STOCHASTIC APPROXIMATION: THEORY AND APPLICATIONS

Dr. Gugan Thope

Department of Computer Science and Automation

Indian Institute of Science

Week 6

Lecture 23

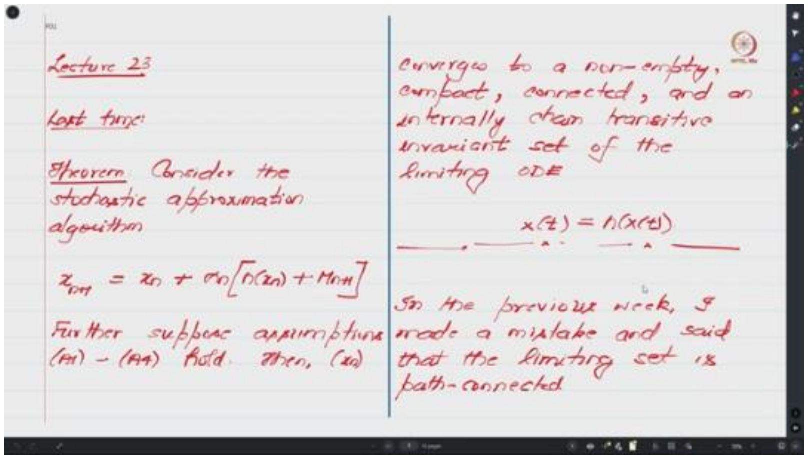
How Stochastic Approximation Iterates Track an ODE: The Key Lemma

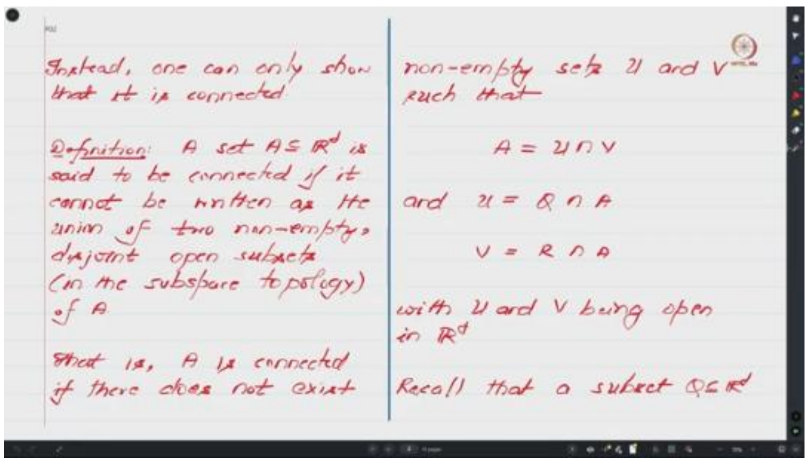
Hello and Namaste everyone. Welcome to lecture 23 of this NPTEL course on Stochastic Approximation. So, in the previous week, we, you know, started combining all the tools that we have studied in order to understand the asymptotic behavior of Stochastic Approximation algorithms, right? This week, we will continue that discussion and try to finish the proof of the main theorem that we started, right? So, let us do a quick recap of what we were trying to do, right?

We were trying to prove this theorem, right? And recall that this theorem said: consider the stochastic approximation algorithm given by this update rule: x_n plus 1 equals x_n plus alpha_n times h of x_n plus m_n plus 1.

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

Further, suppose the four assumptions A1 to A4 hold. Then, the sequence x_n generated by this update rule converges to a non-empty, compact, connected, and internally chain transitive invariant set of the limiting ODE x dot of t equals h of x of t. So, again, I would like to highlight that the stochastic approximation algorithm, by its nature, is stochastic, whereas this limiting ODE is deterministic. Now, in the previous week, I made a mistake and said that the limiting set is actually path-connected. I would like to fix that. You know, I later realized that one can actually only show that this limiting set is connected.





So, now I am going to define this notion of connectedness and then tell you how one can show that this limiting set is actually connected, right? So, a subset of R^D, let us denote it by A, is said to be connected if it cannot—notice that it is 'cannot' and not 'can'—so, a subset A of R^D is connected if it cannot be written as a union of two non-empty, disjoint, open subsets of A, where this, you know, this notion of openness is in the context of subspace topology. So, let me elaborate. A subset of R^D, A, is said to be connected if there does not exist sets U and V. So, I think I made a mistake here. They should not be this.

U and V such that now you require that these sets be disjoint. So, there should be two; you should not be able to write A in this fashion. Which fashion? That you have these two non-empty sets U and V. Such that U intersection V is empty and A equals U union V. And when I say that A, U, and V should be open in the subspace topology, I require that U be Q intersection A and V be R intersection A.

So, I think I should have made a bunch of mistakes here; sorry about that. So, this should be Q and R being open in R^d. Is this okay? So, let me say it again.

