STOCHASTIC APPROXIMATION: THEORY AND APPLICATIONS

Dr. Gugan Thope

Department of Computer Science and Engineering

Indian Institute of Science, Bangalore

Week 8

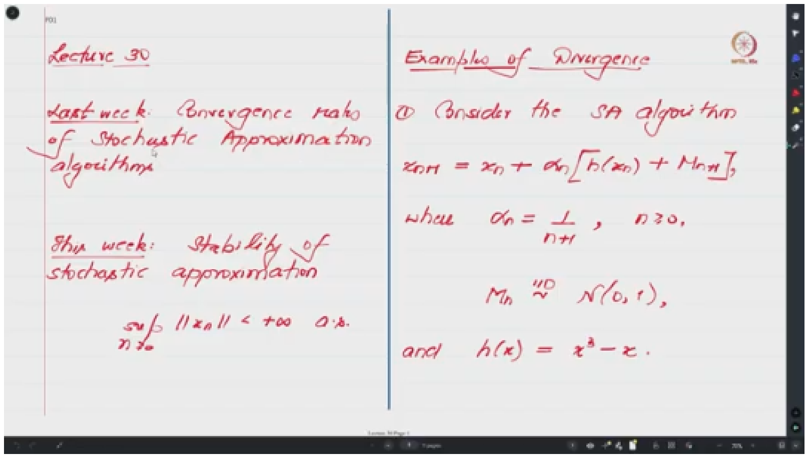
Lecture 30

Stability Requirements in Stochastic Approximation

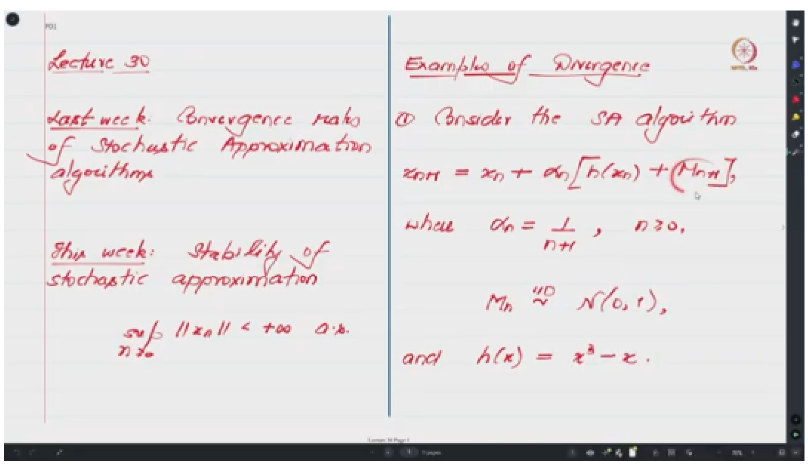
Hello and Namaste everyone. Welcome to lecture 30 of this NPTEL course on stochastic approximation. This is week 8 of this course and during this week we will look at this concept called as stability of stochastic approximation algorithms. So, if you recall in the previous like in the 5th and 6th week we were looking at convergence of almost sure convergence of stochastic approximation algorithms and there there was this assumption labelled A4 where we presume that the iterates are almost surely bounded right and under that presumption we looked at the asymptotic convergence of stochastic approximation algorithms right.

In this week we will discuss the how to you know verify this assumption and also in today's class we will see why you know this condition is in some sense challenging to work with. Is that okay? In fact, in today's class, we will show an example where this assumption does not hold, right? And in the next class, we will discuss one sufficient condition which ensures that this assumption A4 holds almost surely. So, with that in place, let us begin our formal discussion.

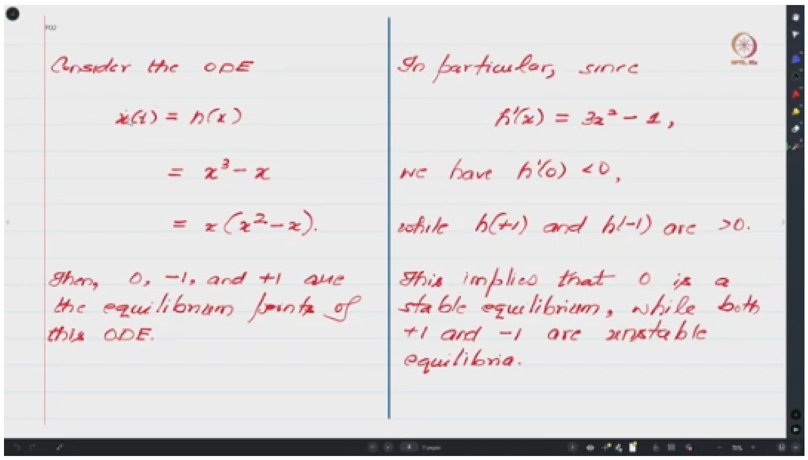
right so last week we looked at convergence rates of stochastic approximation algorithm and during the two weeks before that we were looking at almost sure convergence of stochastic approximation algorithms and in today's class we will discuss this notion of stability of stochastic approximation algorithms by stability we mean that you know the supremum of the norm of your stochastic approximation iterates is almost surely finite. Now, you know to understand why this is something that is not for free, in today's class what we will do is we will discuss an example of a stochastic approximation algorithm for which this condition fails to hold. And in the next class, we will discuss a sufficient condition where this condition is guaranteed to hold. So, towards discussing the example where this condition does not hold, let us look at the stochastic approximation algorithm Xn plus 1 equals Xn plus some step size alpha n times H of Xn plus Mn plus 1, where your step size sequence is of the form 1 over n plus 1.



And this martingale difference sequence that is Mn is very simple. It is just independent and identically distributed Gaussian 01 random variables, right? And what we will do is we will presume that this h of x has the form x cube minus x. So, the key thing to note here is that this function H, right, is a polynomial. In particular, it is a, you know, there is a cubic term over here which makes this polynomial nonlinear.



