

STOCHASTIC APPROXIMATION: THEORY AND APPLICATIONS

Dr. Gagan Thope

Department of Computer Science and Engineering

Indian Institute of Science, Bangalore

Week 10

Lecture 38

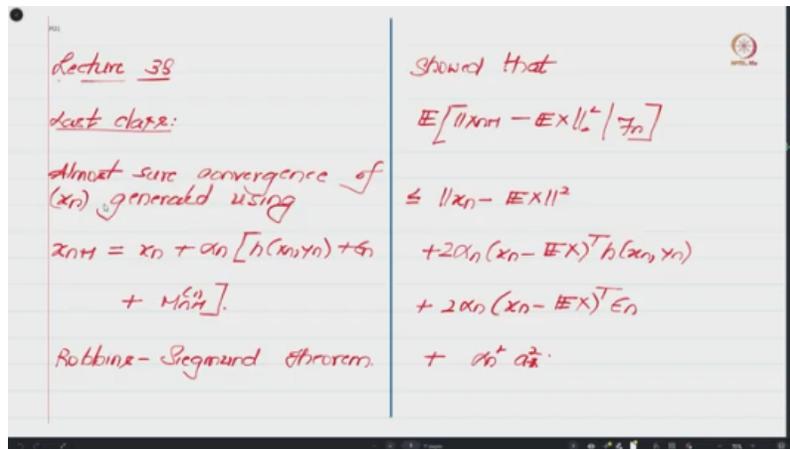
Almost Sure Convergence via Robbins–Siegmund Theorem – Part 2

Hello and Namaste everyone. Welcome to lecture 38 of this NPTEL course on Stochastic Approximation. In the last week and also this week, we have been looking at two variants of the general stochastic approximation algorithms that we had studied in the beginning of this course. So, in particular, we focused on two time scale stochastic approximation and stochastic recursive inclusion, and in particular we looked at the problem of solving a linear system of equations where some of the measurements come from adversaries. Towards that, we had designed an algorithm based on the stochastic gradient descent with respect to the L1 norm of Ax minus the expected value of y .

And in two classes before this, we looked at the differential inclusion that governs its limiting dynamics and showed that all solution trajectories of this differential inclusion indeed converge to the expected value of X . And in the previous class, we said that we are now going to analyze the behavior of the stochastic recursive inclusion algorithm itself. And we said we will make use of the Robin-Siegmund theorem to prove this result. In some sense, this is a different strategy to prove this result. We could have as well used the generalization of the ODE method also to prove convergence. But I felt that the Robin-Siegmund method is easier for the proof of this result.

And that is what we started discussing from the previous class. And in the previous class, we stated the Robin-Siegmund theorem and derived a partial inequality. And in this class, we will take that partial inequality further and, you know, eventually be in a position to apply the Robin-Siegmund theorem and use that to show convergence of X_n . So, let us begin with the formal details. So, as I said, you know, we started looking at the convergence of the following algorithm and we said that you know H has a particular

form, this ϵ_n and $m_n + 1$ have suitable forms, in particular this is a martingale difference sequence.



And we wanted to make use of the Robbins-Siegmund theorem which I had stated in the previous class. And so far what we have done is we have shown that this quantity over here is upper bounded by this term plus this term plus this term plus this. So, this was the bound that we had obtained and there was one more term that involved your $M_{\{n+1\}}$ and that term we managed to show goes away because it is a martingale difference term.

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

$$E[F_n] \leq ||x_n - EX||^2 + 2\alpha_n (x_n - EX)^T h(x_n, y_n) + 2\alpha_n (x_n - EX)^T \epsilon_n + \alpha_n^2 a_n^2$$

And hence we were left with only these expressions and what we will eventually do is we will choose a step size which satisfies α_n superscript 2 is less than infinity and this will take care of this term. So, we do not have to worry about this term.

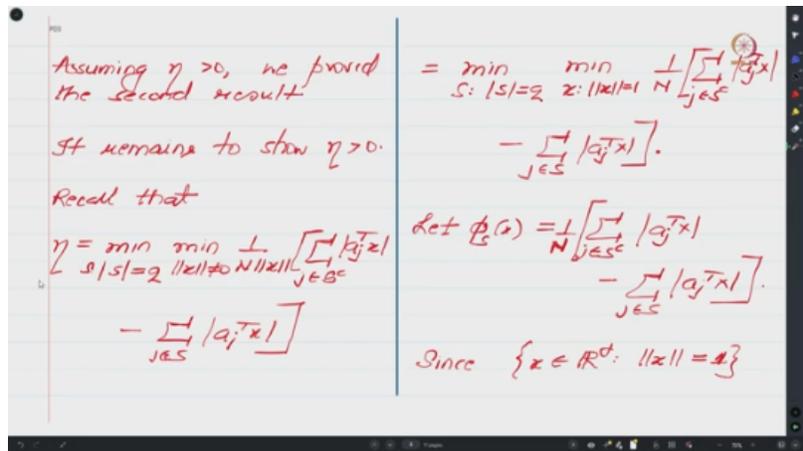
And what remains to take care of are these two terms. And in today's class, we will see how to deal with these expressions. And in the previous class, we had made some initial effort towards obtaining a bound on the second term, which is the term that is present over here. And we had obtained the following intermediate lemma, which we said we

will use to bound the second term. So, what this intermediate lemma said was that suppose η is defined in the following way and let us say K is defined with respect to η in the following way, then for any S which is a subset of the row indices of your matrix A which correspond to the coordinates of your Y vector such that cardinality of S is Q , then such an inequality holds.

