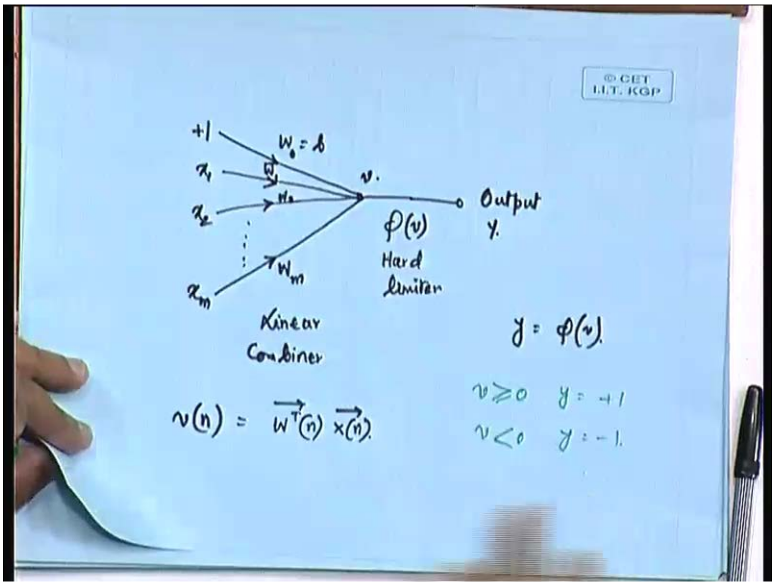
Neural Network and Applications
Prof. S. Sengupta
Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology, Kharagpur

Lecture - 16
Perceptron Convergence Theorem

Today's, lecture is on the Perceptron Convergence Theorem. Now, last class, we were discussing about the least mean square algorithm. And generally, it happens in most of the classes, that even after the class, we have some discussions continuing some doubts cropping up. But, surprisingly for the least mean square algorithm lecture, I did not have much of doubts from the students. So, it is a hope, that the last class was relatively less complicated as compared to things like three dimensions and the shattering, part of it.

So, I assume that the convergence criteria of the least mean square, has been clearly understood and we are now going in for the Perceptron convergence theorem. Basically, the theory part of it is more or less, the same thing that we have been discussing for the past few lectures, related to the single layer Perceptron. So, the architecture that we are considering is as before an m input Perceptron.

(Refer Slide Time: 02:30)



$y = \varphi(v)$

$v \geq 0 \quad y = +1$

$v < 0 \quad y = -1.$

$v(n) = \vec{W}^T(n)\, \vec{x}(n).$

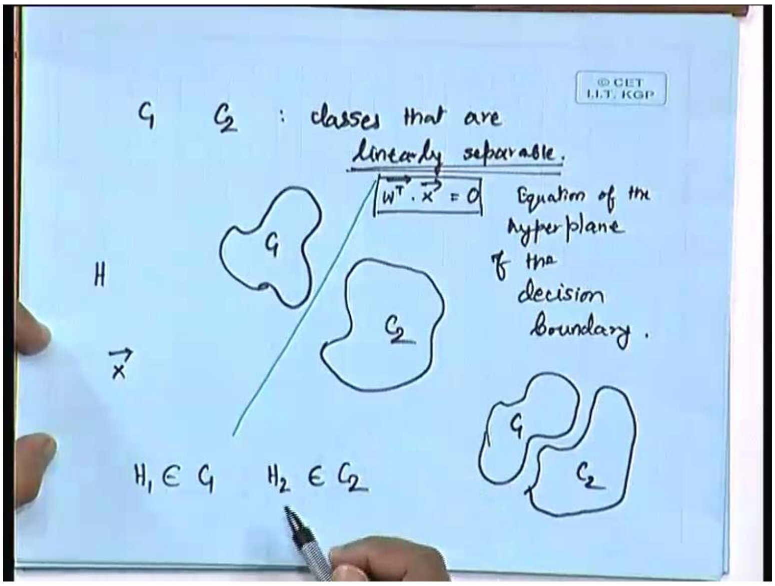
So, we were having a linear combiner that combines the inputs like this, so we are having the inputs as X 1, X 2 up to x m and the associated weights, we are calling as W 1, W 2 and W m. And in this model we are also adding the bias in the linear combiner, directly; that means to say that we are having the input to be of a fixed value of plus 1 and having the weight W 0 equal to the bias b.

So, that effectively, this is a linear combiner, which as you know should add up all this W 0 X 0. In fact, here is equal to 1, plus W 1, X 1 or in other words what we have been writing all the time, W transpose X or x transpose w whatever way. Now, this gives you basically the output v and if this is the output v, that we are getting after the presentation of the nth pattern input, then we write it as W T n X n. Where, W T n is the transpose of the weight vector and X n is the input vector for the nth pattern that we are feeding.

So, this is what the v is and then we are following it up with a herd limiter, where what this herd limiter do is that, this v gets modified into phi v, that the phi v will be transforming it in to herd limiting it in to plus 1 or minus 1. So, this is a herd limiter that will be used after that. So, it is like a binary pattern classification that we are considering, that being the simplest case and this is the output to y; that we are taking. So, basically y is nothing but equal to phi v.

Now, this being the simplest kind of the architecture that we consider, we now attempt to solve a pattern classification problem and let us say that we are solving a two class problem.

(Refer Slide Time: 04:59)


© CET
I.I.T. KGP
C_1 C_2 : classes that are linearly separable.
$\vec{W}^T \cdot \vec{X} = 0$
Equation of the hyperplane of the decision boundary.
C_1
C_2
H
$\vec{X}$
C_1
C_2
$H_1 \in C_1$ $H_2 \in C_2$

In which case, we have got existence of two class's c 1 and c 2, so these are the classes that are linearly separable. In fact, both these classes c 1 and c 2 will be existing in the multidimensional space actually. So, all though I am just drawing some representative diagram, but in general, this is to be considered in a multidimensional sense. So, let us say that, this is a pattern, that we are having c 1 supposing this one is the whole set of c 1, that we have got and let us say that we have got a set of c 2.