Is that okay? And we will see what is the consequence of this. In particular, if your noise has this property, what can we say about the magnitude of your iterates? And it goes without saying that this stochastic approximation algorithm is on the real line which means that every iterate over here is an element in R. So one can look at the limiting OD corresponding to this stochastic approximation algorithm and conclude that the limiting OD is x dot of t equals h of x where h of x is x cube minus x.

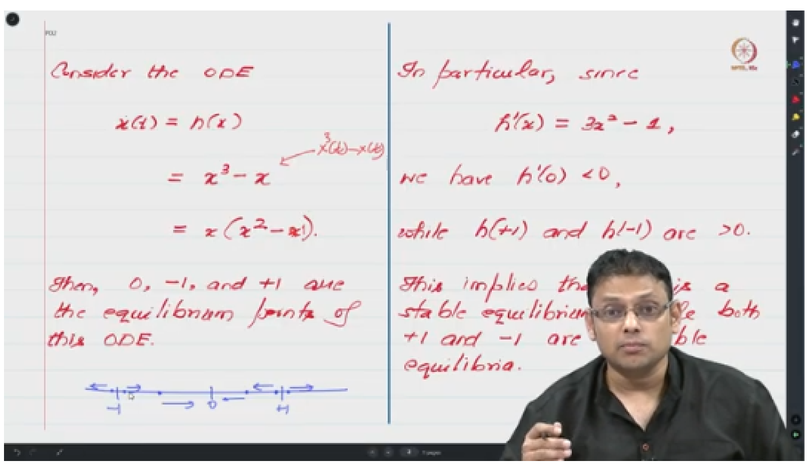


Again, I would like to highlight that, you know, on the right hand side, ideally I should write x of t cube minus x of t. However, for notational convenience, I am dropping this round bracket t over here. Hence, the ODE is x dot of t equals x cube minus x. And I can, you know, take x in common, right? And I would end up with x times t. And from this, we can see that 0, minus 1 and plus 1 are the equilibrium points of this ODE. In particular, if you take the derivative of this expression, we would end up with 3x square minus 1.

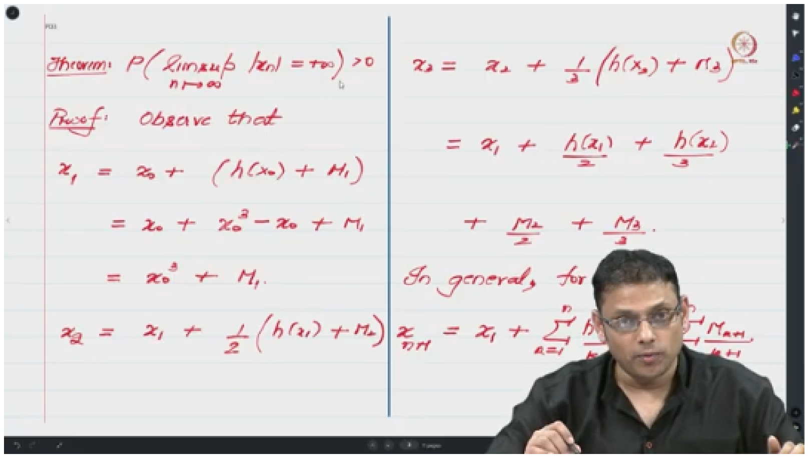
from which we can see that at x equals 0 this expression is negative and at plus 1 and minus 1 this expression is positive and hence one can conclude that among the three equilibrium points 0 is an asymptotically stable equilibrium point with respect to this ODE and while both plus 1 and minus 1 are unstable equilibrium. So what does that mean? If I consider this real line, so here is 0, here is minus 1 and here is plus 1, right? So if I start my ODE solution from somewhere over here, right?

These ODE solutions will actually go towards ODE. the origin. In other words, if you start anywhere in the neighborhood of the origin, the solution trajectories will go towards the origin. On the other hand, even if you start very close to plus 1, then the solution trajectories will go away from plus 1. In other words, plus 1 is not a stable equilibrium point.

Similarly, if you start very close from minus 1, the solution trajectories will go away from minus 1 again indicating that minus 1 is a unstable equilibrium point is that okay so given this situation in particular the fact that if you are above plus 1 or below minus 1 your solution trajectories can you know go away from plus 1 and hence towards infinity we will use this aspect and show that the stochastic approximation iterates can actually blow to infinity in this scenario when the noise is not in the right way right so notice that you know the invariant sets for this OD is 0 minus 1 plus 1 and of course the real line and so on and so forth right and you know the compact invariant sets are minus 1 0 and plus 1 that is something that one can verify separately right so if we could somehow manage to show that the iterates are almost surely stable, then as a consequence of the theorem that we discussed in weeks 5 and 6, we could have concluded that the limit of your xn iterates would be 1 of minus 1, 0 or plus 1, right? Now, in this discussion, what we are going to say is that unfortunately, the iterates are not almost surely bounded.



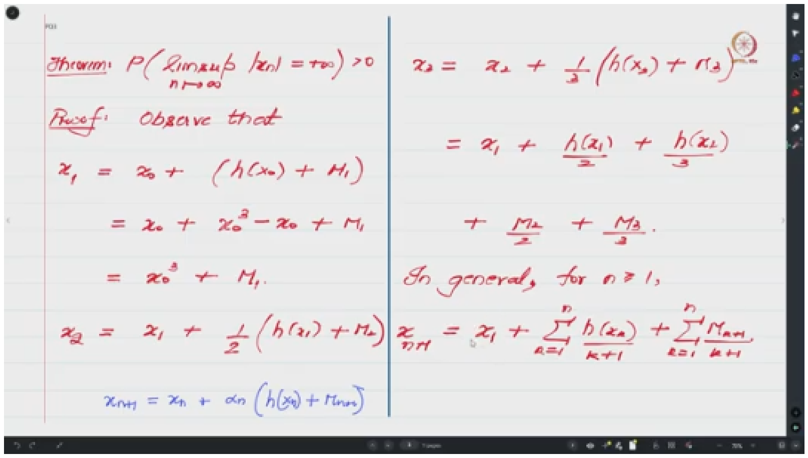
Instead, there is a positive probability that the iterates can race off to infinity, right? So, this is what we are formally going to show. We are going to show that the probability that the magnitude of your Xn is plus infinity, right, has positive probability, right? So, we will now formally show why that is the case. So, recall our update rule.



