STOCHASTIC APPROXIMATION: THEORY AND APPLICATIONS

Dr. Gugan Thope

Department of Computer Science and Engineering

Indian Institute of Science, Bangalore

Week 10

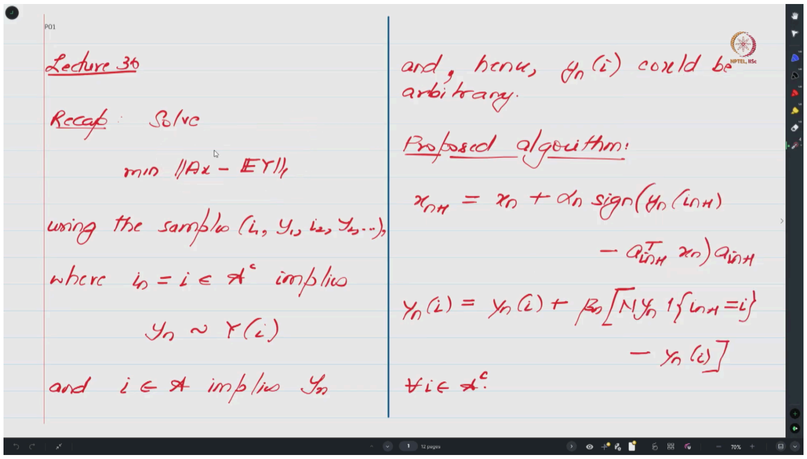
Lecture 36

Convergence of Distributed Robust Mean Estimation Iterates through the Lens of DI

Hello and Namaste everyone, welcome to lecture 36 of this NPTEL course on Stochastic Approximation. So, as you remember, in the past week and this week, we have been looking at some variants of Stochastic Approximation algorithms that we studied in the past 8 weeks. Specifically, we are looking at two-time-scale and, you know, stochastic recursive inclusion variants of stochastic approximation, right? In the previous week, we looked at a particular problem, which is that of solving a linear system of equations when some of the measurements are controlled by adversaries. Towards that, we looked at minimizing the L1 norm of the objective function, right? And I mean, the objective function involved an L1 norm. And then we tried to come up with a stochastic approximation algorithm to solve that objective, right.

So, in the past few lectures, we have tried to come up with a motivation for why we need to look at a two-time-scale algorithm and, in particular, look at a stochastic recursive inclusion interpretation of the algorithm. In today's class, we will continue with that interpretation. In particular, in the past lecture, we saw that the limiting dynamics of this stochastic recursive inclusion interpretation results in a differential inclusion. In the previous class, we sort of showed that the driving set-valued map for this stochastic limiting differential inclusion is Marchaud, which guarantees the existence of solutions.

In today's class, what we will do is we will, you know, show that at least for this idealized DI, the solution trajectories indeed go to the desired solution, and building upon that, we will see how one can extend that convergence result to the original stochastic algorithm itself. So, that is the plan for today. So, let us begin with the formal details. So, if you remember, we wanted to, you know, minimize Ax minus the expected value of y, in particular, the L1 norm of this.



Lecture 36

Recap: Solve

$$\min \|Ax - \mathbb{E}Y\|_1$$

using the samples $(i_1, \mathcal{Y}_1, i_2, \mathcal{Y}_2, \ldots)$,

where $i_n = i \in \mathcal{A}^c$ implies

$$\mathcal{Y}_n \sim Y(i)$$

and $i \in \mathcal{A}$ implies $\mathcal{Y}_n$

and, hence, $y_n(i)$ could be arbitrary.

Proposed algorithm:

$$x_{n+1} = x_n + \alpha_n \operatorname{sign}(y_n(i_{n+1}) - a_{i_{n+1}}^T x_n) a_{i_{n+1}}$$

$$y_n(i) = y_n(i) + \beta_n [N \mathcal{Y}_n 1\{i_{n+1} = i\} - y_n(i)]$$

$\forall i \in \mathcal{A}^c$.

Right, and the samples that we had access to were denoted in the following way. So, at time instance 1, we have one coordinate index and an appropriate sample calligraphic Y1, and similarly, at time instance 2, we have I2 calligraphic Y2, and so on, right? And we were under the setting where if At time instance n, i n equals i, and this i belongs to a complement. Recall that this calligraphic A denotes the set of adversaries. So, if i belongs to A complement, we have that, you know, i corresponds to an honest or a non-adversarial coordinate, right?

And in that case, this calligraphic Y n is presumed to be an independent component. Sample of the ith coordinate of the vector Y, okay? So, the vector Y and the vector Y here are one and the same, right? And on the other hand, If I was adversarial in nature, then this calligraphic Y n and hence the ith coordinate of little y n could be adversarial.

$$||Ax - Ey||$$

$$i_n = i \in A^c$$

$$y_n \sim Y(i)$$

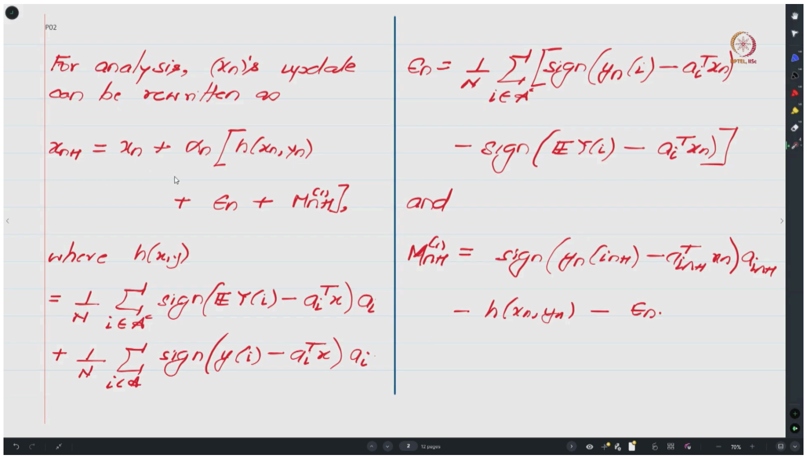
So, recall that this little y n was used as a proxy for estimating—I mean, I should not say proxy—it was used to estimate the expected value of Y, especially for coordinates I which are non-adversarial in nature. So, towards that, we proposed the following algorithm, right.

So, X n had the following update rule, and Y n had the following update rule. In particular, one can view this as—I mean, one can think of this as an update that has been inspired by the gradient with respect to this objective function, right. And if you remember, in the true gradient, we would have the expected value of Y here, and since we do not have the expected value of Y, and since sign and expectation do not interchange easily, so what we did was we had this separate time scale on which we estimated Y n separately. Right, and the update rule, at least for all i which are honest, right, would be given like this.

So, there is a typo here. This should be 'y n plus 1 i equals y n i plus beta n times something like this,' right? And so, I think there is again a typo. This should be calligraphic y n plus 1, alright. So, this was the update rule. So, again, I would like to highlight that this set of adversarial coordinates is unknown. So, if it was known, of course, we could just discard the coordinates. But the challenge is that we do not know which coordinates are adversarial in nature and hence which coordinates are honest.

So, whatever I have written here is for the set of unknown but fixed honest coordinates. So, this is the update rule that will be followed by them. So, here, notice that we do not presume the knowledge of which coordinates are adversarial and which coordinates are honest. We just plug in whatever is i n plus 1 sampled at time instance n. So, we look at the corresponding coordinate of y n. Now, this corresponding coordinate, if it If it comes from an honest worker or honest coordinate, then, of course, this is, in some sense, a proxy for the expected value of y of i. On the other hand, if i n plus 1 could be i, where i is an adversary or i belongs to the set of adversaries, then this could be an arbitrary value.

