STOCHASTIC APPROXIMATION: THEORY AND APPLICATIONS

Dr. Gugan Thope

Department of Computer Science and Engineering

Indian Institute of Science, Bangalore

Week 10

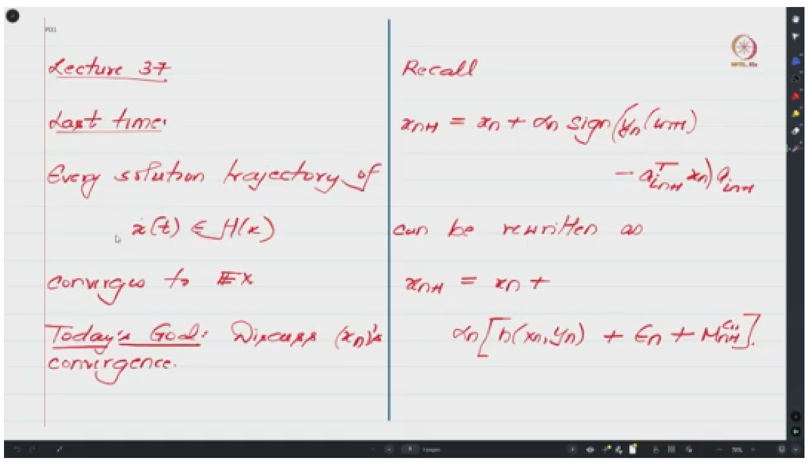
Lecture 37

Almost Sure Convergence via Robbins–Siegmund Theorem – Part 1

Hello and Namaste everyone. Welcome to lecture 37 of this NPTEL course on Stochastic Approximation. So, if you remember, in the last week and this week, we have been looking at, you know, some motivation for two-time scale Stochastic Approximation and Stochastic Recursive Inclusion. Towards that, we looked at solving a linear system of equations where we have access to only the samples of this right-hand side vector.

So, AX equals the expected value of Y, and we have access only to samples of this Y. On top of that, we are presuming that some of these measurements come from faulty sensors or adversaries. And to deal with that, we came up with an algorithm, and we said that that algorithm necessitates the use of two-time scale stochastic approximation because the gradient of this L1 objective that we considered results in a sign function which does not allow the expectation and sign to interchange freely, and that necessitated a two-time scale algorithm and because of the presence of adversaries and also because of the subgradient nature of the algorithm. Right, the subgradient descent nature of the algorithm, right, we were forced to consider a set-valued, you know, interpretation of the update rule, right, and in the previous class, we said that first let us try to see at least in an idealized setting, what is the behavior of the algorithm. In particular, we looked at the limiting differential inclusion, right, and we showed that at least all solution trajectories of the limiting differential inclusion indeed converge to the desired solution, which is the expected value of X. So, that is in the idealized setting. What remains to analyze is the behavior of the original stochastic algorithm itself, which, as I said, has this stochastic recursive inclusion interpretation, and in the next few classes, we will, you know, give a formal convergence analysis, you know, for this stochastic recursive inclusion.

So, let us discuss things in detail. So, as I said, you know, this was the limiting differential inclusion that we looked at, and we showed that every solution trajectory of this limiting differential inclusion converges to the expected value of X. So, our goal right now is to extend this discussion to the level of the convergence of your stochastic recursive inclusion algorithm itself, and recall that Xn had the update rule that is given over here, that is, it is Xn plus alpha n sign of Yn of In plus 1 minus Ain plus 1 transpose Xn times Ain plus 1, right. Here your In plus 1 was sampled uniformly from the set of indices 1 to n. And this little yn, at least for honest coordinates, right, tries to approximate your expected value of y of i, right.



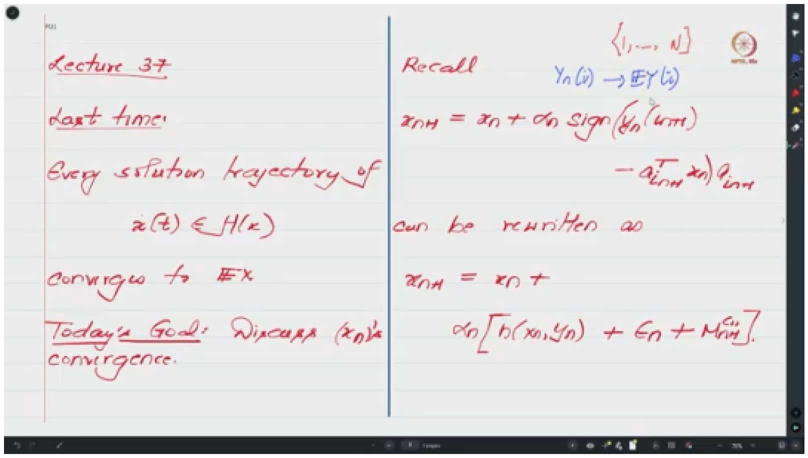
So, it basically takes the running average of the observations for this coordinate y of i. So, you can look at the update rule that we had discussed in the previous class. And then I said that this update rule can be rewritten in this form.