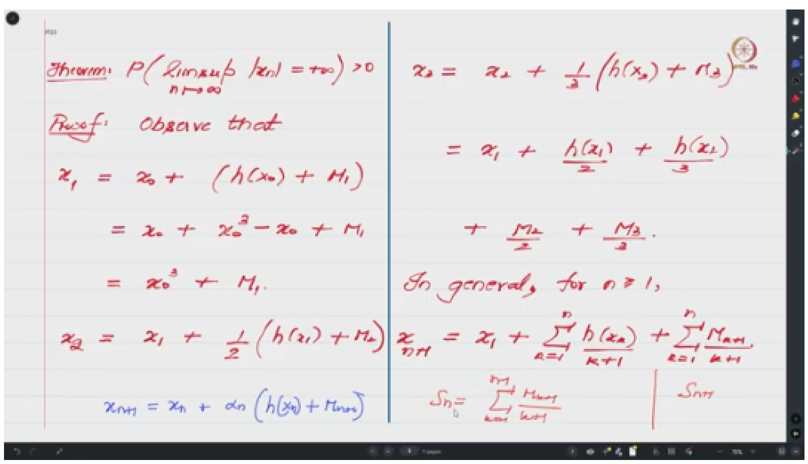
So, our update rule is Xn plus 1 equals Xn plus alpha n H of Xn plus mn plus 1. So, if I substitute n equals 0 in this update rule, one can see that we would end up with X1 equals x0, and alpha n we have presumed is 1 over n plus 1. So, if I substitute n equals 0 there, we would end up with 1 over here, and this is h of x0 plus m1, right? So, let us substitute the form of h, which is x cube minus x.

And in the case of H of X naught, we will end up with X naught cube minus X naught. So, this minus X naught and X naught will cancel off, and we will end up with X naught cube plus M1. So, for n equals 0, this update rule has this form that is given over here. So, let us keep this separate and let us come up with some relation that xn iterates satisfy for n greater than or equal to 2. In particular, if I substitute n equals 1 here, I would end up with x2 equals x1 plus half because this is alpha 1.

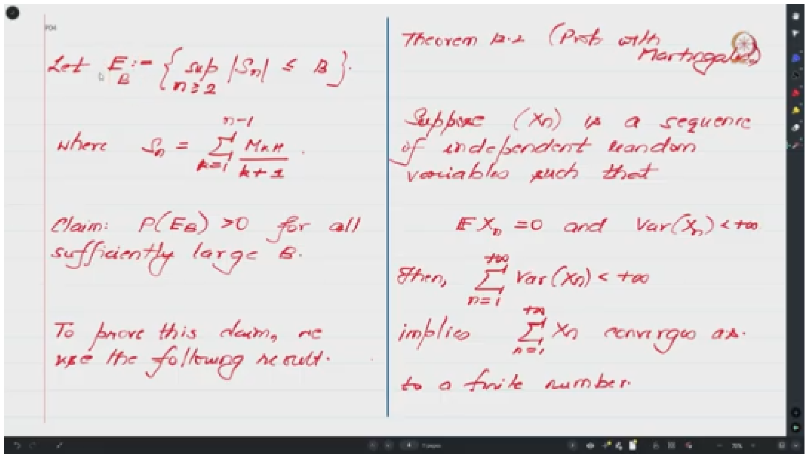
times H of X1 plus M2. Now, keeping this relation as it is, one can see that X3 satisfies this relation, and if I substitute the relation of X2 over here, we would see that X3 equals X1 plus whatever we have, that is H of X1 by 2 plus H of X2 by 3 and M2 by 2, which is what I have written here, and M3 by 3, which is what we get over here. So, in other words, X3 equals X1 plus the sum of these two quantities plus the sum of these two quantities. And if we proceed along similar lines, one can see that for any n bigger than or equal to 1, your Xn plus 1 has this relation, that is, it is H of X—sorry, X1 plus the sum of H of XK over K plus 1 plus the sum of MK plus 1 over K plus 1, right?



So, one can immediately verify this expression, right? So, now what we will do is whatever we have over here, let us call this as SN plus 1, or in other words, let SN be the sum of k equals 1 to n minus 1 of MK plus 1 over K plus 1. So, notice that I have changed the upper index to n minus 1 so that I can call this as SN. And let us define this event Eb.