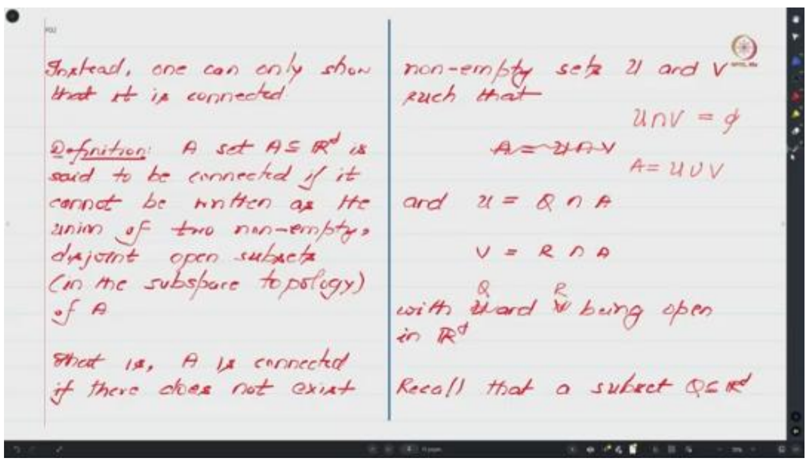
$$\mathcal{U} \cap V = \phi$$

$$A = \mathcal{U} \cup V$$

$$\mathcal{U} = Q \cap A$$

$$V = R \cap A$$

So, A is connected if there does not exist



two non-empty sets U and V such that their intersection is empty and A can be expressed as U union V, where U and V are open in the subspace topology, meaning U can be expressed as Q intersection A and V can be expressed as R intersection A, where Q and R are open sets in R. So, you know, maybe some of you may be a bit confused about what I am saying. So, let me, you know, elaborate. So, let us say you have a set of this form. So, let us say something like this.

So, let us say this is 0, this is 1, and maybe this is 2, and let us say this is 3. So, your set A is 0 comma 1 union 0 comma 1 union 2, 3. So, the point is that a set can be written as a union of disjoint sets. So, ideally, one would expect that such a set is not connected, and this is what this result says.

Now, notice that 0 comma 1, this interval is actually an open interval. However, the interval 2 comma 3, because it is closed from one end, this set by itself is not open. This set by itself is not open. However, this set can be written as, let us say, 3 by 2 comma 3 intersection with A, right?

$$B(x, \mu) \leq Q$$

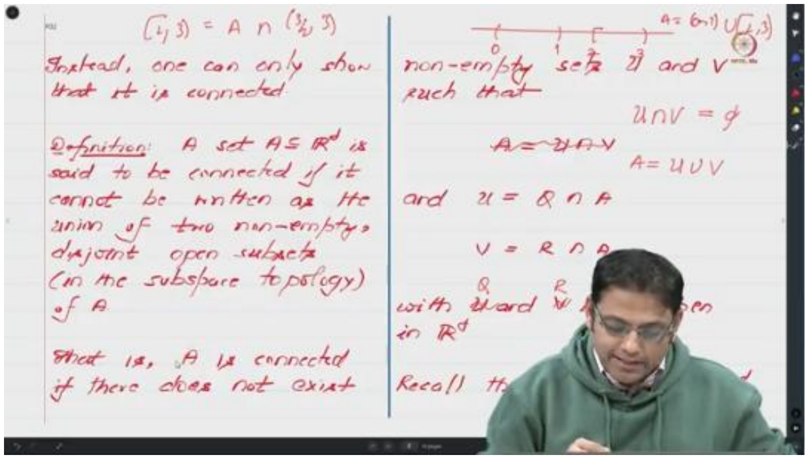
$$A = \cap_{t \geq 0} \overline{\{\bar{x}(s) : s \geq t\}}$$

