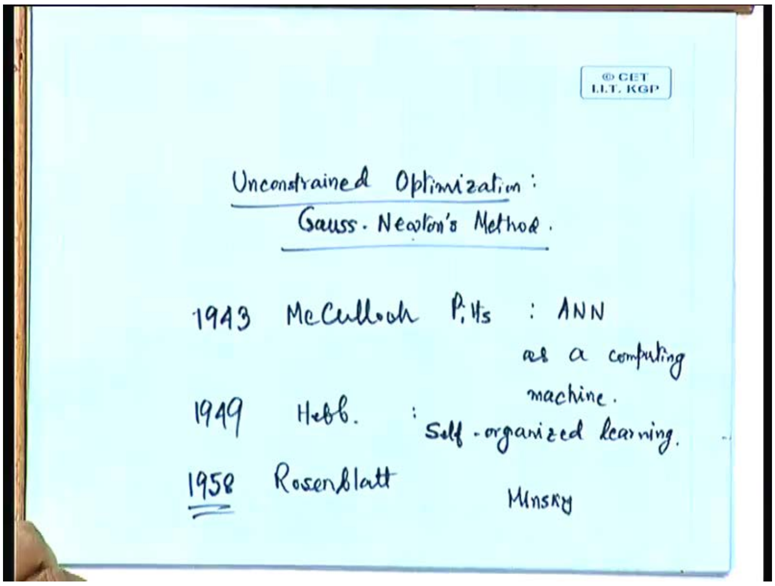
Neural Network and Applications

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

Lecture - 13
Unconstrained Optimization: Gauss: - Newton's Method

Single layer perceptrons, in fact the next few lectures will be related to single layer perceptrons only.

(Refer Slide Time: 01:03)



Unconstrained Optimization :
Gauss. Newton's Method.

1943 McCulloch Pitts : ANN as a computing machine.

1949 Hebb. : Self-organized learning.

1958 Rosenblatt Minsky

So, in particular under the broad heading of single layer perceptrons, we are going to discuss about the third methodology for unconstrained optimization. Because, two of the approaches we all ready covered, which is the steepest descent approach and also Newton's method, these two we have done. And today we are going to consider the third one which is based on the Gauss Newton method.

But, before we actually begin today's topic in general, there are few more things which I would like to do. See, now so far you have gone through I think 11 or 12 lectures in this whole series. And what I have been trying to tell you is different aspects of neural networks, trying to go through the learning mechanism and then, the single layer perceptrons these things we are telling.

But, and also to the distant viewers, I think all these information is getting available. But, what we cannot present to the distant viewers is the kind of discussions which I quiet often have at the end of the class. Because, the students are asking questions, but a lot of times at the end of the class, there are some interesting sessions that I have with the students. And I thought that sometimes sharing such things with the distant viewers might be of help.

Now, one of the things which I would like to point out of course, in a lighter note is that, which was pointed out to me last time. That we were discussing sometimes back about the times in which all the different theories of neural networks, where discovered and we were telling about the time of 1990's when we talked about the theories of Vapnik. So, in fact, going by the historical chronology of neural network, I think we should first begin with 1943 when McCulloch and Pitts.

I think we have all ready mentioned their names McCulloch and Pitts. They first presented the theory of artificial neural network as a computing machine. So, the basic fact that ANN can act as a computing machine that was shown by McCulloch and Pitts, there model presented that. And then, we had in the year 1949, the theory by Hebb, which was the first rule on the self organized learning, unsupervised sort of thing.

Because, that Hebb’s learning theory which we have already seen, that is a self organized learning, which is entirely dependent upon the pre-synaptic and postsynaptic activations. So, that is the second thing and then, also most importantly the one which we are discussing right now, is a theory which was proposed in 1958 by Rosenblatt. In fact, he is the inventor of perceptron, what can surely is give the invention of perceptron to the credit of Rosenblatt.

But, that was in the year 1958, in fact following Rosenblatt’s theory in 1958, in the 60's we had seen lot of turbulence related to the development of perceptron. And I use the word turbulence intentionally, because in 1961 it was another scientist Minsky. He pointed out some of the very fundamental flaws of the perceptron model, which I have all ready told you is that, it can the basic perceptron model can only classify linearly separable problems.

And the example which I had given to you in the last class, that is to say for a simple two input logic gates using the and or these things are very easily linearly separable. But, a simple problem like exclusive or which inherently is not linearly separable, that could not be solved using a perceptron model which I had shown. Now, Minsky, in fact was the first to point out this that the nonlinearly separable classes.

They cannot be classified using the perceptron and that had given a big jolt to the theory of Rosenblatt. And in fact, the growth of neural network, rather the growth of perceptron as a computational unit suffered quite a lot, till the time the multi layered perceptron theory was represented. In fact, we will be discussing little more about it, when we come to the multi layer perceptron discussion, but so much for it now.

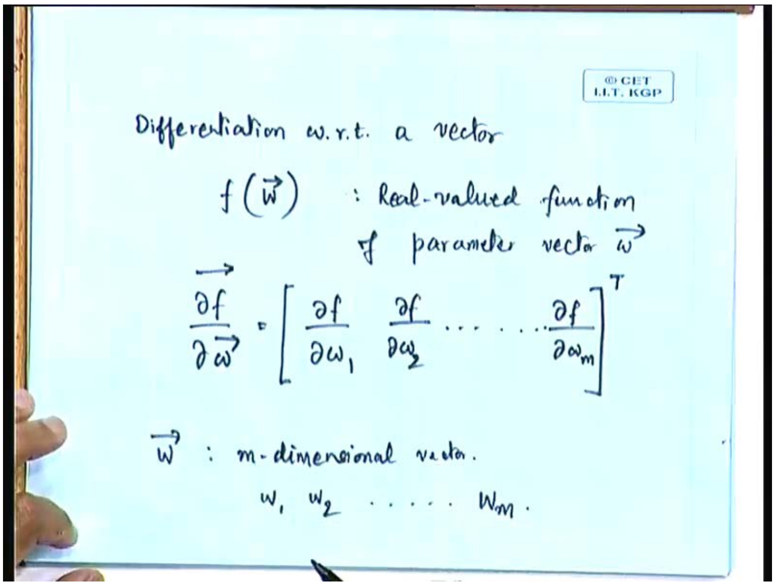
So, the theory is that we are talking of is basically the earlier 40's, late 40's, late 50's that type. Now, very interestingly last time we were discussing about Newton's method. So, the question that came to the peoples mind was that, how is the great scientist Sir Isaac Newton getting associated with neural network, something very interesting. In fact, Sir Isaac Newton has in no way contributed to the neural network as such, but his theories are there everywhere.