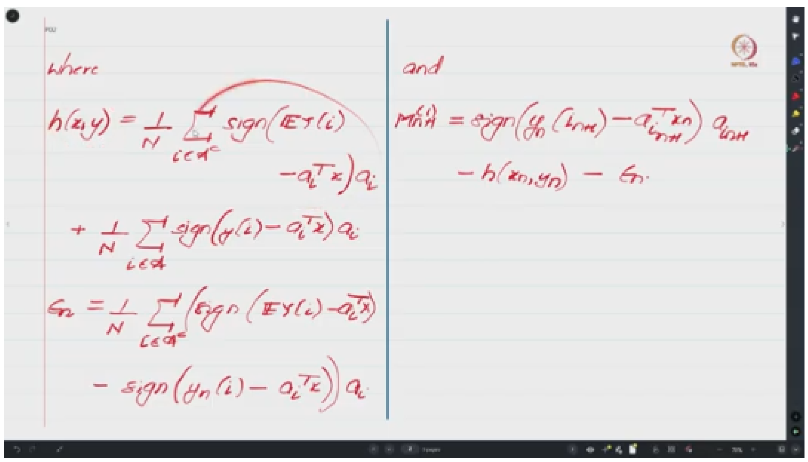
$$x(t) \in H(x)$$

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

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

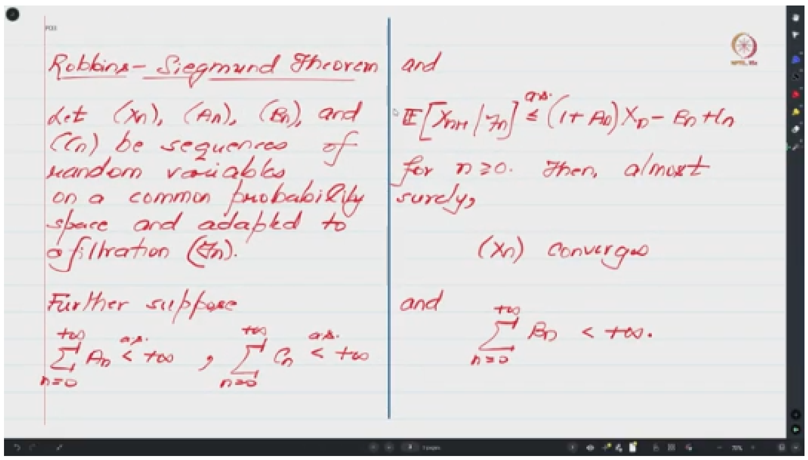
So, we can rewrite this as h of xn yn plus epsilon n plus mn plus 1 1 where h of xy is having this expression. with expected value of Y that is this idealized expression at least for I in the set of you know non adversaries and this is the expression that we cannot in some sense control because here I belongs to the set of adversaries. So we said okay let us define H of XY in this fashion however the update rule does not have H of XN YN.





Instead, it has this expression. So what we do is we add and subtract things that are, you know, not present in the original update. And towards that, we first introduce this epsilon n, where what we do is we first look at the set of non-adversaries. And, you know, wherever your expected value of y is there, we try to replace it y n and sort of subtract it. So I think there is a sign issue here.

When I add epsilon n to this, this term and this term should cancel off and to enable this, I think I should have a minus here and a plus here, sorry about that, okay. So, we have some expression like this and if you plug in H of Xn and Yn you can see that this term and this term right you know will in some sense represent H of Xn Yn plus epsilon n right and if you take the conditional expectation of this expression then one can see that you would end up with this term plus this term over here making this Mn plus 1 as a martingale difference term. So, we want to analyze its you know convergence and towards that we will make use of this Robin Sigmund theorem. So, I had told you that one can actually make use of the generalization of the ODE method in the context of differential inclusions. And, you know, somehow repeat the arguments that we had, you know, discussed in the context of the ODE method in the for convergence analysis of stochastic approximation algorithms.



However, you know, my, you know, TA for this course, who is also doing a postdoc with me, Dr. Anik Kumar, pointed out that maybe we can come up with a simpler argument for showing this result, and I agree with him. So, the simpler argument is by invoking this Robbins-Siegmund theorem. So, what does the Robbins-Siegmund theorem say? Suppose you have a sequence of, you know, random variables on a common probability space. If I am not wrong, I think the Robbins-Siegmund theorem also requires that all these random variables be non-negative random variables.

And let us say they are all defined on the common probability space, and let us say, these random variables are adapted to a filtration Fn, and further suppose that, you know, if you add these An's, they are less than infinity, and similarly if you add the Cn's up, they are less than infinity almost surely. Furthermore, if you take the conditional expectation of Xn+1 given Fn, then this is upper bounded by terms of this form, that is, 1 plus An times Xn minus Bn plus Cn for all n greater than 0. Then what this Robbins-Siegmund theorem says is that almost surely your Xn will converge. Furthermore, your Bn's will also be additive.