However, the nice thing is that whatever the value, they sit inside the sine function. Hence, at least on a per-iteration basis, the worst that an adversary could do is flip the sine. So, this is what an adversary could do in the worst sense. So, now, let us recall how we proposed analyzing Xn's update. We said that Xn's update can be written in this form, where H of X, Y had this interpretation, and this expression depends on the expected value of Y, and if this argument is, you know,



P02

For analysis, (x_n)'s update can be rewritten as

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

where $h(x, y)$

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

$$\epsilon_n = \frac{1}{N} \sum_{i \in A^c} \left[\text{sign}(y_n(i) - a_i^T x_n) - \text{sign}(\mathbb{E}Y(i) - a_i^T x_n) \right]$$

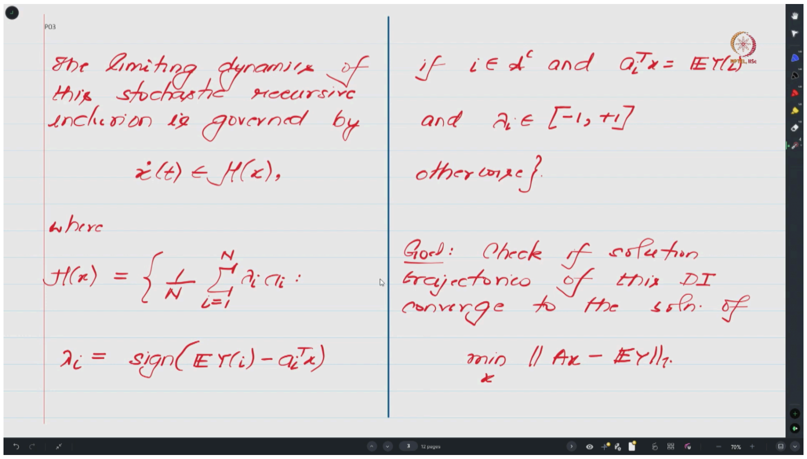
and

$$M_{n+1}^{(i)} = \text{sign}(y_n(i_{n+1}) - a_{i_{n+1}}^T x_n) a_{i_{n+1}} - h(x_n, y_n) - \epsilon_n.$$

Non-zero, then you know this sine function here will either take plus 1 or minus 1, depending on whether this coordinate is positive or negative. On the other hand, if this value is 0, then the sine function is allowed to take any value between minus 1 and plus 1, right? And here, again, the sine function, you know, we allow it to depend on what this value of y is, provided as input over here. In particular, these coordinates are the ones that can be controlled by adversaries. So, the interpretation here is that each time you come to X, the adversary could, you know, choose something over here, right, depending on how they want to derail the algorithm, right.

So, that was the set-valued interpretation of what this H of Xi means. I will, you know, sort of give the formal details on the next slide, and this epsilon n was basically the difference for i in a complement between the sign expression and the sign expression here, and there should be an ai as well, right? And in some sense, this is the gap between what we would have ideally liked—ideal here is, you know, with the knowledge of the expected value of y of i—but this yn is what we have at time instance i, so we ask what would be the difference between these two coordinates. While Yn goes to the expected value of Y, okay, you know, this difference need not go to 0 because the sine function is actually discontinuous. So, there is some challenge involved in the analysis of epsilon n. Nevertheless, I mean, the intuition is that because Yn goes to the expected value of Y, we would be able to show that somehow epsilon n does not impact us that much, and Mn plus 1 is basically what is the true update minus the sum of this quantity and this quantity, right? And one can see that by the way we have defined this Mn plus 1, 1, right, this expression will cancel off with this.

So, we will be left with this and this quantity, and hence one can see that Mn plus 1, 1 is indeed a martingale difference sequence; in particular, its conditional expectation is 0. And here is the set-valued nature that I wanted to describe. One can show that the limiting dynamics of this stochastic recursive inclusion is governed by X dot of T belonging to capital H of X, where H of X is given by this linear sum of these AIs. Recall that this AI transpose is the ith row. of A, right? So, ai over here is the transpose of this vector, right? So, this is a d-dimensional vector, and we have some linear combination of such vectors, and we insist that, you know, for a given x, lambda i be the sign of this quantity if



The limiting dynamics of this stochastic recursive inclusion is governed by

$$\dot{x}(t) \in \mathcal{H}(x),$$

where

$$\mathcal{H}(x) = \left\{ \frac{1}{N} \sum_{i=1}^{N} \lambda_i a_i : \right.$$

$$\lambda_i = \text{sign}(\mathbb{E}Y(i) - a_i^T x)$$

if $i \in \mathcal{A}^c$ and $a_i^T x = \mathbb{E}Y(i)$

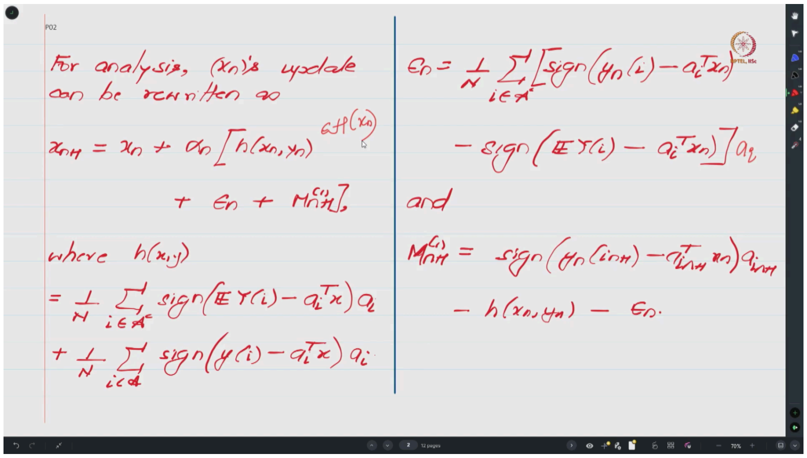
and $\lambda_i \in [-1, +1]$

otherwise$\}$.

Goal: Check if solution trajectories of this DI converge to the soln. of

$$\min_x \| Ax - \mathbb{E}Y \|_1.$$

I am in the set of non-adversaries, right? And AI transpose X, okay, does not equal the expected value of Y of I. So, if both these conditions are satisfied, then we require that lambda I be this sign. On the other hand, if either of these conditions fails to hold, then we allow lambda I to be any number between minus 1 and plus 1. So, we consider all possible linear combinations, and if you go back, one can conclude that this little h of xn comma yn indeed belongs to capital H of xn at every time instance n. So, whatever this yn input may be, one can show that this will belong to capital H of xn, and this is why I said that there is a set-valued interpretation to this update rule over here, which makes it a stochastic recursive inclusion. Right.



P02

For analysis, (x_n)'s update can be rewritten as

$$x_{n+1} = x_n + \alpha_n [\overset{\in H(x_n)}{h(x_n, y_n)} + \epsilon_n + M^{(i)}_{n+1}],$$

where $h(x,y)$

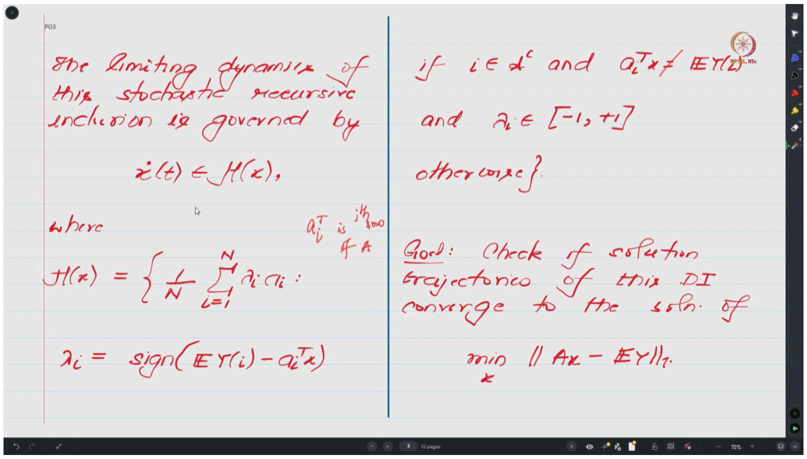
$$= \frac{1}{N} \sum_{i \in \mathcal{A}^c} \text{sign}(\mathbb{E}Y(i) - a_i^T x) a_i + \frac{1}{N} \sum_{i \in \mathcal{A}} \text{sign}(y(i) - a_i^T x) a_i$$

$$\epsilon_n = \frac{1}{N} \sum_{i \in \mathcal{A}^c} [\text{sign}(y_n(i) - a_i^T x_n) - \text{sign}(\mathbb{E}Y(i) - a_i^T x_n)] a_i$$

and

$$M^{(i)}_{n+1} = \text{sign}(y_n(i_{n+1}) - a_{i_{n+1}}^T x_n) a_{i_{n+1}} - h(x_n, y_n) - \epsilon_n.$$

And, you know, in our previous class, we said, you know, this map H over here is merge out. Right. And one of the first goals for today's class is to check if you look at this idealized differential inclusion, will the solution trajectories of this differential inclusion converge to the solution of this objective problem? Right. So, recall that the solution to this objective problem is indeed, you know,



