STOCHASTIC APPROXIMATION: THEORY AND APPLICATIONS

Dr. Gugan Thope

Department of Computer Science and Automation

Indian Institute of Science

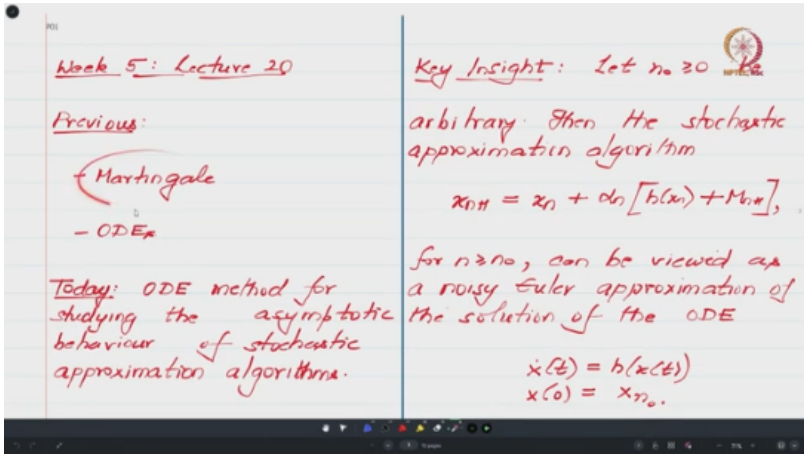
Week 5

Lecture 20

Foundations of Stochastic Approximation – Assumptions and Key Definitions

Hello and Namaste, everyone. Welcome to week 5 of this NPTEL course on stochastic approximation. So, if you have traversed the first 4 weeks of this course, you may have realized that in each week we covered different concepts. Right. And hopefully, you are wondering, OK, when will we combine all of them and study, or use these methods, these different tools, to analyze stochastic approximation algorithms? Well, that combination begins this week.

This week, we will combine all the ideas we have studied so far and use them to analyze stochastic approximation algorithms. In particular, this week, we will look at the asymptotic convergence of stochastic approximation methods. So, let us do a quick recap of what we have done so far. In the third week, we focused on martingales and their convergence. And in the fourth week, we looked at ordinary differential equations.



In today's class, we will discuss something called the ODE method for studying the asymptotic behavior of stochastic approximation algorithms. And the key insight we will rely on is that A stochastic approximation algorithm can be viewed as a noisy Euler approximation of the solution of a certain ODE. So, let us elaborate on this. So, recall that a typical stochastic approximation algorithm has the form given over here.

That is given Xn, Xn plus 1 is alpha of alpha n times H of Xn plus Mn plus 1. Recall that this is to be viewed as a noisy estimate of H of Xn plus 1 where the noise is captured by Mn plus 1. And we will use this update rule for different values of n.

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

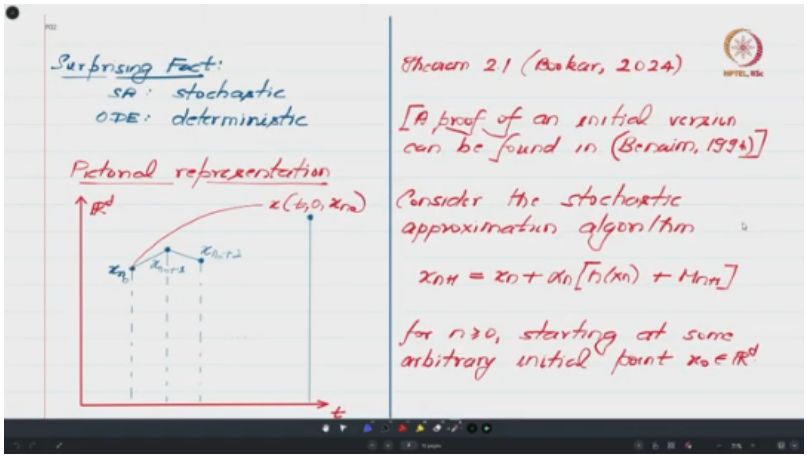
If this expression that is the noise term was not there in the previous class we saw that the resulting update rule can be used as Euler method or an approximate method to find or estimate the solution of this ODE. In particular if you start from n equals n0 that means you start from X n0 for some n0 then this part

That is without the noise can be viewed as an approximation to the solution of this initial value problem given over here.

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

$$X(0) = X_{n_0}$$

So, this ODE will henceforth be referred to as the limiting ODE associated with this stochastic approximation algorithm. And if you notice the surprising fact that a stochastic approximation algorithm is by its very nature stochastic. That means that on different runs of this algorithm, the sequence of iterates that you will get will be different. On the other hand,



The ODE is deterministic in nature. It is deterministic because this H function is well defined. Of course, it is unknown, but it is well defined and there is nothing stochastic about its definition. So, if you start from some any initial point, let us say X n naught and you find the solution to this ODE, then that trajectory will be deterministically defined. So, you know the pectoral representation of what I have just said is the following.

So, let us say you have Xn0, Xn0 plus 1, Xn0 plus 2, and so on. So, these are the iterates generated by your stochastic approximation algorithm. Right, so in this plot, you have time on the x-axis and the value of your iterates on the y-axis. For convenience, we will presume that the y-axis actually illustrates an element of Rd itself. Okay, so at time n0, This is the value of your algorithm, or this is the estimate given by your algorithm. So, this is an element in Rd.