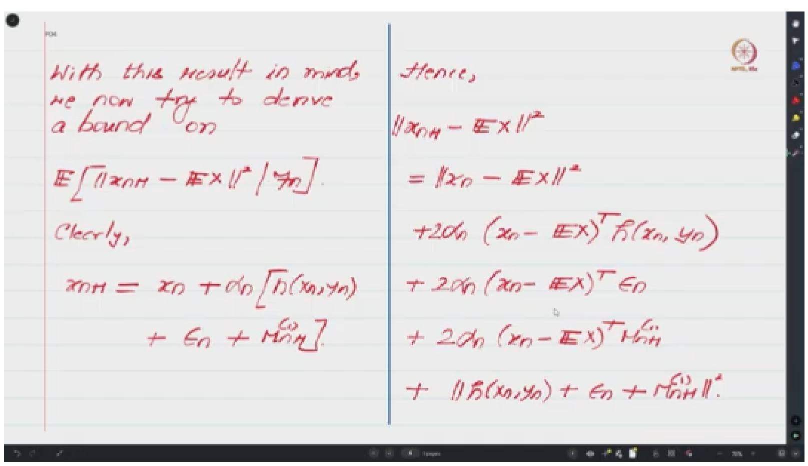
So the Robbins-Siegmund theorem says that if you have a sequence of random variables, non-negative random variables, which satisfy these inequalities, and furthermore these conditions are satisfied, then almost surely your Xn converges and your summation Bn is finite.

$$\sum_{n=0}^{\infty} A_n < +\infty, \quad \sum_{n=0}^{+\infty} C_n < +\infty$$

$$E\left[F_n\right] \leq (1 + An)X_n - B_n + l_n$$

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

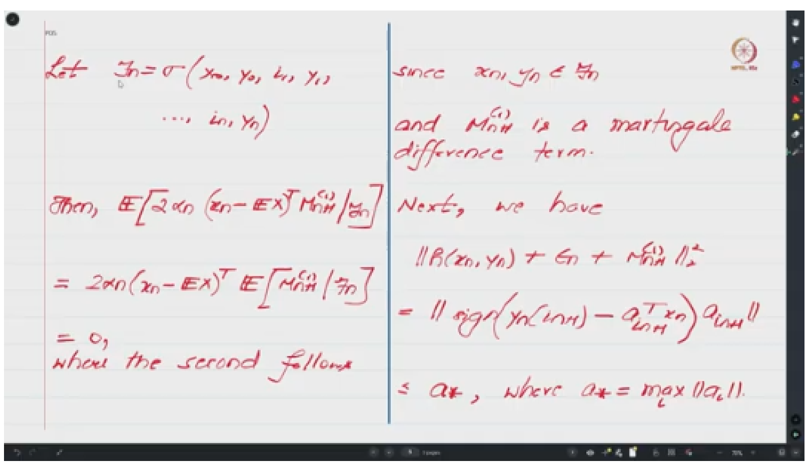
So now with this result in mind, right, we are going to derive a bound on this norm of little xn+1 minus the expected value of x squared, right? So, you know, whatever your capital Xn+1 was over here will now be played by this expression over here. So we want to obtain a bound on this expression which mirrors this form, and you know, whatever An and Cn we construct indeed have these properties so that, you know, we would be able to conclude these two things over here, right?



So, we will worry about what is the conclusion and so on and so forth. At this point, we will try to derive a bound on this expression. Now recall that Xn plus 1 has an update rule like this. So I can subtract the expected value of X from both sides and then square it. So if I square it, the left-hand side will be Xn plus 1 minus the expected value of X the whole square, and on the right-hand side notice that I have, you know, A plus B plus C plus D type expression and I am taking the square of it.

So first I will have, you know, some square here plus 2 times the cross terms. So, the first cross term will be Xn minus the expected value of X transpose H of Xn Yn. The next cross term will be something like this, and the third cross term will be something like this, right? And finally, you know, I would need to take all these terms over here and square them up. I realize there is a typo here.

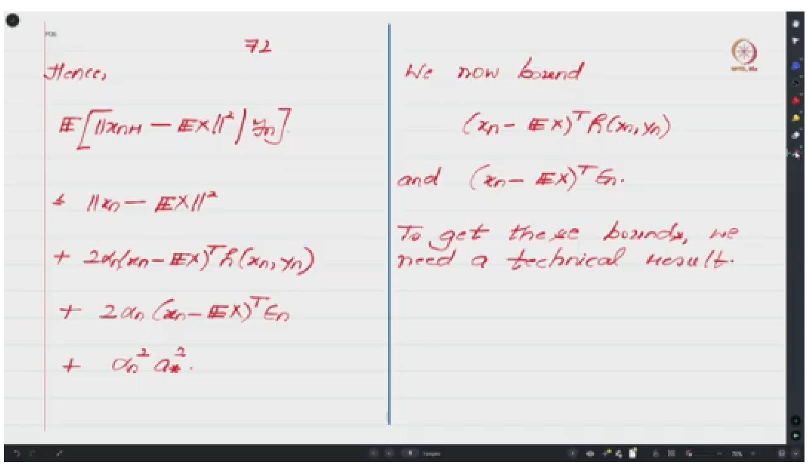
So, this should be alpha n square, right? Is this okay? So, our goal would be to take the conditional expectation of the right-hand side and come up with an expression that mirrors this structure over here, right? So, towards this, what we will do is we will define a sigma field which is basically made up of your initial estimates for X0 and Y0, and furthermore we will include in it I1, Y1 all the way up till IN, YN. In some sense, this



