STOCHASTIC APPROXIMATION: THEORY AND APPLICATIONS

Dr. Gugan Thoppe

Department of Computer Science and Automation

Indian Institute of Science

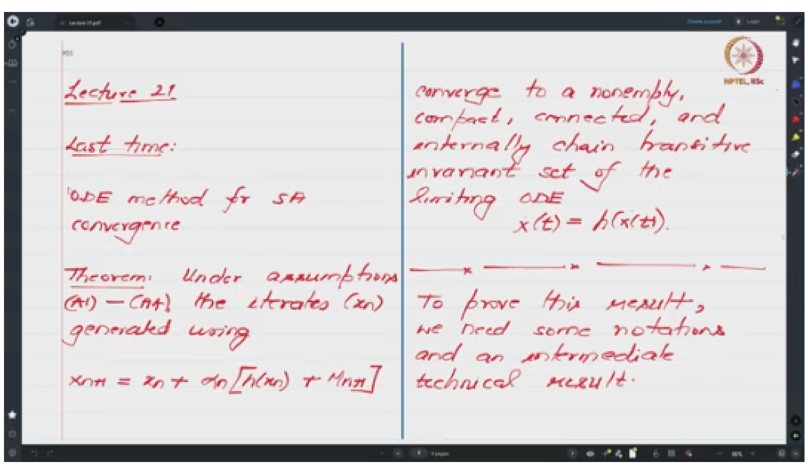
Week 5

Lecture 21

Almost Sure Convergence of Stochastic Approximation Iterates

Hello and Namaste, everyone. Welcome to Lecture 21 of this NPTEL course on Stochastic Approximation. So, we are in Week 5. So, let us do a quick recap of what we did in the previous class. If you remember, we started putting all the basic tools that we have studied, including martingales and ODE theory, right, and tried to analyze the asymptotic behavior of stochastic approximation algorithms.

Towards that, we stated one result and left the proof for future lectures, and we will prove a part of that in today's class. So, let us do a formal recall. So, we wanted to look at what is referred to as the ODE method for stochastic approximation convergence, and this was the theorem that we stated in the previous class. The theorem said that under assumptions A1 to A4, the iterates Xn generated by this generic stochastic approximation algorithm converge to a non-empty, compact, connected, and internally chain transitive invariant set of the limiting ODE $\dot{x}(t) = h(x(t))$, right?

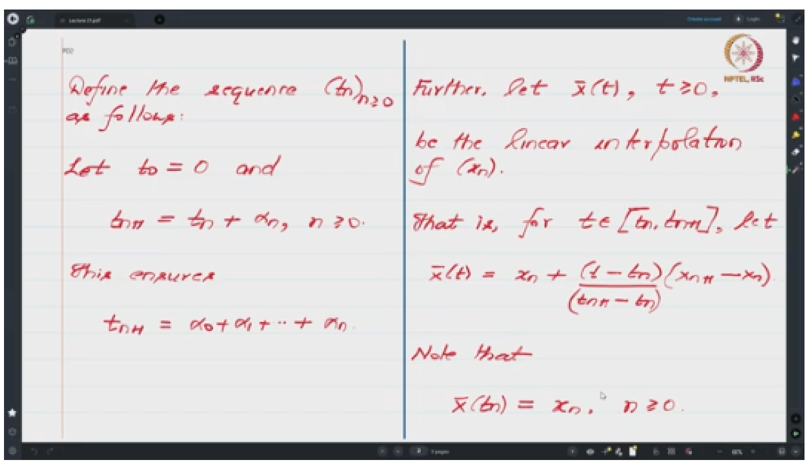


$$X_{n+1} = x_n + \alpha_n\left[h\left(x_n\right) + M_{n+1}\right]$$

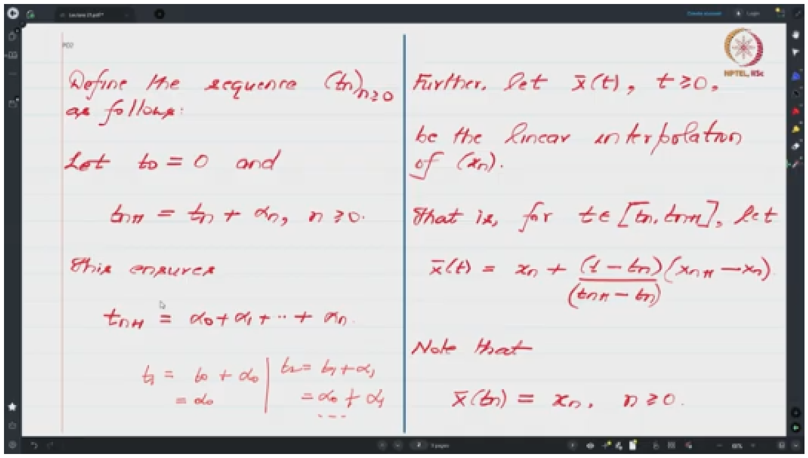
$$\dot{X}(t) = h(X(t))$$

So, again, I would like to highlight that the stochastic approximation algorithm, by its nature, is stochastic—by that, I mean if you run this algorithm on different runs, you will see different iterate sequences, right? On the other hand, this limiting ODE is deterministic in nature. So, its invariant sets will not change on different runs of the experiment. So, what we are saying here is that on different runs of the experiment, the algorithm will converge, and the place where it converges to has some nature that can be described in terms of the invariant sets of the limiting ODE.

So, today we will prove a part of this result, and to prove this part, we need to introduce some notations and an intermediate technical result. So, the first of these notations is that of this time sequence Tn. And the way we will define this is as follows. We will set T0 equal to 0, and for every n greater than or equal to 0, we will set Tn plus 1 equal to Tn plus alpha n. So, this means that your T1 will be T0 plus alpha n. And since T0 is 0, this is basically alpha 0.