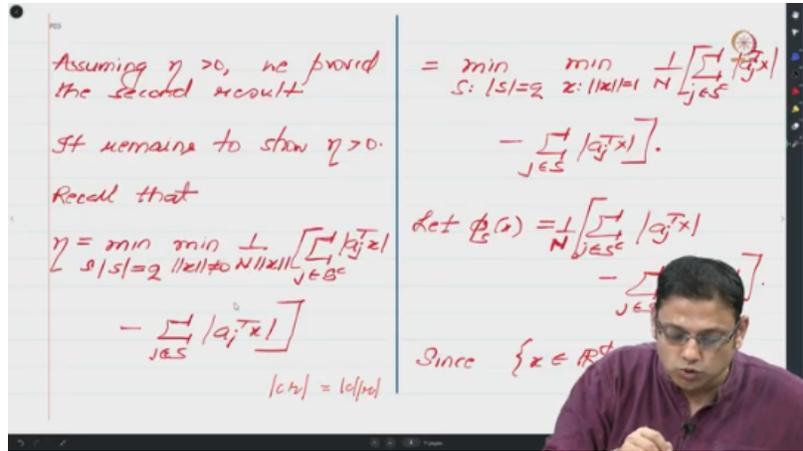
<p>Lecture 38</p> <p>next class:</p> <p>Almost sure convergence of (x_n) generated using</p> $x_{n+1} = x_n + \alpha_n [h(x_n, y_n) + \epsilon_n + M_n^n]$ <p>Robbins-Siegmund theorem</p>	<p>showed that</p> $E[\ x_n - EX\ _2^2 \mathcal{F}_n] \leq \ x_n - EX\ ^2 + 2\alpha_n (x_n - EX)^T h(x_n, y_n) + 2\alpha_n (x_n - EX)^T \epsilon_n + \alpha_n^2 \sum_{n=0}^{\infty} \alpha_n^2 L^2 < \infty$
---	--

<p>Towards bounding the 2nd term, we obtained the following intermediate lemma:</p> <p>Lemma Let</p> $\eta := \min_{S: S =2} \min_{x \neq 0} \frac{1}{N \ x\ } \left[\sum_{j \in S} a_j^T x - \sum_{j \in S} a_j^T x \right]$	<p>then, $\eta > 0$. Further, if</p> $K := \frac{2\alpha \alpha_n + 1}{N \eta}$ <p>then, for any $S \subseteq \{1, 2, \dots, N\}$ such that $S = Q$, we have</p> $(K-1) \sum_{j \in S} a_j^T x \geq (K+1) \sum_{j \in S} a_j^T x $
--	---

And we said that this inequality is saying more than what was presumed about the matrix A . And while we had proved this inequality in the previous class, we had presumed that this η was strictly bigger than 0. And what we will do today is we will first show that this η is greater than 0 and then see how this result of ours can be used to obtain a bound on the second term. So recall that η is defined in the following way. In particular, you can see that there is a norm x that is sitting here and you have a summation that is over here. And we have also presumed that x is not equal to 0.



So we can take this norm x inside, and if you look at x here, I mean also keep in mind that, you know, because of the linearity property of the absolute value, right? You can, I mean, I should not say linearity property, but what I should say is that for the absolute value, you know, for any real number R , C times R is the absolute value of C times R , right? So, this is true. So, the C that is over here is being played by this norm of X . So, I can take it inside, right?



And wherever I have this $A^T X$, I will now have $A^T X$ by norm of x . So because I have this expression over here, one can imagine this to be a unit vector, and hence this definition can be modified in the following way. So observe that here, instead of just considering x not equal to 0, we can restrict our attention to all those x where the norm of x is 1, and we drop this norm of x from the definition. So, we have this expression now.

Now, whatever is there inside, let us call it ϕ_S of x . So, ϕ_S of x is basically 1 over n times this summation minus this summation. Now, since we are taking the min over x such that norm of x equals 1 , let us collect all such x 's, and it is easy to see that this x is a compact set. It is compact because it is closed and bounded.

Assuming $\eta > 0$, we proved the second result. It remains to show $\eta > 0$. Recall that

$$\eta = \min_{S: |S|=2} \min_{\|x\|=1} \frac{1}{N} \sum_{j \in S} |a_j^T x| - \sum_{j \in S} |a_j^T x|$$

Let $\phi_S(x) = \frac{1}{N} \sum_{j \in S} |a_j^T x| - \sum_{j \in S} |a_j^T x|$

Since $\{x \in \mathbb{R}^d : \|x\|=1\}$

Furthermore, this ϕ_S of x is a function that is continuous in x . And because of this reason, I mean the fact that we know that any continuous function on a compact set achieves both its maxima and minima, and for us what is relevant is the fact that this function achieves its minima. Achieves its minima means that there is a specific point x prime. Lying in the set, that is the Euclidean norm of this x prime is 1 and ϕ_S of x is greater than ϕ_S of x prime for every other x . So, for every other x . So, that is what we mean by achieves its minima, right?

is a compact set and $\phi_S(x)$ is continuous, we have that ϕ_S achieves its minimum.

Hence, \exists some $x' : \|x'\|=1$ and $\phi_S(x) \geq \phi_S(x')$.

However, for $S: |S|=2$,

$$\sum_{j \in S} |a_j^T x| > \sum_{j \in S} |a_j^T x|$$

Hence,

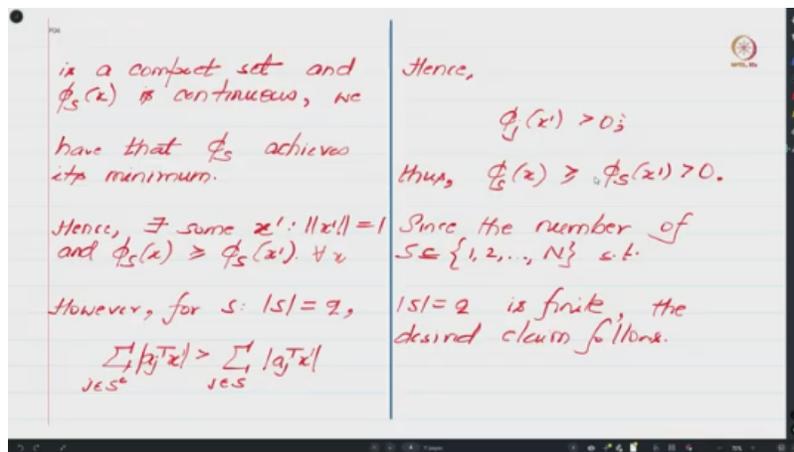
$$\phi_S(x') > 0;$$

thus, $\phi_S(x) \geq \phi_S(x') > 0$.