Filtration or this sigma field right captures the information that we have until time n. So, if you take the conditional expectation of this expression with respect to such an Fn, one can see that you know this expression is measurable with respect to this filtration or this, you know, if you I mean okay I should not repeat it. This is measurable with respect to this Fn. And hence I can you know pull out all these terms from this conditional expectation and we will be left with this. And if you recall the you know description of Mn plus 1 right, you know in this expression and this expression the difference is that you know you have your In plus 1 and you treat it as a random variable, keep imagine everything else to be a constant and only In plus 1 as a random variable, and if you take the conditional expectation you will get an expression like this. So, because of this reason If you take the conditional expectation of this, this will be 0 right. So, let's go back. Our goal was to take the conditional expectation of this expression and on the right hand side we had 1, 2, 3, 4, 5 terms right. Now this term over here right is nice for our analysis because if you take the conditional expectation of this term, we end up with 0. So, this leaves us with 1, 2, 3 and 4 and we will see how to deal with these 4 terms.

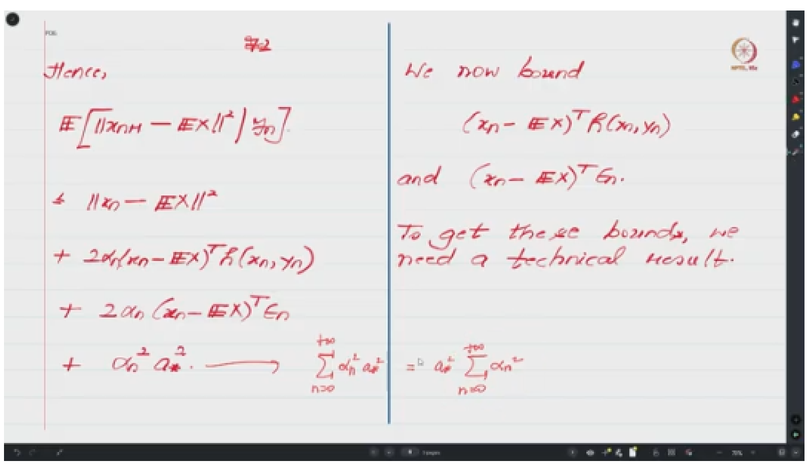
So, now what we will do is we will focus on this term. So, if you remember, if I add these 3 terms, I should get an expression like this because this is what our original update rule was. Now, this sign expression is upper bounded by, I mean the absolute value of this expression is upper bounded by plus 1. So, hence we will be left with norm of just this vector and recall that Ai transpose is the ith row of your matrix A. Right and this is the you know column representation of that row vector.

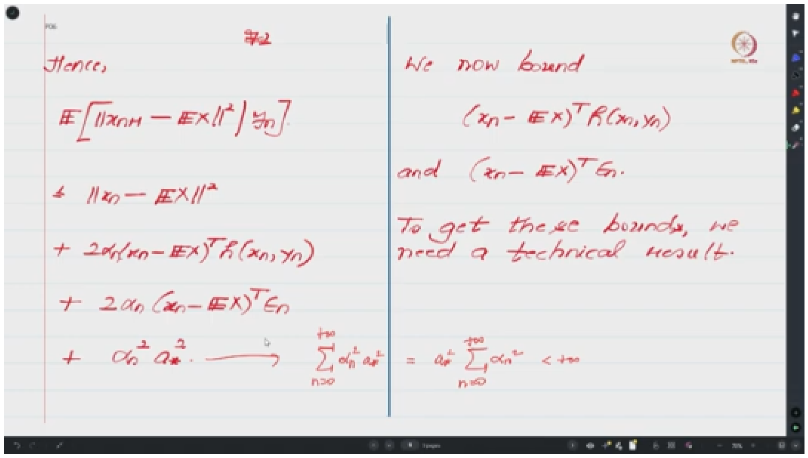
So, if you take the norm of this a i n plus 1 at most it can be a star where recall we had defined a star to be the max of norm a i right. So, we had done something like this. So, from this perspective I should maybe put a square here, sorry about that ok. So, this square is actually a star square ok and I should have a square here as well. So, from this what one can see is that you know forget about this.