Now, your T2 similarly will be T1 plus alpha 1, which is equal to alpha 0. plus alpha 1, and so on. And one can see that if you do this calculation this way, one can conclude that Tn plus 1 equals the sum of your step sizes starting from alpha 0 all the way up to alpha n. Just keep in mind this change in index; you can see that the way we have defined this is indeed true. Now, what we will do is we are going to define a linear interpolation of the iterates that have been generated by your stochastic approximation algorithm. So, keep in mind that this sequence is stochastic in nature, and we are going to work with one



Is that okay? That is the sequence generated on one sample point. And let us look at the linear interpolation of that. We will refer to the linear interpolation by X bar. This is a continuous function of time, and we will presume that time ranges from 0 to infinity.

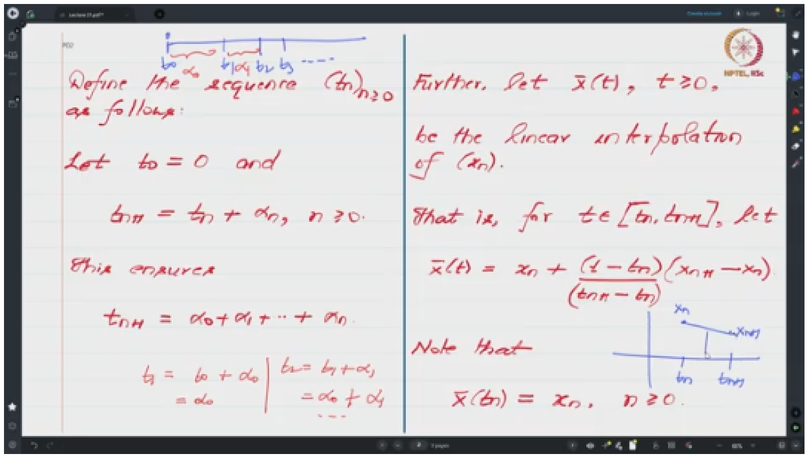
And the way we will define this linear interpolation is as follows. We will split the whole timeline. So, we will split, let us say, the whole timeline. So, let us say this is 0. So, we will call this T0.

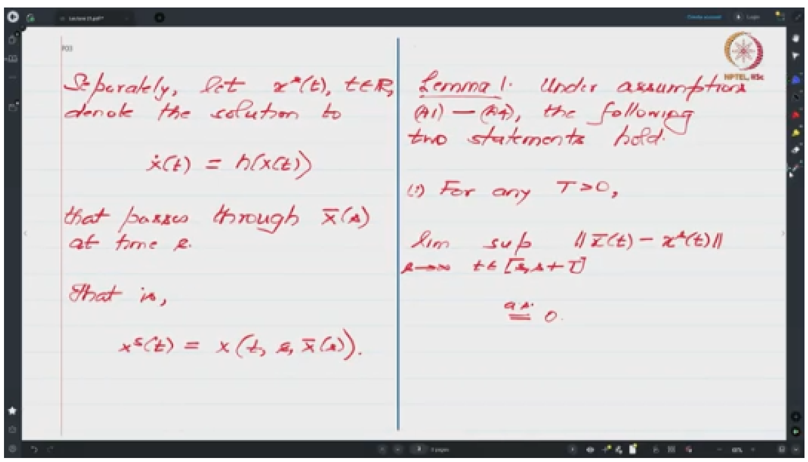
So, then you will have T1, T2, T3, and so on. So, the gap between these two values is alpha 0, between these two is alpha 1, and so on and so forth. So, now what we are going to do is ensure that the value of $\bar{x}(t)$ at the specific time instance tn is xn, right? And for any t between tn and tn+1, we will define $\bar{x}(t)$ to be the linear interpolation between xn and xn+1. Formally, for any t between tn and tn+1, we will set $\bar{x}(t)$ to be xn plus (t − tn) divided by (tn+1 − tn) times (xn+1 − xn).

So, what this is basically doing is that if on the time axis you have tn and tn+1, let us say this is xn and this is xn+1, you are basically connecting them, right?

$$\overline{X}(t) = x_n + \frac{(t-t_n)}{(t_{n+1}-t_n)}\left(X_{n+1} - X_n\right)$$

And this formula basically tells you what the value will be at some time t between tn and tn+1, that is all. We need one more notation to discuss our proof, and that is XST. So, here you see that you have S and T, and both S and T will be some real numbers.





Is this okay? And typically, we will presume S is fixed, right? And T is a variable that can take any value in R, right? So, what is XS of T? Well, XS of T is the solution to your limiting ODE that, at time S, passes through X bar of S. We call that this X bar is your linear interpolation.

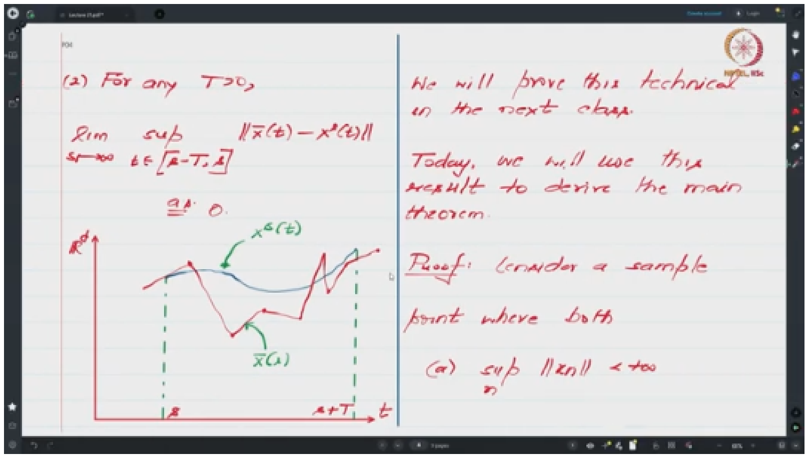
So, we look at the value of the linear interpolation at some time S in R, and whatever this X bar of S is, we force the solution of this ODE to pass through this point at time S. And whatever that solution is, we will refer to it as XS of T. And if you recall, we had developed a notation like this. To, you know, characterize the solution in terms of the initial time and the initial condition. So, in that notation, your XS of T is basically X ts X bar of S. So, this basically emphasizes that this solution passes through X bar of S at time S. Is this okay? Right? So, we will use this shorthand notation.