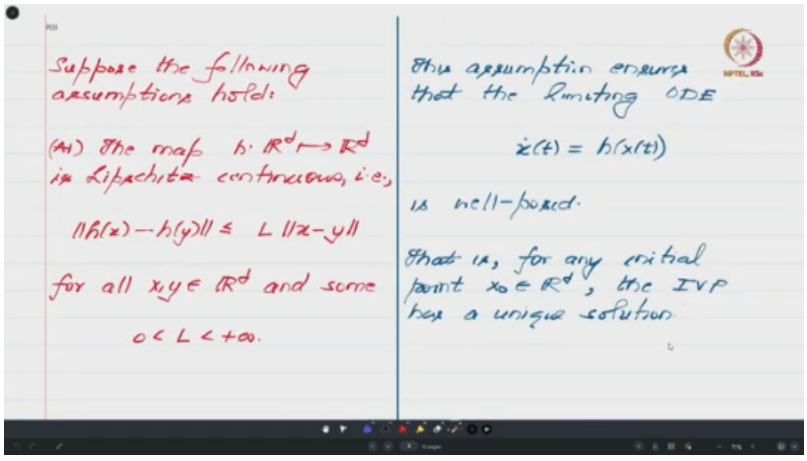
At n0 plus 1, this is the estimate given by your stochastic approximation algorithm, and so on. And let us say we linearly interpolate them, right? Now, because your stochastic approximation algorithm is stochastic in nature, right? This path that I am tracing out will be different on different runs of the algorithm. Is that okay?

On the other hand, let us say you start the solution to this ODE, right? Passing through Xn0, so you start the solution to this ODE, right? Passing through Xn0, so this will be some deterministic trajectory, right? The ODE method's approach, or the insight it relies on, is that this linearly interpolated trajectory can be viewed as a noisy Euler approximation to this solution trajectory of the ODE, right? So, why is it an Euler

approximation? Well, it has this part—that is why it is an Euler approximation. Why is it a noisy Euler approximation?

Well, it is a noisy Euler approximation because of the presence of noise. So, nevertheless, the ODE methods idea is to show that under appropriate conditions, the behavior of this linearly interpolated blue trajectory will very closely resemble the behavior of this red trajectory. And we will formally, you know, discuss this result now, and this result is taken from this book called Stochastic Approximation and Dynamical Systems Viewpoint by Professor Vivek Borkar. In that textbook, from chapter 2, if you look at this theorem called 2.1, you will see the result that we are going to discuss, and this result. by itself, at least in an initial form, appeared in this work which I have referred to as Benign 1996. You can actually find this reference in this textbook itself, right.

So, what does this theorem 2.1 state? It says the following: consider a generic stochastic approximation algorithm of the following form for n greater than or equal to 0, starting at some arbitrary initial point X0 in Rd, right. And suppose the following assumptions hold. So you can see that on this slide there is some text in red and some text in blue. So the text in red actually gives the formal assumption, and the text in blue illustrates or gives an intuitive picture of what this assumption exactly means.



So the first assumption is that the map H that we had over here. This map H, which is a map from RD to RD, should be Lipschitz continuous. By Lipschitz continuous, we require that for all XY in RD, the distance between H of X and H of Y should be less

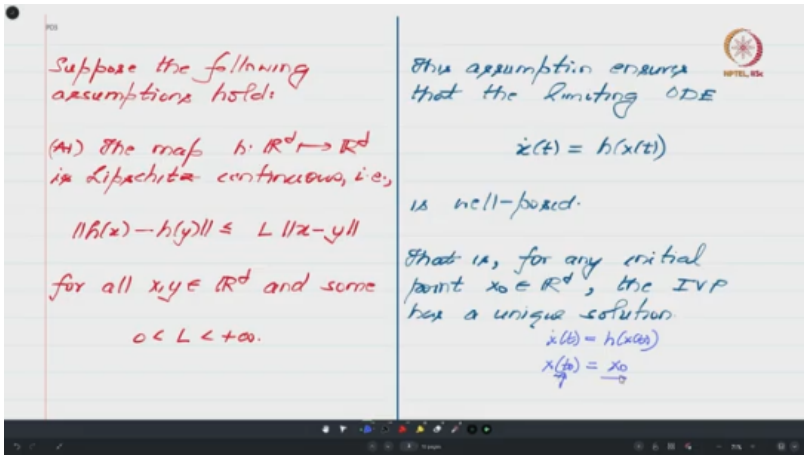
than L times the distance between X and Y, where L is independent of the choice of X and Y and is some finite number.