So, supposing this is the set c 1 and let us say that this is the set c 2, now clearly the way I have drawn it, makes it linearly separable. I think you can just get it, because what we can consider is that, we can simply imagine that, there is a line. Although, in this case, I am showing it as a line, but actually it will be a

Student: Hyper Plane

Hyper plane, that is correct and can you tell me the equation of that hyper plane, because we are in this case, the kind of model that we are using is that for v greater than or equal to 0, we will be having output equal to plus 1 and for v less than 0, we will be having the output to be equal to minus 1. So, if that is the case, then can you tell me that what is going to be the equation of this hyper plane.

What is the equation of that hyper plane, you see v n is this, v n is a scalar quantity, no doubt, but it is are the dot product of these two vectors

Student: ((Refer Time: 07:00))

That is very correct, so here the decision boundary, because this hyper plane that we are imagining, effectively is the decision boundary. And that decision boundary should lie, at v is equal to 0 and what is the equation of v is equal to 0, in this case, because we have to express it in terms of the W's and the x. So, that means, to say that the equation of the hyper plane can be best written as W transpose dot X vector. So, this is the decision boundary.

So, W transpose X vector is equal to 0, this is going to be the equation of the hyper plane of the decision boundary. Now, in this case, it is very clearly linearly separable, but all cases are not linearly separable. Simply, you could take the problem like this that is say that, this is our c 1 and let us say that this is our c 2. And in this case, we certainly cannot find any hyper plane that separates the class c 1 from class c 2.

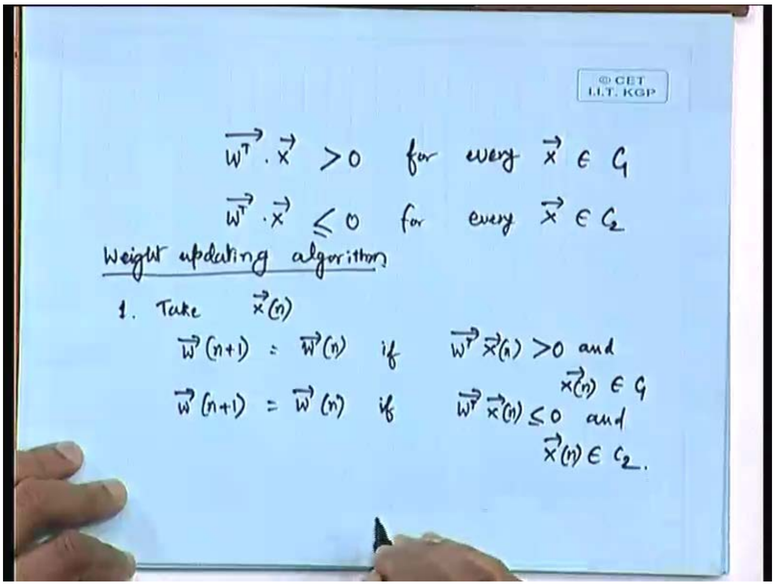
So, the essential point that must be borne in mind is of course, the linearly separable, the fact that we have been harping upon again and again. Now, again we are referring to it to class classification problem, so that means to say that the pattern set, that we are having must be belonging either to the class c 1 or to the class c 2. So, you pickup any pattern, any input vector x that we are going to pickup, from the m dimensional space is either going to lie in c 1 or in c 2.

And in order to train the system, in order to train this Perceptron, we have to take a set of patterns. So, when we when we take a set of patterns for training, naturally we have to derive the training set from both these classes. So, we have to train it with sufficient number of patterns taken out of the class c 1 and sufficient number of patterns to be taken out of class c 2.

So, let us say that for the training purpose, we take H 1 as the training set, which belongs to the class c 1 and H 2 as the training set that belongs to class c 2. So, effectively our training patterns that we are composing, basically is the class H, which is nothing but the union of this classes, H 1 and H 2. And H 1 and H 2 are the subsets of the actual class c 1 and c 2. Because, certainly we are not taking every sample out of the space, we are only taking a subset of that; that is very clear.

So, now that we are classifying it in to a two class problem, then the classification equation, we can simply write, given that the decision boundary is the W transpose X vector equal to 0.

(Refer Slide Time: 10:40)



© CET
I.I.T. KGP

$\vec{W}^T \cdot \vec{X} > 0$ for every $\vec{X} \in C_1$

$\vec{W}^T \cdot \vec{X} \leq 0$ for every $\vec{X} \in C_2$

Weight updating algorithm

1. Take $\vec{X}(n)$

$\vec{W}(n+1) = \vec{W}(n)$ if $\vec{W}^T \vec{X}(n) > 0$ and $\vec{X}(n) \in C_1$

$\vec{W}(n+1) = \vec{W}(n)$ if $\vec{W}^T \vec{X}(n) \leq 0$ and $\vec{X}(n) \in C_2$.

We can definitely write that W transpose is X vector is greater than 0, for every X vector that belongs to the class c 1; that means to say it is drawn from H 1 set. And W transpose X vector is less than or equal to 0, for every X, that belongs to c 2. In fact, you can doubt the case of equal to 0, because where will we fit in equal to 0. Whether, we put it with the greater than sign or with the less than sign, we can put it in any one of them, in fact, equal to 0 is the boundary.

So, whether you club it with the greater than case or club it with the less than case, that is something that you can arbitrarily do, so this is essentially the classification, that we want. So, that means to say, that when we pickup an X vector from the class c 1, we would like to have that W transpose X should be greater than 0. If that is the case, we say that we have classified correctly.

If we are knowing that X belongs to the class c 1 and if we have got W transpose X to be greater than 0, then obviously, X will be classified in to c 1 which means to say it is a correct classification. But, supposing us we pickup X vector from the set H 2, which normally belongs to the class c 2. So, supposing we take X from the class H 2 from the set H 2, then if we get W transpose X, still greater than 0, supposing we get; that means to say, it is correct or incorrect.

It is incorrect classification, because if x belongs to the set c 2, we should have got W transpose X to be less than or equal to 0. So, the classification that the network will ultimately do could be right, could be wrong. If it is wrong, then what we do, if it is wrong, then we needs to update the weights. That means to say that, after find you feeding the nth pattern I find that the W transposes X is wrongly classifying the pattern. Then I should update the weight, so as to get a new W n plus 1.