His theories are there in the sense that, when we try to use the mathematical model, there we try to use the Newton's theory based on which we had already presented the analysis. So, although Newton was born in the 18th century and all this theories are 20th century, late 20th century theories. Sir Isaac Newton in fact, does not have anything to contribute to neural network anyway.

So, can begin with the Gauss-Newton method, but again before that I thought that we should spend little bit of time on the differentiation of a function with respect to a vector. Because, that is what we were discussing in the last class and I do not know if some students over here or if some of the distant viewers, had any problem in getting in to the theory of the differentiation. Because, we were directly using some of the results which I thought that maybe we should spend little bit of time, in order to explain such theories.

So, we will be first talking about some of the theories related to the differentiation with respect to a vector.

(Refer Slide Time: 08:32)



Differentiation w.r.t. a vector

$f(\vec{w})$: Real-valued function of parameter vector $\vec{w}$

$$\frac{\overrightarrow{\partial f}}{\partial \vec{\omega}} = \left[\frac{\partial f}{\partial \omega_1} \ \frac{\partial f}{\partial \omega_2} \cdots\cdots \frac{\partial f}{\partial \omega_m} \right]^T$$

$\vec{w}$: m-dimensional vector.

$w_1 \ w_2 \ \ldots\ldots \ W_m$.

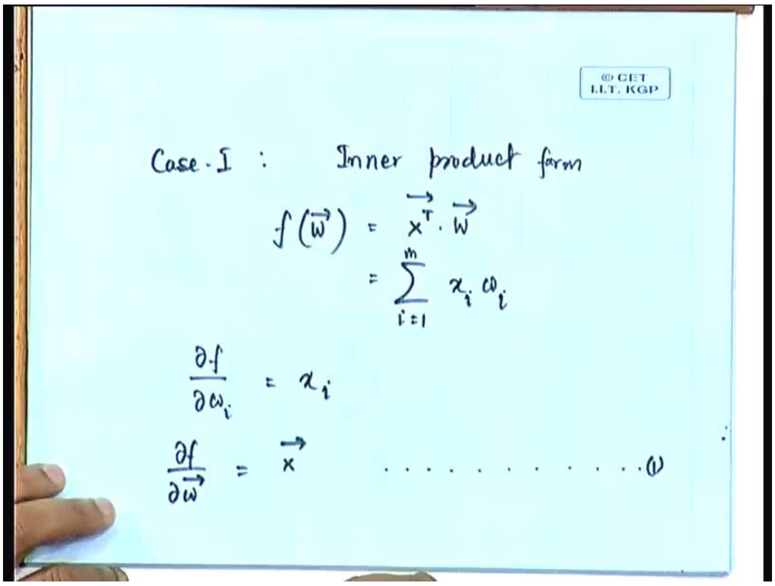
And let us say that we take a real valued function, let us say f of w, but again I take w vector, because there are many free parameters in the system. So, let us take any function real valued, so this is a real valued function of a parameter vector w. In fact, let us talk in a very general terms, we know that in our case w vector means the weight vector. But, any parameter space that is why the function of any parameter space, which we are calling as f of w.

Now, when we define a derivative of this function with respect to a vector. Let us say that this function f we would like to differentiate with respect to a vector, so dou f dou w. Then exactly what we mean by that, is that this w vector that is after all an m dimensional vector. So, w we assume to be an m dimensional vector and whose elements are nothing but, w 1, w 2 etcetera up to w m.

So, by dou f dou w vector, we essentially define another vector which will be of this form, dou f dou w 1, the first element that is the function differentiated with respect to w 1. And then, the second element will be dou f dou w 2 and like this we will be having m number of elements, but the last elements or n th element will be dou f dou w m. And in fact, this being a vector this whole thing acts as a vector.

So, there is no harm in giving an overall vector notation to this, this should be a transpose of what I have all ready written. Now, the thing is that this is the general form of the derivative with respect to a vector. But now, let us consider one or two very special cases of this function, which we will be require, results of which we will be requiring and in fact, we have been all ready using some of this results. So, let us take a few typical cases of this function, in fact two typical cases we will take.

(Refer Slide Time: 11:34)



© CET I.I.T. KGP

Case - I : Inner product form

$$f(\vec{w}) = \vec{x}^T \cdot \vec{w}$$

$$= \sum_{i=1}^{m} x_i \omega_i$$

$$\frac{\partial f}{\partial \omega_i} = x_i$$

$$\frac{\partial f}{\partial \vec{w}} = \vec{x} \quad \ldots\ldots\ldots\ldots (1)$$

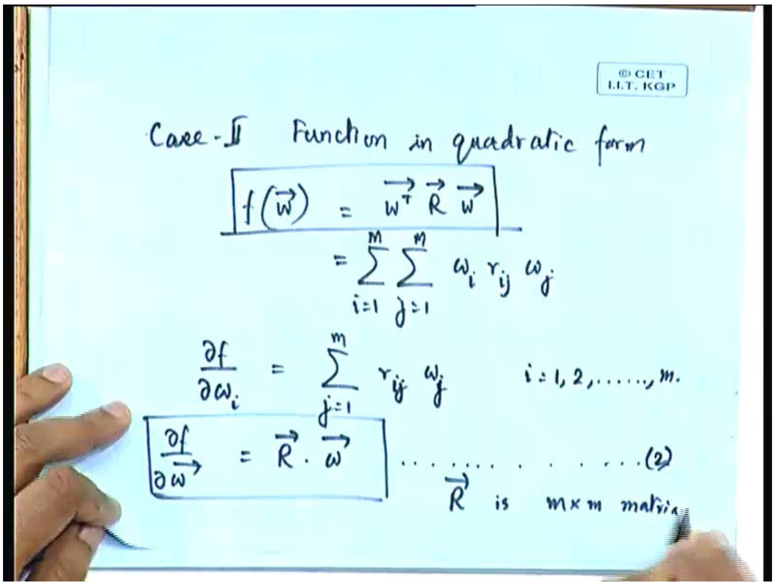
The first case that I am going to take is that, if the function is expressible in the inner product form. In fact, we require such functions very often, because when we are considering the neural network. Then the function that we are realizing is very much expressible as a as an inner product. So, this in fact, becomes x transpose w, so this x transpose w, if we write it this, this way that is in the inner product form algebraically, so this is the matrix form of it matrix vector form of it.

And the algebraic form of this is simply the summation of x i w i, where i is equal to, where i is summed up from 1 to m. Now, if we take the function like this, and if we differentiate this function f expressible in this way with respect to w i, what is it that we are going to get simply x i. So, now if I ask that what is going to be this vector, that is dou f dou w vector, it is going to be x vector. Because, you see when I differentiate with respect to w 1.

Let us say that I different this whole expression with respect to w 1. Then f of w here is x 1 w 1 plus x 2 w 2 plus so on, so on up to x m w m. Now, when I am differentiating this with respect to w 1, then the only non-zero term that remains is x 1. So, dou f dou w 1 is x 1, likewise dou f dou w 2 is x 2, dou f dou w m is x m, so that is why if I take dou f dou w vector that itself will mean, that it is x vector.