So, to prove the theorem that we stated on the previous slide, we will need this intermediate technical result, right. So, let me first state this result. So, the result states that under the assumptions A1 to A4, which we had introduced in the previous class. So, those who have forgotten, I request them to go back to that previous class and look at them, or alternatively, you can look it up in this textbook called Stochastic Approximation by Professor Vivek Borkar. So, under these assumptions A1 to A4, the following two statements hold.

So, the statement says that for any T bigger than or equal to 0, capital T bigger than or equal to 0, some limit is 0 almost surely. Is that okay? So, let us decipher what we have over here. So, the first limit says that you require S to go to infinity. The next part says that for whatever s you work with, you look at this time window that starts at s and goes all the way up to s plus t, right?

And you take a value little t between this time window s to s plus t. And you can see that this window has length capital T, right? And for a t lying in this window, you look at the distance between x bar of t and xs of t. And you take the norm and then you ask what will be the value of this as s goes to infinity. And the lemma says that this distance actually goes to 0. So, this recall is the linear interpolation of your stochastic approximation algorithm iterates.

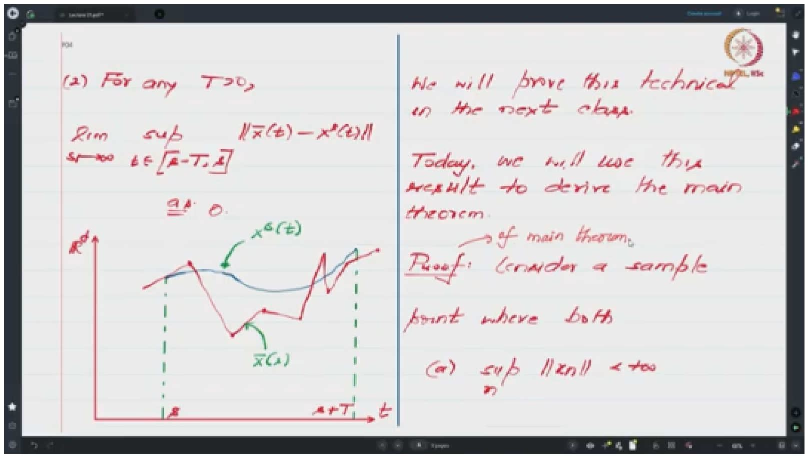
And this is the solution to your limiting ODE, which at time s passes through x bar of s. And what this result says is that this distance in this window, as you shift the initial point of your window to infinity, it goes to 0. And we have an analogous result where instead of looking from s to s plus t, we look at the window from s minus t to s. And again, the claim is that the distance between them actually goes to 0. So, let us pictorially see what this means. So, in this picture, you can see that on the x-axis, you have time and on the y-axis, just for convenience, I put rd. So, let us say you have some fixed time s and then you look at a time window of length t which will take us from s to s plus t. And this red trajectory over here, this is basically your linear interpolation of your stochastic approximation iterates.



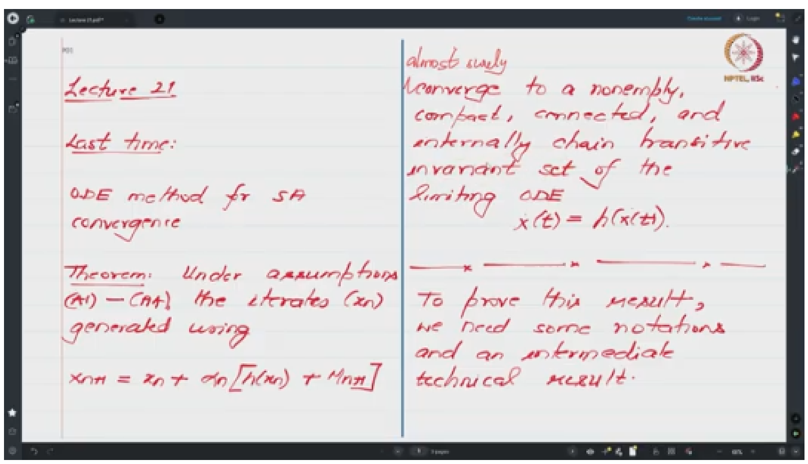
And what we are going to do is, wherever your linear interpolation of your stochastic approximation iterates lies at time s, at that point, you will start a solution of your limiting ODE, right? And this is denoted by this blue trajectory over here. Is this okay? So, you can see that I have denoted it by x(t) over here, right? And then the result says that if you look at the gap between this blue trajectory and the red trajectory in this window of length t, then as the starting point of this window goes to infinity, the distance between these two trajectories actually goes to 0.

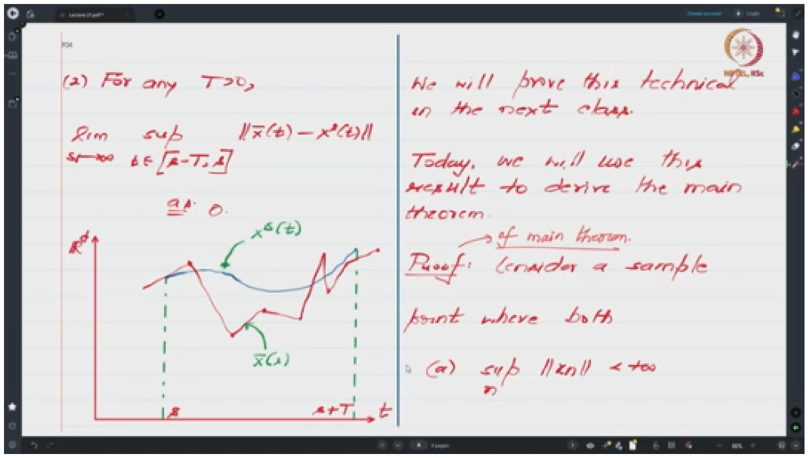
So, we will actually prove this technical result in some future class. Today, what we will do is we will presume that this result is true and derive the main theorem that we stated on the first slide. In particular, we will derive a part of the proof of the main theorem in today's class, and the next class we will derive the second part. And in some future class, we will actually derive this technical result. So, the way I am doing this—the reason why I am doing this—is that you know what purpose this technical lemma is serving with regards to the proof of your main result.