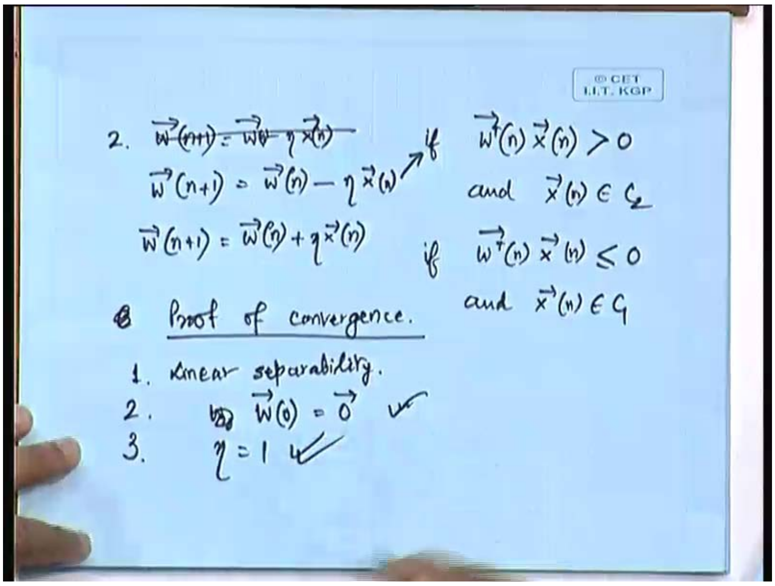
But, if I get a correct classification, then what is it, that I am going to do, supposing I get a correct classification, I pickup x from the H 1 set and I get W transpose X to be greater than 0 or I pickup x from H 2 and I get W transpose X less than or equal to 0. What is the updating strategy, that we are going to take, keep the weights, same, keep the weights same, why are you after all going to disturb a correct thing.

If somebody learns something correctly, you are not going to teach him something more on that particular aspect, you can teach him something else, you can teach him a different pattern. But, you cannot say that no, you have got it correct, but still you will have to relearn the thing. So, if that is the case, if you are already classifying it correctly, then do not update the weight.

So, the weight updating algorithm, that one can state could be, whatever I have just now talked of in the language, can now be just expressed in terms of equation as follows. So, the weight updating algorithm, we can say that, let us take the nth member of the set; that means, to say X n. So, we take the X nth vector as the pattern as the input pattern. So, the updating rule says that W n plus 1 is equal to W n. If W transpose X n is greater than 0 and x belongs to X n belongs to class c 1.

That means to say, that it is greater than 0 and X n belongs to class c 1 means that it is a correct or incorrect, correct classification, it is following in line with this. So, we should have W n plus 1 is equal to W n and likewise we should have W n plus 1 equal to W n. If W transpose X n is less than or equal to 0 and X n belonging to class c 2, that is very correct. So, these are the cases, when correct classification is done and very rightly, we do not update the weight, we keep the weights as the same.

(Refer Slide Time: 16:21)



© CET
I.I.T. KGP

2. ~~$\vec{w}(n+1) = \vec{w}(n) - \eta \vec{x}(n)$~~

$\vec{w}(n+1) = \vec{w}(n) - \eta \vec{x}(n)$ if $\vec{w}^T(n)\vec{x}(n) > 0$ and $\vec{x}(n) \in C_2$

$\vec{w}(n+1) = \vec{w}(n) + \eta \vec{x}(n)$ if $\vec{w}^T(n)\vec{x}(n) \leq 0$ and $\vec{x}(n) \in C_1$

Proof of convergence.

1. Linear separability.
2. $\vec{W}(0) = \vec{0}$ ✓
3. $\eta = 1$ ✓

Now, let us take the different case, when the classification is wrong, so as the second updating strategy, we should say that, when the classification is wrong, then what we do. That means to say, that if W transpose n, X n is greater than 0 and X n belongs to c 2, X n belongs to c 1 would have made it a correct classification. But, I have purposely made it an incorrect classification.

So, I have drawn X n from c 2, but I have got W transpose X n greater than 0, so it is a misclassification, it is a wrong classification. So, wrong classification means we must necessarily update the weight. So, you tell me that, what I should write for W n plus 1

Student: ((Refer Time: 17:08))

I think by now, we are knowing the weight update equation, so I think the students here, the viewers should attempt themselves, should be W n minus eta, yes, that is correct. So, it should be eta X n correct, so it is the learning rate times, eta times X n. So, here we have to write W n, I think if the font does not appear correctly, W n plus 1 is equal to W n minus eta X n, if that is the case.

And if we have W transpose n, X n to be less than or equal to 0 and X n belonging to c 1, that is also a misclassification, in this case what should be my weight updating equation, W n plus 1 is equal to

Student: ((Refer Time: 18:25))

W n, correct, why is it plus here and why is it minus here

Student: ((Refer Time: 18:31))

Yes, yes because we are moving in the direction opposite to the gradient. So, if we have it is very clearly shown by using this concept. If you find that the pattern belongs to c 1, but you have got a c 2, why you have got c 2, may be because your line or hyper plane actually is lying somewhere here. So, naturally you have to push it out; that means to say, that you have to in this case, increase it and in this case, decrease it.

So, that is in line with the direction opposite to the gradient, that we get minus here and we get plus here. So, now with this updating strategy; that means to say, again I repeat that updating strategy here means that for correct classification, no updating, for wrong classification, just update according to the learning rule, this is the standard learning rule. Now, based on this updating policy, we are now going in for the convergence theorem or the proof of convergence.

So, we are now going in for a proof of convergence, but in order to just present the proof of convergence, there are some initial assumptions that we are making. Do not think that these assumptions are directly going to affect the result. One assumption will affect the result surely, that you know, that it is linearly separable class. Naturally, if that is violated, we cannot use the Perceptron for pattern classification.

So, the convergence must necessarily accept the linear separability as the essential criteria. But, there are some extra things which are assumed for just for a simplified mathematical analysis, but we can see, that if they are violated or rather I would suggest that you can examine yourself. That if the extra assumptions, that we are making, if we violate that, that also will lead to a convergence, but may be faster convergence, may be slower convergence, whatever.

But, the kind of assumptions that we are making only for a guideline or only for the ease of computations are as follows. Number 1 is linear separability, which I think you all agree that, we are definitely not going to violate. But, other than that, we are going to have two more assumptions; one is that W 0; that means to say the weight vector that we are using initially is a 0 vector.