So, this is a very important relation that we need to remember. That is to say when the function is expressible in x transpose w, in that case the derivative of that function becomes simply the x vector. So, we will be directly making use of this, when we come to x transpose w, we will be directly making use of this dou f dou w definition. So, this is the simplest of all the forms of the function, that is the inner product form, but we often in fact, we have been expressing the functions also in a quadratic form. So, when we express the function in a quadratic form, let us put a case 2 under it.

(Refer Slide Time: 14:33)



Case-II Function in quadratic form

$$\boxed{f(\vec{w}) = \vec{w}^T \vec{R}\, \vec{w}}$$

$$= \sum_{i=1}^{M} \sum_{j=1}^{M} \omega_i r_{ij} \omega_j$$

$$\frac{\partial f}{\partial \omega_i} = \sum_{j=1}^{m} r_{ij} \omega_j \qquad i = 1, 2, \ldots\ldots, m.$$

$$\boxed{\frac{\partial f}{\partial \vec{\omega}} = \vec{R} \cdot \vec{\omega}} \quad \ldots\ldots\ldots\ldots\ldots (2)$$

$\vec{R}$ is m×m matrix

So, case 2 we are going to represent the function in quadratic form. And for that, what is the representation that we should make f of w should be w transpose...

Student: ((Refer Time: 14:57))

X in to w, but because here we have all ready used x, there is no harm in using a different notation for that. So, let us say that we use a matrix, let us say R matrix we use. So, we use let us say w transpose vector R matrix and w vector and in this case what we take is that this w vector again is an m by m vector. So, very rightly if you take it this way, then in the summation formula this whole thing can be expressed as summation w i r i j w j where we sum it up for i is equal to 1 to m and for j is equal to 1 to m.

Just see that this R matrix is post-multiplied by w pre-multiplied by w transpose, so this is what you are going to get. So, if you just investigate with these expressions you can see that, when we have i is equal to j, that means to say that all the terms where i and j become the same. This in fact, is going to have how many such summations there will be effectively m square number of summations will be there.

So, here you can see that for i is equal to j, there will be all the square terms w i square r i, r i i those type of terms will be there. So, that means to say taking the diagonal elements of the R matrix whereas, all the other terms will be of this nature w i r i j w j. So, if we take the function in this form and then, if we take its derivative let us say that we take a derivative with respective to w i. So, dou f dou w i, now i take can you tell me that what is going to be the expression for that.

So, I am just differentiating with respect to one of these w's, I take w i, I could take w 1, w 2 anything, so I take w i in general. So, what is going to be dou f dou w i it is going to be here.

Student: ((Refer Time: 17:41))

Summation r i j w i that is very good, so here it will be summation for i is equal to 1 to m r i j w j.

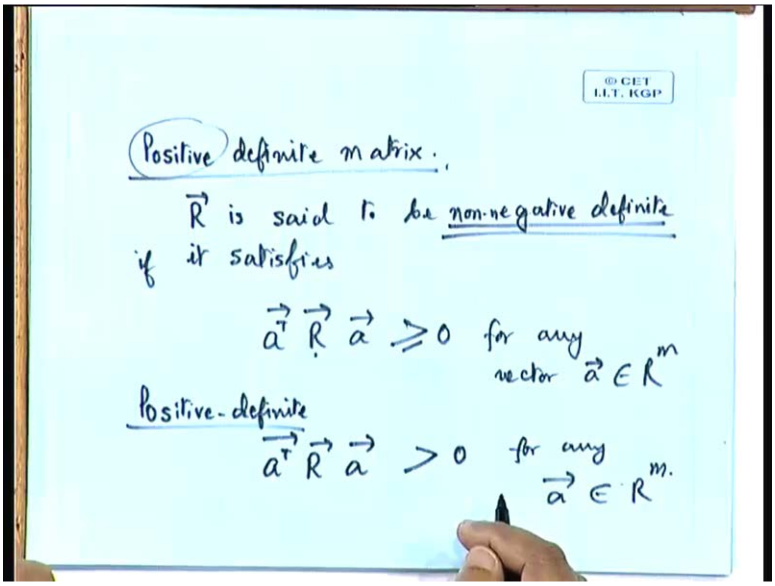
Student: ((Refer Time: 17:56))

J, it should be j is equal to 1 to m, so I should take r i j w j and in this case my i can be from 1 to m. So, like this I will be getting m such expressions this is dou f dou w 1 will be this, dou f dou w 2 will be this. So, effectively again when I write this expression in the matrix form, then what do I get, so when I take dou f dou w vector in that case what is it going to be.

This is simply going to be R matrix times w vector, R times w, so R dot w is going to be the dou f dou w, this is the representation in matrix form. So, we have got two cases, the first case was the inner product form, whereby we got the derivative as simply x, when it is of x transpose w we simply got the derivative as x. And when it is expressible in the quadratic form like this, then if you are differentiating this function, beginning with this, if you are differentiating then you get this, this is very interesting.

So, do not feel confused that if I get a function like this and immediately in the next step, I simply write the derivative of that as R dot w, so it is understood simply.

(Refer Slide Time: 19:51)



Positive definite matrix.

$\vec{R}$ is said to be non-negative definite if it satisfies

$$\vec{a}^T \vec{R} \vec{a} \geq 0 \text{ for any vector } \vec{a} \in R^m$$

Positive-definite

$$\vec{a}^T \vec{R} \vec{a} > 0 \text{ for any } \vec{a} \in R^m.$$

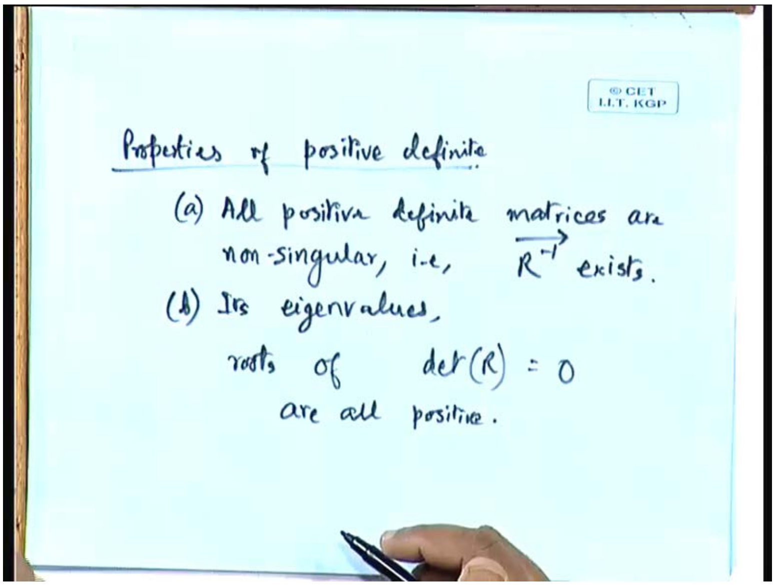
Now, coming to one of the fundamental definitions of matrix, which we will be needing in our discussion very much, that is about the positive definite matrix. So, this definition we will be requiring. Now, after all in this case what is dimensionality of this R matrix, this R matrix is going to be an m by m matrix, so R is m by m matrix. Now, this m by m matrix R is said to be non-negative definite.

