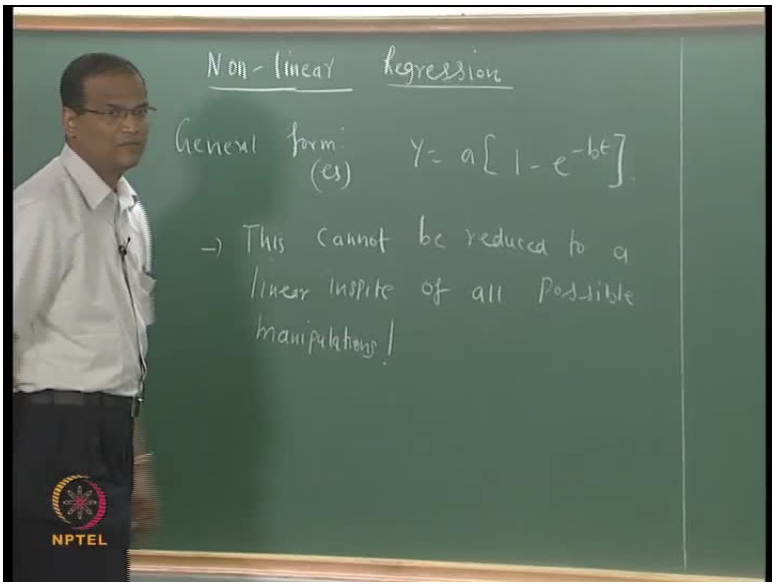
Design and Optimization of Energy Systems

Prof. C. Balaji
Department of Mechanical Engineering
Indian Institute of Technology, Madras

Lecture No. # 20
Non-Linear Regression (Gauss-Newton Algorithm)

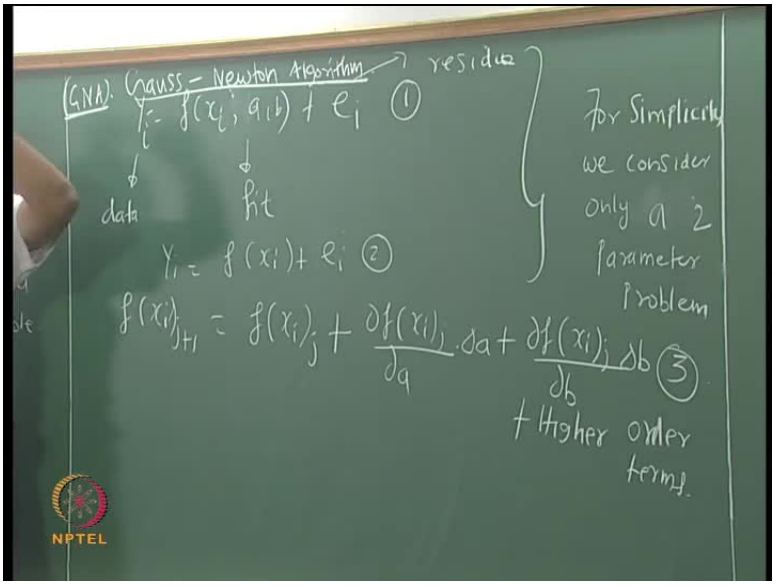
Good morning. We will continue with the discussion on non-linear regression. I will teach you the Gauss-Newton Algorithm, which is very powerful and potent technique to handle non-linear regression. Yesterday, we saw one example namely, the heating of a first order system, where there is a constant heat generation of Q watts. And then, there is a convection afforded by convection heat transfer coefficient of h watts per meter square kelvin and area A. Then, we founded the governing equation is m c p d t by d tau is Q minus hA delta t and so on. And then, there are three regimes: the heating portion of the curve, the steady state and the cooling. And finally, the resulting equation, the resulting solution, was of the form theta is equal to? Theta is equal to theta s into 1 minus e to the power of minus t by tau; where, theta s is the steady state temperature excess; and, theta is the temperature excess; and, tau is the time constant. So, regardless of your mathematical manipulation, it is impossible for you to reduce it to a form, where you can apply linear least square. Therefore, you have to… You have no other choice, but to go for non-linear regression.

(Refer Slide Time: 01:27)



So, the general form is… Suppose you have general form – example – Y is equal to a into 1 minus e to the power of minus bt. This cannot be reduced to a linear form in spite of all possible manipulations. Therefore, we take recourse to non-linear regression. So, that is the idea. Now, how do we write this?