P03

The limiting dynamics of this stochastic recursive inclusion is governed by

$$\dot{x}(t) \in H(x),$$

where

a_i^T is i^{th} row of A

$$H(x) = \Big\{ \frac{1}{N} \sum_{i=1}^{N} \lambda_i a_i :$$

$$\lambda_i = \text{sign}(\mathbb{E}Y(i) - a_i^T x)$$

if $i \in \mathcal{A}^c$ and $a_i^T x \neq \mathbb{E}Y(i)$

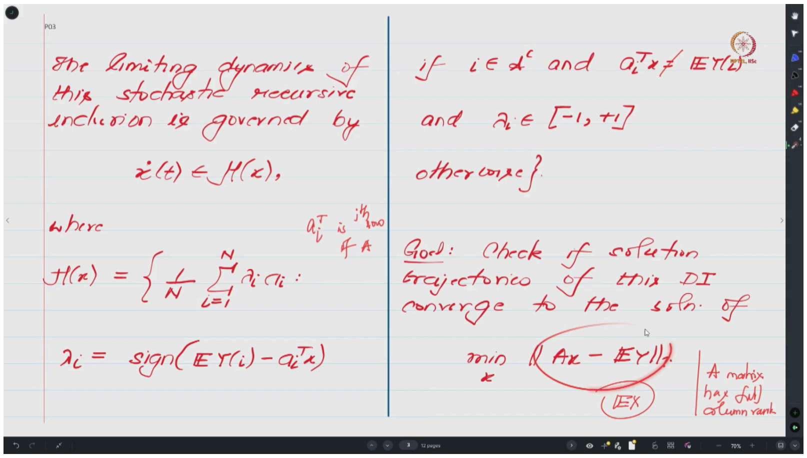
and $\lambda_i \in [-1, +1]$

otherwise$\Big\}$.

Goal: Check if solution trajectories of this DI converge to the soln. of

$$\min_x \| Ax - \mathbb{E}Y \|_1.$$

the expected value of X, which is the true thing, right. So, this is the true thing that we want to estimate, and this is the solution to this objective problem. One can show that, by using the fact that our A matrix is, you know, we presume our main matrix to have full column rank, okay. So, we presume this condition So, under this presumption, one can check that, you know, the unique solution to this optimization problem is indeed the expected value of X. So, far so good.



P03

The limiting dynamics of this stochastic recursive inclusion is governed by

$$\dot{x}(t) \in \mathcal{H}(x),$$

where $\quad a_i^T$ is i^{th} row of A

$$\mathcal{H}(x) = \left\{ \frac{1}{N} \sum_{i=1}^{N} \lambda_i a_i : \right.$$

$$\lambda_i = \text{sign}(\mathbb{E}Y(i) - a_i^T x)$$

if $i \in \mathcal{A}^c$ and $a_i^T x \neq \mathbb{E}Y(i)$

and $\lambda_i \in [-1, +1]$

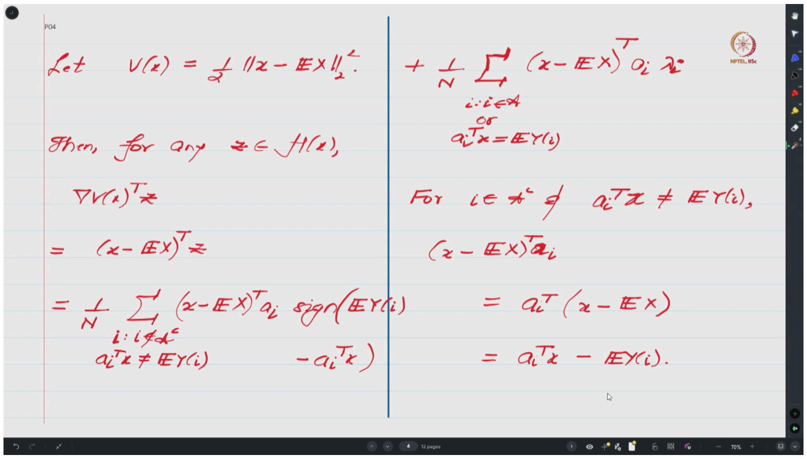
otherwise$\}$.

Goal: Check if solution trajectories of this DI converge to the soln. of

$$\min_x \underbrace{\| Ax - \mathbb{E}Y \|_2}_{\mathbb{E}X}$$

A matrix has full column rank

Now, let us try to see if this goal can be achieved, which is whether we can show that every solution trajectory of this limiting DI indeed converges to the expected value of X or not. Towards that, we make use of a Lyapunov argument, which basically means that we show the existence of a Lyapunov function. A Lyapunov function basically says that along any solution trajectory of this DI, the value of the Lyapunov function keeps decreasing, and it sort of keeps decreasing until you reach the minima of this Lyapunov function. Right, and because of this reason, the existence of a Lyapunov function itself guarantees that all solution trajectories of the limiting DI would indeed converge to the expected value of X, okay. So, let us go over the details more formally. So, we are going to define V in the following way, okay.



P04

Let $V(x) = \frac{1}{2} \| x - \mathbb{E}X \|_2^2.$

Then, for any $z \in \mathcal{H}(x)$,

$$\nabla V(x)^T z$$

$$= (x - \mathbb{E}X)^T z$$

$$= \frac{1}{N} \sum_{\substack{i : i \notin \mathcal{A}^c \\ a_i^T x \neq \mathbb{E}Y(i)}} (x - \mathbb{E}X)^T a_i \, \text{sign}(\mathbb{E}Y(i) - a_i^T x)$$

$$+ \frac{1}{N} \sum_{\substack{i : i \in \mathcal{A} \\ \text{or} \\ a_i^T x = \mathbb{E}Y(i)}} (x - \mathbb{E}X)^T a_i \lambda_i$$

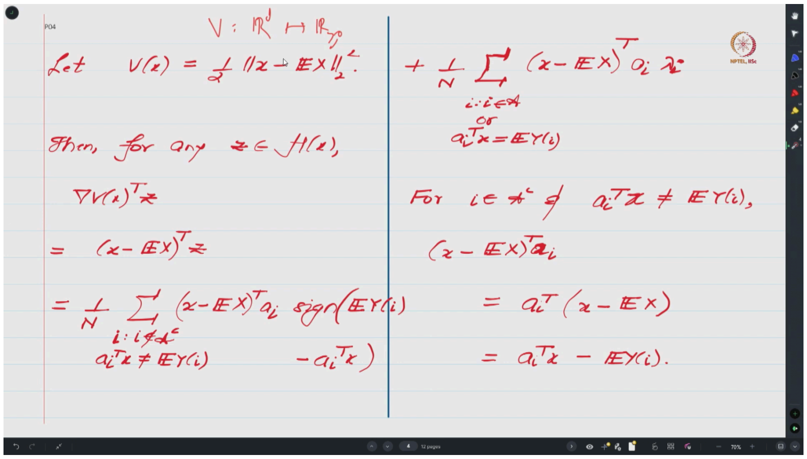
For $i \in \mathcal{A}^c$ & $a_i^T x \neq \mathbb{E}Y(i)$,

$$(x - \mathbb{E}X)^T a_i$$

$$= a_i^T (x - \mathbb{E}X)$$

$$= a_i^T x - \mathbb{E}Y(i).$$

So, V is a map. That actually goes from Rd to R greater than or equal to 0, which means that the input to V is a d-dimensional vector and the output is some non-negative scalar, and it is defined in the following way, right. It is V of x is half times. So, this is some normalization that we use to make our life easy, right, times the norm of x minus the expected value of x, right. And the L2 distance between them and the square of that.