So, before we go in to the definition of positive, we first define what is non negative, this is very rightly understood that. If we understand what is a non negative, then we can always define what is a positive by simply, non-negative means it is greater than or equal to 0. So, if you eliminate the zero case, if you do not consider the zero case it is the positive case. So, let us first take the non negative case, so R is said to be a non-negative definite if it satisfies this condition.

And that condition I am going to write, that is you take R and you pre multiply it by a transpose and post multiply it by a. And if this is greater than or equal to 0, for any vector a that you may find in the m dimensional space. So, when I write R m like this, it means to say in m dimensional space. So, you take any vector a in m dimensional space, and if this relation is always satisfied. Then we will be telling that this matrix R of dimension m by m, that is a non-negative definite matrix, followed.

So, what is positive definite then, positive definite will simply mean that, when this inequality sign is the only thing that remains. That means, to say it will not be greater than or equal to 0, it will be greater than 0. So, positive definite will mean, if it is a transpose R a greater than 0 again the condition is that for any vector a in the m dimensional space. So, that is the definition of positive definite. And now positive definite matrices, they fulfill two interesting properties and what are they.

(Refer Slide Time: 23:01)



Properties of positive definite

(a) All positive definite matrices are non-singular, i.e, $\overrightarrow{R^{-1}}$ exists.

(b) Its eigenvalues, roots of $\det(R) = 0$ are all positive.

So, properties of positive definite matrix, I think some of you must be knowing, because this is normally covered in any elementary matrix theory. The first property is that, all positive definite matrices are non-singular, that is R inverse exists, so that is the first property. And the second property is that it is eigen values, that is to say the roots of determinant of R equal to 0, they are all positive roots of the characteristic equation.

So, this is the characteristic equation determinant R equal to 0 they are all positive. So, these are the two very important properties of positive definite matrix. In fact, we will be needing this very soon, so is there any question related to the theory that we have.

Student: ((Refer Time: 24:36))

If the vector a is 0, a null vector yes

Student: ((Refer Time: 24:46))

Well, what I mean to say is that, do not take the zero case, any non-zero vector that is included, when I take a any vector a belonging to m dimensional space, any vector other than the zero vector, yes.

Student: ((Refer Time: 25:11))

Differentiation with respect to vector, yes case 2, good

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

W i r i j W j it not the omega, that is the W, yes

Student: ((Refer Time: 25:32))

Yes, I was expecting that somebody will point out that. In fact, that two anomaly is existing when the square terms are there. When I have i is equal to 0, you are very correct in you observation that then the terms are W i square r i j those terms. And if you differentiate that then it becomes 2 times W i, so you where very correct about it. So, in fact in that sense , when i is equal to j those cases, it is not accounted, for all i not equal to j cases have been accounted over here by, yes please somebody points out something.

Student: ((Refer Time: 26:28))

All terms will have to, yes

Student: ((Refer Time: 26:35))

There are I was, in fact coming to that that here two is we are getting directly, for the square terms we are getting the two's directly. For all the diagonal elements of this R matrix, they are all ready contributing to two and we get only one of this squared terms, so there will be totally m such terms coming in. So, where this contribution of two will be there, but the contribution of two will not be there for the terms, which are simply of this r i j W j type.

But, remember that when you look at the matrix r i j W i and W j, then effectively there will be two terms that will be contributing. Because, there will be one coming from this side one coming from this side, so effectively those two when added up, so everywhere there will be a uniformity of two. In fact, we could express this as half of W T R, W also in that case, it would have been very rightly written as R W, but this summation of this will always involve two terms.

So, because this is a double summation term going on within the bracket, so I think it is good that you have raise this doubt, and also happen to solve this doubt also, so any...

Student: ((Refer Time: 28:18))

No, second term is coming, second term is coming why you see that

Student: There is a W i for each

There is a W i for each of this W j's, so you see when you are adding it up, you will get that.

Student: ((Refer Time: 28:42))

You will be getting W i r i j W j and you will be getting W j r i j W i, r j i yes.

Student: ((Refer Time: 28:58))

But, here the symmetric matrix, we will be in fact, encountering whatever R form here, there the reason why I brought in R over here is that, R later on we will be always considering the correlation matrices, which will be inevitably a symmetric matrix.

Student: ((Refer Time: 29:30))

Are they in different rows that is what we

Student: ((Refer Time: 29:52))

No, this function has got no rows and column that is true, but you see you are differentiating with respect to W i. Now, use see one ((Refer Time: 30:14)) this ultimately, you are computing this f of w completely. Now, when I say that this f of w what you have computed, you differentiate with respect to W 1, then that will involve the terms, where W 1 is coming this side. And that will also involve the terms where W 1 is coming this side and that is why the factor of two is coming.

Student: ((Refer Time: 30:39))

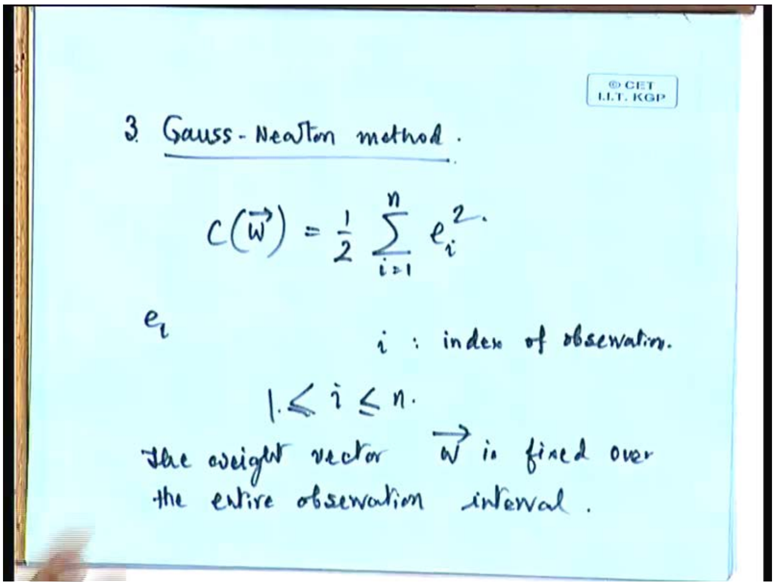
That that is that is accepted, here this should been 1 by 2 in to this, if it is written as 1 by 2 in to this, because any quadratic form. In fact, that is the reason why all quadratic forms we have been writing as half of W, T R W, but here for just for the sake of simplicity we did not write down this 1 by 2. If you were writing it as half you will be getting as R W, if you were writing it as simply W T R W then you will be getting 2 times.