Since the number of $S \subseteq \{1, 2, \dots, N\}$ is finite, the desired claim follows.

So, often there will be the minimum value of ϕ of X , but there could be no value that achieves it. Is this okay? But since we have a continuous function on a compact set, one can conclude that there exists some X prime such that this inequality holds. Is this okay? However, we know that, you know, for S such that cardinality of S is Q , this relationship is greater than this for any X which is not equal to 0.

In particular, this relationship holds for X prime. And hence, one can conclude that your ϕ of x prime is actually strictly bigger than 0, right? And since it is strictly bigger than 0, one can conclude that for any x in this satisfying norm of x equals 1, your ϕ of x is bigger than or equal to ϕ of x prime, right? And this quantity we have already shown is strictly bigger than 0, right? And since the number of S such that cardinality of S is Q and we know that you know if you take the minimum of strictly positive numbers which are finitely many in number then the minimum will also be greater than 0.



From this one can indeed conclude that this η is actually strictly bigger than 0, right. So, what we managed to show is that this quantity over here is upper bounded by some specific quantity that is strictly bigger than 0 and that could depend on S , but since the number of such S 's is finite, this whole quantity is guaranteed to be strictly bigger than 0, which is the definition of η , and that verifies what we wanted to show, which is that η is strictly bigger than 0. So, now let us see how we can use this result to obtain a bound on the second term. So, recall our goal is to obtain a bound on a term of the following form, and towards that, let us first recall what H is, and H is as given over here.

Assuming $\eta > 0$, we proved the second result.

It remains to show $\eta > 0$.

Recall that

$$\mathcal{L} = \min_{S: |S|=2} \min_{\|x\| \neq 0, N \|x\|} \frac{1}{N \|x\|} \left[\sum_{j \in S} |k_j z_j| - \sum_{j \in S} |a_j^T x| \right]$$

$|a_j^T \frac{x}{\|x\|}| \quad |k_j| = |k_j|$

$$= \min_{S: |S|=2} \min_{z: \|z\|=1} \frac{1}{N} \left[\sum_{j \in S} |k_j z_j| - \sum_{j \in S} |a_j^T x| \right]$$

Let $\phi_S(z) = \frac{1}{N} \left[\sum_{j \in S} |k_j z_j| - \sum_{j \in S} |a_j^T x| \right]$

Since $\{x \in \mathbb{R}^d: \|x\|=1\}$

Now, $h(x, y) =$

$$\frac{1}{N} \left[\sum_{i \in A^c} \text{sign}(\mathbb{E}Y(i) - a_i^T x) a_i + \sum_{i \in A} \text{sign}(y(i) - a_i^T x) a_i \right]$$

Then, $(z - \mathbb{E}X)^T h(z, y)$

$$\leq \frac{1}{N} \left[- \sum_{i \in A^c} |a_i^T z - \mathbb{E}Y(i)| + \sum_{i \in A} |a_i^T z - \mathbb{E}Y(i)| \right]$$

From our previous Lemma,

$$(K-1) \sum_{i \in A^c} |(z - \mathbb{E}X)^T a_i| \geq (K+1) \sum_{i \in A} |(z - \mathbb{E}X)^T a_i|$$

Hence, $-\sum_{i=1}^N |a_i^T z - \mathbb{E}Y(i)| \geq K \left[- \sum_{i \in A^c} |(z - \mathbb{E}X)^T a_i + \sum_{i \in A} |(z - \mathbb{E}X)^T a_i| \right]$

In particular, for I in A complement, you presume that you have the ideal update, and for I in A , which is basically the set of coordinates which are governed by adversaries or controlled by adversaries, it can be an arbitrary value, okay? The sign expression can be an arbitrary value and it is controlled by the value of y . In particular, it is controlled by the value of y of i , which is given by your adversaries, right? So now, if you take the inner product of x minus expected value of x transpose h of x, y , right? In one of our previous classes, we had shown that this expression is upper bounded by minus of

the sum of I in A complement the absolute value of A_i transpose X minus expected value of Y of I , right, and plus, okay, so here because the signs are controlled by adversaries, we do not know what this sign value will be; however, if you take this expression and take its inner product with this, one can again show that we would end up with an upper bound like this, okay. So notice that for I in A complement we have a negative sign, but

here we are considering the worst case behavior of your adversaries; hence, this upper bound comes into play, right? And in this intermediate lemma that we had shown right there, we showed that for any X , right? So let me take you to that result, okay? So the result over there showed that for any x , k minus 1 times this summation is bigger than k plus 1 times this summation. So, in place of x , if I replace this with x minus expected value of x , one can see that we will end up with an inequality of the following form, right. So, I have just, wherever x is there, I have replaced it with x minus expected value of x transpose, and we will indeed get

Towards bounding the 2nd term, we obtained the following intermediate lemma:

Lemma Let

$$\eta := \min_{S \subseteq \{1, \dots, N\}} \min_{z \neq 0} \frac{1}{N \|z\|} \left[\sum_{j \in S} |a_j^T z| - \sum_{j \in S^c} |a_j^T z| \right]$$