So, we are starting with all weights to be equal to 0. In fact, it is not very wrong in assuming that, simply because initially you do not have any clue, when are you feeding W 0, before feeding any pattern. So, you do not know that, what sort of pattern, you are going to feed, what the desired output will definitely be, there is no learning at all, it is just an initial assumption.

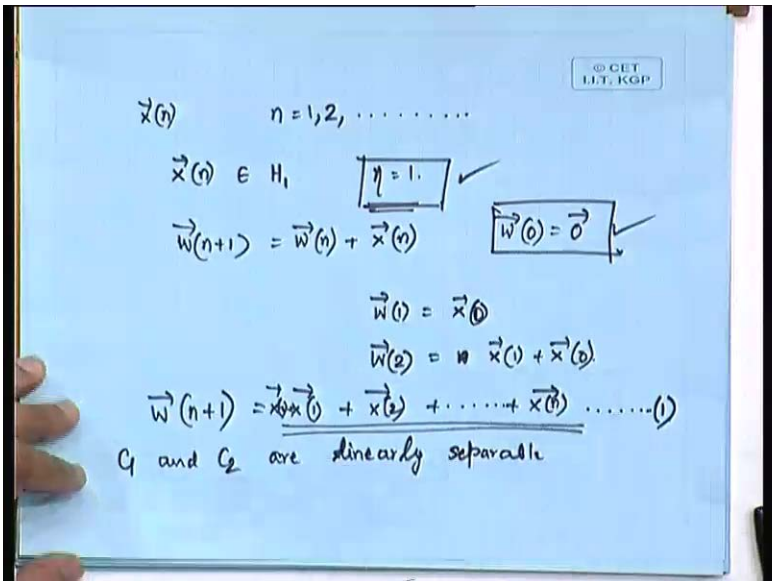
So, initial assumption, because the weights could be positive, because the weights could be negative, so that is why, we are taking it to be 0 vector, nothing wrong. So, if we prove that, for W 0 equal to 0 vectors, it is converging, means that, again if we take any random value of W 0, it should also converge. So, that is 1 and the other assumption that we are making is that, eta is equal to 1; again this is for the ease of our analysis.

And again, eta is equal to 1, I should say, that we are taking more like conservative case, because you know. By now, you have already seen that, when we take a smaller value of eta, the only price that, we are paying for is that the time of speed of convergence is very slow. But, the convergence is possibly more guaranteed for smaller, but in this case, we are taking eta to be equal to 1. So, eta one is sufficiently high value of eta, in fact, normally we take the values of eta in the range of 0 and 1.

So, if we are taking 1 and show that it is convergent for eta is equal to 1, then definitely, there is no reason to believe that, it is not convergent for lower eta values. In fact, I think mathematically also, it can be very easily shown, the approach that we are going to follow, you can yourself show that for smaller etas also this should be remaining valid.

So, these two assumptions with which we are doing going to do the mathematical analysis are not any over simplified assumption or we are not losing the generality in any way by making these two assumptions. So, now let us see that, how do we arrive at the proof.

(Refer Slide Time: 24:17)



© CET
I.I.T. KGP

$\vec{x}(n) \qquad n = 1, 2, \ldots\ldots\ldots$

$\vec{x}(n) \in H_1 \qquad \boxed{\eta = 1.}$ ✓

$\vec{w}(n+1) = \vec{w}(n) + \vec{x}(n) \qquad \boxed{\vec{w}(0) = \vec{0}}$ ✓

$\vec{w}(1) = \vec{x}(0)$

$\vec{w}(2) = \vec{x}(1) + \vec{x}(0)$

$\vec{w}(n+1) = \underline{\underline{\vec{x}(1) + \vec{x}(2) + \cdots\cdots + \vec{x}(n)}} \ldots\ldots (1)$

C_1 and C_2 are linearly separable

Now, let us suppose that we keep on feeding the pattern X n and we start with n equal to 1, 2 etcetera, we keep on feeding the patterns. And initially, let us say that, it is doing misclassifications. So, to start with, we assume that it does misclassification, when I feed X 1, it does misclassification, when I do X 2 and it does misclassification, when I feed x 3. But, misclassification itself is that there is learning, learning is only there when there is error.

So, you keep on doing misclassification, but if it is convergent, then there should be some point, there should be some value of n, where we are going to get a correct classification and am I correct. So, if such value of n exists, where the correct classification ultimately results, then we can definitely say that the system is convergent. So, we are feeding X n with n equal to 1, 2 etcetera.

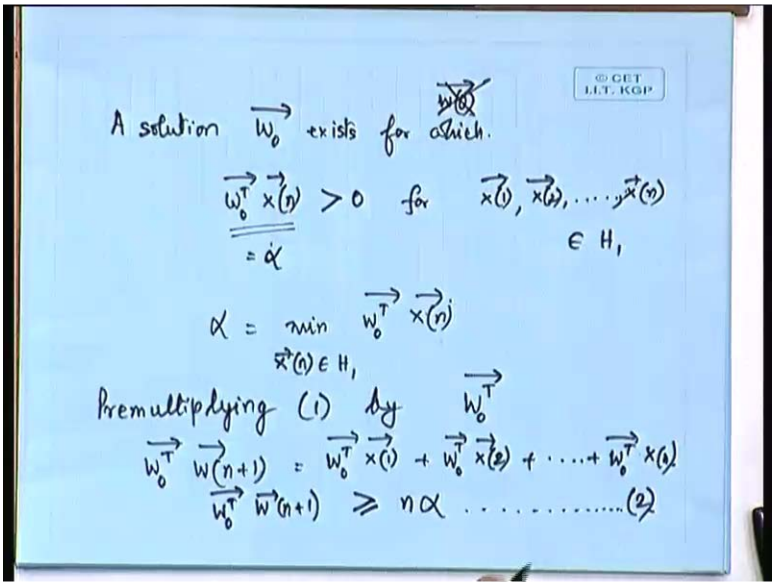
And we assume, that X n belongs to the subset H 1; that means to say that, it should be classified into c 1 and we take eta is equal to unity in this case. So, that means to say, that what will be our weight updating equation. We take eta equal to 1 and X n belonging to H 1 pattern and we have got a misclassification. So, what should be my W n plus 1, plus, yes, so it should be W n plus X n, because eta is equal to 1, so it is simply W n plus X n.