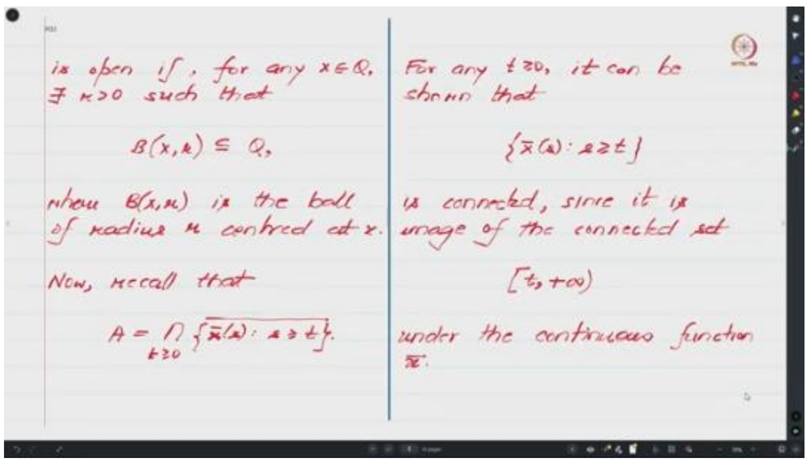
$$\{\bar{x}(s) : s \geq t\}$$

$$[t, +\infty)$$

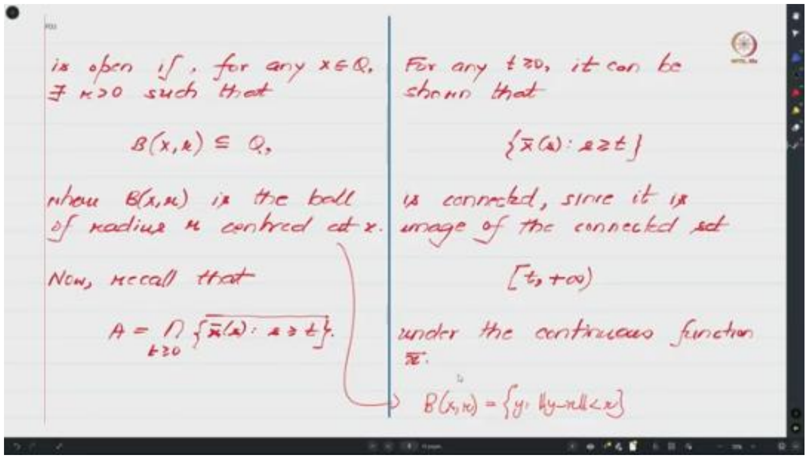
So, if you take A intersection 3 by 2 comma 3, I hope you agree that you will indeed end up with 2 comma 3, and from or by this logic, one can see that you can express 2 comma 3 as A intersection with some open interval. And hence, this 2, 3 is open in the subspace topology relative to A. So, because of this reason, we can say that this union of sets can be expressed as this set union this set. Both of which are non-empty, they have nothing in common, and they are open in the subspace topology. Hence, the union of these two intervals is indeed disjoint, and hence it is not connected.

Is this okay? And so, whenever you are not able to express A in this fashion, we would say that A is connected, and when you are able to express A in this fashion—that is, in terms of U and V—we will say that A is not connected. This and the formal definition of two sets being open in Rd is that, so here I have sort of specified when a subset Q of Rd is open. A set Q of Rd is open if there is some room to maneuver at every point in Q. Specifically, if you consider any point x in Q, there should exist R, which you can think of as the radius, such that the ball centered at X of radius r is fully contained in Q, where B of X r is the ball of radius r centered at X. And so, I will maybe also give the formal definition of this ball.



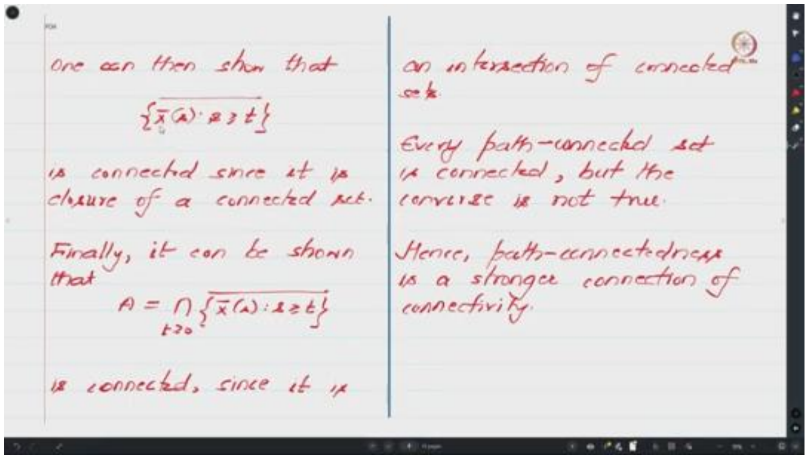


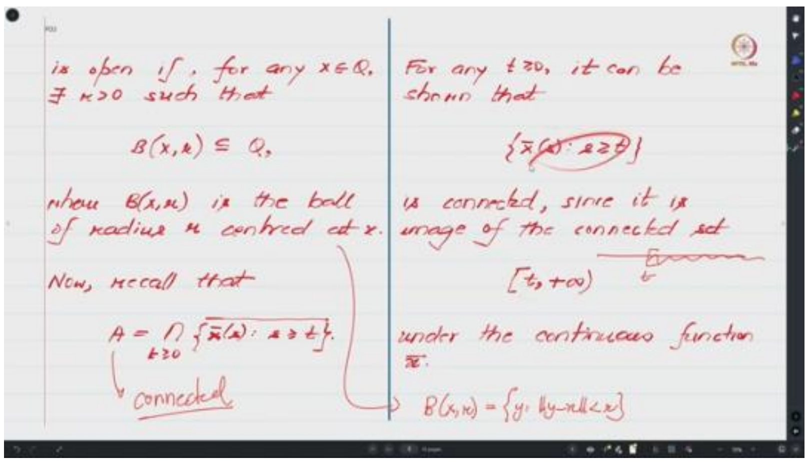
So, this is B of X comma r equals all y such that the distance between y and X is strictly less than r. So, this is referred to as the open ball of radius r centered at X. So, now I hope you have got some intuition of what I mean by a set being connected. Now, let us recall what we were trying to do. If you recall, the limiting set that we were thinking about was defined in the following sense: that is, A equals the intersection of the closures of the tail of the linearly interpolated trajectory obtained from XNs. And, you know, the claim is that this set A is actually connected.