And this factor of half and half here one here or one here, two here that is immaterial.

Student: ((Refer Time: 31:26))

Properties of positive definite that is what you want, so we know go over to Gauss Newton method, that is third approach towards optimization.

(Refer Slide Time: 31:43)



3. Gauss-Newton method.

$$C(\vec{W}) = \frac{1}{2}\sum_{i=1}^{n} e_i^2.$$

e_i

i : index of observation.

$1 \le i \le n.$

The weight vector $\vec{W}$ is fixed over the entire observation interval.

So, the approach three which is the Gauss-Newton method. In fact, the title talk of today is on that only. So, in Gauss-Newton method we are applying this method actually to a cost function, that is expressed as a sum of the squares of the error. Now, I defined a cost function of as a function of W of course, cost function is equal to half of summation i is equal to 1 to n over e i square. In this case what is e i, e i is nothing but, the error, error of the i th observation.

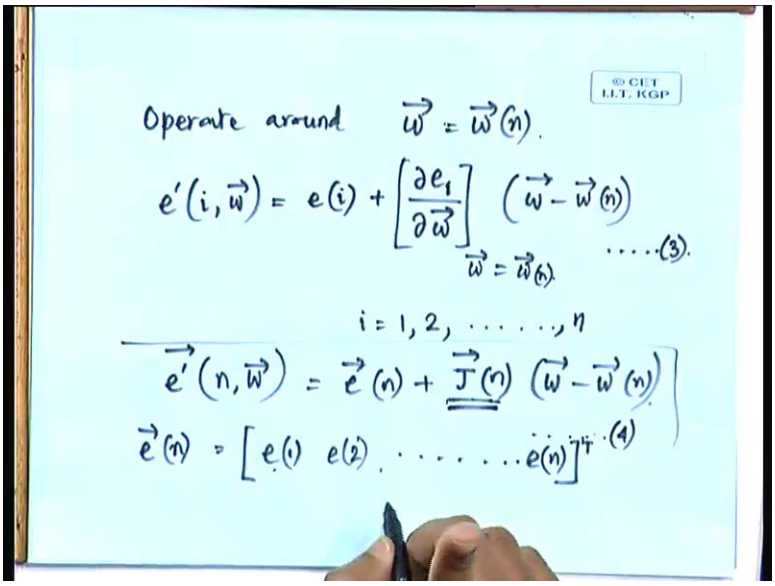
So, we are making an observation n number of times, now please do not neglect this definition please try to understand it very carefully. So, I am making n number of observations, so i is the index of my observation. And I have of course, i something that can vary between 1 to n, so that means, to say from first observation up to n number of observation. And by writing it as summation i is equal to 1 to n, I have summed up the squared error that I have got over n observation.

And one thing which I stipulate here is that, the weight vector W, the weight vector that is the W vector is fixed over the entire observation interval. So, this is fixed over the entire observation interval. Now, I wanted to stipulate this thing, so that people do not get the kind of notion that with every observation we are changing the weights, we are not. We are making n observations and for all the n observations we have kept the weight to be the same.

And then, we are calculating the some of the squared errors over n such observations. And based on the sum of these n squared observation errors, we are going to ultimately find out a new value of W, based on this n observations we are going to find out a new value of W or an updated value of w which we will use after that. So, in order to get a new value of W, what we should do we should find out the new value of W. Such that, this cost function is minimized with the new value of W, the cost function should get minimized.

So, that is why we have to calculate now the gradient of the cost function at this point, what point at the point W is equal to W n, at the end of n observations.

(Refer Slide Time: 34:56)



Operate around $\vec{w} = \vec{w}(n)$.

$$e'(i, \vec{w}) = e(i) + \left[\frac{\partial e_1}{\partial \vec{w}}\right]_{\vec{w} = \vec{w}(n)} (\vec{w} - \vec{w}(n)) \quad \ldots\ldots(3).$$

$i = 1, 2, \ldots\ldots, n$

$$\vec{e}'(n, \vec{w}) = \vec{e}(n) + \underline{\underline{\vec{J}(n)}}\,(\vec{w} - \vec{w}(n))$$

$$\vec{e}(n) = [e(1)\ e(2)\ \ldots\ldots\ldots e(n)]^T \quad \ldots(4)$$

So, when we operate around the point W is equal to W n. Then the error term that we will be having can be expressed as the updated error that is e as a function of i and W. That is very clear, because the error will vary from observation to observation and also the error will vary when you change the weights. So, error is also a factor which is very much dependent upon the W vector.

So, e, i, w, e prime is the updated error that will be nothing but, e i that is the say the error with the n observation. The error e i, the earlier error to say plus what happens to is that at the point W equal to W n, you are going to find out the derivative of this error with respect to W vector. So, derivative of e i with respect to W vector that you are going to compute at the point W is equal to W n.

And then, this derivative you are going to multiply with the incremental weight change. That means, to say that W, then the new W that you are going to get minus the W that you had at the n case, W minus W n. This is something that you will be calculating for i is equal to 1 to with observation you are going to calculate this, that is the error will be this. So, that at the end of n iteration, so when we are having n number of iterations, when we are having n of observations not iterations.

So, in the matrix notation, we can write that e prime n W this will be equal to e n plus, plus what that we will be seeing, this is the Jacobean matrix and how it comes that I will explain very shortly. W minus W n this is effectively the matrix representation of this equation. So, if we call this equation as what is the number that we can give, we give a number 3 pertaining to today's lecture.

Then equation 4 that we are talking of now, that is nothing but, the matrix representation of this. In which case actually this error e n that we are talking of this is an error vector, which will be written as e 1, e 2, etcetera up to e n. So, there are number of observations, so there are n of errors individually. So, this is e n vector is basically the error vector of course, use the transpose notation for that. So, this way if you take this equation you can independently compute e 1, e 2, e 3 all this things.

So, like this you calculate all this things e 1, e 2, up to e n you compute and then, you represent the whole vector by e n vector. So, that if you are describing this equation 3 with i is equal to 1 to n, the way you can describe in the form of a matrix vector representation is this relation four followed, any doubts, yes please.