Now, this is a very interesting form and I have assumed W 0 to be equal to 0. So, what is going to be W 1, W 1 is going to be, when n is equal to 0. So, W 1 will be equal to X 1, X 0, that is correct X 0. So, what is going to be W 2, W 2 is going to be W 1 plus X of 1, so that means to say, that W 2 will be equal to X 1 plus X 0. So; that means to say, that W n plus 1 is going to be equal to yes, so we can write it as X 1 plus X 2 plus so on, so on up to X n, starting from X 0, that is what is leading to starting from X 0.

So, this is simply a summation that we are getting for W n plus 1, but again another assumption that we are making. So, we have made use of these assumptions all ready, we have also made use of another assumption that W 0 equal to 0. And the third assumption that we have started with is that c 1 and c 2 are linearly separable. Now, the basic fact that c 1 and c 2 are linearly separable means that although in this n pattern. So, far we are assumed misclassification.

Although, misclassification has occurred in this n, but a solution definitely exists, because it is linearly separable the solution exists, but only thing is that. So, for we have not reached that solution, so the solution exists and because the solution exists, we can find a solution.

(Refer Slide Time: 28:31)



© CET I.I.T. KGP

A solution $\vec{w}_0$ exists for $\overset{\cancel{w(0)}}{\text{which}}$.

$$\underline{\underline{\vec{w}_0^T \vec{x}(n)}} > 0 \quad \text{for} \quad \vec{x}(1), \vec{x}(2), \ldots, \vec{x}(n) \in H_1$$

$$= \alpha$$

$$\alpha = \min_{\vec{x}(n) \in H_1} \vec{w}_0^T \vec{x}(n)$$

Premultiplying (1) by $\vec{w}_0^T$

$$\vec{w}_0^T \vec{w}(n+1) = \vec{w}_0^T \vec{x}(1) + \vec{w}_0^T \vec{x}(2) + \cdots + \vec{w}_0^T \vec{x}(n)$$

$$\vec{w}_0^T \vec{w}(n+1) \geqslant n\alpha \quad \ldots\ldots\ldots\ldots (2)$$

So, a solution W 0 exists, a solution W 0, here mind you, I have put W as a suffix 0, so it is different from W 0, so let us not make this confusion. So, here this is means W at the iteration 0. So, this is not what I am taking, here I am taking W 0 means, it is some solution, a solution W 0 exists, because they are linearly separable. So, a solution W 0 exists for which we have W transpose, W not transpose, yes, X n as greater than 0 for X 1, X 2 etcetera, etcetera up to X n, all this are vectors.

So, X 1 to X n all of them belonging to H 1, so that means to say, that a solution W 0 definitely exists, which does a correct classification, it is it correct. So; that means to say that, now we can define a quantity. So, that means to say that for W 0 transpose X n to be greater than 0; that means to say that, there is some positive quantity alpha that we can find out for these values.

So, there should be some value for which, we can find a minimum value, minimum positive quantity alpha, which all these quantities W 0 transpose X n, n equal to 1, 2 etcetera. Now, all of them should be greater than some positive quantity alpha, some minimum positive quantity alpha. So, I define alpha as a minimum of W 0 transpose X n and the minimum is defined over X n, n equal to 1, 2 up to n here, X n belonging to H 1. So, that is how alpha is defined.

So; that means to say, that if I now go back to this equation W n is equal to this. In fact, X 0 we can take the initial pattern, in fact there is as such no concept of X 0 because, X 0 is not a feed, it is the initial state that you can take. Now, X 1 onwards to X n; that we are obtaining. Now, for all these things, for all these n terms, that we have got, each of these n terms will be greater than alpha., X 1 will be greater than alpha, X 2 will be greater than alpha X n will be greater than alpha.

So, that I can definitely put forward a case like this, supposing not X 1 or X 2, we will have to multiply it by a W transpose, yes. So, here let us say, that this is equation number 1. So, what I am going to do is that, I am going to multiply the equation 1; I am going to pre multiply, so pre multiplying equations 1 by W 0 transpose. So, what I am going to do in this equation, I am going to pre multiply it this, you strike off.

So, W 0 transpose, W n plus 1 is equal to W 0 transpose X 1 plus W 0 transpose X 2, so on up to W 0 transpose X n. Now, individually, what is W 0 transpose X 1s limit, what is its bound, it must be greater than or equal to alpha. What is W 0 transpose X 2; that is also greater than or equal to alpha. So, what is the bound on W 0 transpose W n plus 1 greater than or equal to n alpha.

So, we can say that multiplying both the sides of equation by W 0 transpose, W 0 transpose W n plus 1 is equal to W 0 transpose, I am just expanding the step, in case you have not got the thing, that I was talking about just now. So, our summation series would look like this W 0 transpose X n and because individually all this terms are greater than or equal to alpha. So, that is why W 0 transposes W n plus 1 is greater than or equal to n alpha.

And this let us understand is a very important bounding expression that we have got, so I will call it as equation number 2, which definitely gives us some limit about this W transpose and let us remember this result that we have got from this analysis. So, what is yes please, any, any?

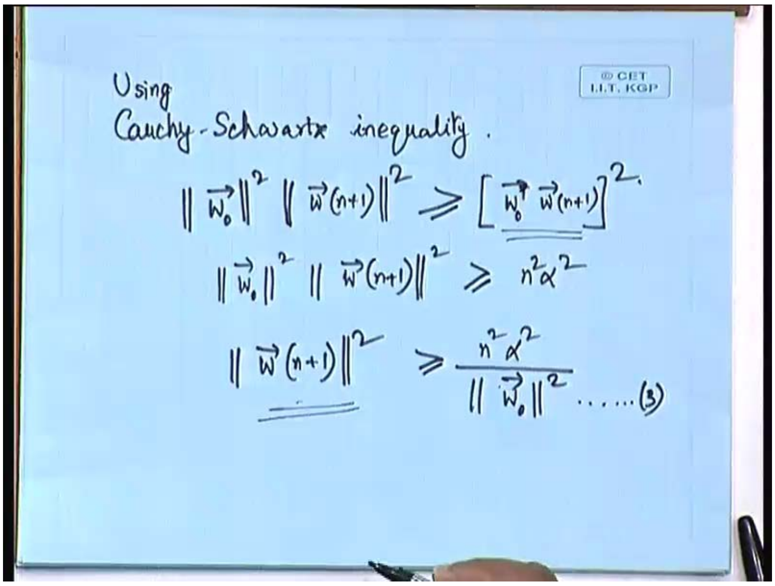
Student: ((Refer Time: 34:26))

0 to n, yes, so I am not considering the case X 0, because X 0 is a trivial, my patterns are n patterns, I am starting with X 1. So, this is W transpose W n plus 1, whose bound I have got to be greater than or equal to n alpha. So, this is the approach that I have got from where, we have basically assumed a misclassification occurring for n number of times and shown. If that is the case, then W 0 transpose W n plus 1 must be greater than or equal n alpha.