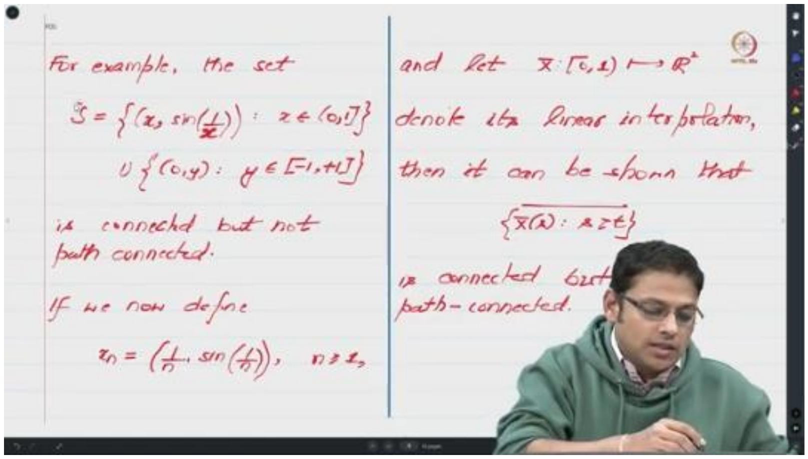
So, the claim is that this set A is actually connected, and the way we show this is that you first show that the tail of this linearly interpolated trajectory is connected for any t bigger than or equal to 0. The way you have to show that this is connected is by showing that this set can be expressed as the image of the connected set t comma infinity. So, if you think of t comma infinity, this is an interval on the real line. So, somewhere you have t, and this is the set. So, I hope you agree that this set—the one that I am highlighting over here—cannot be expressed as a union of two non-empty sets such that they are

Disjoint and both of them are open in the subspace topology. So, this set indeed is connected, and this set over here is the image of this connected set under the continuous function X bar. So, recall that X bar is obtained by linear interpolation and hence it is trivially continuous. And since this is a connected set and this is the image of this connected set under your continuous mapping X bar, one can, you know, apply some concepts from real analysis to conclude that this set must be connected. And since this set is connected, we can thereafter conclude that the closure of that set is also connected. And since A is the intersection of, you know, nested connected sets, one can conclude that A also is connected.



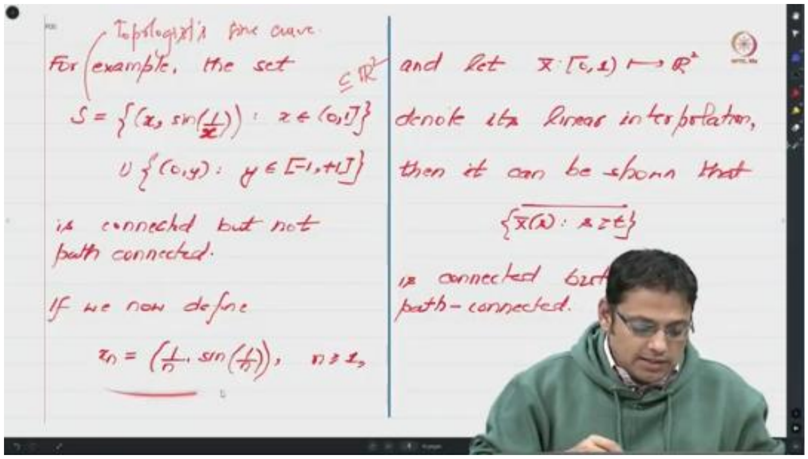


So, last week I made a mistake and said that this set A is path-connected. I would like to highlight that path-connected is a stronger notion of connected, which means that every path-connected set is actually connected, but the converse is not true. That is, you can have sets which are connected but are not path-connected, right? And towards that, I had given one example last time, right? You know, which I have repeated over here. So, if you consider this set S, which is the union of these two subsets.

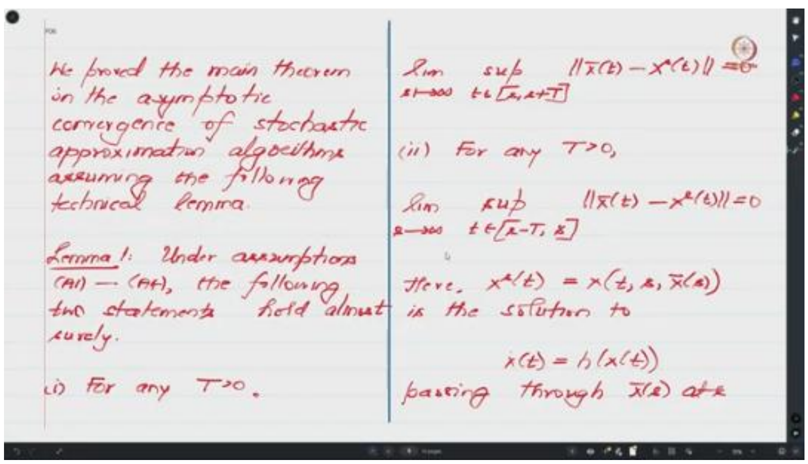


So, the first subset consists of pairwise elements or, you know, two-dimensional coordinates where the first element X comes from the interval (0, 1], which is open at 0 and closed at 1, and the second element of, you know, this tuple is sin(1/x), right? So, this part is a subset of R^2 right? And you take the union of this subset with this subset of R^2, which includes all points of the form (0, y), where the first coordinate is 0 and the second coordinate is some number between -1 to +1. So, if you consider the union of these two sets, one can show that this set is actually connected, but it is not path-connected.