So, the way the main result—I should emphasize that this is not the proof of your technical result, but this is the proof of your main theorem. that we stated before. So, we have to show that, almost surely, the iterates— So, almost surely—I think I did not emphasize this. So, I should emphasize that they almost surely

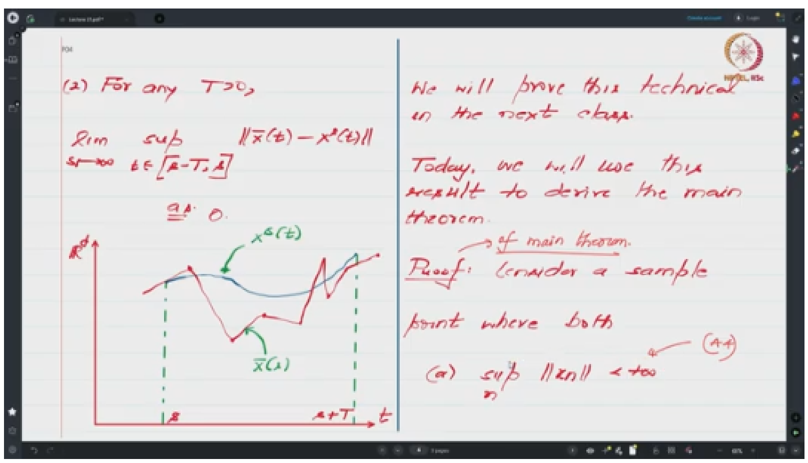


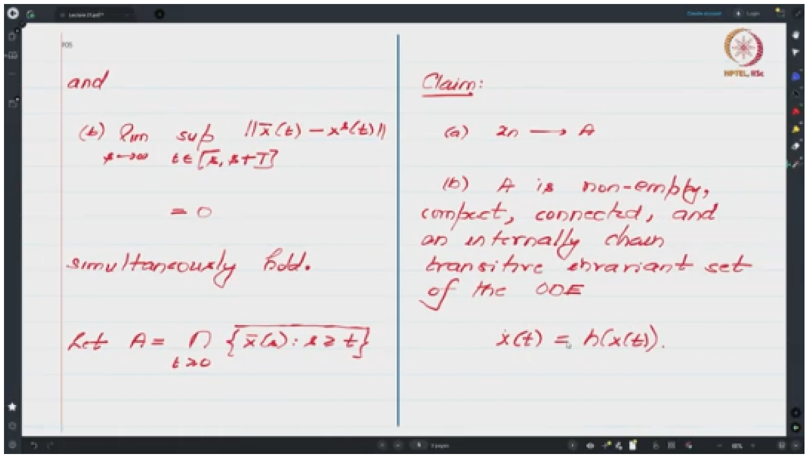
Converts to such an invariant set of the limiting ODE. So, which means that if you collect the set of sample points where this conclusion holds, right, then the collection of those sample points will have probability 1. And the way we prove this result is that we first consider a sample point where both this, you know, statement and another statement, which I will soon describe, hold together. So, first let us understand what this statement is. So, recall that this statement is basically a restatement of



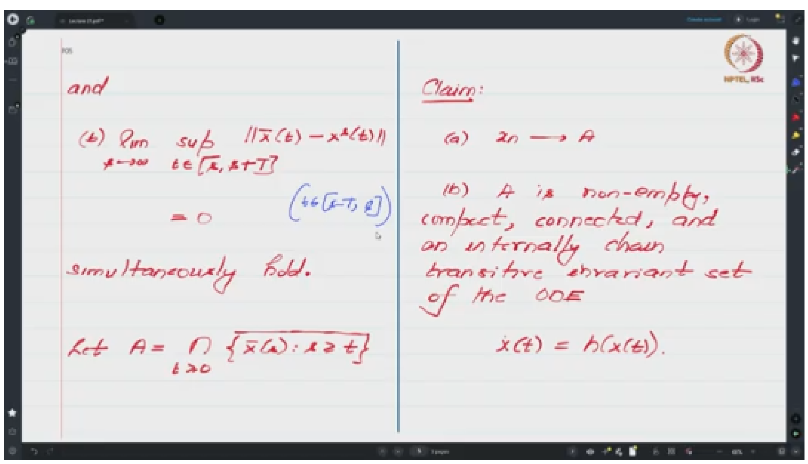


your assumption A4. In assumption A4, we had presumed that this condition holds almost surely. So, what we are doing now is to fix a sample point where this condition holds. And we also need that the conclusions of the technical lemma hold here. So, the technical lemma said that almost surely, okay, for any t greater than or equal to 0, this limit is actually 0.





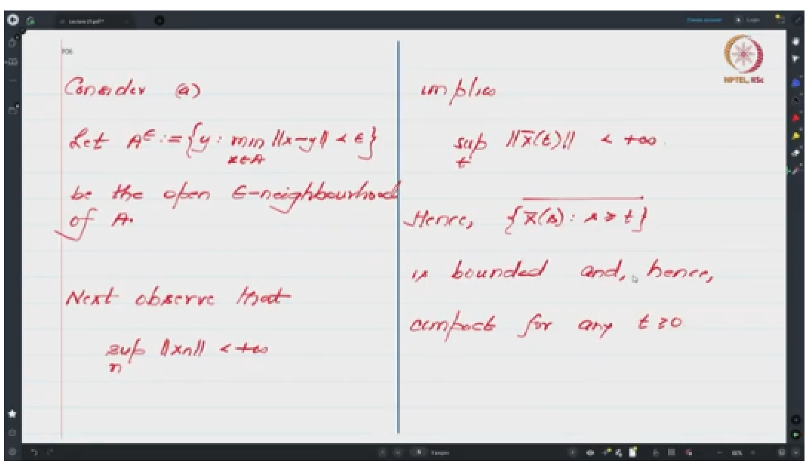
In fact, we also need that, you know, the version where this window is actually s minus t to s, okay, s minus t to s, okay, the variant of that result that is part B of that technical lemma, we also need that to hold, okay. So, let me quickly summarize. We now pick a sample point where the condition in assumption A4 and the conclusion of your technical lemma simultaneously hold. Is this okay? So, now what we are going to define is, we are going to define a set A, right?