(Refer Slide Time: 02:34)



This can be written as Y is equal to f of x i… Y i is f of x i given a, b plus e i. What does it mean? If I happen to get a function f of x i, which has got… for which the parameters a, b are known. If you insert the values of a, b and you substitute it; substitute the value

of x i for that; I will get a value f of x i for that. But, that value of f of x i need not agree with the Y data, because after all, the f x i is a form, which I am fitting. So, the difference between the Y of i minus f of x i with the best values of a and b will result in some error, which is called e of i. So, this is the residue. So, this is the residue. So, this is the fit. This is the data. So, I will simply call it Y i is equal to f of x i plus e i with the understanding that the f has got the parameters a, b.

Now, let us say… It need not be of the form only this; it can be any non-linear form with two parameters: a and b. So, for a general two parameter problem, where linear regression will not hold, we are proposing this. Now, how do you go about doing this – non-linear regression?

Student: Assume a value…

You have to assume a value of a and b; and then, you have to expand this f of x i from the j-th iteration to j plus one-th iteration using Taylor series. And then, you truncate the Taylor series after certain number of terms. And then, find out the difference Y i minus e i; and, find out the difference Y i minus e i for all the data points. Then, in a least square sense, try to minimize this difference. Now, you will get new values of a and b. Compare the old values of a and b with the new values of a and b. That alone will not do. With the old values of a and b, find out what are the values of y you are getting. Then, y data minus y fit for each of the i values; square each of the values; sum it up. So, you will get the s or the sigma of r square.

Now, at iteration 0, for assumed values of a and b, you have a value of s 1 or s 0, that is, the sigma r square. That is the sum of the residue. Now, going through this procedure, you will get a new value of a new, b new. With this a new, b new, you will get the new values of y fit. Find out the y fit minus y data whole square and see how the s, that is, the sigma r square is going down with the iteration. After it has reached sufficiently small value, you can take a call and say that, this is enough. So, basically, the key is we are still using linear least squares; which means we have to approximate the non-linear form by an appropriate linear form; that means we expand it as a Taylor series in the neighborhood of a, b. What is that a, b? That is a naught, b naught. That is the initial values of a, b. And then, truncate with first order terms and then use a least square minimization. When doing the least square minimization, since many points are

involved, it will advantageous for us to use the matrix algebra. That is why I introduced the matrix algebra for the least squares in the last class.

Now, the root is clear, the root we are going to take? Therefore, now, I can say f of x i j plus 1 will be equal to f of x i j plus? I am expanding the vicinity of a and b. Can you tell me?

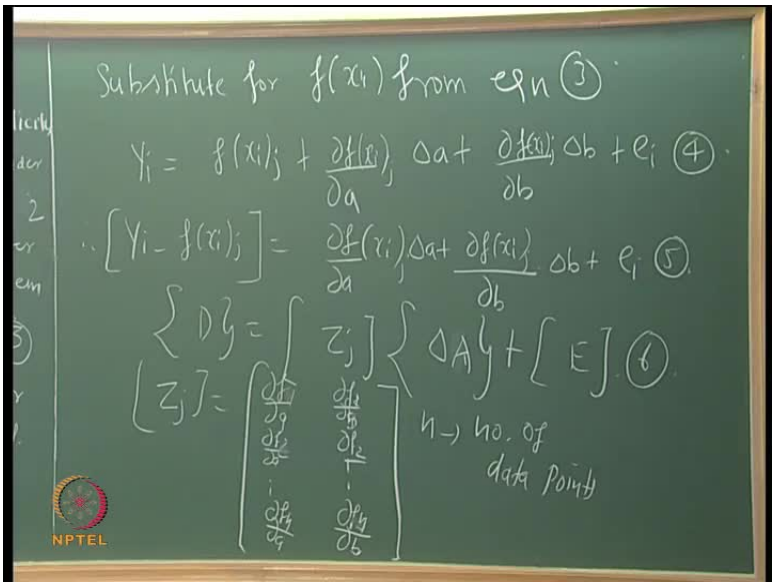
Student: dou f by…

dou f of x i j by dou a into?

Student: Delta a.