P04

$V : \mathbb{R}^d \mapsto \mathbb{R}_{\geq 0}$

Let $V(x) = \frac{1}{2} \|x - \mathbb{E}X\|_2^2.$

Then, for any $z \in \mathcal{H}(x)$,

$\nabla V(x)^T z$

$= (x - \mathbb{E}X)^T z$

$= \frac{1}{N} \sum_{\substack{i : i \notin \mathcal{A}^c \\ a_i^T x \neq \mathbb{E}Y(i)}} (x - \mathbb{E}X)^T a_i \, \text{sign}(\mathbb{E}Y(i) - a_i^T x)$

$+ \frac{1}{N} \sum_{\substack{i : i \in \mathcal{A} \\ \text{or} \\ a_i^T x = \mathbb{E}Y(i)}} (x - \mathbb{E}X)^T a_i \lambda_i$

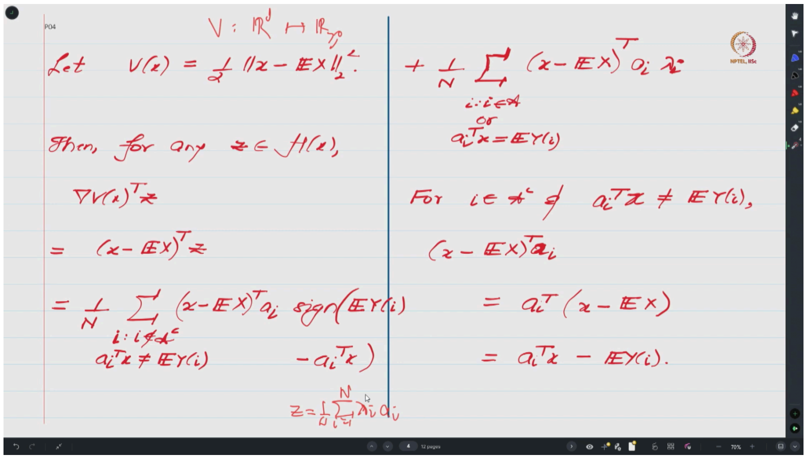
For $i \in \mathcal{A}^c$ & $a_i^T x \neq \mathbb{E}Y(i)$,

$(x - \mathbb{E}X)^T a_i$

$= a_i^T (x - \mathbb{E}X)$

$= a_i^T x - \mathbb{E}Y(i).$

So, this is how we define V of X. Now, if you consider any element in H of X and take the inner product of grad VX transpose Z, let us see what happens over here. So, X minus the expected value of X transpose Z. Now, if you recall H of X, the definition of H of X, one can presume that Z has the following form. So, Z equals summation, 1 over n, I equals 1 to n. lambda I AI, okay, so Z has this form where the lambda I's are signs of suitable quantities when you know your I is non-adversarial and AI transpose X not equals the expected value of I. On the other hand, lambda I could be arbitrary. So, keeping that in mind and keeping this expansion of Z, right, what we do is Okay, what we do is we break this summation into different sets of coordinates.



P04

$V : \mathbb{R}^d \mapsto \mathbb{R}_{\geq 0}$

Let $V(x) = \frac{1}{2} \|x - \mathbb{E}X\|_2^2.$

Then, for any $z \in \mathcal{H}(x)$,

$\nabla V(x)^T z$

$= (x - \mathbb{E}X)^T z$

$= \frac{1}{N} \sum_{\substack{i : i \notin \mathcal{A}^c \\ a_i^T x \neq \mathbb{E}Y(i)}} (x - \mathbb{E}X)^T a_i \, \text{sign}(\mathbb{E}Y(i) - a_i^T x)$

$+ \frac{1}{N} \sum_{\substack{i : i \in \mathcal{A} \\ \text{or} \\ a_i^T x = \mathbb{E}Y(i)}} (x - \mathbb{E}X)^T a_i \lambda_i$

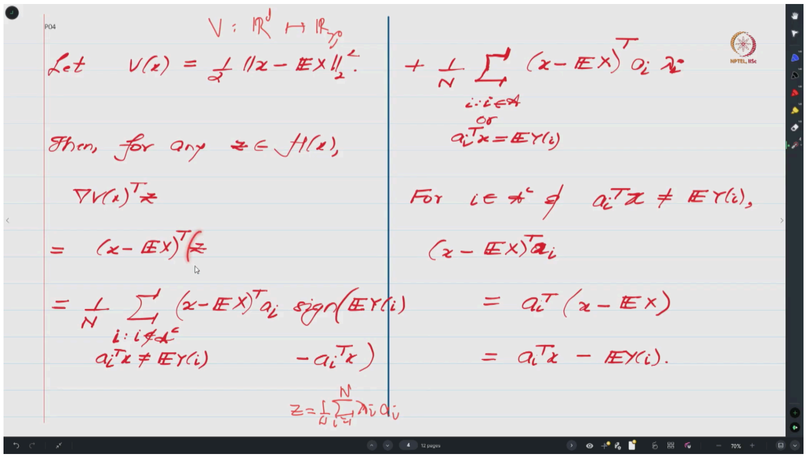
For $i \in \mathcal{A}^c$ & $a_i^T x \neq \mathbb{E}Y(i)$,

$(x - \mathbb{E}X)^T a_i$

$= a_i^T (x - \mathbb{E}X)$

$= a_i^T x - \mathbb{E}Y(i).$

$Z = \frac{1}{N} \sum_{i=1}^{N} \lambda_i a_i$



$$V : \mathbb{R}^d \mapsto \mathbb{R}_{\geq 0}$$

Let $V(x) = \frac{1}{2}\|x - \mathbb{E}X\|_2^2$.

Then, for any $z \in \mathcal{H}(x)$,

$$\nabla V(x)^T z$$

$$= (x - \mathbb{E}X)^T z$$

$$= \frac{1}{N}\sum_{\substack{i : i \notin \mathcal{A}^c \\ a_i^T x \neq \mathbb{E}Y(i)}} (x - \mathbb{E}X)^T a_i \, \mathrm{sign}(\mathbb{E}Y(i) - a_i^T x)$$

$$z = \frac{1}{N}\sum_{i=1}^{N} \lambda_i a_i$$

$$+ \frac{1}{N}\sum_{\substack{i : i \in \mathcal{A} \\ \text{or} \\ a_i^T x = \mathbb{E}Y(i)}} (x - \mathbb{E}X)^T a_i \lambda_i$$

For $i \in \mathcal{A}^c$ & $a_i^T x \neq \mathbb{E}Y(i)$,

$$(x - \mathbb{E}X)^T a_i$$

$$= a_i^T (x - \mathbb{E}X)$$

$$= a_i^T x - \mathbb{E}Y(i).$$

One is where I do not belong to A. So, I should say I do not belong to A and not I do not belong to A complement. I do not belong to A and A I transpose X is not equal to the expected value of Y. Collect all those I's and there this X minus the expected value of X transpose is as it is and this lambda I A I whatever we have is basically A I and this lambda I should equal this sign expression over here. So that is exactly what I have written. And for all other i's, I have written x minus the expected value of x transpose times Ai times lambda i and this lambda i could be some arbitrary number between minus 1 and plus 1. So, what we will do is let us consider some i which satisfies these two conditions that is it is not an adversarial coordinate and Ai transpose x not equals the expected value of y.

In that case, one can see that X minus the expected value of X transpose AI, right? You know I can sort of since X transpose, sorry, X transpose Y equals Y transpose X, so using this relation I can sort of interchange the position of transpose, right, and then I can use the linearity property and say that AI transpose X minus the expected value of X is basically AI transpose X minus the expected value of Y of I, so notice that this expression. Is basically what this expression is and whatever you have over here appears inside this sine function as well. So the argument to the sine function and what you have here are similar in spirit except for a sign change.

$$V(x) = \frac{1}{2}||x \; - \; EX||_2^2$$

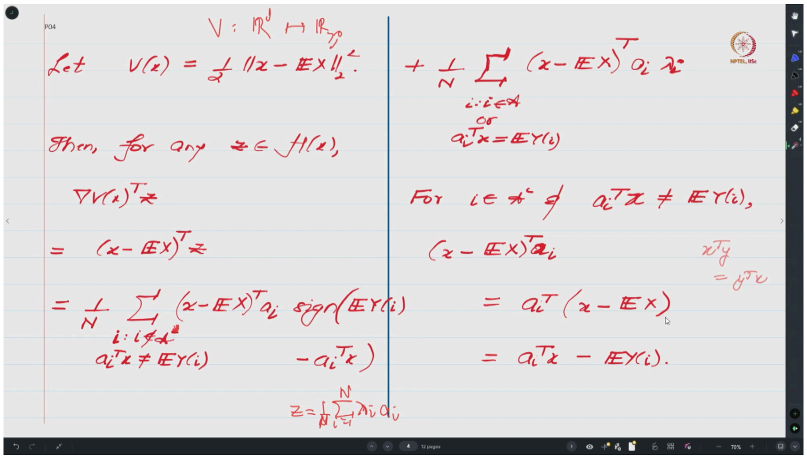
$$\nabla V(x)^T Z \; = (x \; - \; EX)^T Z$$

$$= \frac{1}{N} \sum_{i: i \in A^c \; a_i^T x \neq EY(i)} (x - EX)^T a_i \, sign\left(EY(i) - a_i^T x\right)$$

$$(x - EX)^T a_i = a_i^T (x - EX)$$

$$= a_i^T x - EY(i)$$

So notice that this expression equals this and the negative of this is what is present inside the sine expression over here.