And we are going to define this set A to be the following. So, let us understand what this definition is, right? So, X̄(S), recall this is the linear interpolation and the value of this linear interpolation at time S. So, what we are going to do is, we are going to collect the value of your linear interpolation for all s greater than or equal to t for some fixed t. So, this will form a set, and this overline over here implies that we look at the closure of this set. Is that okay?

So, this is a closed set, and we look at the intersection of all such sets for t greater than or equal to 0, and we will refer to this as A. And what we are going to do in order to prove or get to the conclusion of the main theorem that we stated, we are going to prove these two claims. The first claim is to show that your X□ actually converges to this A, the A that we have defined over here. In some sense, that is obvious, but we will formally show that. And the next part is to show that this A, the A that we have defined over here, indeed has the desired properties, right?

That is, A is non-empty, it is compact, it is connected, and it is internally chain transitive with respect to the limiting ODE, right? So, again, I would like to emphasize that because your $\bar{X}$ is generated by your X□s, which are stochastic in nature, this A set would change from one run of your experiment to another run. So, what we are saying is that on this sample point that we have fixed, let this be A, and we are going to claim that whatever this A is, which is a function of your sample point, on one hand, X□ converges to this set, and on the other hand, it has the desired properties. And in this class, what we are going to do is, we are going to basically prove this part, and in the following lecture, we are going to prove the B part of this claim.

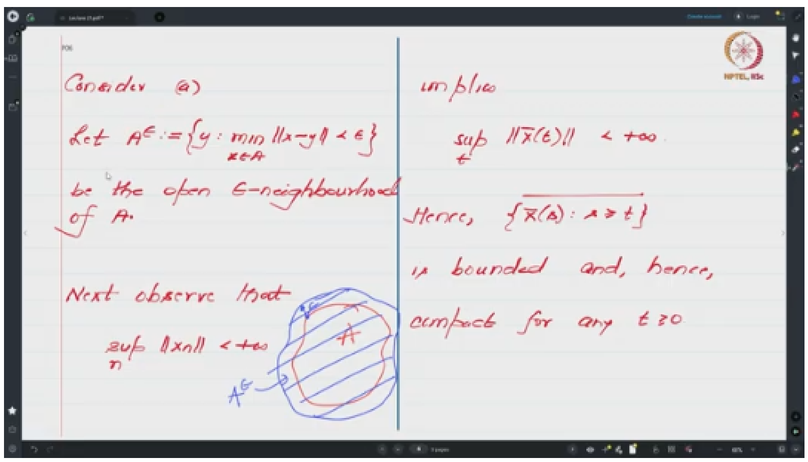


So, towards proving A, What we are going to do is, we are going to first introduce some set called A epsilon. So, what is A epsilon? Well, A epsilon is the open epsilon neighborhood of A. Formally, it is the collection of all y's whose distance to A is strictly less than epsilon. So, one can see that this collection will indeed include every element in A.

On top of that, it will also include elements which are epsilon away from A. So, formally, if you think this is your set A, then your epsilon neighborhood will basically be an expansion of this. So, this will be like an epsilon neighborhood. So, the blue shaded region is what is your A epsilon.

$$A^{\epsilon} := \{y: ||x - y|| < \epsilon\}$$

So, now we are going to make use of this A epsilon set to show that Xn converges to A, and the way we go ahead is we are going to introduce or define some sets. So, towards that, first observe that because of our A4 assumption, your iterates are bounded.



And because your xn's are bounded, one can easily conclude that the linear interpolation is also bounded. The difference over here is that n runs over positive integers, whereas t runs over all values between 0 to infinity. So, if you look at the supremum of norm x bar of t, this will also be less than infinity. So, if you look at this closed set, that is the collection of X bar of S for S greater than or equal to t for any fixed t bigger than or equal to 0, and if you look at its closure, this set will be bounded because of this fact. And since this is bounded and closed, one can immediately conclude that this set is actually compact.

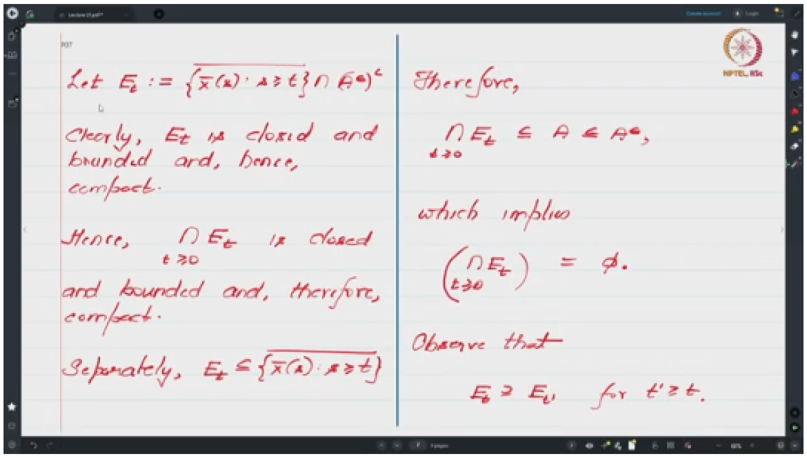
$$||X_n|| < +\infty$$

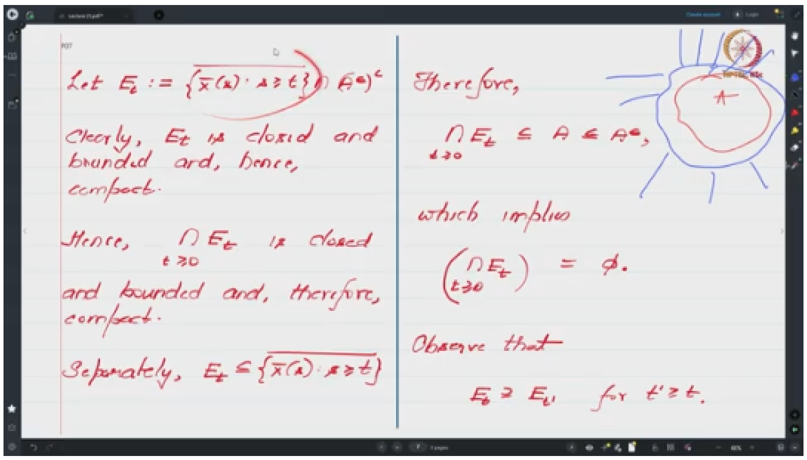
$$||\bar{X}(t)|| < +\infty$$

$$\overline{\{\bar{X}(s): s \geq t\}}$$

And since T over here is arbitrary, our conclusion that this set is compact holds for any T bigger than or equal to 0.