Then, $\eta > 0$. Further, if

$$K := \frac{2\eta}{\eta + 1},$$

then, for any $S \subseteq \{1, 2, \dots, N\}$ such that $|S| = \eta$, we have

$$(K-1) \sum_{j \in S} |a_j^T x| \geq (K+1) \sum_{j \in S^c} |a_j^T x|.$$

Now, $h(x, y) =$

$$\frac{1}{N} \left[\sum_{i \in \mathcal{A}^c} \text{sgn}(EY(i) - a_i^T x) a_i + \sum_{i \in \mathcal{A}} \text{sgn}(y(i) - a_i^T x) a_i \right].$$

Then, $(x - EX)^T h(x, y)$

$$\leq \frac{1}{N} \left[- \sum_{i \in \mathcal{A}^c} |a_i^T x - EY(i)| + \sum_{i \in \mathcal{A}} |a_i^T x - EY(i)| \right]$$

From our previous Lemma,

$$(K-1) \sum_{i \in \mathcal{A}^c} |(x - EX)^T a_i| \geq (K+1) \sum_{i \in \mathcal{A}} |(x - EX)^T a_i|.$$

Hence, $-\sum_{i \in \mathcal{A}^c} |a_i^T x - EY(i)|$

$$\geq K \left[- \sum_{i \in \mathcal{A}^c} |(x - EX)^T a_i| + \sum_{i \in \mathcal{A}} |(x - EX)^T a_i| \right]$$

You know the inequality of the following form, right? And now what we are going to see is, you know, what does this inequality tell for an expression of this form, right? So what we are going to do is from this inequality we are going to rearrange terms so that we get an expression that looks something like this. So towards that, you can see that there is a

minus some i in A complement absolute value, right? So that minus one is sitting over here, so we retain it as it is. Right, and what we—sorry, I should be careful—I mean this term will appear on the right-hand side. So what we do is towards getting the term on the right-hand side, the K times this expression, we move towards the right-hand side, and you can see that this expression mirrors what you have over here. Right, and this minus 1 led to this expression, and you can see that there is a plus 1 with this expression with i in A . So if you bring it to the left-hand side, everything will be with a negative sign. So the left-hand side will basically be i in A and i in A complement, which basically, you know, forms your complete index set. Hence, the left-hand side, because of this minus 1 and this— Plus 1 being moved to the left-hand side, we will end up with minus summation i equals 1 to n the absolute value of this quantity over here, right? And the k times this expression, I have moved to the right, so you have k times this negative expression, and whatever is the plus 1 over here, I retain it as it is. And now this k I have taken out in common—is this okay? And one can now see that this expression that you have—

This expression that you have is precisely what you have over here, and if I multiply by 1 over n on both sides, one can see that I will end up with minus 1 over n and k over n and whatever is this expression, and as I told you, this expression is precisely X minus expected value of X transpose H of X , Y . Now, the question is, what happened over here? Well, if you see here, I have i equals 1 to n , the absolute value of this. Well, this is precisely the $L1$ norm of AX minus expected value of Y . You can see that this expression is precisely this expression, and hence we end up with a relation of the following form.

Now, $h(x,y) = \frac{1}{N} \left[\sum_{i \in A^c} \text{sign}(EY(i) - a_i^T x) a_i + \sum_{i \in A} \text{sign}(y(i) - a_i^T x) a_i \right]$.

Then, $(z - EZ)^T h(z,y) \leq \frac{1}{N} \left[- \sum_{i \in A^c} |a_i^T z - EY(i)| + \sum_{i \in A} |a_i^T z - EY(i)| \right]$.

From our previous Lemma, $(K-1) \sum_{i \in A^c} |z - EZ)^T a_i| \geq (K+1) \sum_{i \in A} |z - EZ)^T a_i|$.

Hence, $-\sum_{i=1}^N |a_i^T z - EY(i)| \geq K \left[- \sum_{i \in A^c} |z - EZ)^T a_i| + \sum_{i \in A} |z - EZ)^T a_i| \right]$.

<p>Therefore,</p> $-\frac{1}{N} \ A_n - EY\ _1$ $\geq \frac{K}{N} \left[-\sum_{i=1}^n (x - EX)^T a_i + \sum_{i=1}^n (x - EX)^T a_i \right]$ $\geq K (x - EX)^T \beta(x, y).$	<p>This implies that</p> $E \left[\ x_n - EX\ ^2 \right] \leq \ x_n - EX\ ^2$ $- \frac{2\alpha_n}{K} \ A_n - EY\ _1 + 2\alpha_n (x_n - EX)^T \epsilon_n + \alpha_n^2 \sigma^2$
---	---

That is, K times this inner product is less than or equal to minus of this expression. So if you remember, you know, we were working or trying to derive an upper bound on this quantity, right? So this, we had this expression on the right-hand side, this expression on the right-hand side, and this expression. Furthermore, we had this expression on the right-hand side. Now what we are saying is that we are going to replace this quantity

by this expression and with a k in the denominator. So, α_n times this expression was there initially. Now, we are replacing it with a negative of this L_1 norm by k . So, this is the expression that we have. So, whatever is the second term, we have now managed to obtain a bound. How this bound will be useful, we will soon discuss.

In particular, if you remember the Robbins-Siegmund theorem, the Robbins-Siegmund theorem said that suppose you have something like expected value of X_{n+1} given F_n . This is less than $1 + \alpha_n x_n$ okay minus b_n plus c_n so you can see that there is a negative b_n term that sits over here so that role is being played by this quantity over here right and now it remains to bound an expression of the following form now let us see how we can bound an expression of the following form Now towards obtaining a bound on this first let us recall the description of ϵ_n . So ϵ_n recall is $\frac{1}{n}$ plus you know the difference of I mean sum of differences of signs. In particular this sign included the idealized you know expected value of y and this included what we have at time n . So, we tried comparing this sign expression with this sign expression and multiplying this whole expression by A of I . And the nice thing over in ϵ_n is that this I is in A complement which means that every I over here corresponds to an honest

sensor which means that as n becomes larger and larger indeed this Y_n will converge to expected value of Y .