So, this set is formally known as the topologist's sine curve. The sine curve, and you can read more about it online and, in fact, figure out why this set is connected but not path-connected, right? So, in that sense, if you consider a sequence made up of (1/n, sin(1/n)), where each X_n is a two-dimensional coordinate, and if you look at the linear interpolation of this sequence, right, one can show that the closure of this sequence, you know, is connected but not path-connected. So, this is the reason why, in general, we would not be able to show that such sets are path-connected but would only be able to show that they are connected.



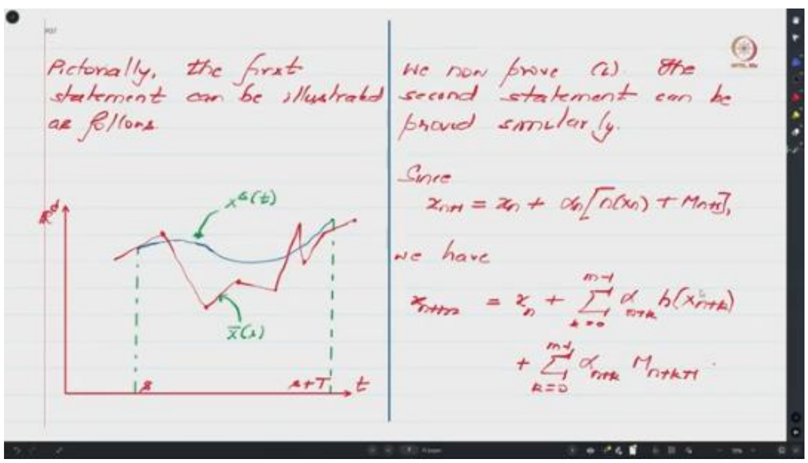
So, this, in some sense, fixes the issue that I had unfortunately made in the previous week. I hope this is now resolved. Now, we go back to proving the main theorem. And if you remember, what we had done last week was to establish the main theorem assuming the following technical lemma. So, we had said that



Suppose this Lemma 1 holds; then the main theorem holds. Now, what we are going to do is prove this Lemma 1. In particular, throughout this week, we will be proving bits and pieces of this lemma, and putting them together will give us the full proof of Lemma 1. So, what does Lemma 1 state? It states that, under assumptions A1 to A4, the following two statements hold almost surely. So, what are these two statements? For any t greater than or equal to 0.

This quantity over here goes to 0. So, what is this quantity? Well, it is the distance between your linearly interpolated stochastic approximation trajectory and a suitable solution of your limiting ODE. Right, and you look at the distance between these two trajectories within the window that starts at s and goes all the way up till s plus t. Thus, we are looking at a window of size capital T, and this limit ensures that the starting point of this window goes to infinity, right? And then we have a similar result where instead of looking at the window s to s plus t, we look at the window s minus t to s, right? So we sort of look t time units before s, and here we are looking at capital T time units after s. And what is this excess of T?

Well, formally, excess of T is this quantity over here. So, this is the solution of this ODE which at time s has to pass through X bar of s. Is this okay? And pictorially, you know what this lemma is trying to tell us is that if you think of this red trajectory over here as the linear interpolation of your XNs. So, on your X-axis you have time, and on the Y-axis you have RD in an illustrative sense.



So, these are your different, you know, elements generated by your stochastic approximation algorithm. Now, you linearly interpolate them, and in that way, you get this red trajectory. So, now you pick some s on the timeline below and look at the window s to s plus t. So, what you do is you start the solution of your limiting ODE, which is denoted by blue color over here, from where your linearly interpolated trajectory is at time s, right. So, you can see that at time s, the blue trajectory and the red trajectory are at the same place, right? And thereafter, the blue trajectory is governed by the deterministic limiting ODE, whereas the red trajectory is governed by your stochastic approximation iterates.

And what this Lemma 1 says is that if you look at the supremum distance between, you know, these two trajectories within this window, then as s goes to infinity—that is, as your initial time instance goes to infinity—this gap between, or the gap between these two trajectories, will go to 0. Is this okay? And now we are going to formally prove this lemma, right? So the proof of this statement and the proof of this statement, you know, more or less follow similarly. So what we will do is we will only focus on proving the first part, and the second part will be left as an exercise.