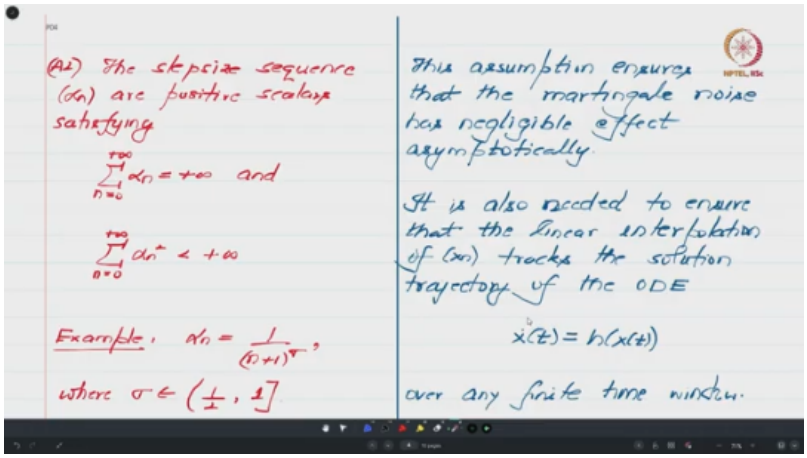
$$h: R^d \mapsto R^d$$

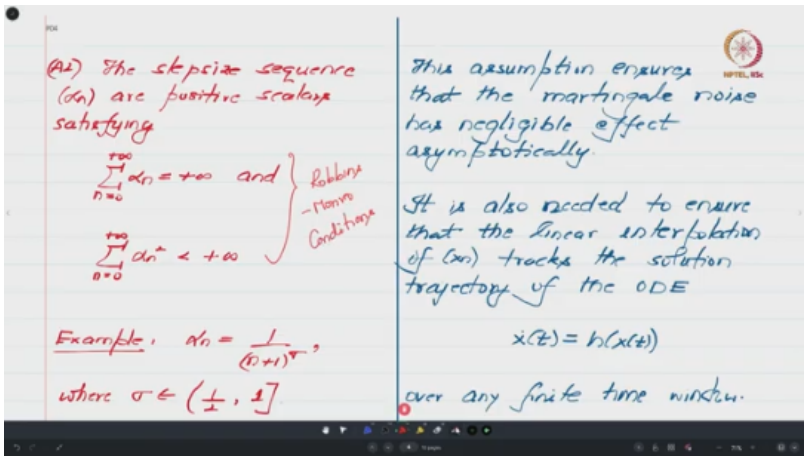
$$||h(x) - h(y)|| \leq L||x - y||$$

Right, and this metric over here could be a metric of your choice, for example, it could be the Euclidean metric. So, now one can ask why is this assumption needed? Well, you know, based on our discussions from the previous week, I hope you are able to guess that one of the reasons this assumption is needed is to ensure that this limiting ODE is well-posed, that is, for any initial point X naught, the initial value problem has a unique solution. By the initial value problem, I mean the following.

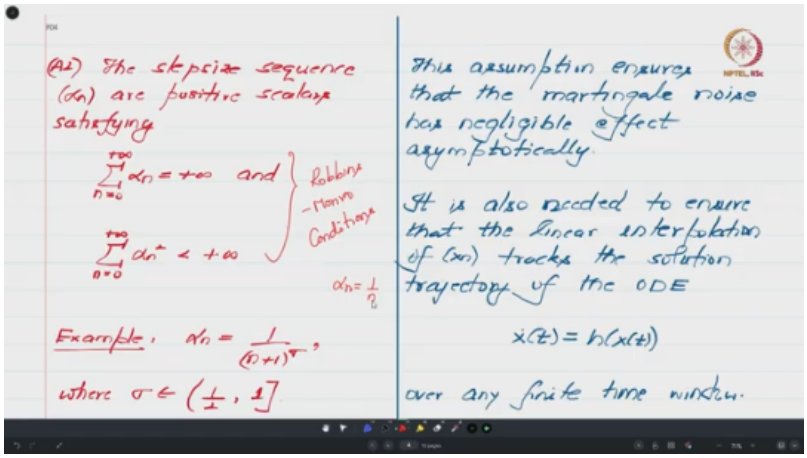
So, if this is your initial point, you can pick your favorite initial point, and let us say I give you an arbitrary initialization X0. Then, I ask you what the solution of this would be. One can show that if H is Lipschitz continuous, then this initial value problem has a unique solution. Then, this result requires the next assumption, which is that the step size sequence defined by these alpha n's should be positive scalars satisfying these two properties. These properties are popularly known as the Robbins-Monro condition. This condition requires that these step sizes







should add up to infinity—that is, they should not be summable. On the other hand, if you take their squares, then they should become summable. That is, the summation of alpha n should be infinity, while the summation of alpha n squared should be strictly less than infinity. An example of such a step size sequence is alpha n equals 1 over n. Now, we know that if you add up 1 over n for different values of n, that sum will be infinity. On the other hand, if you take their squares and add those squares, then that will be less than infinity.



More generally, if you take a step size of this form—that is, alpha n (I should put a plus 1 here), just to ensure that for n equals 0, alpha 0 is a finite number. So, more generally, if you define alpha n to be 1 over (n plus 1) to the power sigma, where sigma is some number between half and 1, then these conditions will actually be satisfied.