Now, there is another route by which we can try to attempt the bound, before we go in for that, let me also discuss more on this. So, this is a very interesting bound that we have got, now on this equation 2, if we apply the Cauchy Schwarz inequality. If we apply the Cauchy Schwarz inequality, then you see on the left hand side of it, we are having W 0 transpose W n plus 1, which is itself a dot product, a scalar quantity.

And going by the Cauchy Schwarz inequality, we should say, that if we individually take the norms of these two vectors. If we take the squared norms of W 0 transpose and W n plus 1 and if those squared norms are multiplied together. Then, that should be greater than or equal to the W 0 transpose W, this dot product that you are getting squared norm of that, square of that rather.

(Refer Slide Time: 36:26)



Using
Cauchy-Schwartz inequality.

$$\|\vec{W}_0\|^2 \, \|\vec{W}(n+1)\|^2 \geq \left[\underline{\underline{\vec{W}_0^T}} \, \vec{W}(n+1)\right]^2.$$

$$\|\vec{W}_0\|^2 \, \|\vec{W}(n+1)\|^2 \geq n^2\alpha^2$$

$$\underline{\underline{\|\vec{W}(n+1)\|^2}} \geq \frac{n^2\alpha^2}{\|\vec{W}_0\|^2} \ldots\ldots(3)$$

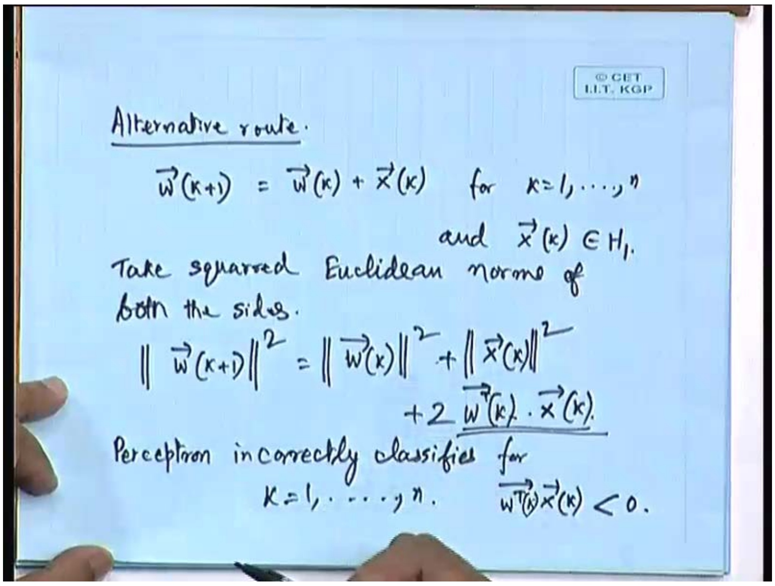
So, in other words, by using Cauchy Schwarz inequality, we can write norm of W 0 square times, norm of W n plus 1 square is greater than or equal to this quantity, this being as a scalar quantity, W 0 transpose W n plus 1. So, it is W 0 transpose W n plus 1, this square, so this is Cauchy Schwarz inequality and we know the bound on W 0 transpose W n plus 1, what is that, that is greater than or equal to n alpha.

This term is greater than or equal to n alpha, means the square term is greater than or equal to n square alpha square. So, we can definitely write W 0 square W n plus 1 norm square is greater than or equal to n square alpha square. So, that now we can get W n plus 1 norm square to be greater than or equal to n square alpha square by norm of W 0 square, pretty simple.

So, this is what we have got as a bound for W n plus 1, because ultimately, that is what we are looking for, that what is going to be the bound on this W n plus 1. And just see, as if it is going to suggest, something very alarming; that means to say that, we keep on increasing n, supposing we keep on increasing n; that means to say that the weights will keep on increasing in magnitude and exponentially. It is basically n square and it must be greater than or equal to this, so this is something to feel alarmed off.

So, we were thinking of showing convergence, but are we really going in for a convergence like this. But, we have got this expression correct and let us remember this that we have got an equation 3, which we treated as an alarming equation.

(Refer Slide Time: 39:09)



© CET I.I.T. KGP

Alternative route.

$$\vec{w}(k+1) = \vec{w}(k) + \vec{x}(k) \quad \text{for } k=1,\dots,n$$

and $\vec{x}(k) \in H_1$.

Take squared Euclidean norms of both the sides.

$$\|\vec{w}(k+1)\|^2 = \|\vec{w}(k)\|^2 + \|\vec{x}(k)\|^2 + 2\,\underline{\vec{w}^T(k)\cdot\vec{x}(k)}$$

Perceptron incorrectly classifies for

$k=1,\dots,n$. $\vec{w}^T(k)\vec{x}(k) < 0$.

And now, let us take a second development route, an alternative route that we explore. So, the alternative route, we take as follows, again we consider a series of misclassifications to start with. So, we take that W k plus 1 is equal to W k plus x of k for k equal to 1 to n and X k belonging to what, you are not attentive belonging to H 1. This is basically the misclassification update, same thing that we had written last time also; last time also we began with this.

So, there is nothing new that I have written, only thing is that I just put the index to be k, k is equal to 1 to n and X k belonging to H 1, but there is a misclassification, because there is a classification, that is why, I could write that. Now, in the earlier case, what I did, earlier case I iteratively took, k is equal to 1, 2 etcetera, etcetera, I kept on taking. In fact, one of the things that we did little wrong is, we took k is equal to, W 0 we started with and but ultimately the patterns are k is equal to 1 to n, that we are feeding, anyway.