P04

$V : \mathbb{R}^d \mapsto \mathbb{R}_{\geq 0}$

Let $V(x) = \frac{1}{2} \|x - \mathbb{E}X\|_2^2.$

Then, for any $z \in \mathcal{H}(x)$,

$\nabla V(x)^T z$

$= (x - \mathbb{E}X)^T z$

$= \frac{1}{N} \sum_{\substack{i : i \notin \mathcal{A}^c \\ a_i^T x \neq \mathbb{E}Y(i)}} (x - \mathbb{E}X)^T a_i \; sign(\mathbb{E}Y(i) - a_i^T x)$

$z = \frac{1}{N} \sum_{i=1}^{N} \lambda_i a_i$

$+ \frac{1}{N} \sum_{\substack{i : i \in \mathcal{A} \\ \text{or} \\ a_i^T x = \mathbb{E}Y(i)}} (x - \mathbb{E}X)^T a_i \lambda_i$

For $i \in \mathcal{A}^c$ & $a_i^T x \neq \mathbb{E}Y(i)$,

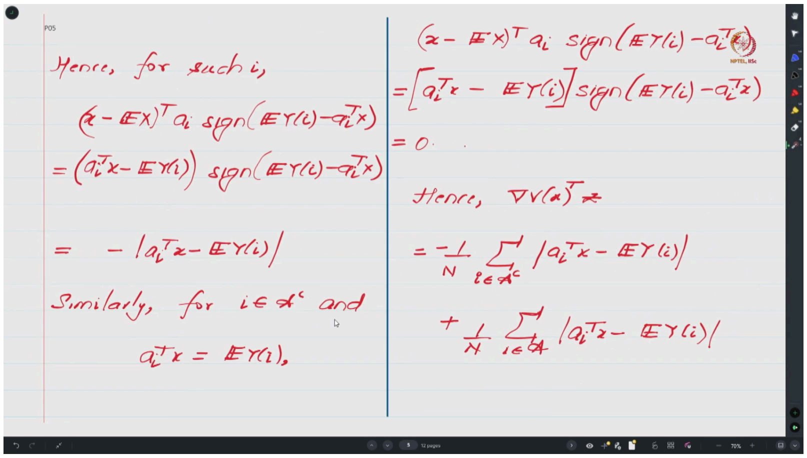
$(x - \mathbb{E}X)^T a_i$

$= a_i^T (x - \mathbb{E}X)$

$= a_i^T x - \mathbb{E}Y(i).$

$x^T y = y^T x$

And hence, one can conclude that for such i's, for such i meaning those i's where these two conditions hold, the inner product between x minus the expected value of x and ai times the sine of this will equal the negative of the absolute value of this difference. So, let me elaborate. So, if you take x and multiply it with sine of x, My claim is that this equals the absolute value of x. So this is easy to check. If x is positive, then sine of x will be plus 1.



Hence, for such i,

$$(x - \mathbb{E}X)^T a_i \, \text{sign}(\mathbb{E}Y(i) - a_i^T x)$$
$$= (a_i^T x - \mathbb{E}Y(i)) \, \text{sign}(\mathbb{E}Y(i) - a_i^T x)$$
$$= -|a_i^T x - \mathbb{E}Y(i)|$$

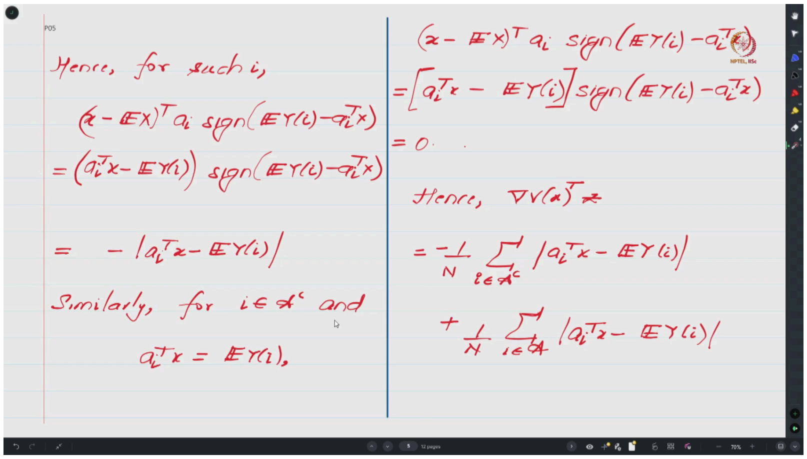
Similarly, for $i \in \mathcal{A}^c$ and

$a_i^T x = \mathbb{E}Y(i)$,

$$(x - \mathbb{E}x)^T a_i \, \text{sign}(\mathbb{E}Y(i) - a_i^T x)$$
$$= [a_i^T x - \mathbb{E}Y(i)] \, \text{sign}(\mathbb{E}Y(i) - a_i^T x)$$
$$= 0.$$

Hence, $\nabla V(x)^T x$

$$= -\frac{1}{N} \sum_{i \in \mathcal{A}^c} |a_i^T x - \mathbb{E}Y(i)| + \frac{1}{N} \sum_{i \in \mathcal{A}} |a_i^T x - \mathbb{E}Y(i)|$$


P05

Hence, for such i,

$$(x - \mathbb{E}X)^T a_i \, \text{sign}(\mathbb{E}Y(i) - a_i^T x)$$
$$= (a_i^T x - \mathbb{E}Y(i)) \, \text{sign}(\mathbb{E}Y(i) - a_i^T x)$$
$$= -|a_i^T x - \mathbb{E}Y(i)|$$

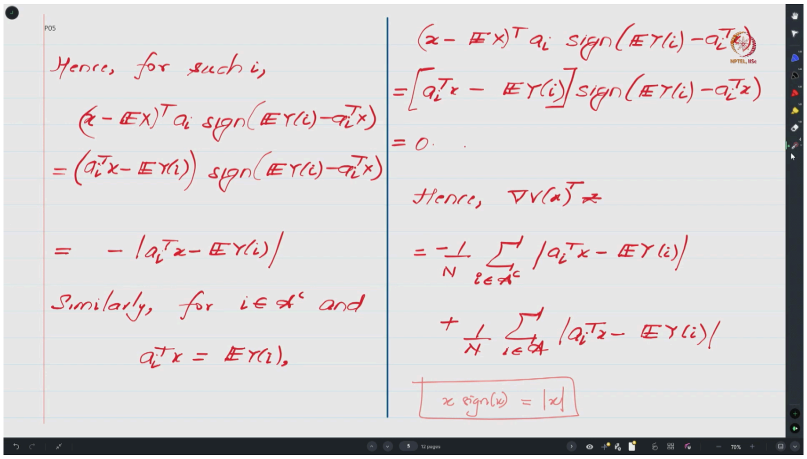
Similarly, for $i \in \mathcal{A}^c$ and

$a_i^T x = \mathbb{E}Y(i)$,

$$(x - \mathbb{E}x)^T a_i \, \text{sign}(\mathbb{E}Y(i) - a_i^T x)$$
$$= [a_i^T x - \mathbb{E}Y(i)] \, \text{sign}(\mathbb{E}Y(i) - a_i^T x)$$
$$= 0.$$

Hence, $\nabla V(x)^T x$

$$= -\frac{1}{N} \sum_{i \in \mathcal{A}^c} |a_i^T x - \mathbb{E}Y(i)| + \frac{1}{N} \sum_{i \in \mathcal{A}} |a_i^T x - \mathbb{E}Y(i)|$$


P05

Hence, for such i,

$$(x - \mathbb{E}X)^T a_i \, \text{sign}(\mathbb{E}Y(i) - a_i^T x)$$
$$= (a_i^T x - \mathbb{E}Y(i)) \, \text{sign}(\mathbb{E}Y(i) - a_i^T x)$$
$$= -|a_i^T x - \mathbb{E}Y(i)|$$