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

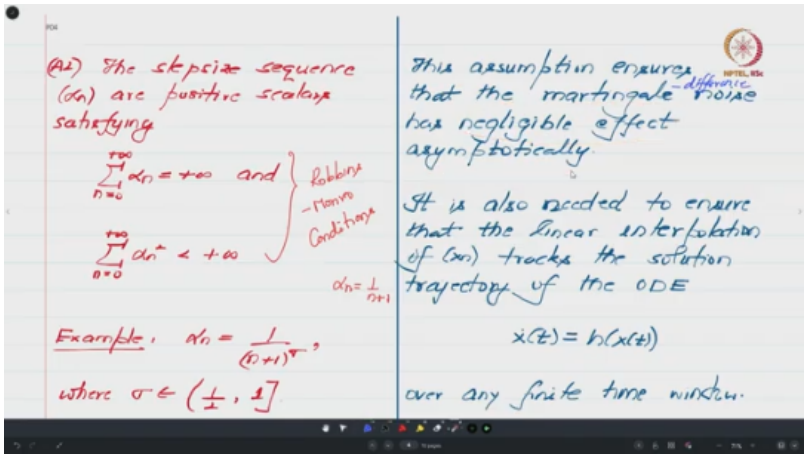
$$\sum_{n=0}^{\infty} \alpha_n^2 < +\infty$$

$$\alpha_n = \frac{1}{(n+1)^{\tau}}$$

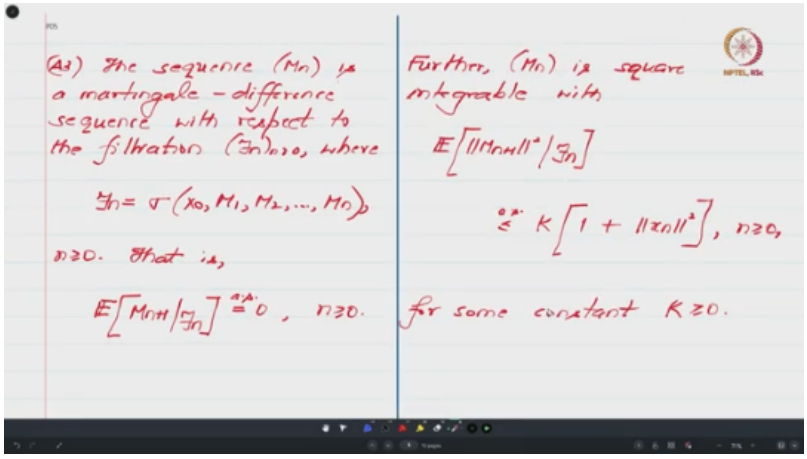
Where,

$$\sigma \in (\frac{1}{2}, 1] \qquad \alpha_n = \frac{1}{n}$$

So, why is this assumption needed? Well, as we will see in later classes, this assumption is needed to ensure that the martingale difference noise (I should say martingale difference noise) has a negligible effect asymptotically. We will also see that this condition

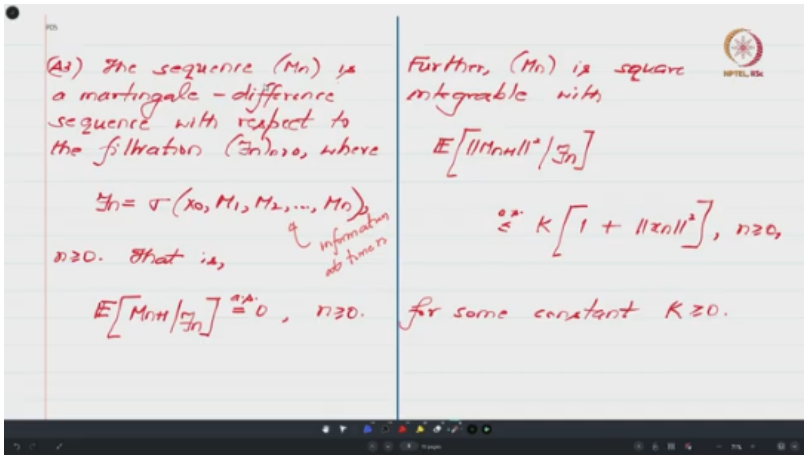


Ensures that the linear interpolation of XNs tracks the solution trajectory of the ODE given by this as your time goes to infinity. So, we will see what all these statements mean soon, but this is at a broad level why this assumption is needed. The next assumption that we need is that this noise sequence be a martingale difference sequence. This is what ensures that certain terms will form a martingale, and we would be able to invoke some of the results that we studied as part of our martingale convergence discussions. So, what this assumption requires is that the sequence Mn be a martingale difference sequence with respect to the filtration Fn.



where the nth term in this sequence is the sigma field defined by these elements, that is X0, M1, M2, all the way up till Mn. So, M1 is the Mn for n equals 1, M2 is the value of Mn for n equals 2, and so on and so forth. So, you can view this as the information at time n. So, we require that this sequence be a martingale difference sequence with respect

to this filtration, that is, we require that the conditional expectation of Mn plus 1 given the information until time n to be almost surely 0, and this should be true for all n greater than or equal to 0. So, that means we are sort of imposing some conditions on the noise.