Correct. Plus dou f of x i j by dou b into delta b. So, when you try to regress by starting this approach, where in, you take f of x i j plus 1 is equal to f of x i j plus this and truncate with the first order terms; you are using the Taylor series to linearize the non-linear form. So, this is called the Gauss-Newton method – GNA – Gauss-Newton algorithm. So, now, I can call this as 3. Here the f of x i – right side, I can treat it as f of x i of j plus 1. I can substitute this f of x i j plus 1 on the right-hand side, because I am try to look at the residual e; I am trying to minimize the residual e. What is the use of this Taylor series expansion? This Taylor series expansion has to be used in this equation 2. Equation 2 and equation 1 are one and the same, except that I drop the a and b, so that we do not have notation confusion. So, once you have a Taylor series… This is not a Taylor series expansion. This says that, when a data is fit with an approximate function f, there will a residue associated with that. Please remember that, ultimately, you want to minimize a residue. We do not want to minimize e 1, e 2, e 3; we want to minimize the total residue when there are n data points. Therefore, now, we substitute for f of x i from equation 3.

(Refer Slide Time: 10:05)



Y of i is equal to f of x i j plus dou f of x i by dou a into delta a plus dou f by dou b into delta b plus e i. There are some subtle things I have done – i, j plus 1; you should be able to follow that. Only j values are known when you start with. From j, you are going on to j plus 1. Correct? Therefore, we can bring the f of x i j here – Y i minus f of x i j is equal to dou f of x i by dou a into delta a plus dou f of x i – this should be j – by dou b into delta b plus e i. So, D is equal to… What do you want to call that as? What was the notation we used in yesterday's class? Z. So, Z of j into delta a plus E. Why am I using matrix form? There are several values of Y I – Y 1 to Y n, where there are n data points. Are you getting the point? Now, Z j will be…

Student: What about…

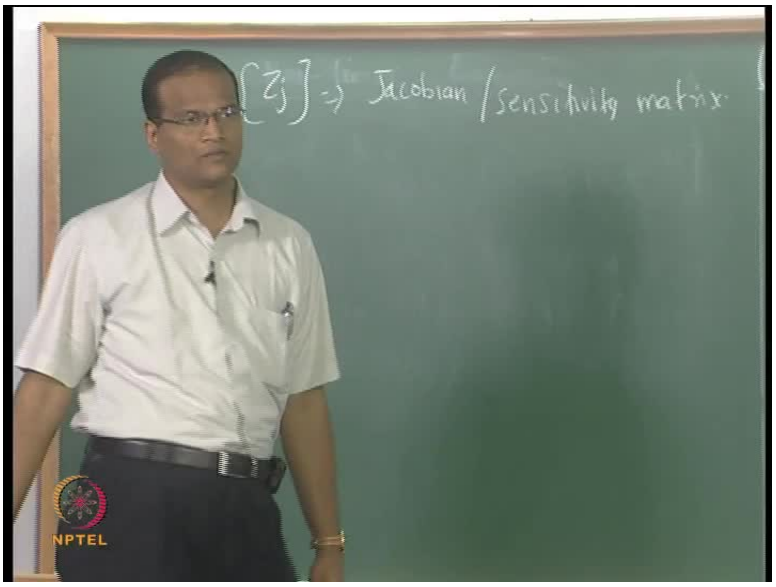
What?

Student: Delta…

No, I will call it delta capital A. So, that is a small notation problem. a and b are there, of course. See you can call it a general a; a itself can be made of a, b, c, d, e, f and all that. Sorry about that. Just change it. So, this will be dou f 1 by dou a, dou f 2 by dou a up to dou f 4 by dou a; dou f 2 by dou b up to dou f 2 by dou b; same. There is only f; f 1, f 2; where, n is the number of data points. So, dou f 1, dou f 2. That f 1, f 2 is the notation – the number of points. So, in the parlance of optimization or in the parlance of matrix

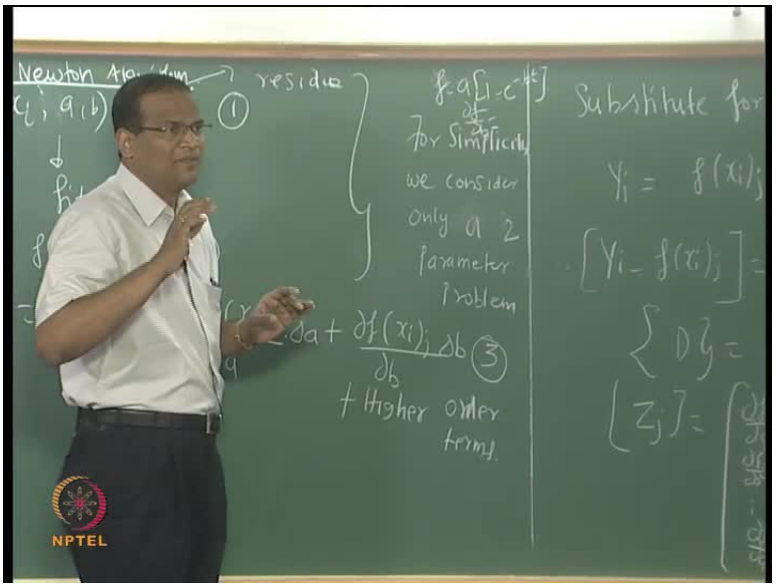
algebra, if you get a matrix, where there is a function f and you are taking the partial derivative with respect to the parameters a and b; and then, you are denoting it for each of these points, and finally, you will get a matrix like this; such a matrix is called the? It is called the Jacobian matrix.