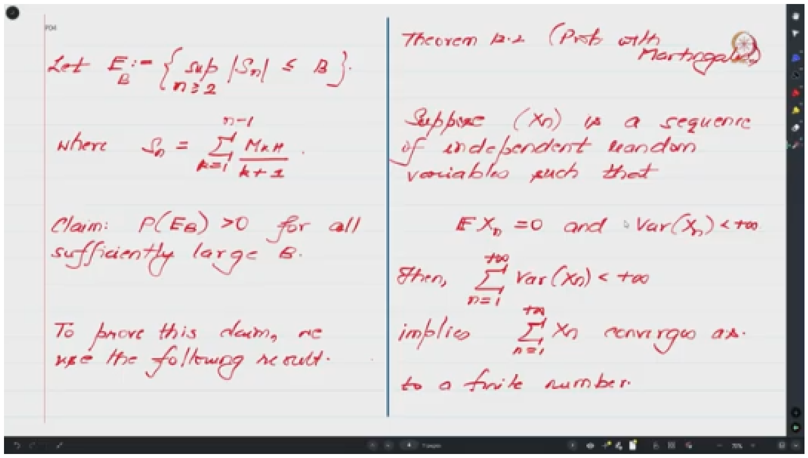


where Eb is the event where the supremum of the absolute value of SN for n bigger than or equal to 2 is less than or equal to b. So, we will presume that b is some real number, a fixed real number, and we will define Eb to be the event where the supremum of SN over n bigger than or equal to 2 is less than b. So, as I change b, this event will change. And we will work with this definition in the rest of the analysis. And the goal now is to show that the probability of this event is actually strictly positive for all sufficiently large B. So now we will focus on showing that this event has positive probability for all sufficiently large B, and you will soon see why such an event having positive probability—I should not say large probability, but having positive probability—is good enough to conclude our main theorem, which is that our iterates, you know, race off to infinity with positive probability. So, to prove this claim,



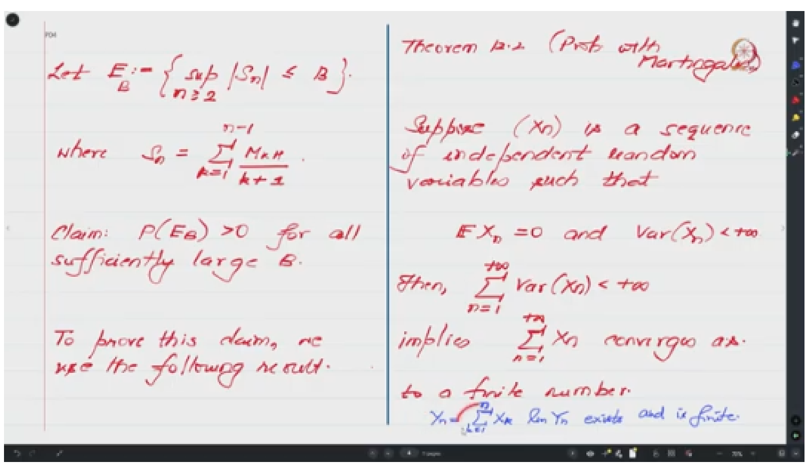
Okay, we will make use of the following result. You know about sequences of independent random variables, and the details and the proof of this result can be found in this textbook called Probability with Martingales, which we were referring to when we were going over probability. In particular, if you look at Theorem 12.2 in this textbook, then you will find the details. So, what does this theorem say? It says that suppose Xn is a sequence of independent random variables such that the mean of these random variables is 0 and

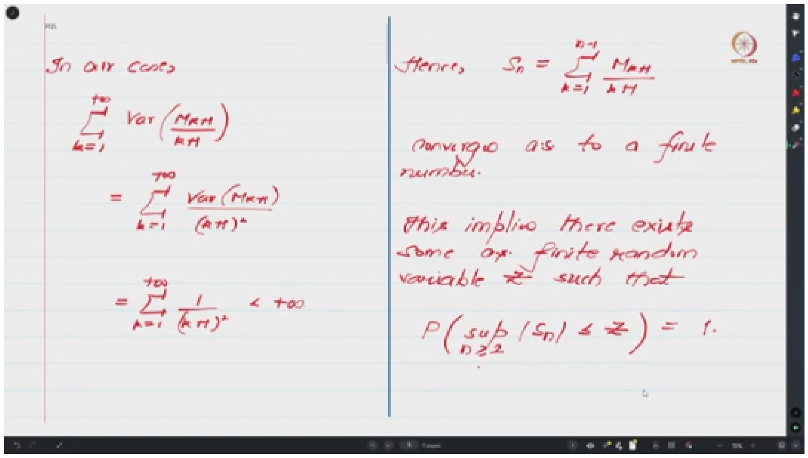
each random variable has finite variance, right? So, we are not presuming here that they are identically distributed, right? So, suppose each of these random variables has finite variance. Then the result says that if the sum of the variances is finite, right? Right, then the sum of these random variables will converge almost surely, right? So, what does that mean?



It means that if you look at the partial sums—okay, partial sums, summation x1, so let us say xk, k equals 1 to n, right? And let us call this, you know, yn. The claim is that the limit yn exists and is finite. This is the claim.

So, let us go over the claim one more time. It says that if the sum of these variances is finite, then if you look at the partial sums, then as n goes to infinity, the limit of these partial sums exists and is finite. So, now let us see if in our case, we can invoke this result, right? So, recall that in our case, this Sn was, you know, this sum of mk plus 1 by k plus 1, k equals 1. 0 till n minus 1.





So, let me just confirm if this is what we had, sorry. So, we had actually started—we were starting from this x1 over here. Hence, this also starts from x1 here. So, let me correct this. So, this also starts from k equals 1.

So, we had some expression like this. So, you can see that you have a sum of terms like this. So here, whatever your Xn—capital Xn—is, that role is played by this mk plus 1 over k plus 1. And let us look at the sum of variances of these expressions. So, the variance of a random variable divided by a scalar is the square of this scalar—1 over the square of this scalar—times the variance of this quantity.

And if you recall, we had presumed that your mk plus 1 or in general mk is Gaussian with mean 0 and variance 1. Hence, from this fact, one can conclude that the variance of mk plus 1 is actually 1. And hence, this sum translates to sum of 1 over k plus 1, the whole square, right? And we know that, you know, the sum of squares of the inverses of integers is actually finite. And hence, one can conclude that this sum is actually finite, okay?

And from the previous result, one can then conclude that if you look at these partial sums, then these partial sums Converge almost surely to a finite number. So, what does this imply? So, of course, when I say it converges to a finite number, it does not mean that on different sample points, the limit will be the same. All it means is that on one sample point, maybe the limit is, you know, some finite number, let us say A and on some other sample point, the limit is some other finite number, let us say B, right?