So, here instead of going in that iterative way, that we first compute W 1, then compute W 2, then compute W 3 and then expressing the last W n plus 1, which we have already done and seen. Instead of doing that, let us take this equation, on it is own and let us obtain the square Euclidean norm of both the sides. So, take squared Euclidean norm of both the sides, easy.

What, we are going to do is the left hand side, you are going to multiply by W transpose k plus 1, the right hand side you are going to pre multiply by W k plus x k whole thing transpose or rather to say W k transpose. W transpose k plus X transpose k; you are going to pre multiply with that, so you know, what you are going to get. So, we can directly expressed it, I thing that we did not have to expand it, if you have doubts please do not hesitate to ask me.

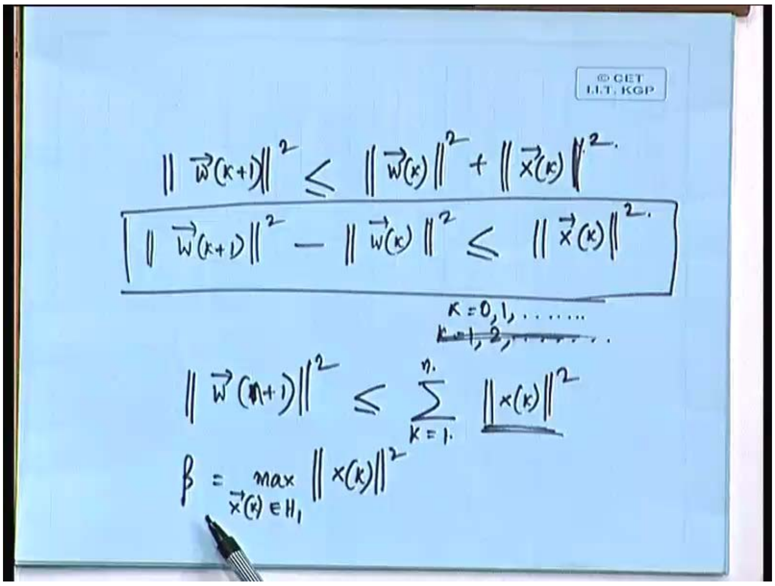
W k plus 1 is going to be equal to W k square plus, yes please, please go on, X k norm square plus twice correct twice W k transpose dot X k. Now, we are assuming that the Perceptron incorrectly classifies, so our basic assumption is that Perceptron incorrectly classifies for k equal to 1 to n. So, because it incorrectly classifies, what can we say about this W k transpose X k less than 0, yes, less than 0. So, w transpose k, X k is less than 0.

So, if w transpose k, X k is less than 0, in that case what can we say about the relationship between the left hand side and the first two terms of the right hand side.

Student: ((Refer Time: 43:13))

Smaller than that definitely, because W k X k is less than 0, that is why, it is subtracting from what I have got as W k X k. So, definitely W k plus 1 should be less than or equal to W k plus 1 norm square, should be less than or equal to this square plus this square.

(Refer Slide Time: 43:38min)



© CET
I.I.T. KGP

$$\|\vec{w}(k+1)\|^2 \leq \|\vec{w}(k)\|^2 + \|\vec{x}(k)\|^2.$$

$$\boxed{\|\vec{w}(k+1)\|^2 - \|\vec{w}(k)\|^2 \leq \|\vec{x}(k)\|^2.}$$

$k = 0, 1, \ldots\ldots$

~~$k = 1, 2, \ldots\ldots$~~

$$\|\vec{w}(n+1)\|^2 \leq \sum_{k=1}^{n} \underline{\|x(k)\|^2}$$

$$\beta = \max_{\vec{x}(k) \in H_1} \|x(k)\|^2$$

So, we can definitely write, because misclassification has happened, we can definitely write that W k plus 1 square is less than or equal to W k square plus X k square. Or equivalently, we can say that W k plus 1 square minus W k square is less than or equal to X k square, now this very interesting expression. You see now, we are going to apply this equation, we are going to just apply this in equation rather for k is equal to 1, 2 etcetera.

You put k is equal to 1, what you get W 2 square minus W 1 square less than or equal to X 1 square. You take k is equal to 2, what do we get W 3 square minus W 2 square less than or equal to X 2 square, add them together. So, you have X 2 square here, minus X 2 square there, all cancelling out, it is only the last term minus the first term, that you are going to get.

And if you are taking initial weight assumption W 0 is equal to 0, so definitely that finally, leads to the fact, that if we go on iterating this. Then, W k plus 1 square, that should be less than or equal to summation of X k square, where we sum up for k is equal to 1 to n, W 0 square is in this case 0, because

Student: ((Refer Time: 45:50))

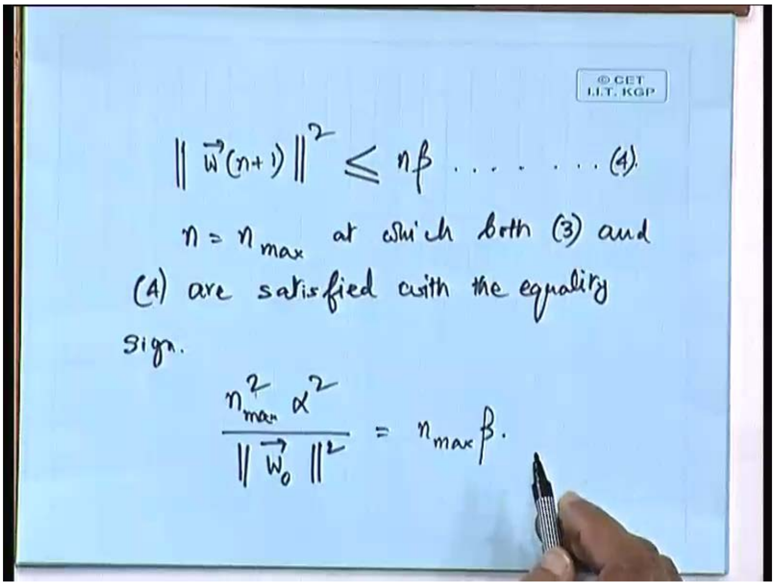
k is equal to yes, always we land up in a this index 0 is always a problem, if we go on for k is equal to 0, 1 etcetera, but the pattern is not feed, for k is equal to 0. So, that is why, for the pattern, we sum up for k is equal to 1 to n, but take that as 0. So, here we are getting W 0 square to be equal to 0. So, here we can write that W k plus 1 square is equal to summation X k square.