(Refer Slide Time: 14:54)



Z j is called as the sensitivity of the Jacobian matrix. Do not get lost here. Do not get lost in matrix algebra.

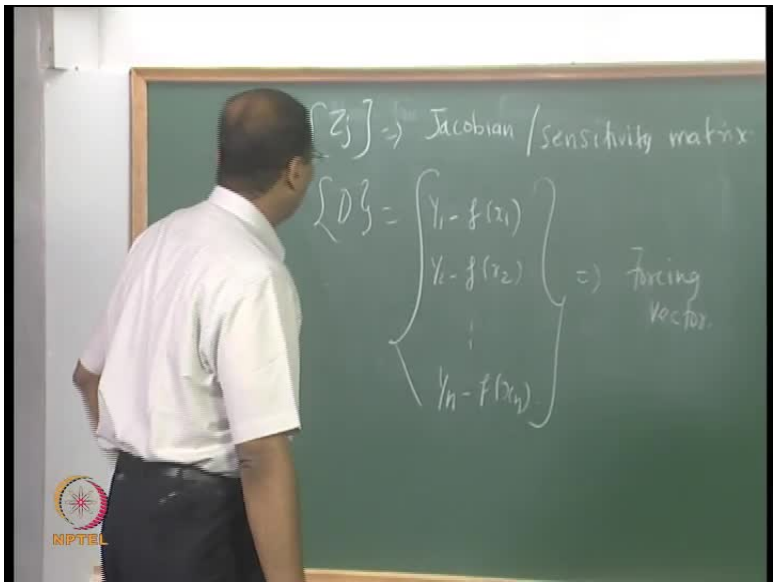
(Refer Slide Time: 15:22)



For example, it is very straightforward. Y is equal to a into 1 minus e into power of minus bt. For example, Y is equal to a into 1 minus e to the power of minus bt; dou f is this. So, f is this. So, dou f by dou a… Watch carefully; dou f by dou a is 1 minus e to the power of minus bt. So, you have to start with some values of a and b. So, in initial guess values of b you use; this dou f by dou a is 1 minus e to the power of minus bt. But, this can be worked out for each of the data points. If I give you a table, it can be worked out for the first point, second point, third point. Similarly, you can work out dou f by dou b. Correct? Which is a into minus b into minus e to the power of minus b t; a into minus t. So, at e to the power of minus bt. So, you can work out at e to the power of minus bt (Refer Slide Time: 16:17). Substitute the individual values of t and you can work out this matrix. Is that fine?

Student: Iteration for a and b.

(Refer Slide Time: 17:50)



It is an iteration. You start with… See the beauty is you cannot evaluate the Jacobian matrix, unless you have the values of a and b. a and b are the things, which we want to find. You cannot go to the next step, unless you assume a and b. This was not the case for the linear least squares. That is called the non-linearity. So, you have to assume a value of a and b; work out all these and get the new values of a and b. Are you getting the point? This situation did not arise in the linear least squares case; we straight away got e i sigma x i, y i minus sigma a by b, sigma… b i equal to… And, you just got sigma

x i, y i and sigma x square and all that. This is not the case here, because it is not possible for us to reduce to the straight line form.

Now, if D is given like (Refer Slide Time: 17:23) this; from what I taught in the last class, can you write the representation using linear least square; if you want to minimize the residue, how will the final result look like? I already gave you a hint in the last class – linear least squares. So, what is the D matrix? D vector? What is D?

Student: Residue.

How do you calculate the D? Y 1 is equal to f of x 1, Y 2 is equal to f of x 2 up to Y n is equal to f of x n. This is basically called the forcing vector. Why am I calling it Z j (Refer Slide Time: 18:40)?