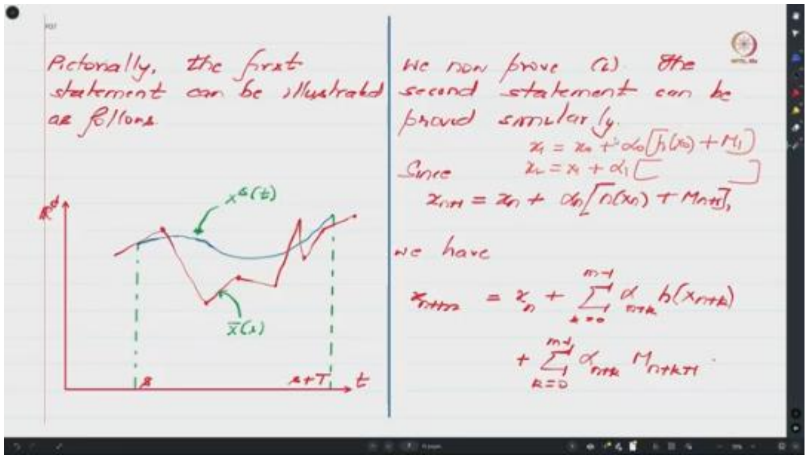
Is this okay? So, how to prove the first part? The first thing you can observe is that, you know, from our update rule, we have x_{n+1} equals x_n plus α_n times h of x_n plus m_{n+1}. So, if you look at this update rule over, let us say, m many time steps. So, here you have m.

So, if you look at this update rule over m many time steps, then one can see that x_{n+n} equals x_n plus, you know, the sum of quantities of this form plus the sum of quantities of this form. So, let me just quickly elaborate why I got this.

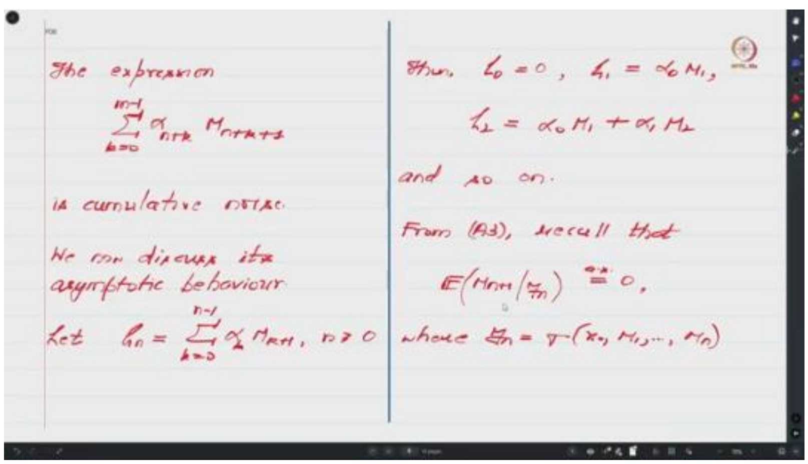
$$x_{n+1} = x_n + \alpha_n[h(x_n) + M_{n+1}]$$

$$x_{n+m} = x_n + \sum_{k=0}^{m-1} \alpha_{m+k} h(x_{n+k}) + \sum_{k=0}^{m-1} \alpha_{n+k} M_{n+k+1}$$

For example, you would have x_1 equals x_0 plus α_0 times h of x_0 plus m_1. And similarly, you would have x_2 equals x_1 plus α_1 times something. So, wherever you have x_1, you replace this quantity, and one can see that one can express x_2 equals x_0 plus this quantity plus this quantity.

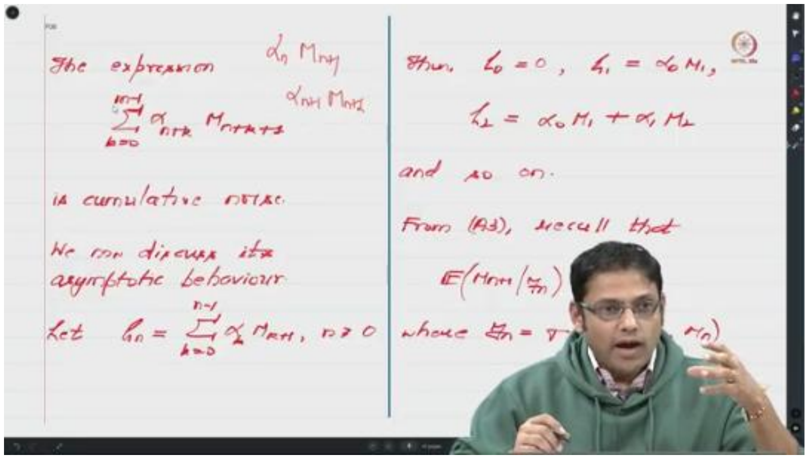


That is exactly what we have over here. So, now we are going to take a look at this quantity and this quantity separately and see how this behavior compares with the solution trajectory of an ODE, which, let us say, started from X1. So, we are going to do this comparison. And in this class, what we are going to do is focus on this part, which is what we had over here. So, this expression is the cumulative noise in the update rule.