So, on different sample points, the limit could be different, but all we are saying is that except for a probability, except for a set of probability 0 other than that this partial sum actually converges to a finite number. So this can be formally restated as follows that there exists or I should not say restated I can say that whenever the partial sums converge almost surely One can conclude that because this convergence is to a finite number, there exists an almost surely finite random variable Z such that the supremum of these SNs is less than equal to Z with probability 1. So let me just elaborate a bit over here.

So if you have a sequence that converges. So imagine on the x axis you have n and on the y axis you have the value of the sequence. So you have a sequence let's say that converges to this place. So you could have some large values, then maybe some small values and eventually you sort of converge to this place. So which means that for all large enough n, the values will be very close to your limiting value.

And before that, since you have only finitely many values, max of those finitely many values is trivially finite. Hence, if you combine both these statements together, one can conclude that on different sample points because you are converging to a finite number, the value along the trajectory will be bounded and you can collect all these bounds on different sample points and call that as this random variable Z and together one can conclude that supremum of these partial sums or supremums of the absolute value of the partial sums is bounded by this almost sure finite random variable Z.

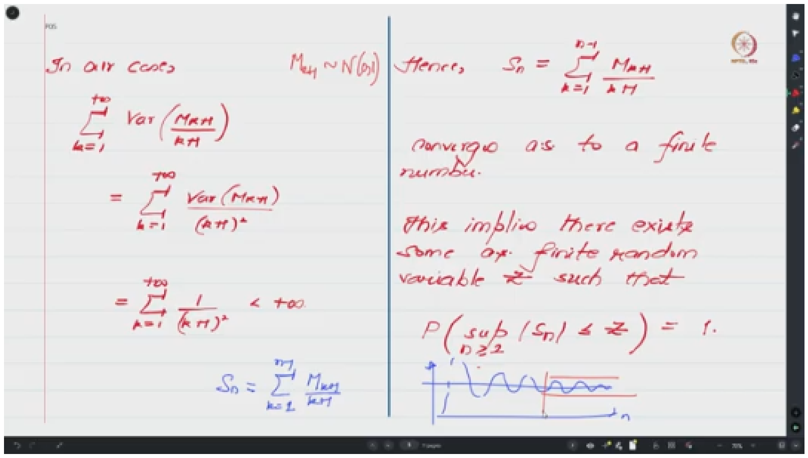
$$\sum_{k=1}^{+\infty} Var\left(\frac{M_{k+1}}{k+1}\right) = \sum_{k=1}^{+\infty} \frac{Var\left(M_{k+1}\right)}{(k+1)^2}$$

$$= \sum_{k=1}^{+\infty} \frac{1}{(k+1)^2} < + \infty$$

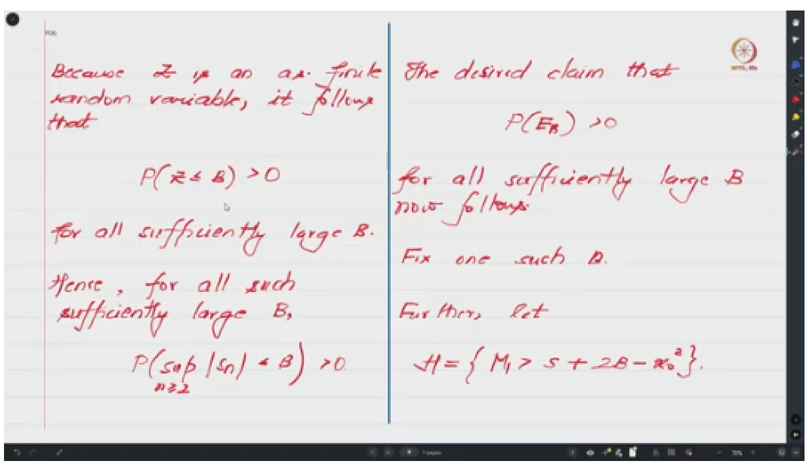
$$S_n = \sum_{k=1}^{n-1} \frac{M_{k+1}}{k+1}$$

$$P\left(\left|S_n\right| \leq Z\right) = 1$$

Now, since Z is almost surely finite, One can conclude that for all sufficiently large B, probability Z less than equal to B is greater than 0. Is this okay?



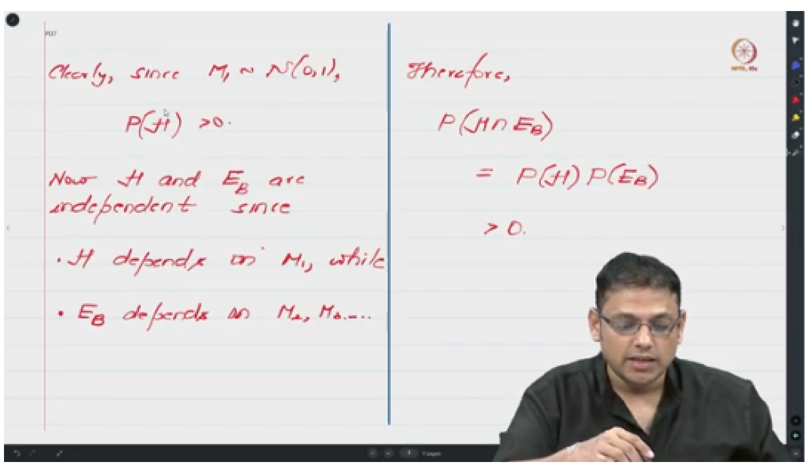
So one can, you know, use the continuity of probability, that property to make such a statement. So, let us make use of such a statement and you know the fact that this is less than b and this is having some positive probability. One can conclude by putting both these statements together that for all sufficiently large b, probability that supremum of Sn is less than equal to b has positive probability. So, what we will do is we will if you remember we had used EB as the notation for this event. Hence, from this statement one can conclude that probability of EB is actually strictly positive for all sufficiently large B. So, in the rest of the analysis let us fix sufficiently large B such that probability of EB is greater than 0.