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

$a_i^T x = \mathbb{E}Y(i)$,

$$(x - \mathbb{E}x)^T a_i \, \text{sign}(\mathbb{E}Y(i) - a_i^T x)$$
$$= [a_i^T x - \mathbb{E}Y(i)] \, \text{sign}(\mathbb{E}Y(i) - a_i^T x)$$
$$= 0.$$

Hence, $\nabla V(x)^T x$

$$= -\frac{1}{N} \sum_{i \in \mathcal{A}^c} |a_i^T x - \mathbb{E}Y(i)| + \frac{1}{N} \sum_{i \in \mathcal{A}} |a_i^T x - \mathbb{E}Y(i)|$$

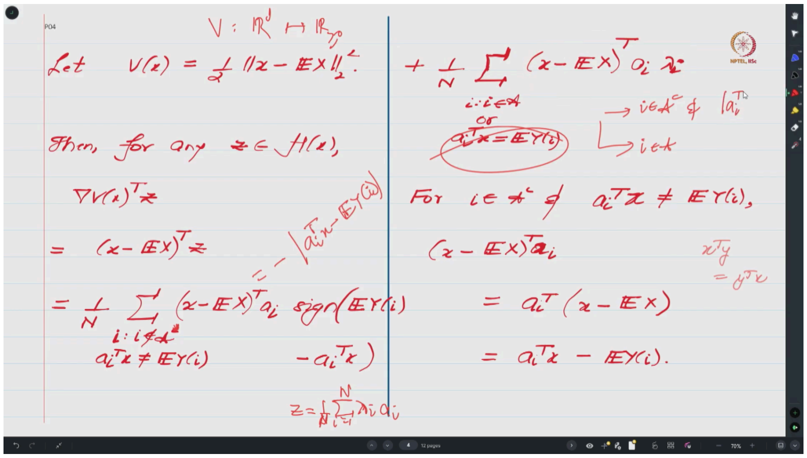
$$\boxed{x \, \text{sign}(x) = |x|}$$

So, x times plus 1 will be x itself, and because x is positive, it will equal the absolute value of x. On the other hand, when x is 0, then this will be 0, and sine of x can be any number between minus 1 and plus 1. But whatever value sine x takes, x times sine of x will be 0, and hence in that case also this expression will equal the absolute value of 0, which is 0. On the other hand, when x is negative, So, let us say this is minus 5. So, this will be minus 5 times minus 1, which will again be equal to 5, right?

So, if you understand this part, you can immediately see that, you know, one can immediately see that x times sine of minus x equals minus the absolute value of x, right? So, from this, one can immediately see that the relation that we have over here leads to this thing that is minus the absolute value of A I transpose expected value of Y of I. And in the same way, one can see that if I belongs to A complement and A I transpose X equals the expected value of Y of I, which means the argument to your sine expression is actually 0. So, in this case also, this expression

times this expression will equal 0. This will equal 0 because, you know, Ai transpose X equals the expected value of I whenever such a thing holds. And again, one can show that this equals minus Ai transpose X minus the expected value of Y of I. So, in this way, one can see that for every X, your grad VX transpose Z for any Z in H of X. So, recall that Z belongs to H of X. So, this inner product can be written as minus 1 over N X. summation over i in A complement the absolute value of Ai transpose minus the expected value of Y of I plus 1 over N summation i in A Ai transpose X minus the expected value of I. So, let me elaborate what happened over here.

So, recall that this is the true equality. So, what we have done is, for all such terms, we have shown that this expression is equal to minus ai transpose x minus the expected value of i here. What we have done is we have broken this sum into those which, you know, belong to i in a complement and this condition is satisfied, or i belongs to a. So, we have sort of split it up into these two cases. And for this case also, we have shown that this expression must equal ai transpose x minus the expected value of y of i with a negative sign. I mean, for this case and this being true, right, this being true, one can see that this expression will basically be 0.



P04

$V : \mathbb{R}^d \mapsto \mathbb{R}_{\geq 0}$

Let $V(x) = \frac{1}{2}\|x - \mathbb{E}X\|_2^2$.

Then, for any $z \in \mathcal{H}(x)$,

$\nabla V(x)^T z$

$= (x - \mathbb{E}X)^T z$

$= \frac{1}{N} \sum_{\substack{i : i \notin A \\ a_i^T x \neq \mathbb{E}Y(i)}} (x - \mathbb{E}X)^T a_i \, \mathrm{sign}(\mathbb{E}Y(i) - a_i^T x)$

$= -|a_i^T x - \mathbb{E}Y(i)|$

$z = \frac{1}{N} \sum_{i=1}^{N} \lambda_i a_i$

$+ \frac{1}{N} \sum_{\substack{i : i \in A \\ \text{or} \\ \cancel{a_i^T x = \mathbb{E}Y(i)}}} (x - \mathbb{E}X)^T a_i \lambda_i$

$\rightarrow i \in A^c$ & $|a_i^T$

$\hookrightarrow i \in A$

For $i \in A^c$ & $a_i^T x \neq \mathbb{E}Y(i)$,

$(x - \mathbb{E}X)^T a_i$

$= a_i^T (x - \mathbb{E}X)$

$= a_i^T x - \mathbb{E}Y(i)$.

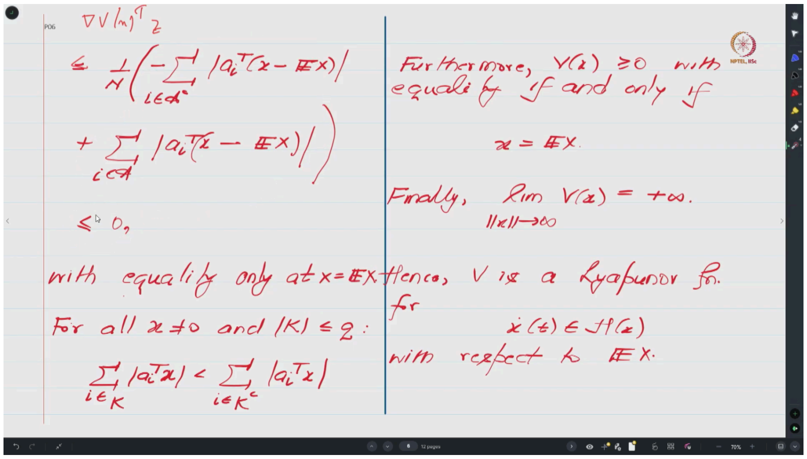
$x^T y = y^T x$

And for all other i, all I am saying is that look, whatever this, you know, quantity over here is. Whatever this quantity is, this lambda i is a scalar. Hence, this whole expression is upper bounded by X minus the expected value of X transpose AI. So, it is upper bounded by this quantity.

It is upper bounded by this quantity because lambda I's absolute value is at most 1. Hence, this whole expression is upper bounded. Let me write it cleanly one more time. So, it is upper bounded by X minus the expected value of X transpose So, it is upper bounded by this quantity, and again, if I take Ai inside, I will get Ai transpose X minus the ith coordinate of the expected value of Y. And this is what I have written here.

So, in other words, for the honest coordinates, I have a negative sign and the absolute value. I should actually say that this is not equal to but rather less than or equal to. And for the, you know, value corresponding to the adversarial coordinates, this expression is an upper bound to whatever we had before, right? That is, this is an upper bound to this Ai, you know, so the expected value of Y of i minus Ai, okay. So, let me be careful.

x minus the expected value of x transpose a i times lambda i. So, one can see that for this expression, whenever i is in a, this quantity that we have is an upper bound to this quantity, and hence we have this upper bound over here, right? So, finally, one can, you know, by some rearrangement, one can see that for any X, your grad V of X transpose Z is actually less than this quantity that we have over here. And my claim is that this quantity, that is the one within the round bracket, is less than or equal to 0, with equality if and only if. So, I should emphasize that it is if and only if your x, that is the argument that you have here, equals the expected value of x. So, why is that the case? So, first, let us understand, you know, what was the assumption on A that we had made? I think this was made a few lectures back, right? In particular, when we looked at a simpler variant of this problem where we had, you know, presumed that there was no noise and so on and so forth.