<p>Therefore,</p> $\frac{1}{N} \ Ax - EY\ _1$ $\geq \frac{K}{N} \left[-\sum_{i=1}^L (x - EX)^T a_i + \sum_{i=L+1}^L (x - EX)^T a_i \right]$ $\geq K (x - EX)^T \beta(x, y).$	<p>This implies that</p> $E \left[\ x_{n+1} - EX\ ^2 \mid \mathcal{F}_n \right]$ $\leq \ x_n - EX\ ^2$ $- \frac{2\alpha_n}{K} \ Ax_n - EY\ _1$ $+ 2\alpha_n (x_n - EX)^T \epsilon_n$ $+ \alpha_n^2 \sigma_\epsilon^2$ $E[X_{n+1} \mid \mathcal{F}_n] \leq (1 + \alpha_n) X_n - B_n + C_0$
---	--

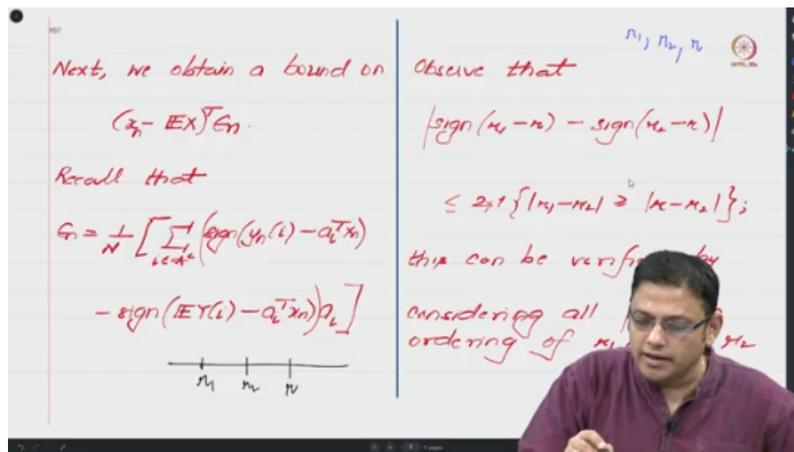
<p>Next, we obtain a bound on</p> $(x_n - EX)^T \epsilon_n.$ <p>Recall that</p> $\epsilon_n = \frac{1}{N} \left[\sum_{i=1}^L (\text{sign}(y_n(i) - a_i^T x_n) - \text{sign}(EY(i) - a_i^T EX)) a_i \right]$	<p>Observe that</p> $ \text{sign}(x_1 - x_2) - \text{sign}(x_2 - x_1) $ $= 2 \cdot \mathbb{1}\{ x_1 - x_2 \geq x_2 - x_1 \};$ <p>this can be verified by considering all possible ordering of x_1, x_2, and x_2.</p>
--	--

However, just because this converges, the sign expressions need not converge, right? Obviously, because, I mean, it is not an obvious matter of fact because sign is a discontinuous function, right? And we need to take care of it and we will see how to deal with that, okay? So, towards obtaining a bound on the sign expression, what we will do is, We will first make use of a relation that we discovered that exists between sign of numbers of such form.

So let us first understand what we have over here. Suppose R_1, R_2 and R , let us say, are three real numbers. So, since they are three real numbers, I hope you agree that if I try to draw the real line, you could have maybe R_1 here, maybe R_2 here, maybe R here, and maybe in a different scenario, your R could be sitting over here, or maybe in a different

scenario, R could actually be here. So, in other words, your R1, R2, R, right, they could come up in different orders.

Now, whatever these possible orders may be, one can see that whatever the arbitrary value of R1, R2, and R, it always satisfies, you know, the inequality sine of R1 minus R minus sine of R2 minus R is upper bounded by 2 times, okay, 2 times R1. 2 times the indicator that R1 minus R2 is bigger than R minus R2. So, you can, you know, consider maybe such an ordering and verify that such an inequality indeed holds. Is this okay? And recall that this sine expression, you know, can either be minus 1 or plus 1, and if the argument is 0, right, in which case it can take an arbitrary value between minus 1 and plus 1.



So one can see that in all those scenarios, one can, you know, indeed verify this condition over here. And because this inequality holds, one can show that this X_n minus expected value of X transpose epsilon n is upper bounded. So this 2 comes from that inequality that we had. times summation i in a complement and if you recall, you know, we had, you know, so x minus x transpose times a_i . So this expression will lead to something like this, and I should perhaps put an absolute value over here.

Hence, $(z_0 - EX)^T G$

$$\leq \frac{2}{N} \sum_{i \in \mathcal{A}^c} |(z_0 - EX)^T a_i|$$

$$\times \left\{ |y_0(i) - \mathbb{E}Y(i)| \right.$$

$$\left. \geq |a_i^T z_0 - \mathbb{E}Y(i)| \right\}$$

$$\leq \frac{2}{N} \sum_{i \in \mathcal{A}^c} |y_0(i) - \mathbb{E}Y(i)|$$

Now, for $i \in \mathcal{A}^c$,

$$y_{n+1}(i) = y_n(i) + \beta_n \left[\mathbb{1}_{\{L_{n+1}=i\}} - y_n(i) \right]$$

Let $\beta_n = \frac{1}{n+1}$, $n \geq 0$.

$$\text{Then, } y_n(i) = \frac{1}{n} \sum_{k=1}^n y_{k+1} \mathbb{1}_{\{L_{k+1}=i\}}$$

So this absolute value is upper bounded by this, and this is why you have an absolute value over here, and then you have the difference of signs, and the difference of signs leads to an indicator of the following form. So you can check from the previous slide why this indicator sort of makes sense. And we end up with an indicator of the following form. And then my PhD student Swetha Ganesh realized that this expression that you have, since it is upper bounded by this, we can use this inequality that is sitting within the indicator to upper bound this quantity over here.

