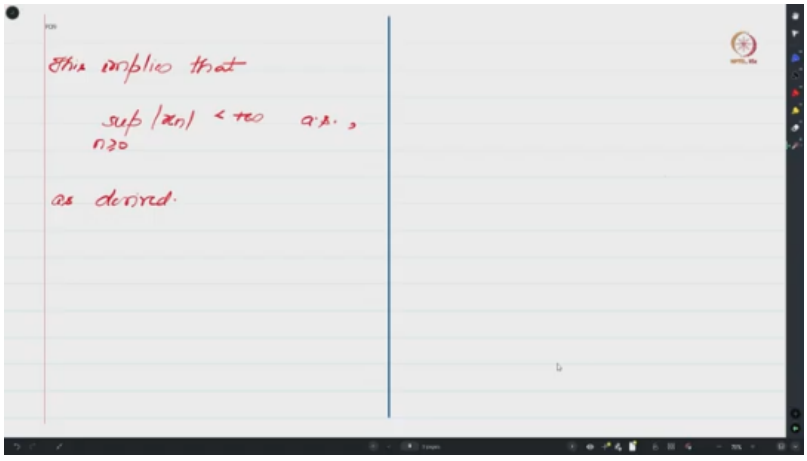
again, you can see that your h function has some constant, right? p lambda over lambda plus p minus some linear in x term, right? So this is the linear part, right?



So if you look at hc of x and take it to infinity, again this part which does not have x, when you divide it by c and take c to infinity, this term will get killed and we will be left with this. So accordingly one can see that in this particular case your H infinity of X will be minus P over lambda plus P of X, and recall that P was the known rate at which we visit the web pages and this lambda was the unknown rate at which the web page changes. And since p and lambda are both strictly positive, even though we do not know lambda, we know that this ratio is actually strictly positive. And because this ratio is strictly positive, one can see that the ODE, x dot of t equals h infinity of x of t, which in this case will be minus p over lambda plus p of x of t, can be shown to have origin as its unique globally asymptotically stable equilibrium. Is this okay?



And now from this new theorem that we discussed during this week, one can show that the supremum of the iterates' magnitude will actually be less than infinity. In other words, your iterates will be almost surely bounded. So, this brings us to the end of this class. Let us do a quick summary of what we have studied. So, in today's class, we discussed a sufficient condition under which your iterates can be shown to be almost surely bounded.



Then we looked at two specific cases of H function. One was the case where H was B minus AX, and then we showed that if this matrix A has the property that minus A is Hurwitz, then indeed if you look at this H infinity ODE, which we refer to as the scaled ODE, right, that ODE indeed will have origin as its globally asymptotically stable equilibrium. And in a similar spirit, we then looked at this, you know, often studied algorithm as part of this course, which is the estimation of this page change rate, web page change rate. And for that algorithm also, we observed that the driving function H is linear in X, and hence whatever ideas we had previously used in the context of B minus AX.

The same set of ideas can be used to show that the scaled ODE in this case, as well, has the origin as its unique globally asymptotically stable equilibrium. So, in some sense, this is the aspect of stability that we will be studying this week. Next week, we will look at some additional concepts. Until then, thank you, and Namaste.