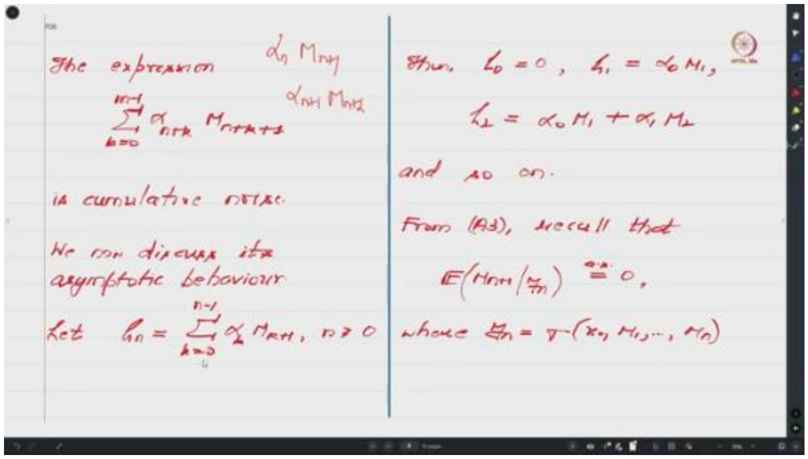
Cumulative noise means that at time step n, we add the noise term alpha n mn plus 1. At time step n plus 1, we will add alpha n plus 1, mn plus 2, and so on. So, at each time step, we will be adding certain noise to our update rule. So, this can be viewed as the cumulative noise that is there in our update rule. So, what we are going to ask is: what can we say about this cumulative noise?

Now, you have to keep in mind that cumulative noise, in general, will be good for us for the reason that At any particular time instance, the noise is actually going to push you in a certain direction. So, let us say at time instance n, the noise pushes you this way. At time instance n plus 1, the noise pushes you this way. And maybe at n plus 2, it pushes you backward, and subsequently, it pushes you forward, and so on and so forth.



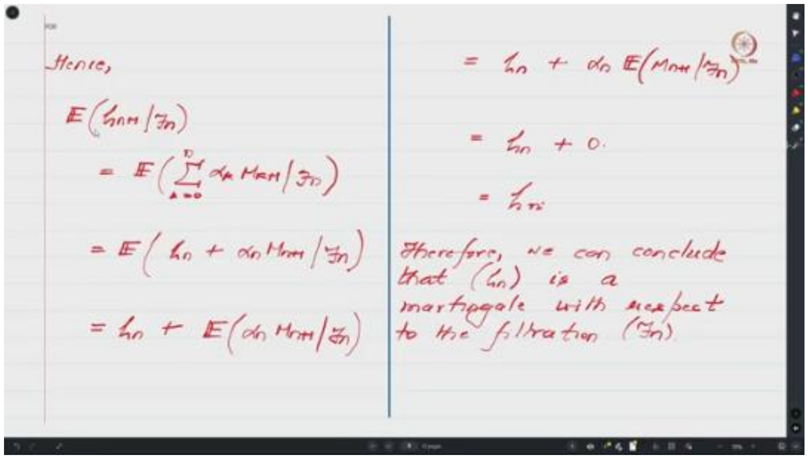
So, at any given time instance, the noise is going to push you in some arbitrary direction. But because of its very nature that its conditional expectation is 0. what we would hope is that if you add enough of these noise terms the cumulative effect sort of cancels out because you know at some point you are pushed forward maybe at some other point you are pushed backward and so on and so forth so that the cumulative effect is negligible okay so that's why often we work with cumulative noise expressions and And now we are going to formally see why a term like this has negligible effect asymptotically. So, you will see what do I mean by all these terms as we move along.

So, what we will do is as a first step, we are going to introduce this notation zeta n, which is the cumulative noise from time instance 0 to discrete time instance n minus 1. So, one can see that From this definition, if you put n equals 0, this is n minus 1, so which is minus 1. So, whenever the upper index is smaller than the lower index, we will presume that the sum is 0. Accordingly, I have said zeta 0 is 0.



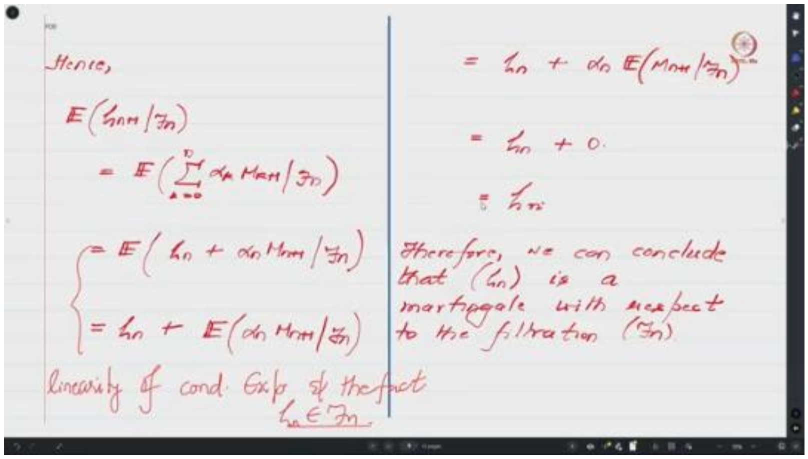
Now, if you put n equals 1 over here this expression will be 0. So, we only have to look at one term with k equals 0 and hence we will end up with alpha 0 m1 and similarly one can see that zeta 2 is alpha 0 m1 plus alpha 1 m2 and so on. Recall from our assumption A3, which was done in the previous week. So, you can look at that lecture and explicitly see what is A3. So, one of the statements