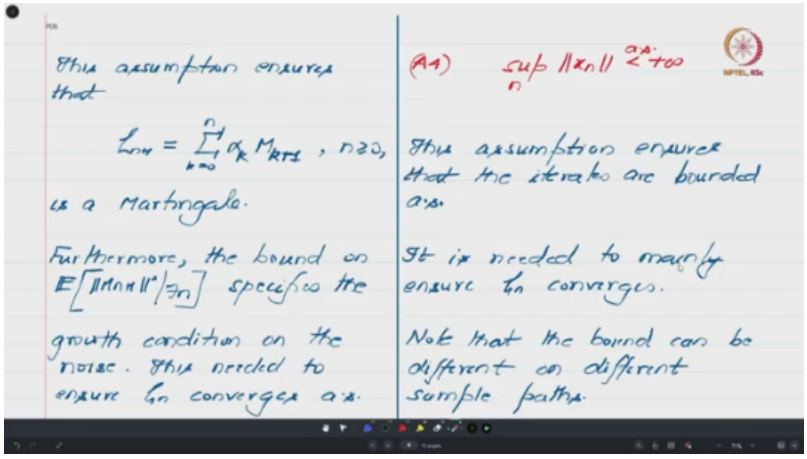
So, one can ask why do we need this assumption and how practical this assumption is. So, we need this assumption to be able to show that the cumulative effect of noise is asymptotically negligible; that is why we need it, and one can view this assumption to be, in some sense, a generalization of zero mean i.i.d. random variables, right? So, the simplest noise assumption that you can make is that your noise is zero mean and i.i.d., that is, they are independent across the values of n. So, a martingale difference sequence is actually a generalization, meaning it allows for more sophisticated noise sequences as well that could have dependencies across N. We only require that the conditional expectation be 0 for all values of N. Additionally, we require that this Mn noise sequence satisfy some growth condition, that is given over here.

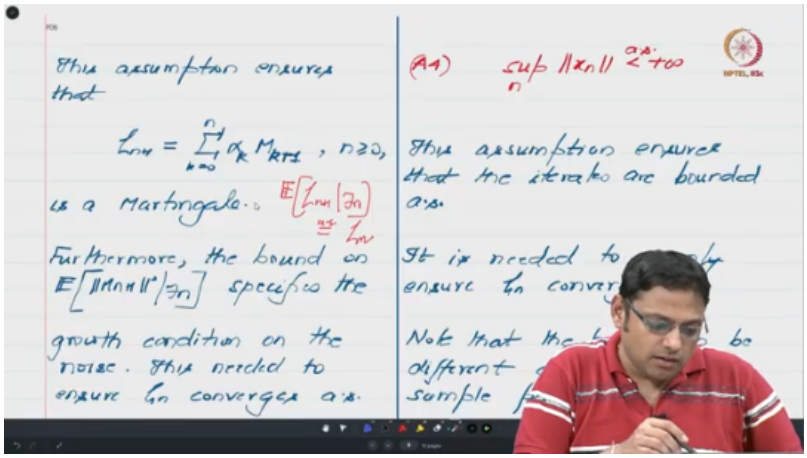
So, let us try to understand this growth condition. So, this condition says that the square of the norm or the expectation of the norm square of mn plus 1 given fn should be upper bounded by some constant k, which is independent of n, times 1 plus norm xn square. So, what this condition allows us to do is that When Xn is large, okay, this noise can have a large conditional expectation. On the other hand, when Xn is sufficiently small, we require that the conditional expectation be upper bounded by a constant, right?

So, I mean, in particular, if your noise for whatever reason was bounded, then that would satisfy this condition. However, this, you know, condition also allows for noise sequences

that actually grow with the value of Xn. So, you can think of, you know, you make a bigger error—you are allowed to make a bigger error when the value of Xn is large. So, in some sense, that is what is captured by this growth condition. And again, why do we need this assumption?

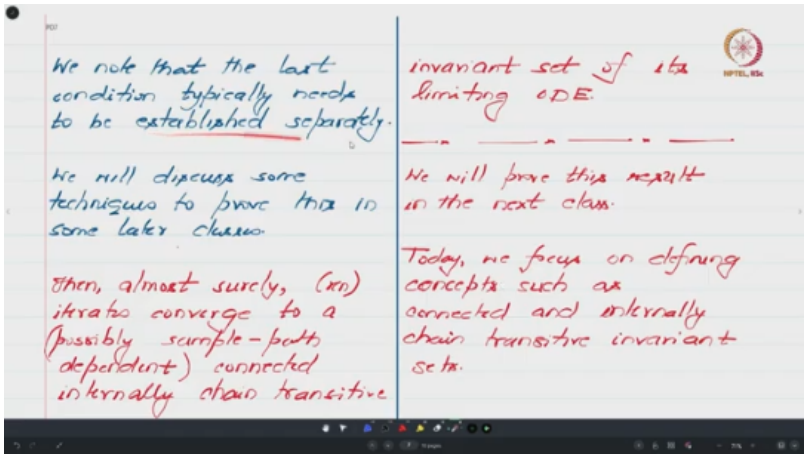
Well, you will see that if I look at this cumulative noise effect. So, alpha k mk plus 1 is the noise that we inject at time instance k. So, if you think of the cumulative noise that has been injected from time 0 to n, that will be summation alpha k mk plus 1. So, because you know this MN sequence is a martingale difference sequence, one can show that this Zeta N sequence actually is a martingale, meaning its conditional expectation does not change with N. In other words, the conditional expectation of Zeta n plus 1 given Fn can be shown to be Zeta n, right? And because this condition holds, one can conclude that your Zeta n is actually a martingale sequence, not a martingale difference—it is a martingale sequence, right. And we will see that, you know, this growth condition on this, you know, square of this Mn plus 1 square, this will be needed to show that your Zeta n converges almost surely.