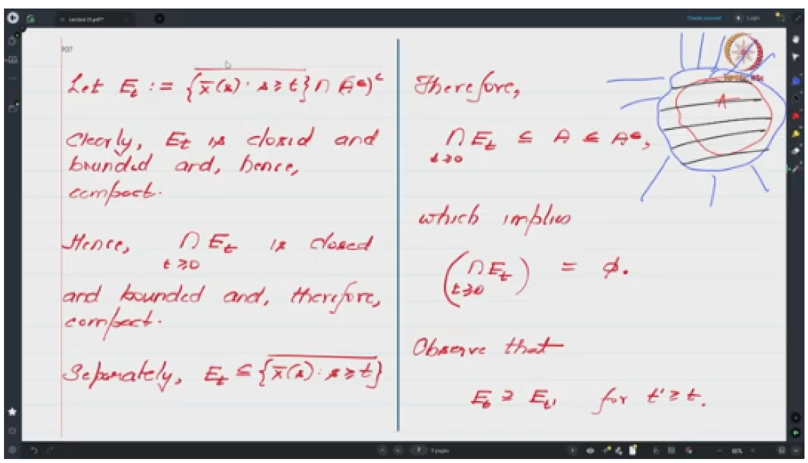
So, recall that A is an intersection of such types of sets, and we have taken one element in that collection and have shown that it is compact. We will define another set, which is ET, which is whatever we saw before, and its intersection with the complement of A epsilon. So, recall that your A is something like this, your A epsilon is something like this, and A epsilon complement will be everything that is outside this. So, what we are doing here is looking at the intersection of this set with the complement of A epsilon. And what we are going to show is that this ET is empty.





For all sufficiently large t. So, if we somehow manage to show that ET is empty for all sufficiently large t, we will then be able to show that this set has nothing in common with A epsilon complement. Hence, this set must actually be within A epsilon. So, A epsilon is this part over here. A epsilon is the black part. So, this trajectory here will completely lie in this black region eventually.

And since epsilon is arbitrary here, one can conclude that eventually, your linearly interpolated trajectory lies very close to A, and we can use that to formally show that your X bar of t converges to A and hence your Xn converges to A. So, the goal over the next few minutes would be to show that this ET is actually empty for all sufficiently large t. And the way we will go about showing this is to first observe that because this is a closed set and A epsilon is open, and hence A epsilon complement is closed. So, we have an intersection of two closed sets, and hence we can conclude that this set ET is also closed. And since this is bounded, one can conclude that your ET is also bounded, and since we are in RD, that is a Euclidean space, closed and bounded implies that ET is also compact.



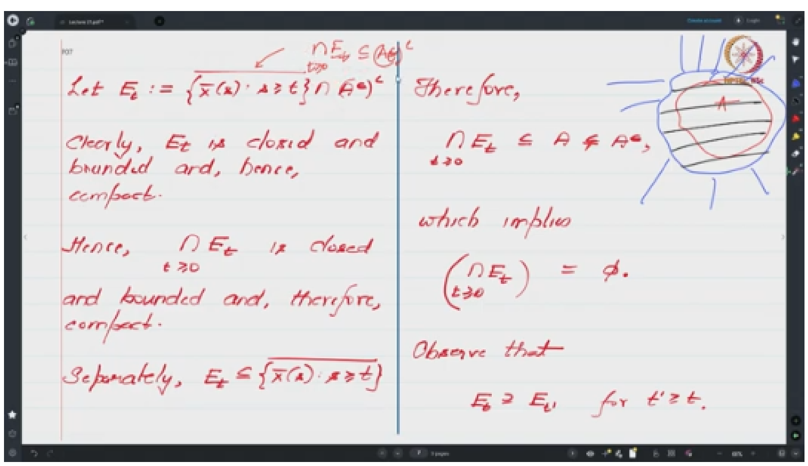
Is that okay? So, now let us look at the intersection of ET, right? Because, you know, ET for every T is compact, right? One can conclude that the intersection of ET is also closed and bounded and hence compact, right? And then there is an easy fact that one can see because ET is an intersection of two things, ET will actually be a subset of one of the sets in this intersection.

Hence, ET is a subset of this. And since A is an intersection of such sets for all t bigger than or equal to 0, one can conclude that the intersection of ET must be a subset of A. So, the intersection of ET is a subset of A, and A is actually a proper subset of A epsilon.

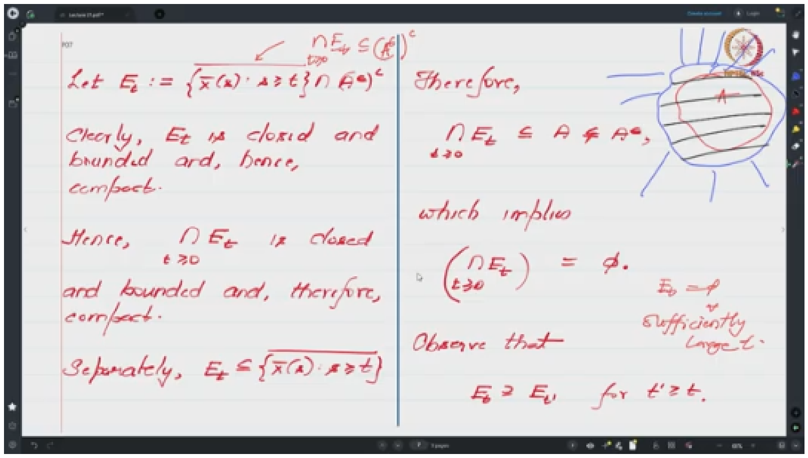
And from this fact, one can actually conclude something very, very interesting, which is that the intersection of ET must then be empty, right? So, why is that?