Right, and in parallel, what we will do is we will define this set H where the value of M1 is greater than little s. Plus, so this should only be b here, little s plus b minus x0 cube. So recall that M1 is the noise that is added when we go from x0 to x1 computation in our stochastic approximation update rule. So let M1 be greater than this, where your s is equal to 1 plus epsilon. So this 1 plus epsilon over here

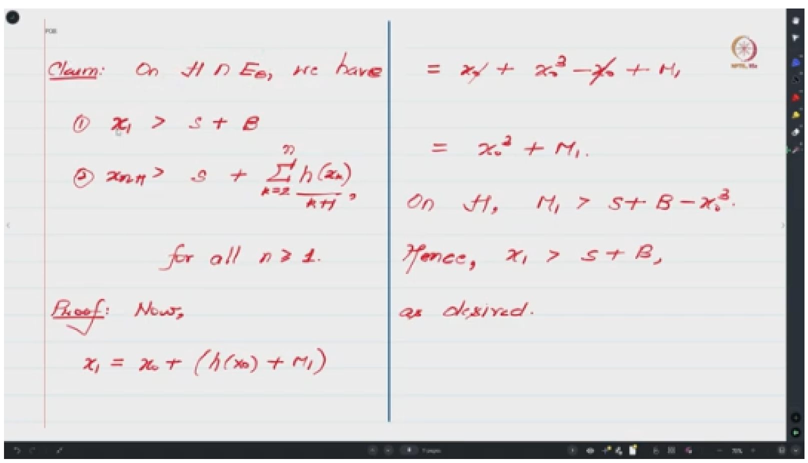
is basically saying that, recall that this plus 1 was one of your equilibrium points and it was unstable. So, let us presume that we are slightly above this unstable equilibrium point for some epsilon strictly bigger than 0. So this just means that we are slightly above this, and you will soon see that this condition, in some sense, ensures that the value of X1 is actually on the right side of your unstable equilibrium point. And because of this unstable nature of this equilibrium point, we will see that the subsequent noise, if it has this property, then along such sample points, your limit or lim sup of Xn will actually be infinity. So, let us verify that part.

Now, one can ask: should such an event have positive probability? Well, recall that M1 is a Gaussian random variable, and once some initialization is given, whatever this value is, this will be some finite number, and because M1 can take any value—recall it is Gaussian—so it can take any value over the whole real line. Hence, once you have such a bound or some finite number, the probability that M1 is bigger than this has positive probability, right? So, that is what I have written over here: that M1 is Gaussian, and hence whatever the value of X0 cube and B, right?



The probability of H will actually be greater than 0, right? Now, I would like to additionally claim that this event H and this event EB are independent. Why is that? Because H depends on M1, while EB depends on M2, M3, and so on. Recall that the definition of this event EB included the supremum of this absolute value of essence, and the supremum went from N greater than or equal to 2.

And if you go back to the definition of Sn, you will see that we started from K equals 1, and hence, in that expression for Sn, we only have expressions involving M2, M3, and so on. So because H and EB are independent, the probability of H intersection EB will equal the probability of H times the probability of EB. We have separately shown that the probability of H is positive and the probability of EB is positive, and since the product of two positive numbers is also strictly positive, one can conclude that this joint event also has strictly positive probability. Now let us see how this fact—that the probability of this joint event has strictly positive probability—helps us conclude that on this event, the lim sup of Xn actually grows to infinity. So the claim is that On this event H intersection EB, that is, you pick any sample point which lies in both H and EB.

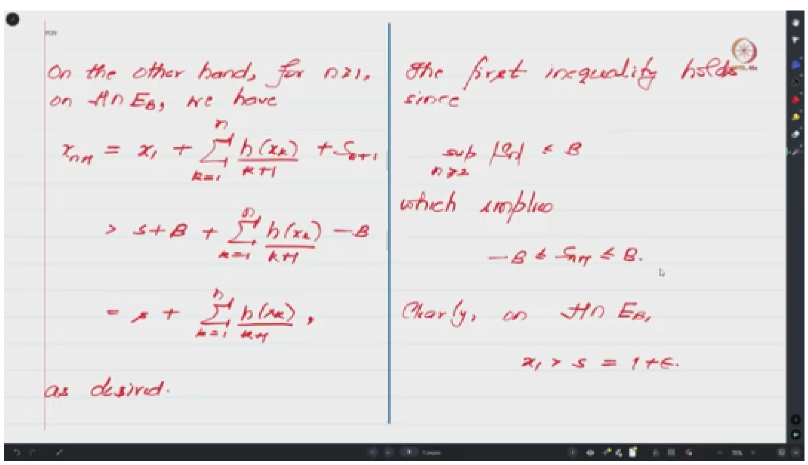


Recall that H says that the first noise that was added was large enough. And the claim is that if the first noise was sufficiently large, then the first X1 will actually be strictly bigger than S. Recall that S was slightly on the right side of 1, which was an unstable equilibrium. So one can conclude that because the noise was large enough, we actually landed on the bad side of the unstable equilibrium, right? That is, the right side of this unstable equilibrium 1, and then we will use this fact to conclude that xn plus 1 consequently will satisfy some relation like this.

And we will then use this fact to conclude that on this event, your Xn's will actually rise off to infinity. So first, let's verify these two claims. The first claim is that X1 is greater than or equal to S plus B on this event. So let's verify that. Recall that X1 has this relation with X0.