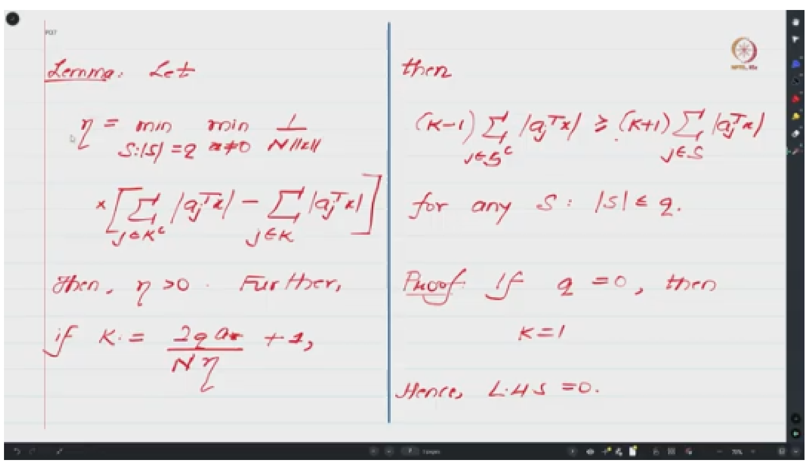
Okay, what one can see is that the left-hand side, whose upper bound we want to find, is upper bounded by this term. We have not touched this term; also we have not touched right, and this term also we don't touch. But there were two other terms: so one of the terms' conditional expectation was zero, so that vanished, and then there was this last term. Where we showed that it is now upper bounded by alpha n squared times a star squared. So let us look at this expression for the time being. And I hope you agree that if I sum it up, that is from n equals 0 to infinity. Alpha n squared times a star squared, and because a star does not depend on n, I can pull this off, right, and I would end up with something like this.

So if I presume my step sizes are square summable, then indeed this expression that I have here will be less than infinity. So, if you go back to that Robbins-Siegmund theorem, whatever is the last term, I hope you agree that I can subsume it within Cn because in Cn, we require that this summation Cn should be less than infinity. So, at least for one of those terms, we can see that such a condition indeed holds. So now you know, in today's class and the next class that follows, we will try to bound these expressions; that is, the expressions that we have over here. So again, we want to write it in a form so that we end up with some minus bn type cn type expression, so that this whatever is there within cn, if you add it up, that will be less than infinity.





So, whatever we end up doing should enable us to apply this Robbins-Siegmund theorem. So, to derive bounds especially on this expression, we need a technical result, and in today's class, we will mainly focus on this technical result, and in the class that follows, we will make use of this technical result and try to derive a bound on this expression over here. Is this okay? So what is the technical result? It says that suppose eta is defined in the following way.



I will go over the details. So eta is some number. Then the result says that whatever this eta is, this is strictly bigger than 0. Furthermore, if you define the constant K in this fashion, so let us go over the terms here. There is 2 here, which is the number 2.

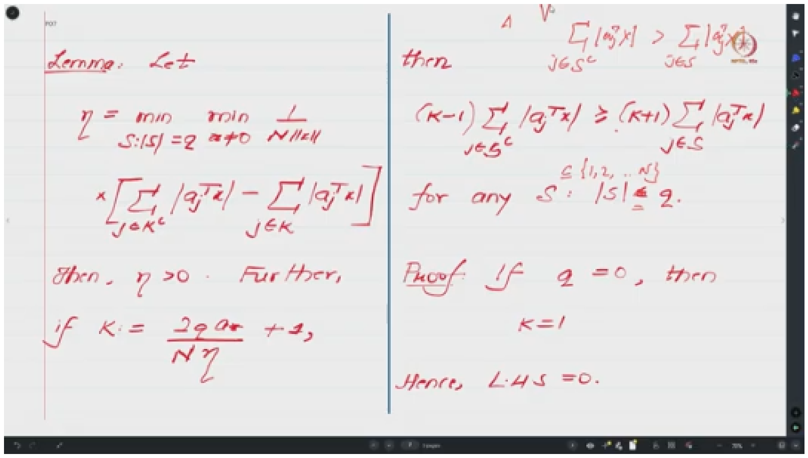
There is Q here. Q, recall, is an upper bound on the number of adversaries. Then you have... And we presume that Q is known and A star is again an upper bound on the norm of your A j's and divided by N times eta. So, N is the number of rows in your matrix A.

In other words, this is the number of coordinates of your vector Y, and you can think of this as the total number of measurements that we have.

or measurement sensors that we have, and this eta is what you have over here. And since eta is greater than 0, this quantity over here is finite, and what we do is we take this quantity and to it add plus 1. So, this is your K, and the claim is that you give me any S which is a subset of 1 to N right, you give me any set S which is a subset of this, and its cardinality S okay so at this point I want to keep it equal, and you give me any S such that its cardinality is equal to Q, then the claim is that such an inequality holds that is K minus 1, notice that K is something plus 1 so K minus 1 will be strictly positive, so this quantity is strictly positive right, times this expression And we are claiming that this should be greater than K plus 1 times this.

And recall that S over here is of cardinality Q. So, you know, presuming that Q is less than N over 2, one can see that we are looking at the sum of A j transpose X. on a small set of coordinates and we are looking at the sum of A j transpose X over the remaining coordinates. So this expression is over a larger set of indices and this expression is over a smaller set of indices. In particular, I can take S to be We will actually do that, but here I am insisting that this inequality should hold for every S, the reason being that I do not know which set of adversaries that is something I do not know.