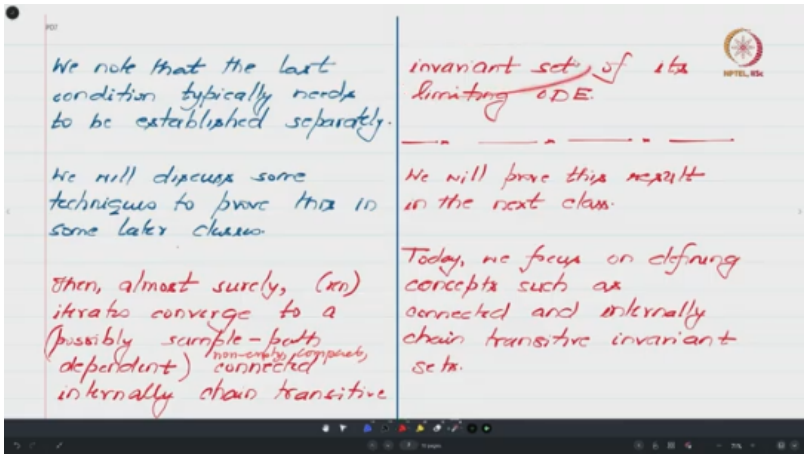
How we will show that will unfold in subsequent lectures, but you know this is just to give you a heads up on why this assumption is needed. And the last assumption that the result needs is that your iterates should be almost surely bounded, right. So, I would like to highlight that, you know, here we are saying that it should be almost surely bounded. This does not mean that there is a uniform bound across different sample points.

On one sample point, the bound could be 5; on another sample point, the bound could be 1 trillion, and so on and so forth. All we require or insist is that on the collection of sample paths where this supremum is bounded, that collection has probability 1. Again, this assumption will be needed mainly to show that your noise sequence Zeta n converges almost surely. Now, typically, this last assumption will need to be established separately; that is, you know, we will have to rely on the problem settings and so on and exploit some additional knowledge that you may have about that specific problem in order to conclude that A4 holds, right? And in some later classes, we will see or we will discuss some techniques to establish this last condition. So, let us quickly summarize the four assumptions.



The first assumption states that this function H should be Lipschitz continuous. The next condition states that the step size sequence alpha n should not be summable but should be square summable. The third condition requires that your noise sequence Mn be a martingale difference sequence and should satisfy some growth condition of the following form, and the fourth condition requires that your iterates be almost surely bounded. Then the conclusion of the result is that if all these four conditions hold, then the iterate sequence Xn converges to a possibly sample path-dependent, connected internally chain transitive invariant set of its limiting ODE.

Maybe I should add a few more things here. It converges to a non-empty, compact, connected, internally chain transitive invariant set of its limiting OD. So, let us just pause for a few seconds to understand what this result is trying to say. So, it had a bunch of assumptions. Based on those assumptions, what this result is saying is that your Xn sequence will converge.

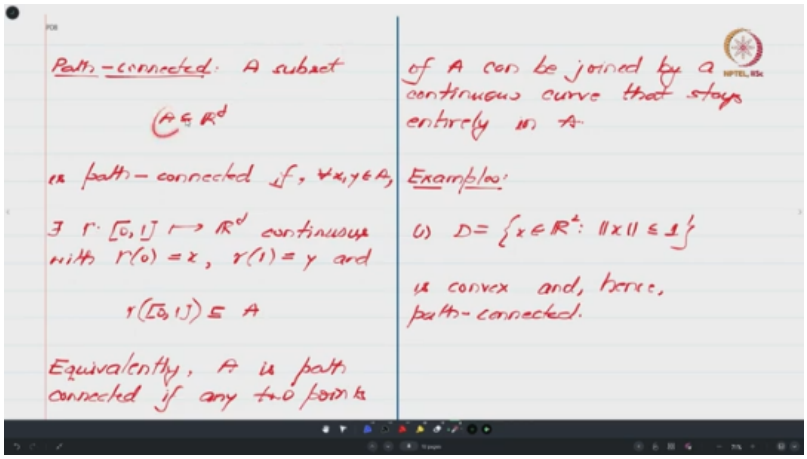


That is the first conclusion that this result is stating. The next thing is that it is characterizing the nature of the limiting set. In particular, it is saying that the limit to which it will converge—first of all, notice that this limit need not be a point, right? It can be a set. And what this result says is that this set is a special set from the perspective of the limiting OD.

In what sense is it special? Well, it is an invariant set with regard to the limiting OD, right? On top of that, this invariant set has some additional properties: first, it is guaranteed to be non-empty; second, it is guaranteed to be compact (that is, closed and bounded); third, it is guaranteed to be connected—in particular, path-connected; and lastly, it is guaranteed to be internally chain transitive. This concept I will soon define. So, all these properties together ensure or help us in identifying or, in some sense, discarding most of the invariant sets of your limiting OD. So, we want a set that is internally chain transitive, we want an invariant set that is non-empty, we want an invariant set that is compact, we want an invariant set that is connected.