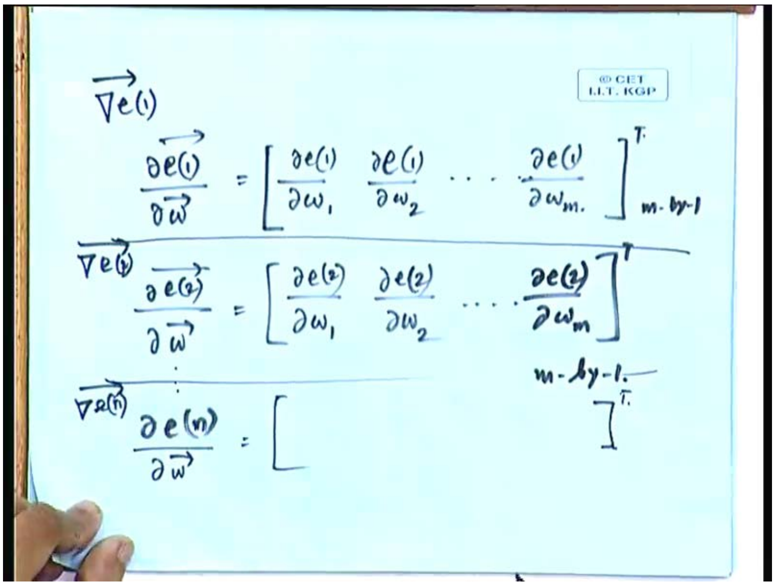
Student: ((Refer Time: 39:04))

E i, del e 1 del w vector, now instead of that I have written new matrix which is a J matrix and

Student: ((Refer Time: 39:21))

That is nothing but, the del e i only, so I am coming to this. So, when I transform from 3 to 4 and make in the matrix form that time this del e i del W is coming out to be the J matrix. And what is this J matrix, now you try to have a look at it. You see you have e 1 or e prime 1 you have, e prime 1 when you have in that case you have it as dou e 1 dou W vector. And what is dou e 1 dou W vector, there are n elements of such W vectors.

(Refer Slide Time: 40:02)



$\vec{\nabla} e(1)$

$$\frac{\partial e(1)}{\partial \vec{w}} = \left[\frac{\partial e(1)}{\partial \omega_1} \quad \frac{\partial e(1)}{\partial \omega_2} \quad \cdots \quad \frac{\partial e(1)}{\partial \omega_m} \right]^{T}_{m\text{-by-}1}$$

$\vec{\nabla} e(2)$

$$\frac{\partial e(2)}{\partial \vec{w}} = \left[\frac{\partial e(2)}{\partial \omega_1} \quad \frac{\partial e(2)}{\partial \omega_2} \quad \cdots \quad \frac{\partial e(2)}{\partial \omega_m} \right]^{T}$$

m-by-1.

$\vdots$

$\vec{\nabla} e(n)$

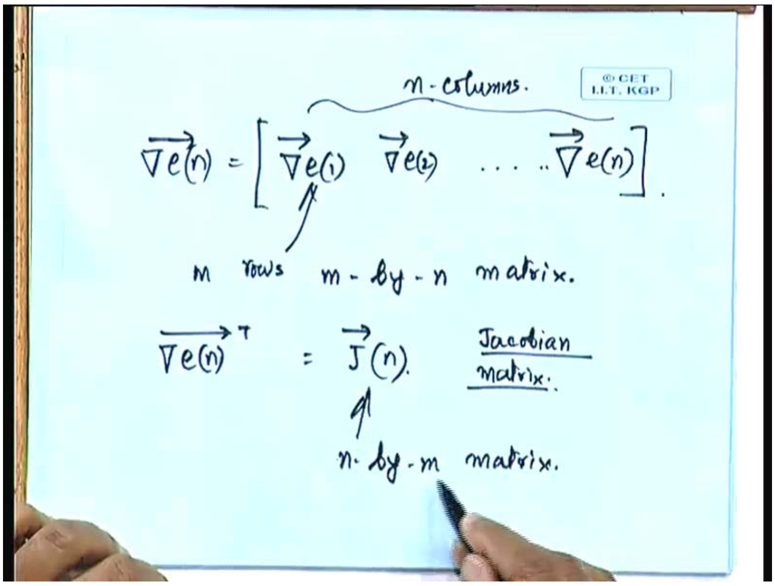
$$\frac{\partial e(n)}{\partial \vec{w}} = \left[\qquad \right]^{T}$$

So, that means, to say that when I have dou e 1, let me write it this way dou e 1 dou W vector that will be nothing but, I will write it as a row vector only, so and then make it transpose. So, it should be dou e 1 dou W 1 dou e 1 dou W 2 up to dou e 1 dou W m. Now, this is yes, this is are transpose of this will be the column vector of this. Just a minute, this is our dou e 1 dou W's definition will be this.

So, likewise when we make this second observation, when we take the second observation that is e 2. That time we will be having dou e 2 dou W as simply dou e 2 dou W 1 dou e 2 dou W 2 upto dou e 2 dou W m. Again a transpose of this, so that means to say that is m by 1 vector, so by taking this transpose I am ensuring that these are m by 1, this is also m by 1. So, likewise we will be considering dou e n dou W vector that also will be this, I am not writing down.

So, here actually what happens is that, when we are considering these transpose, these derivatives together in this vector form, then combinedly I am going to write the gradient of the error vector e n. Let me complete this you see that, I have defined this e n vector, e n vector is what e 1 to e n. Now, each of these errors e 1, e 2 etcetera are differentiable with respect to W, that means to say when I differentiate with respect to W that time it leads to m dimensional vector. So, what I do is that, now I define the gradient matrix.

(Refer Slide Time: 43:05)



$$\overrightarrow{\nabla e(n)} = \Big[\overbrace{\vec{\nabla} e(1) \quad \vec{\nabla} e(2) \quad \dots \dots \vec{\nabla} e(n)}^{n\text{-columns.}}\Big].$$

m rows   m-by-n matrix.

$$\overrightarrow{\nabla e(n)}^{T} = \vec{J}(n).$$

Jacobian matrix.

n-by-m matrix.

Now, I define a gradient matrix dou e n, where this dou e n will be nothing but, dou e 1 del, that is to say the gradient of e 1 and what is gradient of e 1. Gradient of e 1 is nothing but, the m dimensional vector of this is first one. So, this itself will be written as dou e 1, del e 1 that is the grade of e 1, grade of e 1 will be this. Similarly, this term will be nothing but, grade of e 2 followed like this, this will be grade of e n.

Now, I have to write this grade of the whole e n as grade of e 1, grade of e 2 up to grade of e n. And what is it, that I have got here you see each of these, this is a column, this is an m dimensional column and I am going to have n such rows. So, this is an m by n matrix, why because each of these terms are having m rows and there are totally n number of columns over here. So, it is an m by n matrix and this matrix that we have got will be described as grade e n matrix.

Now, if I take the transpose of this matrix, if I take grade e n transpose of that, that is defined as the Jacobian matrix. And so what is dimensionality of Jacobian matrix, Jacobian matrix in this case is n by m matrix. And you can now see that here, we require n by m matrix I think that something what, somebody was trying point out over here.