So that leads us to a bound of the following form: that is 2 over n—we retain it as it is—and wherever you have this expression, because we have an indicator here, I will replace this expression with what you have over here. That is what I have written here, and hence we have this upper bound. Now, I drop the indicator because the indicator is upper bounded by 1. I drop that indicator, and finally, one can show that the absolute value of this quantity is upper bounded by the quantity that we have over here. Now, what we will do is to get an upper bound for this quantity over here, and towards that, we will make use of what is known as the law of iterated logarithm. I encourage you to look it up, and now let me first try to tell you why the law of iterated logarithm sort of becomes natural here. So notice that here we have i in \mathcal{A} complement, which means that your i is not related to your, you know, adversarial measurement, right. It is actually, you know, given by an honest or a non-adversarial sensor.

And hence, for any i in \mathcal{A} complement, recall that y_{n+1} of i has the following update rule. y_{n+1} of i equals y_n of i plus β_n times some expression like this. Now, if I

substitute for beta n, so now I am actually choosing a step size choice. If I choose something like 1 over n plus 1, So, if I choose a step size of the form 1 over n plus 1, then one can see that your y n i has an update rule of the following form.

I think there is some typo here. This should be n times some, you know, so there is an issue here. So, this should be k. I mean there are some mistakes here I am sorry about that this should basically be I this should be K. So you should have N times summation K equals 1 to N calligraphic YK indicator IK equals I the whole divided by N. So what I am basically saying is that if you have an update tool like this and your beta n is actually of the form 1 over n plus 1, then one can show that this update rule is basically averaging the samples that you have over here across different time instances.

So, this is your little y and this is your calligraphic y which corresponds to the sample that you have obtained. In other words, your little y_n is basically doing an averaging of the samples that you have obtained at every time instance. Now, you know, because of our assumption, okay, these random variables that we are using over here are independent and identically distributed, right? And in particular, if you take Y_n of i and subtract it with the expected value of Y of i, Then one can show that you know because of the presence of n over here okay you have 1 over n and this sum of you know random variables of the following form and this random variable one can easily show is of 0 mean.

Therefore,

$$y_n(i) - EY(i)$$

$$= \frac{1}{n} \sum_{i=1}^n (N Y_n \{i_n = i\} - EY(i))$$

Let $\sigma^2 = \text{Var}(N Y_n \{i_n = i\} - EY(i))$

Then, from the law of iterated logarithm,

$$\limsup \frac{|S_n|}{\sigma \sqrt{n \ln \ln n}} = 1 \text{ a.s.}$$

where $S_n = \sum_{i=1}^n (N Y_n \{i_n = i\} - EY(i))$

Thus, for a sample-path-bound $C(i)$,

$$|y_n(i) - EY(i)| \leq C(i)$$

Therefore,

$$y_n(\omega) - EY(\omega)$$

$$= \frac{1}{n} \sum_{i=1}^n (N_{Y_k} \{i_n = i\} - EY(\omega))$$

Let $\sigma^2 = \text{Var}(N_{Y_k} \{i_n = i\} - EY(\omega))$

Then, from the law of iterated logarithm,

$$\limsup \frac{|S_n|}{\sigma \sqrt{n \log \log n}} = 1 \text{ a.s.}$$

where $S_n = \sum_{i=1}^n (N_{Y_k} \{i_n = i\} - EY(\omega))$

Thus, for a sample-path-based $C(\omega)$,

$$|y_n(\omega) - EY(\omega)| \leq C(\omega) \frac{\sqrt{2n \log \log n}}{\sqrt{n}}$$

And if you define sigma square to be the variance of this quantity, then from the law of iterated logarithm, one can show that if you take the absolute value of S_n . So S_n is basically the sum that is sitting over here and divide it by sigma and in particular you divide it by the square root of n times log of n. So here is where you see this iterated law of logarithm. So, if you take the lim sup of this quantity then you know the law of iterated logarithm says that the lim sup of this quantity is going to be 1 almost surely which means that you know eventually these two numbers will go towards I mean will be upper bounded by 1 because of the lim sup that is present over here. And from this result and the fact that here you have S_n over N . So this is basically S_n over N . So if I take S_n over N and write it as S_n over square root N . So, this n I have written it as square root n times square root n and then what I will do is I will multiply and divide by square root log log n . And what we have been told is that this expression.

Therefore,

$$y_n(\omega) - EY(\omega)$$

$$= \frac{1}{n} \sum_{i=1}^n (N_{Y_k} \{i_n = i\} - EY(\omega))$$

Let $\sigma^2 = \text{Var}(N_{Y_k} \{i_n = i\} - EY(\omega))$

Then, from the law of iterated logarithm,

$$\limsup \frac{|S_n|}{\sigma \sqrt{n \log \log n}} = 1 \text{ a.s.}$$

where $S_n = \sum_{i=1}^n (N_{Y_k} \{i_n = i\} - EY(\omega))$

Thus, for a sample-path-based $C(\omega)$,

$$|y_n(\omega) - EY(\omega)| \leq C(\omega) \frac{\sqrt{2n \log \log n}}{\sqrt{n}}$$

$$\frac{S_n}{\sqrt{n} \sqrt{\log \log n}}$$

So, let me put it in red. So, this expression. will basically be upper bounded by a constant asymptotically. So, that is what we have been told over here. And this constant could depend on the sample path.

That is, on one sample path, the constant could be 5. In another sample path, it could be 1000, and so on and so forth. So, on different sample paths, you would have different constants, but nevertheless, it will be a finite constant. And one can, you know, make use of this result to show that your $y_{n,i}$ minus expected value of y_i , which is your S_n over n , is upper bounded by c_i times the square root of $\log \log n$ over square root n . So, from the law of iterated logarithm, one can show that this indeed holds.