Well, recall that, you know, ET, right, has some elements which are, you know, contained over here and some elements which are outside A epsilon, right? So, there are some elements here and some elements outside A epsilon, right? Now, this intersection of ET, right, similarly will have some elements in A, right? Right, and some elements outside A epsilon, but we have managed to show that this intersection of ET, right, actually lies within A. So, just one minute, I just want to make sure why this is right, okay. So, the idea is the following that in some sense this fact implies that your intersection ET right T greater than equal to 0 is a subset of A epsilon complement right because

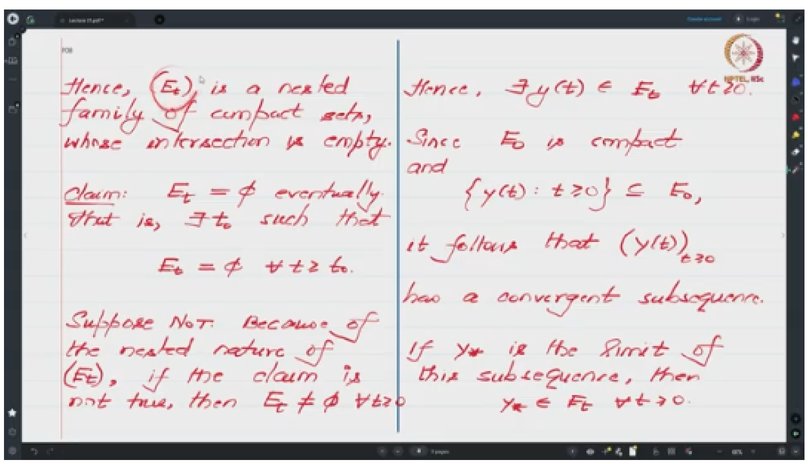
In every ET, you have this set over here, and hence ET is a subset of A epsilon complement, and hence the intersection of ET is also a subset of A epsilon complement. Maybe I will write this properly here. Right? And here we have separately shown that the intersection of ET is a subset of A epsilon. Now, if you have a set which is a subset of both A epsilon and A epsilon complement,



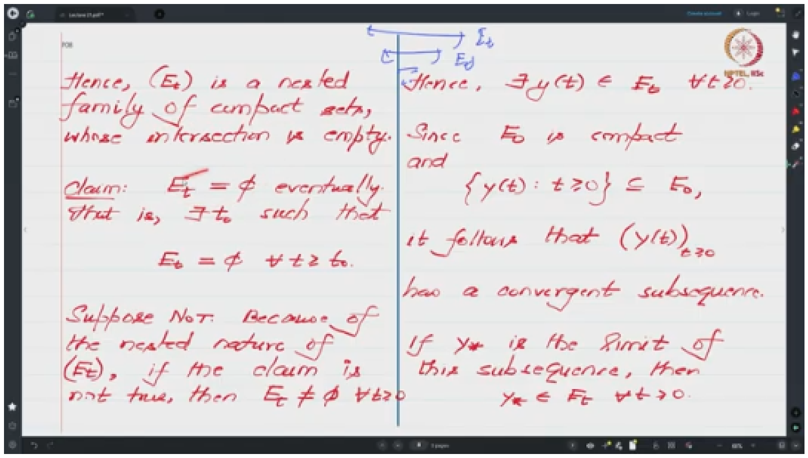
one can then immediately conclude that that can only happen when the intersection of ET is actually empty. So, what we have shown so far is that the intersection of ET is empty. Now, what we will do is we will use this fact to conclude that ET is empty for all sufficiently large T. In other words, if the intersection is empty, we are going to use that fact to conclude that individual elements in the intersection themselves are empty for all sufficiently large T. So, towards that, observe that the ETs that we have defined over here are special. They are special in the sense that if you pick any T prime which is larger than T, then your ET prime is a subset of ET.



That is, you know, your ET contains ET prime, right? And hence, one can conclude that this sequence ET is actually a nested family. Nested family means, you know, if you sort of draw your ET in this fashion, right? Right? Your ET prime is



Is actually contained within this. So, in a set-valued sense, this is contained within this, and as you increase your t values, you will get smaller and smaller sets. So, in that sense, you have a nested sequence. So, you have a nested sequence, and we have separately shown that each element in this family is actually a compact set. We have also shown that the intersection of this nested family of compact sets is actually empty. And we are going to use all these facts—that it is a nested family of compact sets whose intersection is empty—to conclude that ET must be empty eventually.



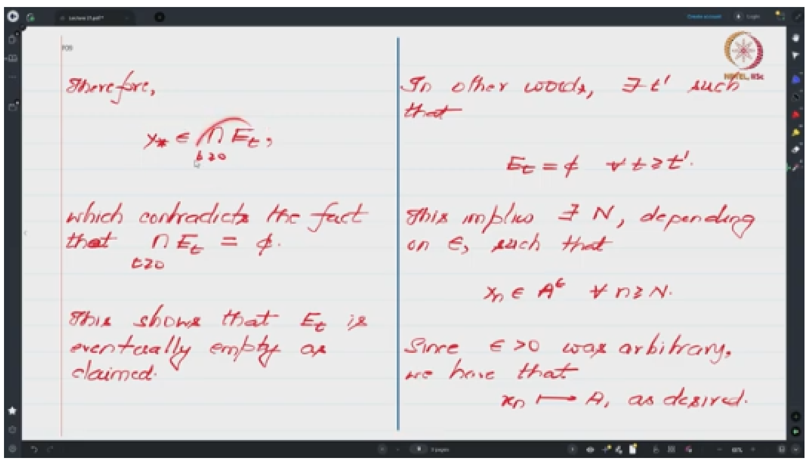
That is, there exists some time t0 such that for all t bigger than t0, ET is actually empty. That is what 'ET equals empty eventually' means. And to prove this claim, we will use the idea of proof by contradiction. That is, suppose this claim is not true. In other words, your ET is not eventually empty, right?