And this is something that we can tell that here W k plus 1 is definitely less than or equal to here, less than or equal to this quantity, so which means that if we define, yes any doubts, W n plus 1. So, w is n plus 1 square, we have already got one expression of W n plus 1, mind you, we remember this. And now we are going to get something new over here.

And one new thing that we are getting is, earlier it was greater than or equal to something, whereas here it is less than or equal to something and something how conveniently, we can express. So, let us define a positive quantity, norm of x square is definitely going to be positive. So, let us define beta as a positive quantity, which is going to be the maximum of this X k square. And where do we define X k, this X k all belong to the space H 1.

So, X k is picked up from the space H 1 and we are taking maximum value of that as beta. If we take the maximum value of this to be equal to beta, then naturally X 1 square is going to be less than or equal to beta, X 2 square less than or equal to beta, like every X k square is going to be less than or equal to beta. So, that in effect, what is the bound that we are getting on W n plus 1 square less than or equal to n beta.

(Refer Slide Time: 48:33min)



$$\|\vec{w}(n+1)\|^2 \leq n\beta \quad \ldots\ldots\ldots (4).$$

$n = n_{max}$ at which both (3) and (4) are satisfied with the equality sign.

$$\frac{n_{max}^2 \alpha^2}{\|\vec{w}_0\|^2} = n_{max}\beta.$$

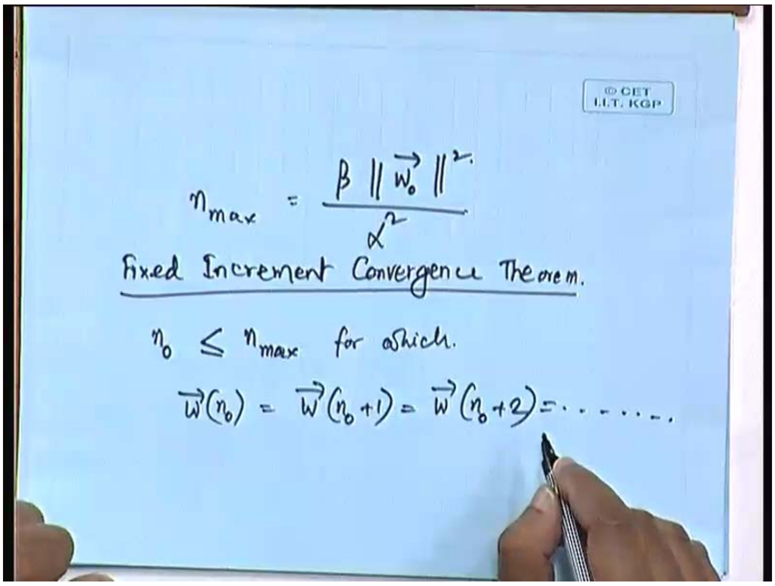
So, what we are getting is that W n plus 1 square is less than or equal to n beta and let us call this as equation number 4. So, we have got equation number 3 and we have got equation number 4. The interesting observation is that, equation number 4 puts a bound, which says very interesting thing that W n plus 1, as you increase n, the W n plus 1 square is linearly decreasing. You, not that, it is linearly increasing with, it is bound is, it is it is going to be less than or equal to n beta.

So, it is specifying a lower limit according to equation 4 and according to equation 3, it is specifying an no, it is sorry, other way, it is specifying a lower limit from equation 3 and it is specifying and upper limit from equation 4. And is it going to be to these equation number 3 and equation number 4. This two that, I have shown are they going to be consistent for very large value of n, let us say.

It will be impossible to satisfy both equation 3 and 4 simultaneously at very large value of n. Because, at very large value of n this lower bound is going to be very high, whereas the upper bound is going to be low. So, that is why, we must have some point at which is the limiting case that both 3 and 4 gets satisfied. And let us say that there is some n, which we can say that n equal to n max, n equal to n max at which both 3 and 4 are satisfied with equality sign.

So, that means to say that n max is the solution of what, let us again have a look at this, so we must put n max in place of n over here. So, here this leads to n max square, alpha square by W 0 square and that should be equal to n max times beta. So, we must have n max square, alpha square by W naught square, equal to n max times beta. If we satisfy both 3 and 4 at the point n equal to n max with the equality sign.

(Refer Slide Time: 53:11)



$$n_{max} = \frac{\beta \, \|\vec{W}_0\|^2}{\alpha^2}$$

Fixed Increment Convergence Theorem.

$n_0 \leq n_{max}$ for which.

$$\vec{W}(n_0) = \vec{W}(n_0+1) = \vec{W}(n_0+2) = \cdots\cdots$$

So, from here, we can obtain the limit on the n max, so we obtain what, so we obtain here that n max is going to be equal to beta W 0 square by alpha square, this is the limiting value of n. So; that means to say that, after these n is reached or this is the upper limit on this n, beyond which, there will not be any further updating of the weight. Meaning that by the point n equal to n max, we have achieved a correct classification.

We began with incorrect classification and at n equal to n max or may be earlier than that also, we are going to have a situation, where the incorrect classification will be replaced by a correct classification and that is the convergence that we are aiming for. So; that means to say that the Perceptron, so this is called as the fixed increment convergence theorem. Beta, no, here see we are satisfying 3 and 4 simultaneously, so 3 and 4 can be only satisfied up to a limiting value of n and that is what we are taking in to be n max.

Now, we just state this fixed increment convergence theorem, now what it says is that, we have some n 0, which is less than or equal to n max, n max is an upper bound that we have got. But, we have got n 0 less than or equal to n max for which we have W n 0 equal to W n 0 plus 1 equal to W n 0 plus 2 so on. So, this is ultimately the solution vector.

So, here all these are equal means, it is a convergence, it has reached a convergence, so that is what a for today, because we are running short of time, the convergence theorem aspect more or less, we have covered just 1 or 2 small items remaining to that. And then we are going to take up another very interesting case.

Where, we will be taking up one of the classical classifiers, which is the Bayes classifier and we are going to show an interesting relationship between the Bayes classifier and the Perceptron that will be in the coming class.

Thank you very much.