That is, X1 equals X0 plus H of X0 plus M1. And after substituting the expression for H of X0, which is X0 cubed minus X0, we can cancel off this X0 and we will end up with X0 cubed. And on the event H, we had presumed that M1... is strictly larger than S plus B minus X0 cubed. And hence, your X1 will be greater than S plus B, as desired.

So, the first thing we observe is that X1 is actually bigger than S plus B. Now, what we will do is we will separately use this fact to say something about Xn plus 1. In particular, for any n bigger than 1, we had previously shown that Xn plus 1 satisfies this relation, that is, Xn plus 1 equals X1. plus the sum of H of XK over K plus 1, and then we also have SN plus 1. Now, my claim is that on this event H intersection EB, we have already shown that X1 is S plus B, and because your SN plus 1's, or in particular your SN's, we have shown that their supremum is less than or equal to B on this event, one can conclude that SN plus 1 should be sandwiched between minus B and plus B, right?



So, if the supremum is less than or equal to b, every element here should be less than b, and if the absolute value is less than b, one can conclude that this Sn plus 1 value should be between minus b and plus b. In particular, Sn plus 1 should be greater than or equal to minus b. And that is, I mean, the expression that we have substituted over here. So, on H, X1 is bigger than S plus B, and on EB, your SN plus 1 is greater than or equal to minus B. So, these expressions will cancel off. And what we will be left with is that XN plus 1 is greater than S plus this sum. Okay.

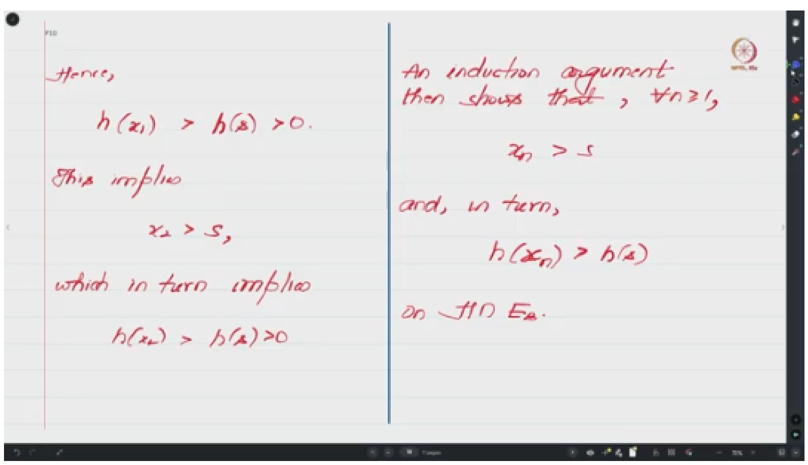
$$x_{n+1} = x_1 + \sum_{k=1}^{n} \frac{h(x_k)}{k+1} + S_{n+1}$$

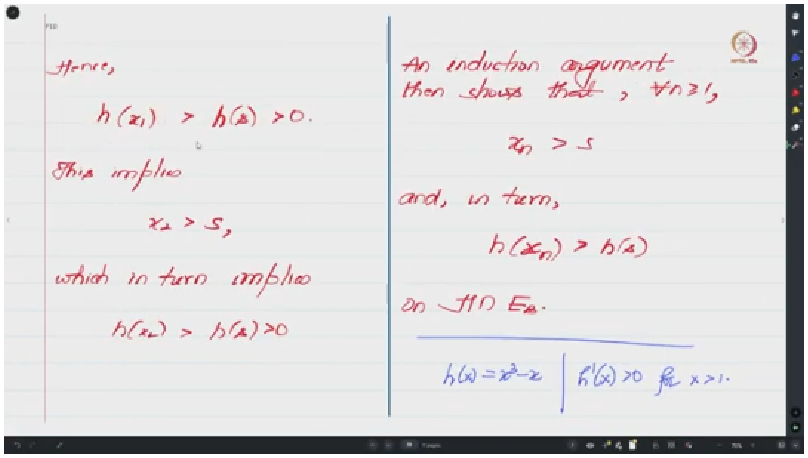
$$> S + B + \sum_{k=1}^{n} \frac{h(x_k)}{k+1} - B$$

$$= s + \sum_{k=1}^{n} \frac{h(x_k)}{k+1}$$

And this is what we wanted to show previously. And this is what we have managed to show. Is this okay? Now, what does this imply? Like the two facts that we have shown here, what does this imply?

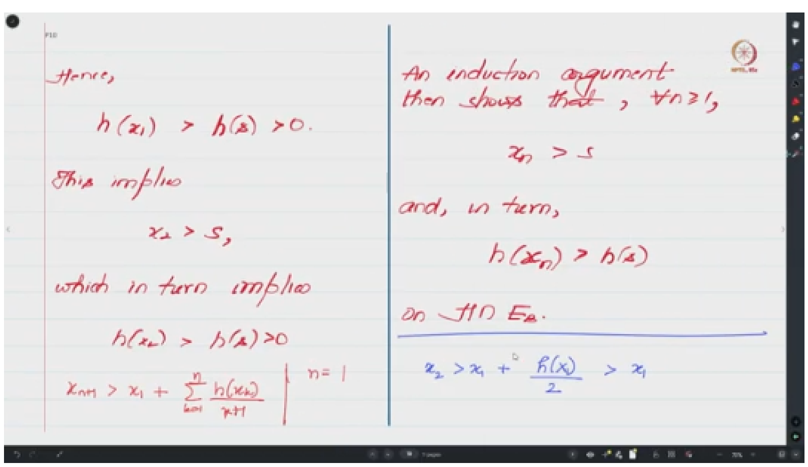
The first thing we know is that because B is non-negative, we can conclude that X1 is greater than S, which is slightly to the right of 1 plus epsilon. And recall that your H is actually... x cube minus x, and in particular, x prime x is greater than 0 for x bigger than 1, right? And because your h prime, that is the derivative, is bigger than 0 for x bigger than 1, one can conclude that the function h is monotonically increasing right after 1, and hence we can conclude that since your x1 is bigger than s, h of x1 should also be bigger than h of s, right? And since s is slightly to the right of this unstable equilibrium point plus 1, we can conclude that h of s must be strictly bigger than 0.