And because of the nested nature of these ETs, one can conclude that if this claim is not true, then ET must not be empty for all T greater than or equal to 0, right? So, for all t greater than or equal to 0, ET is not empty, which means that for every t, there must be an element in ET. The fact that they are non-empty implies that there must be an element in ET, right? Now, if you collect all these elements together, right, and because of this nested nature, one can conclude that this collection is actually a subset of E0, and we know that this E0 is compact, right?

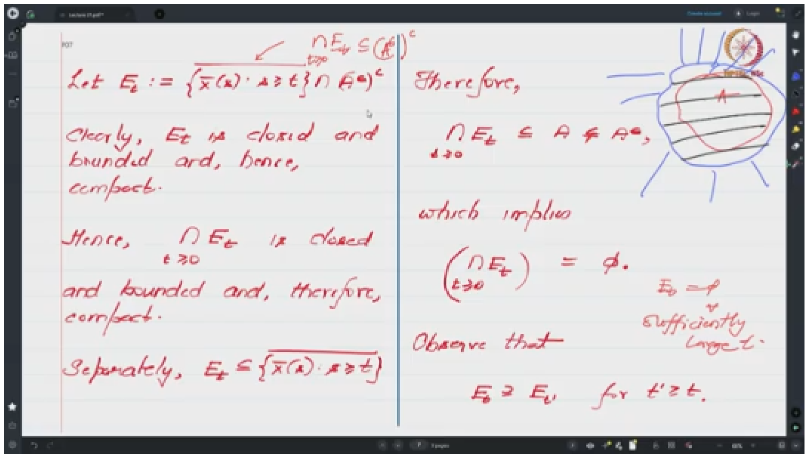
This is a trajectory in a compact set, and hence one can conclude that this trajectory will have a convergent subsequence. I mean, this is a fact based on a compact set. So, if you are not aware of this, I request you to look it up. So, the fact that this trajectory lies in a compact set allows us to conclude that this trajectory will have a convergent subsequence. So, because it has a convergent subsequence, that subsequence will have a limit.

So, let us refer to this limit as Y star. So, one can conclude from all these facts that this limit must actually belong to ET for all t greater than or equal to 0. The reason being, Y star is the limit of a subsequence, and the tail of the subsequence will lie in ET for all t greater than 0. I mean, different parts of the tail—that is, the tail starting at different points in time—will lie in ET for different t's. So, since the tail lies in ET fully and ET itself is compact, your Y star will lie in ET for all T greater than or equal to 0.

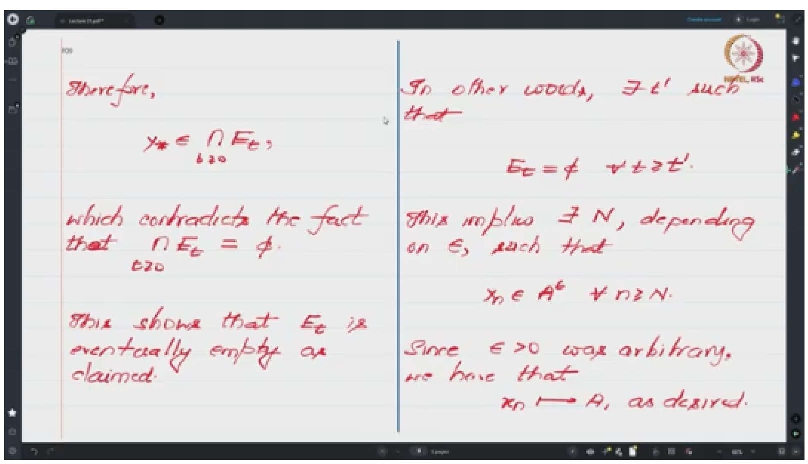
Now, this Y star does not depend on T, and we are saying that this Y star lies in ET for all T greater than or equal to 0. So, that allows us to conclude that Y star must be an element in the intersection of ET. However, we previously showed that the intersection of ET is empty, which means that you cannot have any element in the intersection of ET. But we just showed that there is an element in ET, which allows us to conclude that our supposition is incorrect—that is, we have a contradiction. In other words, the claim indeed must be true, and hence ET must be eventually empty as claimed.



So, which basically means that there is some time t' such that for all t greater than or equal to t', your E_t is empty. So, E_t being empty means that So, recall the definition of this. So, E_t being empty means that there is no element which is common to this and this. So, there is no element that is common to these two sets, and hence it implies that, you know, every element here must actually be contained within A_epsilon.



Every element here must be contained within A_epsilon. for any t greater than or equal to t', and because these x_t's are linear interpolations of your x_n's So, that allows us to conclude that, you know, there exists some n depending on your choice of epsilon such that x_n belongs to A_epsilon for all n greater than or equal to n. So, in other words, you know the tail of your stochastic approximation iterate sequence is in the epsilon neighborhood of A. And since this choice of epsilon was arbitrary, one can conclude that X_n gets closer and closer to A in its tail portions, and that allows us to conclude that X_n goes to A as desired. So, this brings us to the end of this class.



Let me quickly summarize what we have done in this class. So, we have discussed the linear interpolation of X_n, stated one technical result which we will be proving in one of the future classes, and then we began the proof of the main theorem assuming that this technical result is actually correct. And then, in our proof, we are going to actually prove two subclaims. The first of these subclaims involved showing that, you know, identifying a definition of a limit set A and then showing that this X_n actually converges to A, which is what we have finished so far. In the next class, what we will show is that this A that we have has the desired properties.

That is, it is non-empty, compact, and connected, right? And it is internally chain transitive and an invariant set with respect to the limiting ODE. Thanks for now. Hope you will join for the next discussion.