Therefore,

$$y_n(i) - EY(i) = \frac{1}{n} \sum_{i=1}^n (NY_n(i) - EY(i))$$

Let $\sigma^2 = \text{Var}(NY_n(i) - EY(i))$

Then, from the law of iterated logarithm,

$$|y_n(i) - EY(i)| \leq \frac{c_i}{\sqrt{n}} \sqrt{\log \log n}$$

Thus, for a sample-path-based $C(i)$,

$$\limsup \frac{|S_n|}{\sigma \sqrt{\log \log n} / \sqrt{n}} = 1 \text{ a.s.}$$

where $S_n = \sum_{i=1}^n (NY_n(i) - EY(i))$

And I would like to again emphasize that this constant could be sample path dependent. And one can then show that $2 \times \alpha_n \times n$ minus expected value of x transpose ϵ_n . You know, if you remember, we had shown it to be 2 over n , α_n times this summation. And now, using the law of iterated logarithm, we have shown that this quantity over here is, you know, upper bounded by some constant times the square root of $\log \log n$ divided by the square root n . Now, if you define c to be the max of... i in A complement c_i , you can pull this off, right? And since there are n minus q quantities in A complement, right, we will have A minus n minus q over here, and then we will have the square root of $\log \log n$ divided by the square root n , right. So, the key thing that we have to keep in mind is α_n times the square root of $\log \log n$ divided by the square root n , right.

<p>This implies that</p> $2\alpha_n (x_n - EX)^T \leq 0$ $\leq \frac{2\alpha_n}{N} \sum_{i=1}^N y_n(\omega) - EY(\omega) $ $= \frac{2}{N} (N-q) C \alpha_n \frac{\sqrt{\ln k n n}}{\sqrt{n}}$ <p>where $C = \max_{i \in A^c} C(i)$.</p>	<p>Finally, we have that</p> $E[\ x_n - EX\ ^2] \leq \ x_n - EX\ ^2$ $- 2\alpha_n \ AX_n - EY\ _1$ $+ \frac{4}{N} (N-q) C \alpha_n \frac{\sqrt{\ln k n n}}{\sqrt{n}}$ $+ \alpha_n^2$
---	--

What we have managed to show is that this conditional expectation is upper bounded by X_n minus the expected value of X square minus $2\alpha_n$ times the L1 norm of this quantity, and we had $2\alpha_n$ which was sitting over here. I think I should say that this will lead to a 4. This also should be replaced by a 4, and because of this 4 over n , n minus q c α , and this I have written that expression over here, and then we had this A star times α_n square. Is this okay? A star square times α_n square.

And now one can see that this expression that we have is indeed of the form that is desired to invoke the Robbins-Siegmund theorem. And in the Robbins-Siegmund theorem, we require that this quantity also be finite. So, I will treat this whole quantity as the term that we require to be finite. So, we now require that summation n equals 1 to infinity α_n square root $\log \log n$ by root n to be finite.

<p>This implies that</p> $2\alpha_n (x_n - EX)^T \leq 0$ $\leq \frac{2\alpha_n}{N} \sum_{i=1}^N y_n(\omega) - EY(\omega) $ $= \frac{2}{N} (N-q) C \alpha_n \frac{\sqrt{\ln k n n}}{\sqrt{n}}$ <p>where $C = \max_{i \in A^c} C(i)$.</p>	<p>Finally, we have that</p> $E[\ x_n - EX\ ^2] \leq \ x_n - EX\ ^2$ $- 2\alpha_n \ AX_n - EY\ _1$ $+ \frac{4}{N} (N-q) C \alpha_n \frac{\sqrt{\ln k n n}}{\sqrt{n}}$ $+ \alpha_n^2$
---	--

Is this okay? So if I want this expression to be finite, one can see that if I pick α_n to be, you know, α_n to be $n^{-1/2 + \epsilon}$, I think α_n may be, I mean, I think there is an $n^{-1/2}$ missing over here, maybe one can do the calculation appropriately. So if I choose this α_n to be $n^{-1/2 + \epsilon}$, then one can see that if I take, you know, on the one hand, this summation α_n^2 will be less than infinity, which is the term that is sitting over here. On the other hand, this $\alpha_n \sqrt{\log \log n}$, this will also, you know, this α_n over \sqrt{n} will lead to a term of the form $n^{-1/2 + \epsilon}$, right, this half plus epsilon, and you know, you had this \sqrt{n} already sitting, so this already contributes $n^{1/2 + \epsilon}$, so together we will have $n^{1/2 + \epsilon}$, right? And because we have chosen this epsilon to be greater than 0, one can show that this sum, because it is of the form $n^{1/2 + \epsilon}$, this is less than infinity, and hence one can use the Robbins-Siegmund theorem to conclude that, you know, x_n minus the expected value of x converges.

This implies that

$$\alpha_n^2 (x_n - EX)^T \leq 0$$

$$\leq \frac{\alpha_n^2}{N} \sum_{i=1}^N |y_i(\omega) - EY(\omega)|$$

$$= \frac{\alpha_n^2}{N} (N-2) C \alpha_n \frac{\sqrt{\ln \ln n}}{\sqrt{n}}$$
 where $C = \max_{i=1, \dots, N} C_i$.

Finally, we have that

$$E[\|x_n - EX\|^2] \leq \|x_n - EX\|^2$$

$$= 2 \alpha_n \|AX_n - EY\|_1$$

$$+ \frac{1}{N} (N-2) C \alpha_n \frac{\sqrt{\ln \ln n}}{\sqrt{n}}$$

$$+ \alpha_n^2 \sum_{i=1}^N \frac{\sqrt{\ln \ln n}}{\sqrt{n}}$$

Pick $\alpha_n = \frac{1}{(n+1)^{1+\epsilon}}$ for some $\epsilon > 0$.