Student: ((Refer Time: 45:39))

Previous equation

Student: ((Refer Time: 45:43))

Del, transpose of that, but here what happens is that in this matrix notation this J n is going to be n by m matrix and W's are all m vector. So, W minus W n is also m dimensional vector, so this product becomes n dimensional vector. So, this expression, this expression 4 that I have written out over here, that is actually valid for an n dimensional vector over here.

Yes please

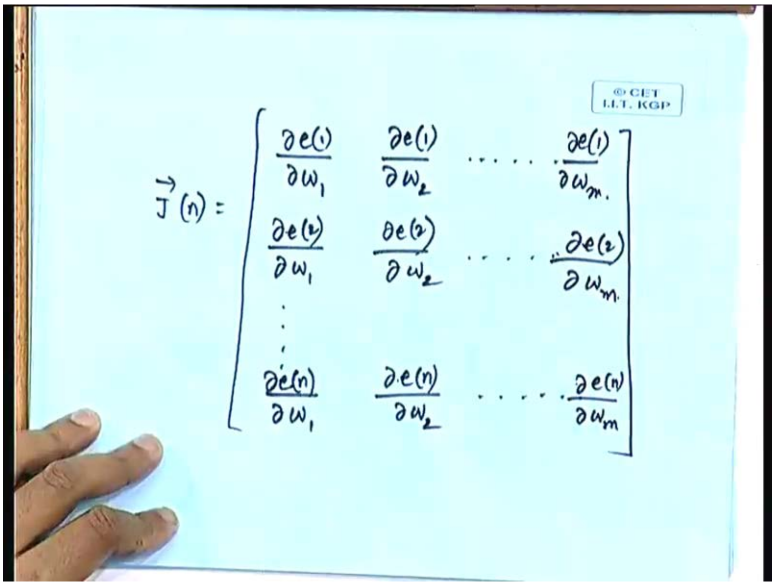
Student: ((Refer Time: 46:28))

That means, to say that I make the first observation i is equal to 1 and I do not change the weight of that, I keep the weight constant at W. So, I make n number of observation, I sum up all the squared errors that is making a summation of e i square and then, only I am changing the weight. That means, to say that I am minimizing the sum of squares followed, sum of squares of the observation. So, now having defined this e vector e n vector.

So, in fact, this is what we are back again to, so I have defined what this matrix is, so this matrix J matrix is called as Jacobian matrix, so this is the Jacobian matrix. So, Jacobian matrix will be of n by m dimension, where basically you are considering here the m dimensional input you are considering. That is to say the error, the error that you are considering to be of, each of the error terms are m dimensional vectors and we are having n such errors.

So, totally it is n by m Jacobian matrix that we are getting, which is basically the Jacobian matrix composition, now we can simply see.

(Refer Slide Time: 48:12)



$$\vec{J}(n) = \begin{bmatrix} \frac{\partial e(1)}{\partial w_1} & \frac{\partial e(1)}{\partial w_2} & \cdots\cdots & \frac{\partial e(1)}{\partial w_m}. \\ \frac{\partial e(2)}{\partial w_1} & \frac{\partial e(2)}{\partial w_2} & \cdots\cdots & \frac{\partial e(2)}{\partial w_m}. \\ \vdots & & & \\ \frac{\partial e(n)}{\partial w_1} & \frac{\partial e(n)}{\partial w_2} & \cdots\cdots & \frac{\partial e(n)}{\partial w_m} \end{bmatrix}$$

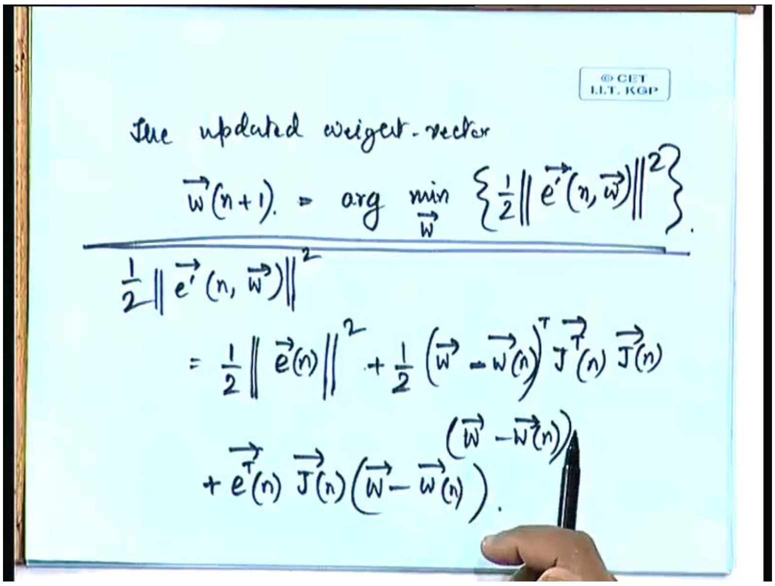
So, J n the Jacobian matrix is in fact tell me dou e 1 with respect to W 1 dou W 1, then

Student: ((Refer Time: 48:28))

Correct, e 1 by W 2, here e 1 by W m and what will be the second term over here, second row over here e 2 by W 1 and here it will be e 2 by W 2 up to e 2 by W m. And what will be the n th row over here, n th row will be dou en by dou W 1 dou e n by dou W 2 up to dou e n dou W m. So, this is the composition of the Jacobian matrix, n by m Jacobian matrix.

Now, using this Jacobian matrix we have formed this equation, so ultimately the error expression that we have got is that at the point W is equal to W n. Around this point the error e n that we are going to get is equal to the e n plus this Jacobian term, Jacobian multiplied by this weight term. And this has got some significance which I will be coming to very soon.

(Refer Slide Time: 50:06)



© CET I.I.T. KGP

The updated weight-vector

$$\vec{w}(n+1) = \arg\min_{\vec{w}} \left\{ \frac{1}{2} \left\| \vec{e'}(n, \vec{w}) \right\|^2 \right\}.$$

$$\frac{1}{2} \left\| \vec{e'}(n, \vec{w}) \right\|^2 = \frac{1}{2} \left\| \vec{e}(n) \right\|^2 + \frac{1}{2} \left(\vec{w} - \vec{w}(n) \right)^T \vec{J}^T(n)\, \vec{J}(n) \left(\vec{w} - \vec{w}(n) \right) + \vec{e}^T(n)\, \vec{J}(n) \left(\vec{w} - \vec{w}(n) \right).$$

Now, the updated weight vector that we are going to have, that is to say that we have let us say already W n we have, so now I am talking of W n plus 1. Now, this W n plus 1 will be what, that will be the argument of minimum over W half of norm e prime n W this square, this is easily understood. This is to say, this is the error term e n W is our error term and we are, that is the combined error term, sum total of all the errors and we are taking the square of that error.