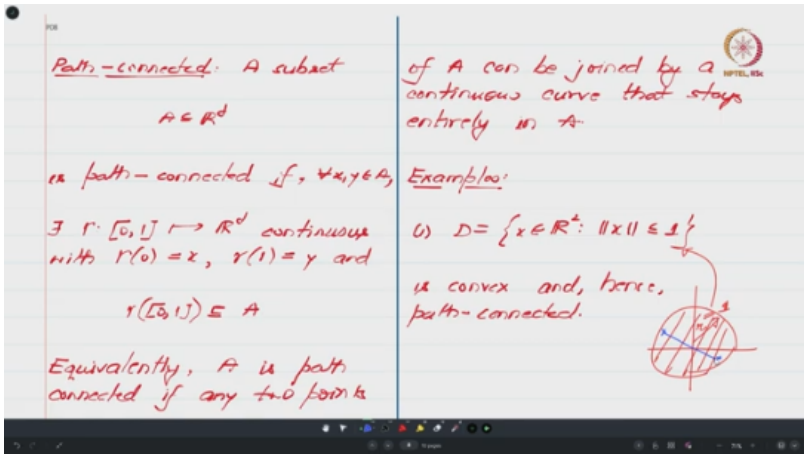
So, if you have an invariant set of your limiting OD which does not satisfy—even if it does not satisfy even one of these properties—then we can, you know, discard that invariant set from the potential limit sets. Is that okay? So, that is, in some sense, helping us narrow down the list of invariant sets to which stochastic approximation algorithm iterates can converge. Right, so we will, you know, discuss the proof of this result from the next class onwards. Today, what we will do is focus on defining these concepts, such as connected and internally chain transitive. Right? So, these other concepts—that is,

non-empty and compact—okay, these are things that you should know. If you do not know, I request you to look it up, you know, from any favorite source of yours. Right? So, what do I mean by path-connected? So, a subset A of Rd is said to be path-connected if, for any two points in A, there exists a function gamma which goes from [0, 1] to Rd, which is continuous, and at time 0 takes the value x and at time 1 takes the value y.

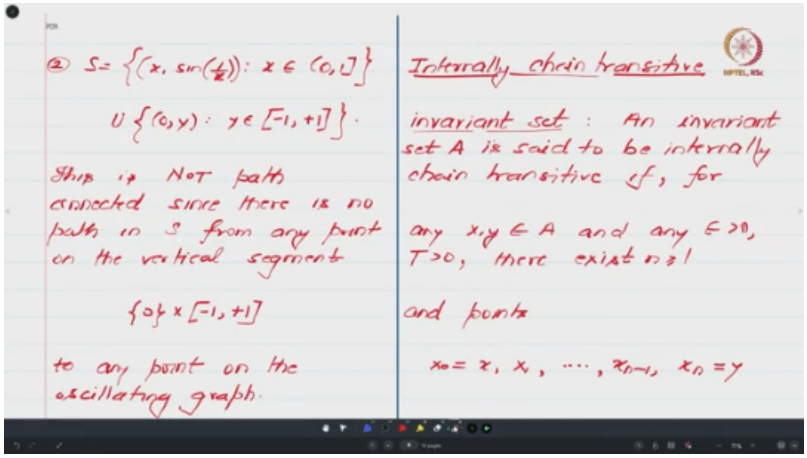


For every t between 0 and 1, gamma takes the value in A, right? So, this is like the technical definition of what path-connected means. At a very high level, A is said to be path-connected if any two points of A can be joined by a continuous curve that stays entirely in A. So, let us look at an example to decipher this part a bit more in detail. So, let us say we are in R2 and let us collect the set of points in R2 whose length is less than or equal to 1.

So, for example, if this is your X and Y axis, this is the collection of points where the length is less than or equal to 1. So, this is R equals 1. So, you collect all these points which lie within this disk—that is what will be part of this set. Now, this set is also actually convex, and hence, if I take any two points within this set, I hope you agree that I can actually use a straight line and guarantee that such a straight line will actually lie within this set. So, because of this reason, this set is actually path-connected.

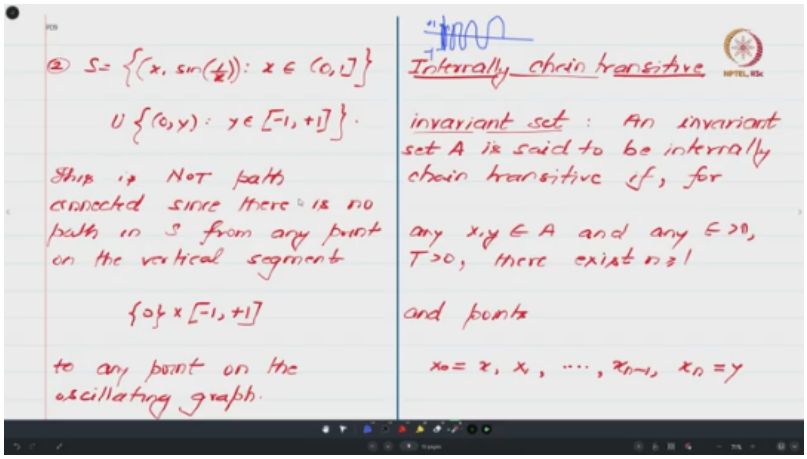


And here is an example of a set which is not path-connected. So, what is this set? This set basically includes all points of this form. So here, x is actually a number between the interval that is open at 0 and closed at 1. And the second coordinate, or the y-coordinate, is sine of 1 over x. So, you basically collect all such points for different values of x which lie between 0 and 1.



And to this collection, you add all elements of the following form. That is, all elements whose x-coordinate is 0 and the y-coordinate lies between -1 and +1. So, if you have to pictorially draw this, this curve will look something like this. So, this is like the sine part. Okay, and as your x gets closer and closer, the frequency of your sine function will actually keep increasing, so it will have this, you know, nature. And very close—I mean, at 0—this part which goes between -1 and +1, that portion is also included in this set S.