Then, $\sum_{n=0}^{\infty} \alpha_n < \infty$

Further, $\sum_{n=1}^{\infty} \frac{\alpha_n \sqrt{\lambda_n \lambda_{n+1}}}{\sqrt{n}} < \infty$ a.s.

From Robbins-Siegmund theorem, we get

$\|x_n - EX\|$ converges.

Further, $\sum_{n=0}^{\infty} \alpha_n \|Ax_n - EX\| < \infty$ a.s.

If A has full column rank, then

$\|x_n - EX\| \rightarrow 0$ a.s., as derived.

Pick $\alpha_n = \frac{1}{(n+1)^{1+\epsilon}}$ for some $\epsilon > 0$.

Then, $\sum_{n=0}^{\infty} \alpha_n < \infty$

Further, $\sum_{n=1}^{\infty} \frac{\alpha_n \sqrt{\lambda_n \lambda_{n+1}}}{\sqrt{n}} < \infty$ a.s.

From Robbins-Siegmund theorem, we get

$\|x_n - EX\|$ converges.

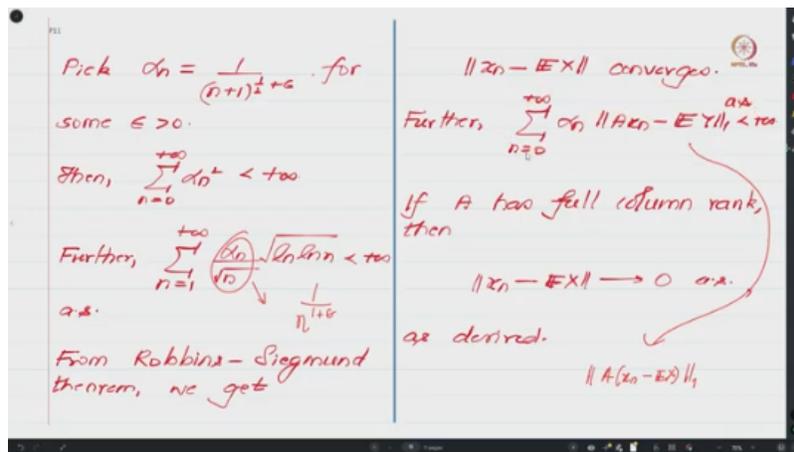
Further, $\sum_{n=0}^{\infty} \alpha_n \|Ax_n - EX\| < \infty$ a.s.

If A has full column rank, then

$\|x_n - EX\| \rightarrow 0$ a.s., as derived.

Right, which is what you have on the right-hand side. I mean, the square converges, and hence the square root of that, which is this quantity, converges. Furthermore, whatever you had over here, if you take their sum, okay, that sum should be less than infinity, right? And this quantity, I hope you agree, okay, can be written as A times X_n minus the expected value of X . We can write it this way, and then you will have this L1 norm. Now if we presume A has full column rank, which means that this product is 0 if and only if, you know, this quantity over here is 0. So by making use of this fact, one can show that, you know, this being finite and this converges together imply that the limit should be equal to 0. And the way to verify this is the following: that if X_n minus the expected value of X does not go to 0, then that would imply that your X_n minus the expected value of X goes to some positive value. And because it goes to a positive value, one can show that this quantity will also be a positive value.

But we have presumed that your summation alpha n is actually infinity. So if this converges to a positive value, one can then show that this sum will be infinity, violating this condition. So, in other words, because you have that summation alpha and this quantity is less than infinity and the fact that this converges together, you know that can only imply that, you know, this difference should go to 0, which in particular implies that X_n goes to the expected value of X almost surely. So, this brings us to the end of our discussion.



So, let me quickly summarize what we have done. So, in these two weeks, we looked at two-time-scale stochastic approximation, stochastic recursive inclusion. In particular, we looked at a scenario where we wanted to solve a linear system of equations where some of the measurements come from adversaries. And because they come from adversaries, instead of trying to minimize the L2 norm, we focused on minimizing the L1 norm of AX minus the expected value of Y . And that led us to some update rule. And then what we said was that if A , your matrix A , satisfies some conditions,

Then, by, you know, looking at the, you know, this behavior of the stochastic recursive inclusion from the perspective of this Robbins-Siegmund theorem, one can show that if A satisfies that condition, right, which is that of full column rank and the fact that you are, you know, for any k right, your summation i in k absolute value of $A_i^T x$ is less than summation i in k complement absolute value of $A_i^T x$. So, if that condition is satisfied then one can show that, you know, your stochastic recursive inclusion will also converge to the expected value of x almost surely. So, one can see that this algorithm

then becomes robust; that is, even in the presence of adversaries we are able to show that this algorithm converges to the expected value of X , which was what we wanted to recover. And just to highlight, the set-valued nature of the algorithm came into play because each time your algorithm visits a specific value X , the adversary has a choice of values, right, to play. And in general, okay, this leads to the set-valued nature, right. And if I want to quickly remind you why we wanted to look at X_n and Y_n separately, right, in particular why we wanted to look at different updates, one with X_n and the other with Y_n , was that, you know, we wanted the expected value of Y to sit within the sign function. And unfortunately, if we plugged in directly the samples of Y , and because the expectation and sign, you know, cannot interchange naturally with each other, we had to make use of a two-time-scale algorithm.

And again, the last thing I would like to highlight is the choice of step size that we made. Observe that β_n is $1/n + 1$, and the step size choice α_n that we made was $1/n + 1$ to the power of half plus epsilon. So, you can see here that, you know, α_n and β_n are of different orders, which is what, you know, brings in the two-time-scale nature. So, this completes our discussion on two-time-scale stochastic approximation and stochastic recursive inclusion. In the next couple of weeks, we will be looking at various applications of stochastic approximation.

In particular, we will look at some applications in the context of reinforcement learning. Until then, goodbye, and thank you, and namaste.