So, this is this is actually a vector, so we take squared norm of that vector. And the value of W that minimizes this error term, it is that W that we are going to use as the updated weight. We are minimizing this error and the corresponding value of W is the updated weight vector. Now, we again come back to our this equation, from this equation where the new error vector is expressible as the old error vector plus this term, here it is possible for us to calculate the error square very easily.

You see that, if the equation 4 is like this, that is e vector is equal to e n vector plus J n in to this, then the transpose of the left hand side is also equal to the transpose of the right hand side. Now, you just multiply the transpose with the vector, then what you get, you get the squared term. So, you can form the square of the left hand side and then, this transpose multiplied by this term will give you a square of this whole term and what does it lead to.

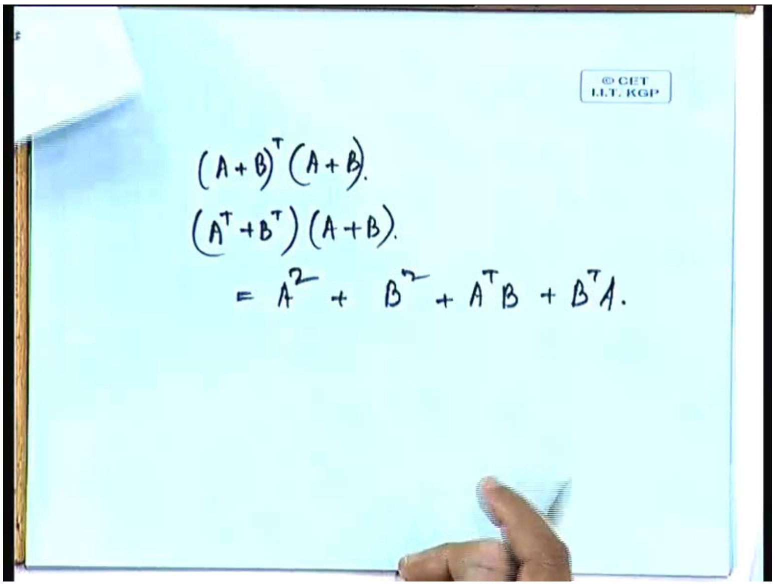
So, using this equation 4, in order to evaluate the squared Euclidean norm of this e vector, this is what you will be getting. We will be getting square of e prime n W that will be equal to half, in fact again square term I am deliberately multiplying by half just, so that during derivative I do not have any problem. So, half of the squared of e n square, that is the norm of e n square plus anybody can have a look at this expression and predict the rest of the terms.

So, this is very clear, this will be squared of this which will be this plus this, transposed in to this plus this. So, what is it going to be the first term is definitely going to be e transpose n, e transpose which leads to e n square this is very clear. And then, we will be having two such cross terms, where it will be e transpose J n this term then there will be another term which will be.

Student: ((Refer Time: 54:11))

There will be in fact, two such terms why because, this plus this transpose in to this plus this, effectively it is of this form.

(Refer Slide Time: 54:23)



$$(A+B)^T (A+B).$$

$$(A^T + B^T)(A+B).$$

$$= A^2 + B^2 + A^T B + B^T A.$$

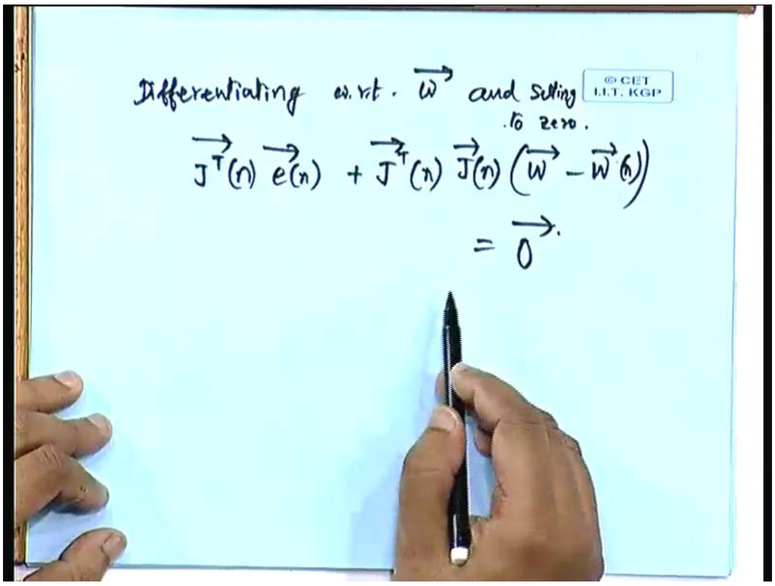
Effectively it is of this A plus B transpose into A plus B that is what we are doing. So, effectively it is A transpose plus B transpose into A plus B, which will be A square plus B square B square term also will come. And what happens is that we will be getting A transpose B plus, we will be getting a B transpose A. And in fact, in this case what happens is that these two will be, if in this particular case, where we are expressing it in this form.

We will be getting ultimately this plus the second squared term will be the B B transposed part. The B B transpose part will be half of W minus W n this transpose J transpose, this multiplied by J n in to W minus W n thus second term simply, so there is no doubt about this term. And then, the other two terms, they will be equal and they become e transpose n J n W minus W n this, so this is the updated weight vector.

So, this is the square norm of the equation that we have got and what do we need to do now, we need to differentiate this with respect to W. So, differentiate what, differentiate what differentiated the left hand side with respect to W, means that individually all these terms we will have to differentiate with respect to W. Now, what happens to the first term, first term if you differentiate with respect to W 0, because it is not dependent on W over here.

What will be the second terms contribution, if you are differentiating with respect to W in that case this is with respect to W. So, W J transpose and you will be getting here, for this, no this whole thing transpose is there W minus W n and transpose. So, it is of the form W transpose, it is of the form x transpose W and there your derivative is simply going to be x. So, in this case the derivative is going to be simply the x term that is to say by differentiating you will be getting...

(Refer Slide Time: 57:42)



Differentiating w.r.t. $\vec{W}$ and setting to zero.

$$\vec{J}^T(n)\,\vec{e}(n) + \vec{J}^T(n)\,\vec{J}(n)\left(\vec{W} - \vec{W}(n)\right) = \vec{0}.$$

Here J transpose n e n plus, you will be getting another expression that is this one, this one also will be differentiable with respect to W, just a minute we will solve your doubt. Here this is J transpose n in to J n into W minus W n equal to 0 vector. So, this is differentiating with respect to W vector and setting the results to zero, setting to zero this is what we will be getting. We will be getting W minus this term, so please verify this expression yourself. In fact, we are already running out of time, so we will continue from this point in the next class.

Thank you very much.