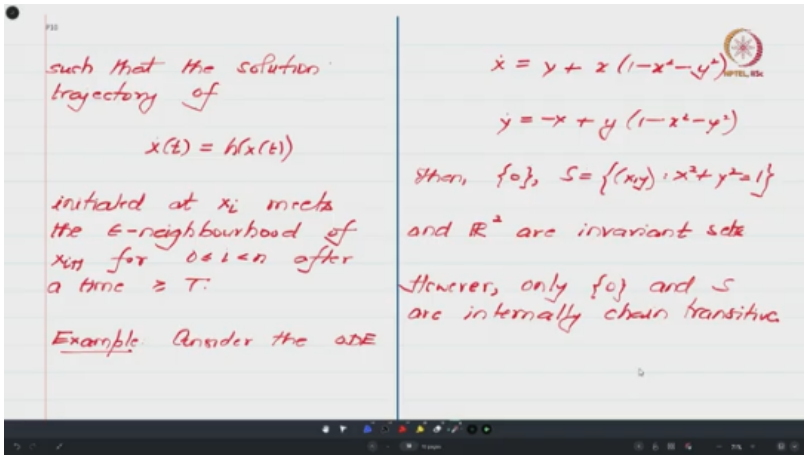
And one can show that such a set is actually not path-connected. It is not path-connected because there is no path from a point on this segment that you have written over here. There is no path that starts from here and goes to the sinusoidal part of the graph. Or the oscillating part of the graph. So, you can check why that is the case, and because of that reason, one can conclude that this set S over here is actually not path-connected.



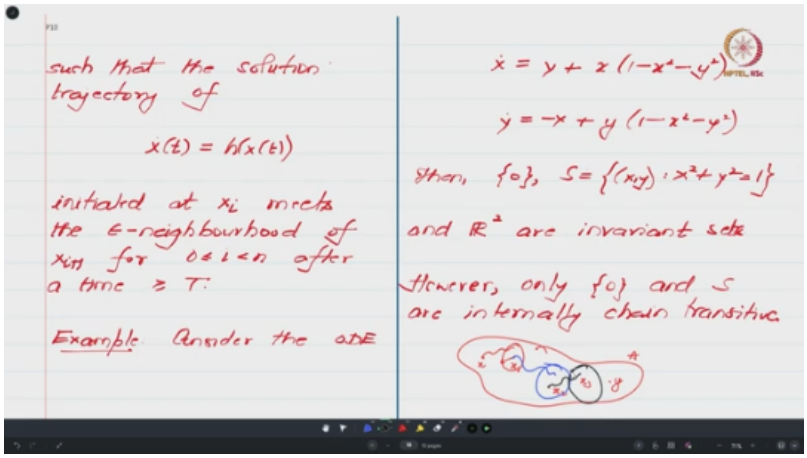
So, this was one of the characterizing features of the limiting set, which is that it should be path-connected. The other characterizing feature that the theorem guarantees is that of being internally chain transitive. So, what is the meaning of an internally chain transitive invariant set? Well, an invariant set A is said to be internally chain transitive. If for any x, y in A and any epsilon greater than 0 and t greater than 0, there exists

a certain set of points that satisfy some condition. First, let us look at this condition in a technical detail fashion, and then we will try to interpret it intuitively. So, what this definition says is that you give me any two points in A and you give me an epsilon and t greater than 0, then we require that there be a little n greater than or equal to 1 or a finite number and a collection of n plus 1 many points, where the 0th point in this collection is actually at x, which is this x, and the last point over here is y. The in-between points, we will label them as x1 to xn minus 1, and this collection of points should satisfy the following property: that is, the solution trajectory of this ODE initiated at xi should meet the epsilon neighborhood of xi plus 1 for every i. This thing should hold, and this solution trajectory should meet this epsilon neighborhood for a time greater than or equal to t after a time that is greater than or equal to t. Okay, so let me sort of pictorially explain what is

happening. So, let's say you have a set A right, and let's say we have been given these two points x and y. What we require is a bunch of points, in some sense, between x and y. So, let's say we have these three points, and what this collection of points should satisfy is that if I start a solution trajectory of the limiting ODE from x after a time which is greater than or equal to t



this solution trajectory should reach the epsilon neighborhood. Similarly, if I start the solution trajectory from x1, then after a time which is greater than or equal to t, this blue solution trajectory should actually reach the epsilon neighborhood of x2. Similarly, if I start a solution trajectory from x2, then after a time larger than or equal to capital T, the solution trajectory should be in the epsilon neighborhood of x3, and so on and so forth. So, if you are able to connect x to y in this very loose sense for every x and y, we will say that the set A is internally chain transitive. And one can check that for this ODE, which we had discussed in the previous class, the origin



The points on the unit circle and R2 are the only invariant sets, and among them, only this set and this set. These two non-empty sets are the only ones that are internally chain transitive. In other words, this set R2 is not an internally chain transitive invariant set. Right. And hence, in that result, one can actually discard R2 from the potential limit set.

So, we can say that on certain sample paths, maybe the iterates go to 0, and on certain sample paths, we can show that the iterates actually go to this unit circle. So, with that, we come to the end of today's class. Hope to see you in the next class. Thank you and bye.