P06

$\nabla V(x)^T z$

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

$$\leq 0,$$

with equality only at $x=\mathbb{E}X$

For all $x \neq 0$ and $|K| \leq q$:

$$\sum_{i\in K} |a_i^T x| < \sum_{i\in K^c} |a_i^T x|$$

Furthermore, $V(x) \geq 0$ with equality if and only if

$$x = \mathbb{E}X.$$

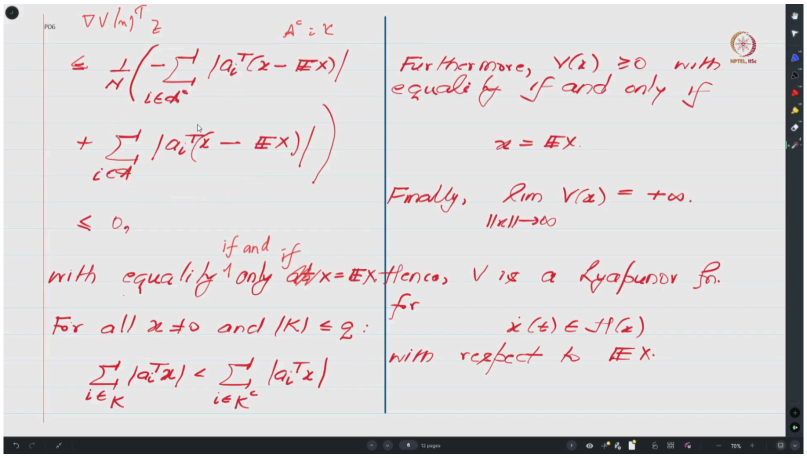
Finally, $\lim_{\|x\|\to\infty} V(x) = +\infty.$

Hence, V is a Lyapunov fn. for

$$\dot{x}(t) \in \mathcal{H}(x)$$

with respect to $\mathbb{E}X$.

So, in that case, from that Fauzi, Tabauda, and Digave paper that I had shown, this was the assumption on A. So, what was the assumption? That for every x which is non-zero and for every k, which is the subset of your row indices, right, whose cardinality is upper bounded by Q, right, we require that your sum of AI transpose X, where I runs over K, right, and the absolute value of AI transpose X, right, where I runs over K complement. This left-hand side should be strictly less than the right-hand side, so this was the property of your A matrix, that is, this condition should be true for every X which is not 0 and for every K for which this condition holds, right? Now, while your, you know, calligraphic A, which is the set of adversarial coordinates, is unknown, right, we have presumed that the maximum number of adversaries is at most K, so wherever your calligraphic A occurs, you can think of right, your A complement being K, so for that particular choice of K, right, and substituting X minus the expected value of X instead of X, one can see that this quantity will be negative, right? This quantity will be strictly negative because of this condition, right, and whenever this is not equal to 0, right, we will have this strict inequality, and hence this quantity will be less than or equal to 0 and



$\nabla V(x)^T z$ $\quad A^c = K$

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

$\leq 0,$

with equality if and only ~~at~~ if $x = \mathbb{E}X$

For all $x \neq 0$ and $|K| \leq q$:

$$\sum_{i\in K} |a_i^T x| < \sum_{i\in K^c} |a_i^T x|$$

Furthermore, $V(x) \geq 0$ with equality if and only if

$x = \mathbb{E}X$.

Finally, $\lim_{\|x\|\to\infty} V(x) = +\infty$.

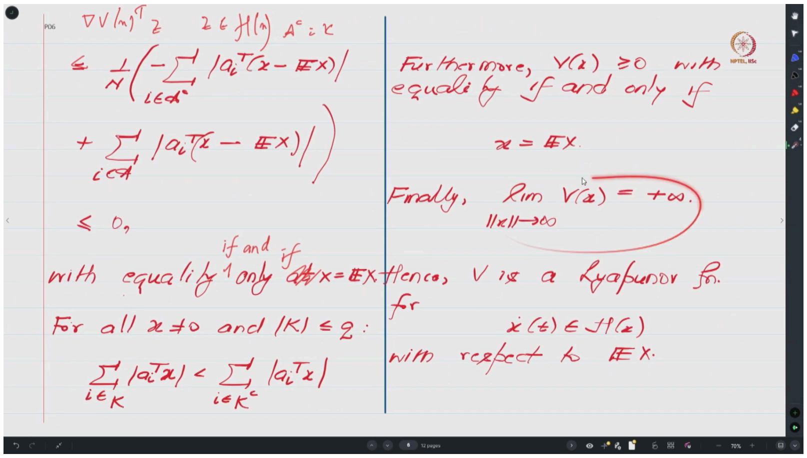
Hence, V is a Lyapunov fn. for

$\dot{x}(t) \in \mathcal{H}(x)$

with respect to $\mathbb{E}X$.

With equality if and only if X equals the expected value of X. Of course, when X equals the expected value of X, both these quantities are 0, in which case this whole sum will be 0. So, one can see why this conclusion holds true. So, from this, what we have managed to conclude is that For any X and any Z in H of X, capital H of X, so let me elaborate. So, Z belonging to capital H of X, one can conclude that this inner product is less than or equal to 0 with equality if and only if X equals the expected value of X.

Furthermore, one can see that the objective function itself is non-negative for all x, and this equality holds if and only if x equals the expected value of x. So, recall the definition of V that is given over here, and from this is where we are making that conclusion that x minus the expected value of x is greater than or equal to 0 with equality if and only if x equals the expected value of x. And furthermore, one can see that from this definition itself, as the norm of X goes to infinity, this quantity will also blow up to infinity. And since we have verified this condition, this condition, and the fact that this condition holds true, one can conclude that V is a Lyapunov function with respect to this differential inclusion. In particular, it is a differential—I mean, a Lyapunov function with respect to this point, the expected value of x. So, you can see that V of x is 0 if and only if x is the expected value of x. That is what this point over here means. And from that, one can conclude that your expected value of X is actually a globally asymptotically stable equilibrium for this differential inclusion.



P06

$\nabla V(x)^T z \quad z \in \mathcal{H}(x) \quad A^c = K$

$$\le \frac{1}{N}\Big(-\sum_{i\in A^c} |a_i^T(x-\mathbb{E}X)| + \sum_{i\in A} |a_i^T(x-\mathbb{E}X)|\Big)$$

$\le 0,$

with equality if and only if ~~at~~ $x=\mathbb{E}X$

For all $x \neq 0$ and $|K| \le q$:

$$\sum_{i\in K} |a_i^T x| < \sum_{i\in K^c} |a_i^T x|$$

Furthermore, $V(x) \ge 0$ with equality if and only if

$x = \mathbb{E}X.$

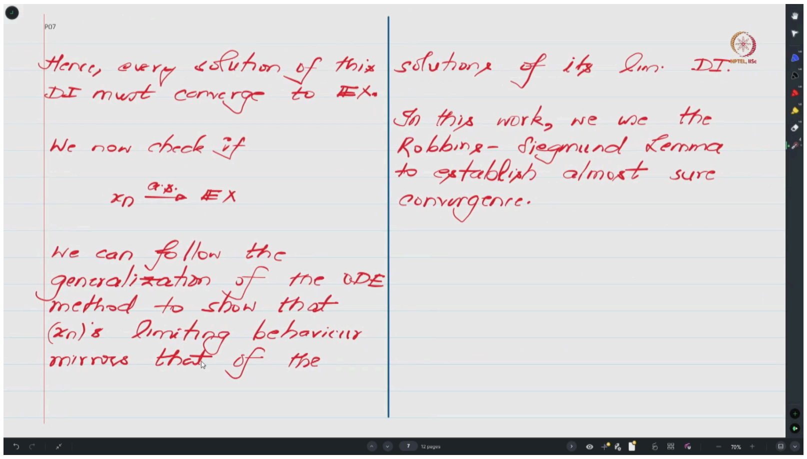
Finally, $\lim_{\|x\|\to\infty} V(x) = +\infty.$

Hence, V is a Lyapunov fn. for

$\dot{x}(t) \in \mathcal{H}(x)$

with respect to $\mathbb{E}X$.

And what this means is that, you know, if you take every solution trajectory of this DI, then it has to necessarily converge to the expected value of X as T goes to infinity. So, that is the conclusion. So, what this means is that at least for this idealized differential inclusion—I say idealized because there is no noise over here. For this idealized differential inclusion, whenever we presume that A satisfies a condition like this, then indeed your solution trajectories actually converge to the expected value of X. This is as desired.



P07

Hence, every solution of this DI must converge to $\mathbb{E}X$.