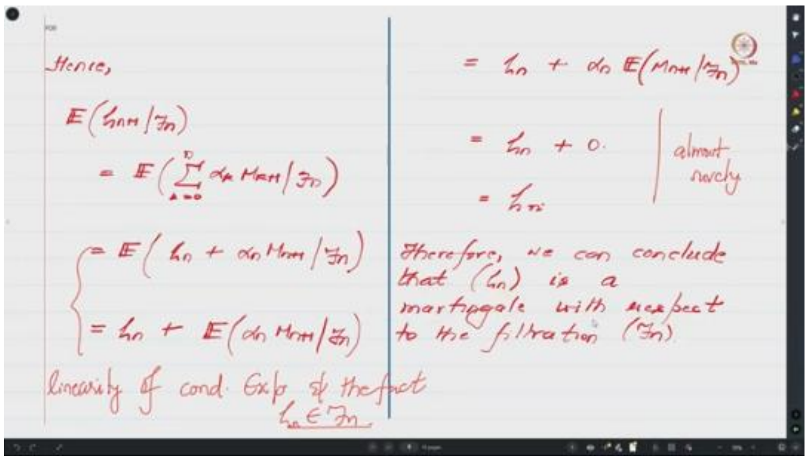
That we had presumed under A3 was that this noise sequence is a martingale difference sequence, which means that almost surely the conditional expectation of Mn plus 1 is 0, where Fn is the sigma field generated by X0, M1 all the way up till Mn. So, one can view this as the information that is available until time n. So, what this says is that given all the information that you have until time n, the conditional expectation of this is 0. And from this, one can see that if you take the conditional expectation of zeta n plus 1, which is the sum of terms of this form, you can break this sum into the first n many terms and then you can look at the last term over here. And zeta n, if you go back to the definition of zeta n, this is made up of terms like alpha 0 M1, alpha 1 M2, all the way up till alpha n minus 1 Mn.



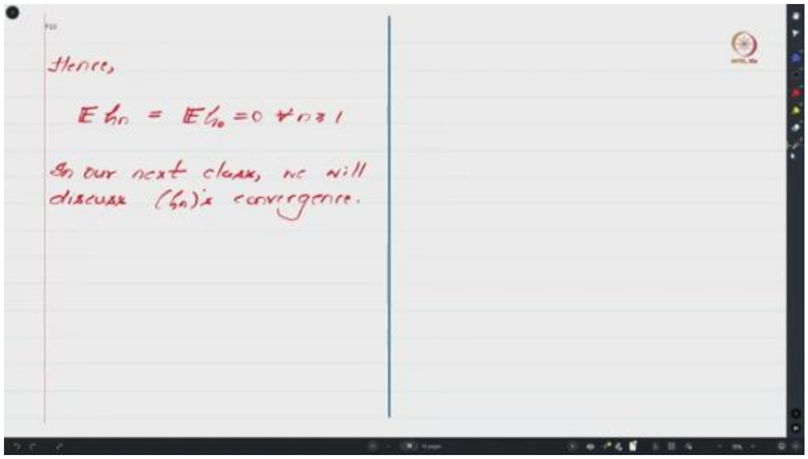
So, in other words, zeta n is a sum of quantities whose value is available at time Fn in some sense, or, you know, formally one can show that zeta n is measurable with respect to Fn. And since zeta n is measurable with respect to Fn, the conditional expectation of zeta n with respect to Fn is zeta n itself, and then we are left with the conditional expectation of alpha n Mn plus 1. So, you know, we are making this jump by using the linearity of conditional expectations, right, and the fact that your zeta n is measurable with respect to Fn.

So, by using these two facts, one can show that the conditional expectation of this is zeta n times this. And thereafter, by making use of the linearity property of conditional expectation, we can write this expression as zeta n plus alpha n times the conditional expectation of Mn plus 1 given Fn. And by our assumption in A3, we know that this is 0, and hence alpha n times this will be 0, and hence this whole term will basically be zeta n. So, what we have managed to show is that the conditional expectation of zeta n plus 1 is zeta n, where all these equalities are presumed to hold almost surely. So, from this, you know, relation one can conclude that the sequence zeta n is actually a martingale with respect to the filtration of sigma fields Fn. So, recall that your Mn sequence was a sequence of martingale differences, whereas your zeta n is actually a martingale.





Now, as soon as you see martingale, one can ask, oh, can we make use of the martingale convergence theorem that we had studied before? So, the answer is yes. And we will now see how to make use of this martingale convergence. Is this So, the first thing we have to observe is that because it is a martingale, the expectations will remain the same throughout. In particular, one can show that the expectation of zeta n is actually equal to the expected value of zeta 0. Recall that zeta 0 is 0 almost surely, and hence the expectation is 0 for all n greater than or equal to 0.



So, let me just say 0 here. So, you know, I will end the class here. But in the next class, we will see how we can make use of the fact that zeta n is a martingale sequence, along with our assumptions in A3, to show that zeta n converges. And once we establish that zeta n converges, we will use that fact to show that the tail of the cumulative terms of the noise arising due to this MK plus 1 will have a negligible effect on the behavior of your XNs compared to the behavior of the solution trajectories of your limiting ODE.

So, we will see all these things formally in the next class. Until then, bye and thank you.