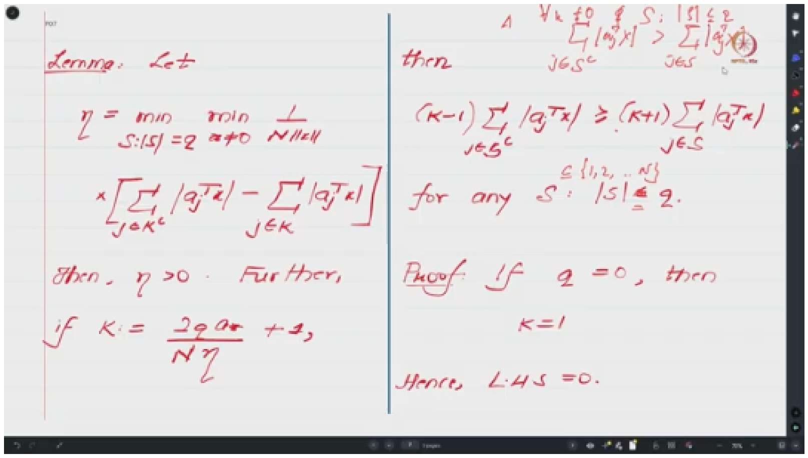
So, hence I am at this point trying to ensure that this inequality holds for every S, and I mean if you think about it, what this inequality is trying to say is the following. So, remember we had some assumption on the A matrix. So, the assumption on the A matrix was that for any such subset S, if I take A j transpose X and take this sum over S complement, this should be strictly bigger than summation j in S A j transpose X. This was the assumption that we had about A. In particular, we said that for all X, not equal to 0 and any S such that the cardinality of S is less than Q, I require that some inequality like this holds. Is this okay?



So I require that some inequality like this holds, and here if you notice, we have similar expressions to what I have here. So whatever this expression I have here and whatever this expression I have over here, but I am not just saying that this is strictly bigger than this. I am instead saying that this expression divided by this expression is greater than k plus 1 over k minus 1. So this inequality over here is basically saying that if I take this left-hand side and divide it by the right-hand side, this ratio will be strictly bigger than 1. Here on the other hand, I am sort of quantifying how big this ratio is guaranteed to be; that is, if I take this expression divided by this expression, then it is at least bigger than k plus 1 over k minus 1, right? And since the numerator is bigger than the denominator, that is k plus 1 is bigger than k minus 1, one can see that this ratio, that is k plus 1 over k minus 1, is going to be strictly bigger than 1.

So, in other words, whatever is the given assumption, We are sort of expanding on that assumption and trying to quantify: is there some lower bound on this ratio that I can get? And the nice thing is, this lower bound that we have over here universally applies. That means, for any x, and S for which such a condition holds, right?

Indeed, this, you know, relationship holds, right? So, this K does not depend on the choice of X, right? So, you will see how this helps. In particular, an inequality like this only says that this is bigger than this, but it is possible that, you know, maybe if I take a very crazy X, right? Maybe this ratio, that is left-hand side by right-hand side, is very, very close to 1.

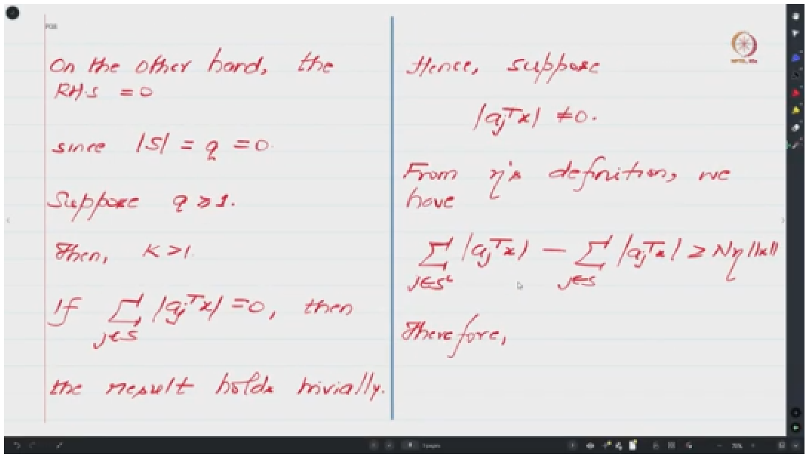


And maybe I can get a number which is as close to 1 as desired. So, what this result of R tries to tell us is that that won't happen. This ratio, irrespective of the choice of X, will be lower bounded by K plus 1 over K minus 1. So, that is the conclusion that we are trying to derive over and above what is the assumption that was given to us. So now we are going to prove this inequality, and you will see in our subsequent class how this inequality enables us to derive bounds on such an expression over here, and separately we will see how we can derive bounds on this expression as well.