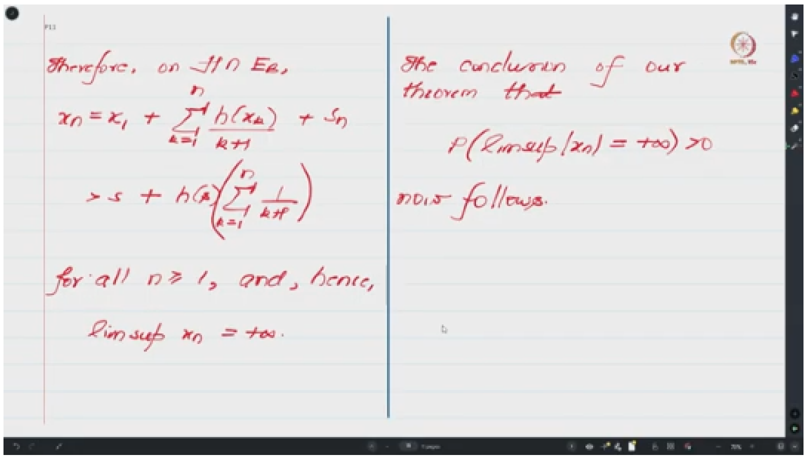


Is this okay? And hence, one can conclude that h of x1 is greater than h of s, and h of s is greater than 0, right? And we will use this fact. To conclude that X2 is actually bigger than S. So, recall that, in this result, what we had managed to show is that your X n plus 1 is greater than X1 plus some k equals 1 till n H of Xk. By k plus 1, okay.

So, if you substitute k—sorry, if we substitute n equals 1 in this relation, right, what we will end up with is What we will end up with is X2 is greater than X1 plus H of X1 by 2, right? And we have shown here that H of X1 is greater than 0, and hence this quantity will be bigger than X1. And we have separately shown that X1 is bigger than S, and hence one can conclude that because X1 is bigger than S, X2 is also bigger than S. Now, by a similar analysis like this, one can then conclude that H of X2 must also be bigger than H of S, and hence this is also positive. Now, by a similar inductive argument, one can conclude that

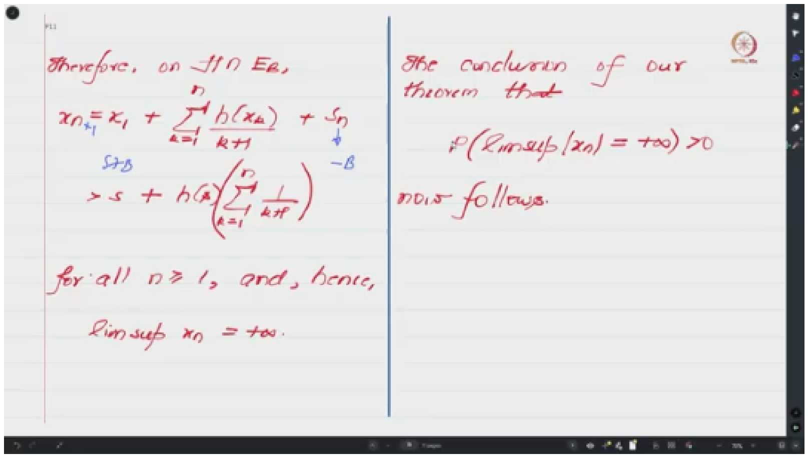


For every n bigger than or equal to 1, Xn must be strictly bigger than S, and hence H of Xn must be bigger than H of S, which is greater than 0 on this event H intersection EB, right? And one can make use of this fact to conclude that on H intersection EB, right, this expression is bigger than Equal to, right? So this expression is bigger than s plus b. This expression is bigger than minus b, and here we have shown that each of these quantities is actually strictly bigger than h of s. So, we can pull that outside the sum and we will be left with this sum of k plus 1's, right? And in conclusion, what we can say is that for every n bigger than or equal to 1, xn—so I should perhaps write plus 1 over here.



Xn plus 1 is actually bigger than S plus B. So this B and minus B will cancel out. And hence, Xn plus 1 is strictly bigger than S plus H of S times this sum. And we have also shown that H of S is strictly bigger than 0. And this quantity races off to infinity. And from this fact, one can conclude that if you look at the limit supremum of Xn's, this will actually be plus infinity, right?

And because your H intersection EB had positive probability, one can now conclude that the probability of the limit supremum of the absolute value of Xn equals plus infinity is also strictly bigger than 0, right? So, this brings us to the end of this class. Let us quickly summarize what we have done so far. What we have done is we looked at a stochastic approximation algorithm whose limiting ODE had a few unstable equilibrium points. In particular, it had an unstable equilibrium point 1, which had the property that if you are to the right of that, the ODE solutions actually go to infinity.



So, we used that fact right along with some properties of your noise sequences to show that, you know, the same fate also holds for the stochastic approximation algorithm. By same fate, I mean that its iterates, you know, can also race off to infinity if the noise is not good enough, right? And we used that to conclude that there is a positive probability that your iterates can go off to infinity. So, why is this important? Well, when you run or design an algorithm for practice, we often require that the iterates of the algorithm, you know, be bounded, you know, otherwise, you know, they can cause damage to wherever you are deploying them, right? And what this example suggests is that if you do not ensure enough care, then with some probability, your iterates could race off to infinity, right? And one can actually cook up worse examples as well where iterates of the algorithm could race off to infinity with very high probability.

And in the next class, what we will see is we will come up with a sufficient condition that ensures this does not happen. Until then, bye. See you then. Bye.