We now check if

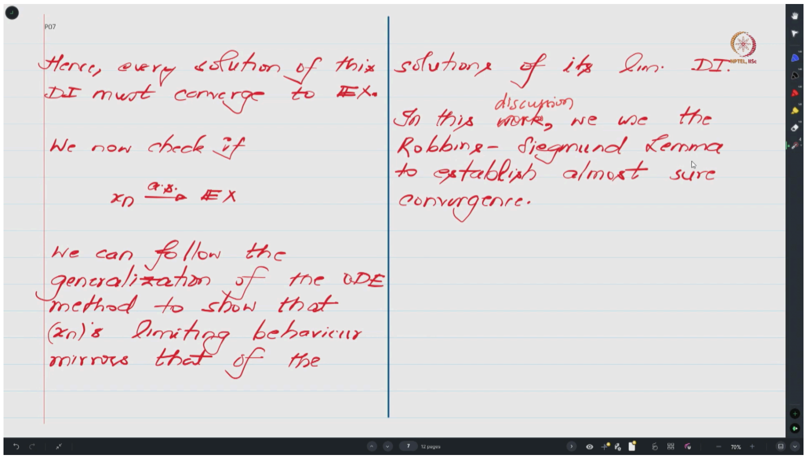
$x_n \xrightarrow{a.s.} \mathbb{E}X$

We can follow the generalization of the ODE method to show that (x_n)'s limiting behaviour mirrors that of the solutions of its lim. DI.

In this work, we use the Robbins-Siegmund Lemma to establish almost sure convergence.

Now, what we need to check is whether Xn converges almost surely to the expected value of X or not. So, recall Xn's are your iterates of your stochastic approximation algorithm. So, you know, of course, whatever we have studied so far, we can build upon that. That is, you know, previously we studied what is known as the ODE method. And that ODE method, what we showed was that, you know, under some conditions, the limiting behavior of your update rule is governed by a suitable ODE.

So, there is a similar generalization in the context of stochastic recursive inclusions as well, which says that if your Xn update is a stochastic recursive inclusion, then its limiting behavior is also governed by, you know, solution trajectories of this limiting DI. However, in this, you know, in our discussion, we will actually make use of a different result, which we will refer to as the Robbins-Siegmund Lemma. So, I will discuss this in the next class. However, you know, for those who want to, you know, make use of some result like this, I request you to, you know, perhaps refer to this research paper by Professor Shalabh Bhatnagar and his former PhD student Vinayak Yaji, who sort of gives some sufficient conditions under which the limiting behavior of stochastic recursive inclusions is governed by a suitable differential inclusion.



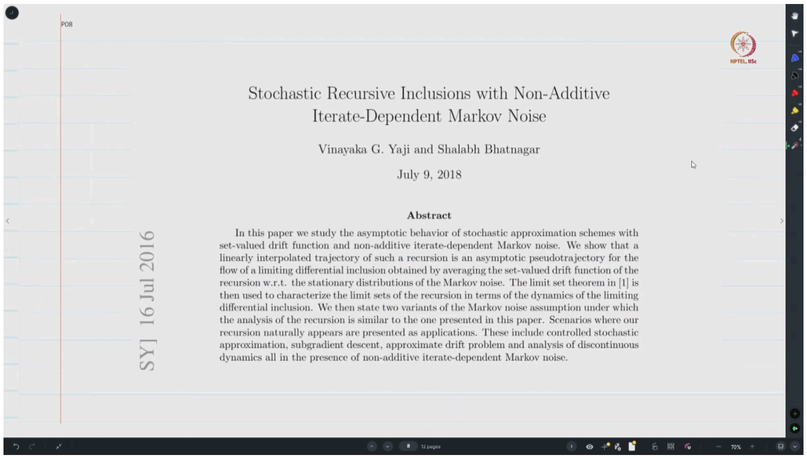
Hence, every solution of this DI must converge to $\mathbb{E}X$.

We now check if

$$x_n \xrightarrow{a.s.} \mathbb{E}X$$

We can follow the generalization of the ODE method to show that (x_n)'s limiting behaviour mirrors that of the solutions of its lim. DI.

In this ~~work,~~ discussion we use the Robbins- Siegmund Lemma to establish almost sure convergence.



Stochastic Recursive Inclusions with Non-Additive Iterate-Dependent Markov Noise

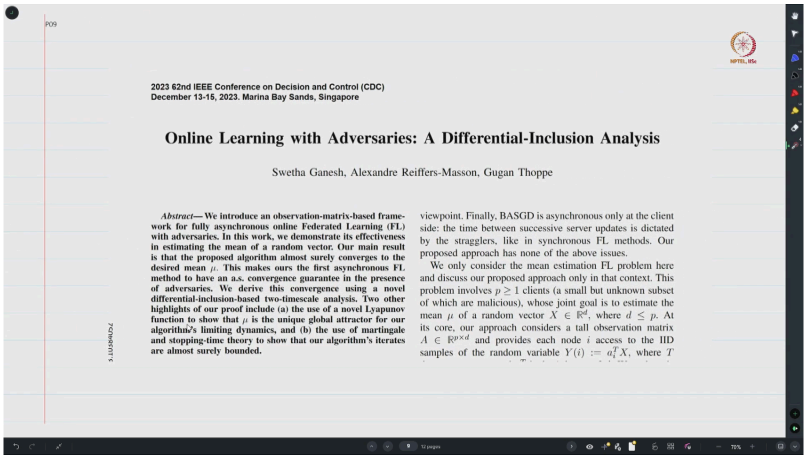
Vinayaka G. Yaji and Shalabh Bhatnagar

July 9, 2018

Abstract

In this paper we study the asymptotic behavior of stochastic approximation schemes with set-valued drift function and non-additive iterate-dependent Markov noise. We show that a linearly interpolated trajectory of such a recursion is an asymptotic pseudotrajectory for the flow of a limiting differential inclusion obtained by averaging the set-valued drift function of the recursion w.r.t. the stationary distributions of the Markov noise. The limit set theorem in [1] is then used to characterize the limit sets of the recursion in terms of the dynamics of the limiting differential inclusion. We then state two variants of the Markov noise assumption under which the analysis of the recursion is similar to the one presented in this paper. Scenarios where our recursion naturally appears are presented as applications. These include controlled stochastic approximation, subgradient descent, approximate drift problem and analysis of discontinuous dynamics all in the presence of non-additive iterate-dependent Markov noise.

And we had actually used this idea—that is, the connection between a stochastic recursive inclusion and its limiting DI—to actually discuss the almost sure convergence of this update rule that we have been discussing so far. However, that discussion is slightly more involved, and thanks to the TA for this course, Dr. Anik Kumar Paul, who suggested that we can perhaps use this, you know, this Robbins-Siegmund lemma to instead, you know, prove the almost sure convergence of—I mean, the result of this form, right? So, we will actually follow that because I realize that is actually very easy to explain and teach in class. So, in the next class, what we will do is we will actually make use of this Robbins-Siegmund Lemma to show the almost sure convergence of your stochastic recursive inclusion, which we designed over here.



2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Online Learning with Adversaries: A Differential-Inclusion Analysis

Swetha Ganesh, Alexandre Reiffers-Masson, Gugan Thoppe

***Abstract*— We introduce an observation-matrix-based framework for fully asynchronous online Federated Learning (FL) with adversaries. In this work, we demonstrate its effectiveness in estimating the mean of a random vector. Our main result is that the proposed algorithm almost surely converges to the desired mean μ. This makes ours the first asynchronous FL method to have an a.s. convergence guarantee in the presence of adversaries. We derive this convergence using a novel differential-inclusion-based two-timescale analysis. Two other highlights of our proof include (a) the use of a novel Lyapunov function to show that μ is the unique global attractor for our algorithm's limiting dynamics, and (b) the use of martingale and stopping-time theory to show that our algorithm's iterates are almost surely bounded.**

viewpoint. Finally, BASGD is asynchronous only at the client side: the time between successive server updates is dictated by the stragglers, like in synchronous FL methods. Our proposed approach has none of the above issues.

We only consider the mean estimation FL problem here and discuss our proposed approach only in that context. This problem involves $p \geq 1$ clients (a small but unknown subset of which are malicious), whose joint goal is to estimate the mean μ of a random vector $X \in \mathbb{R}^d$, where $d \leq p$. At its core, our approach considers a tall observation matrix $A \in \mathbb{R}^{p \times d}$ and provides each node i access to the IID samples of the random variable $Y(i) := a_i^T X$, where T

So, until then, goodbye, namaste, see you then, bye.