So let us try to prove this. I want to prove this whatever be the value of Q. So to begin with, let us presume Q is 0. So if Q is 0, which means there are no adversaries, then in that case one can see that K is equal to 1. Only in the case where Q is 0 is K equal to 1, and for any positive number of integers, K will be strictly bigger than 1. So when Q is 0,

this left-hand side is 0 because K is 1. Now, let us see what happens on the right-hand side. Because Q is 0, whatever s you pick, this will be an empty set. And because this is an empty set, the right-hand side will also be 0. And hence, one can conclude that this inequality holds for the case where Q is 0.

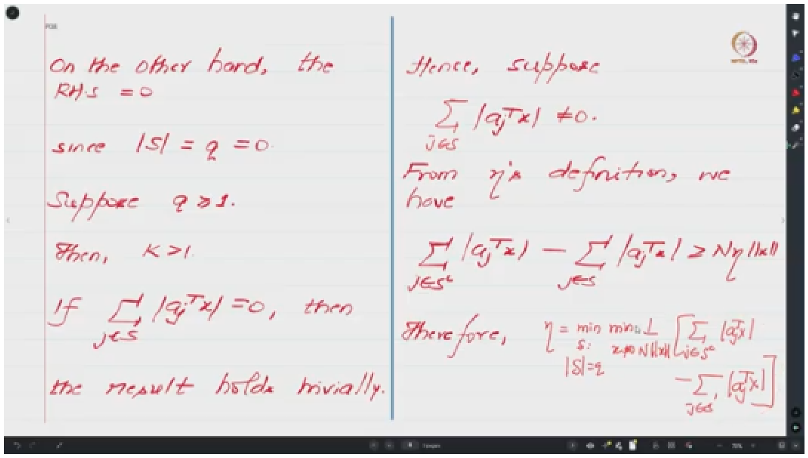
So, hence, it only remains to see what happens when Q is bigger than or equal to 1. Now, when Q is bigger than or equal to 1, from the definition of K, one can see that K will be strictly bigger than 1. That is what I have written over here. So, now we want to prove this, you know, inequality, right? So, towards that,



So, what we will presume is that you know this left-hand side to begin with is sorry the right-hand side this expression is 0. So, let us presume that this expression is 0. So, this expression is 0 well the right-hand side will be 0. So, since the right-hand side is 0 whatever be the left-hand side since I have already shown k is bigger than 1. Whatever be the left-hand side, this inequality will trivially hold.

So, that is what I have mentioned here. If the summation on the right-hand side of the inequality that we wish to prove is 0, then the result trivially holds. Hence, it only remains to consider the case where this summation is actually not 0. So, I should say summation j in S, you know, absolute value of a j transpose X is not equal to 0. So, now if you look at eta's definition.

So, let us recall eta's definition. So, eta was min over s such that cardinality of s equals q min over x not equal to 0 1 over n times norm x right norm x times summation j in s complement A j transpose X minus sum j in S A j transpose X. So, this is the definition of eta. So, since eta is the min of all those things, so if you pick an arbitrary S, then whatever is this expression that you have, so since eta is the min of all these quantities, so if you pick arbitrary S right this expression that you have over here will be bigger than eta right so if I just rearrange the terms now and keeping in mind that your X is presumed to have nonzero norm one can see that because eta is the min of this eta will be less than this quantity for any arbitrary choice of X and S and



And hence, whatever is your s and x, this expression will be greater than n times eta norm of x, right? So, now, if I take this quantity that I have and divide on both sides, I hope you agree that I will end up with summation j in s complement absolute value of aj transpose x divided by summation j in s absolute value of aj transpose x is greater than 1 plus whatever n eta norm x is divided by the sum of j and s absolute value of aj transpose x, which is exactly this expression over here, right?

$$|s| = q = 0$$

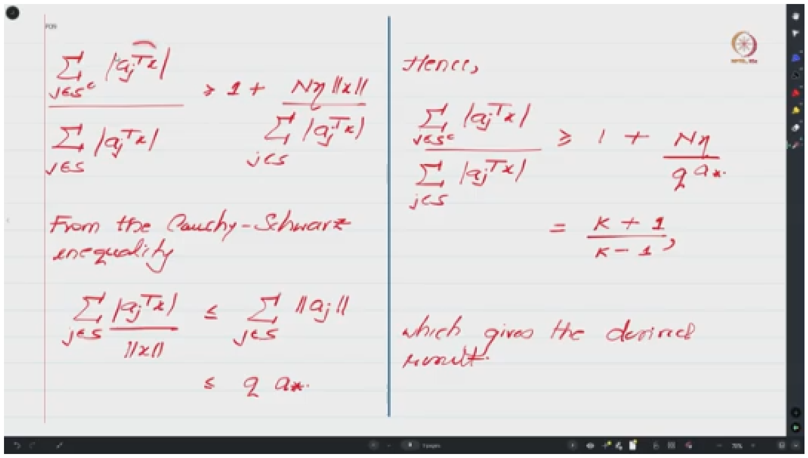
$$q \geq 1$$

$$\sum_{j \in s} \left| a_j^T x \right| = 0$$

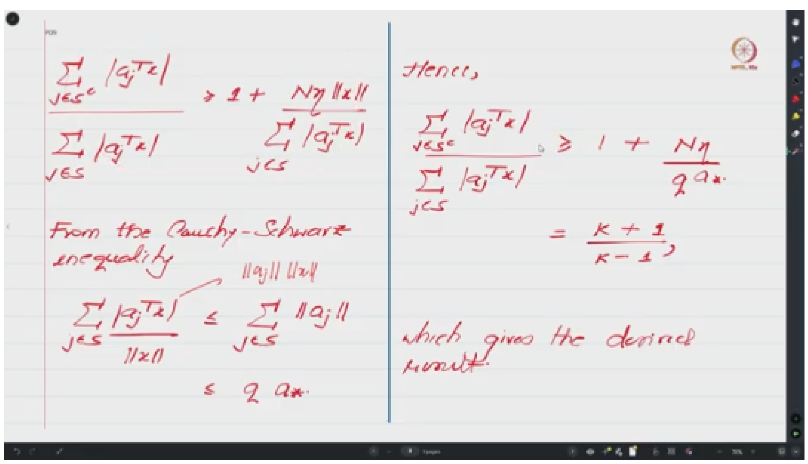
$$\sum_{j \in s} \left| a_j^T x \right| \neq 0$$

$$\sum_{j \in s} \left| a_j^T x \right| - \sum_{j \in s} \left| a_j^T x \right| \geq N_\eta ||x||$$

Now, I would also like to highlight that I can actually divide by this quantity because in this sub case that I am considering, I have presumed that this expression over here is not 0 and hence is strictly positive. Now, from the Cauchy-Schwarz inequality, we know that this expression that we have over here, and if I divide it by norm x, from the Cauchy-Schwarz inequality, I know that this expression is upper bounded by norm of aj by norm.

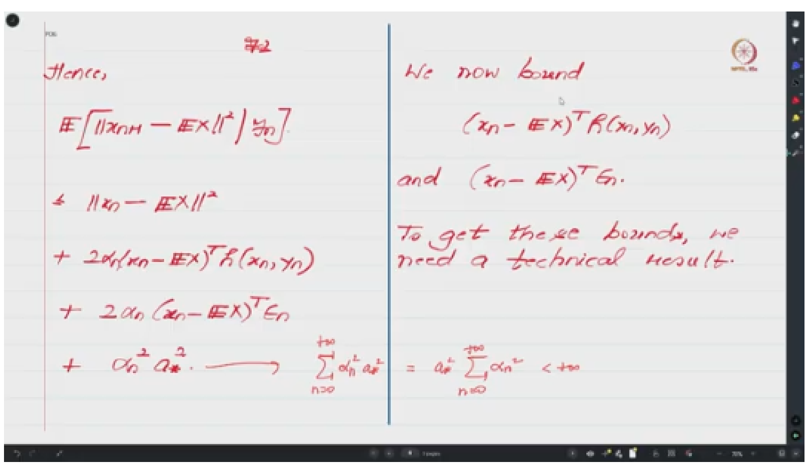


norm of x. So, this x and this x will cancel off and hence I will be left with sum of norm aj and since we have presumed that A star is the largest norm and since cardinality of s is q, one can see that this sum is upper bounded by q times A star. So, this expression is upper bounded by q times A star irrespective of the value of x and in particular, we have presumed that x is having, you know, does not have norm 0. So, from this calculation, that is, this is less than this, one can show that this whole expression is greater than 1 plus n eta, which comes from here. And we have shown that this expression is less than q A star.



And hence, one can conclude that this expression over here is greater than this, which is greater than 1 plus n eta by q A star. And from the definition of K, one can immediately see that this expression equals K plus 1 over K minus 1 and since this expression is greater than K plus 1 over K minus 1, one can see that the desired result holds. So, let me take you back to the desired result. was basically something like this okay so this is what we have indeed managed to show for K which is defined in this fashion right now there is one part that is remaining to be shown that is to show that this eta is actually strictly positive I proved this result presuming eta is strictly positive in the next class I will first begin by showing that indeed this eta which is defined over here is strictly positive and then I will use this result to obtain a bound on this expression.

So, let me conclude today's talk. In today's talk, we looked at this notion of the Robbins-Siegman theorem, which somehow said that if you had these random variables Xn, An, Bn, and Cn, and if they satisfied some inequality, then one can use certain conditions to talk about the convergence of Xn. And we sort of said, okay, let us see if we can use this result to discuss the convergence of the stochastic recursive inclusions iterates; that is, your little xn's, and that is what we started doing. Towards that, we wanted to obtain an upper bound on the conditional expectation of xn plus the square of the norm of xn plus 1 minus the expected value of x. And we did some preliminary calculations and we were able to see that at least two of the terms on the right-hand side of this bound can be easily dealt with, right? And towards deriving bounds on the remaining one, we discussed an intermediate result which somehow tried to quantify the mass that is there on the Q many coordinates to the mass that is present in the remaining set of coordinates related to AJ transpose X for any X.



So, we will be using these results to derive bounds on the remaining terms in the next class. Until then, goodbye and namaste.