Student: (( )) for the j-th value…

It is for the j-th value of iteration, because Z j is not constant. But, the Z was constant for the linear least squares; that is invariant, because the Z contains 0, 1, 2, 3, 1, 1, 1 in the yesterday's problem. Yes; you missed yesterday's class?

Student: Presume the y here, we have j plus 1-th iteration.

Y is data, no? j plus 1-th iteration, we do not know the Y. Y data is not associated with any iteration. Y data is given to you. Y fit varies with iteration. Correct? Now, applying linear least squares…

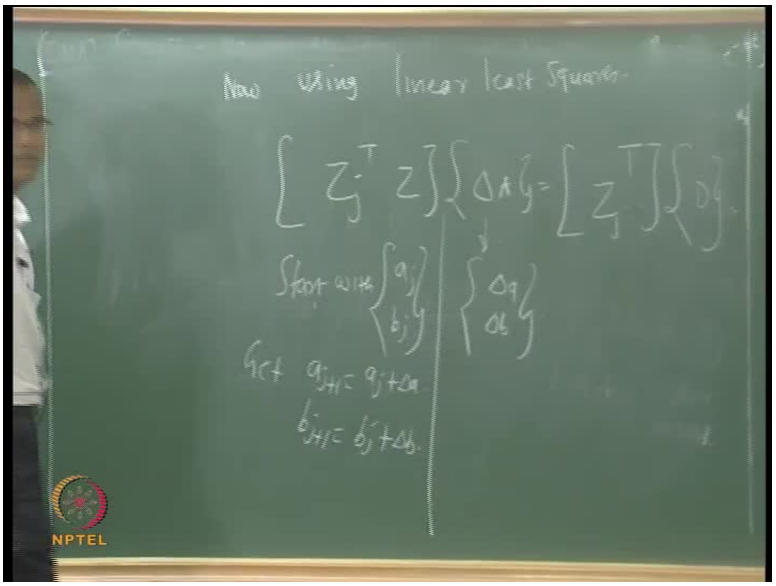
Student: Delta e…

Delta e?

Student: Sir, we are finding delta a.

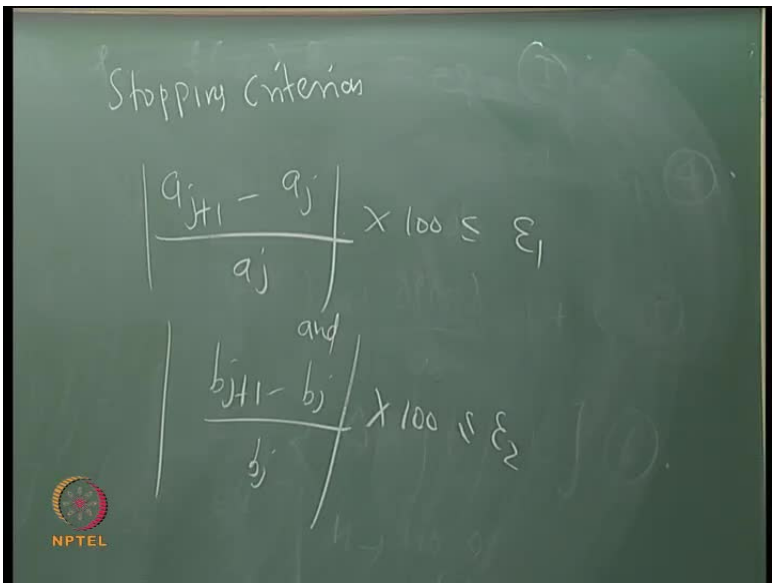
Yes, correct. To minimize the e.

(Refer Slide Time: 19:52)



Now, Z j transpose T Z into delta A is equal to Z j transpose T operating on? Correct; good; D; that is it. So, this delta A… So, you will get delta a, delta b; start with a j, b j. Get a j plus delta a, b j plus delta b. When do you stop?

(Refer Slide Time: 21:21)



When epsilon 1 and epsilon 2 are decided by you; they can both be equal, but it could be different. There are other criteria, for example; theta minus theta, fit sigma, r square; how it gets minimized and so on. So, the key points in a non-linear regression are as follows: the residual e gets absorbed here (Refer Slide Time: 22:32). The residual e; the sum total

of residual e for all the values… In a least square sense, if it has to get minimized; that is the theory of least square.

Student: (( ))

That we have seen minimax (( )) We already did so much for the least square. Even for the least square, Y is equal to f of x plus e i, which we want to minimize? e 1 or e 2 or e 3 or e 4?

Student: (( ))

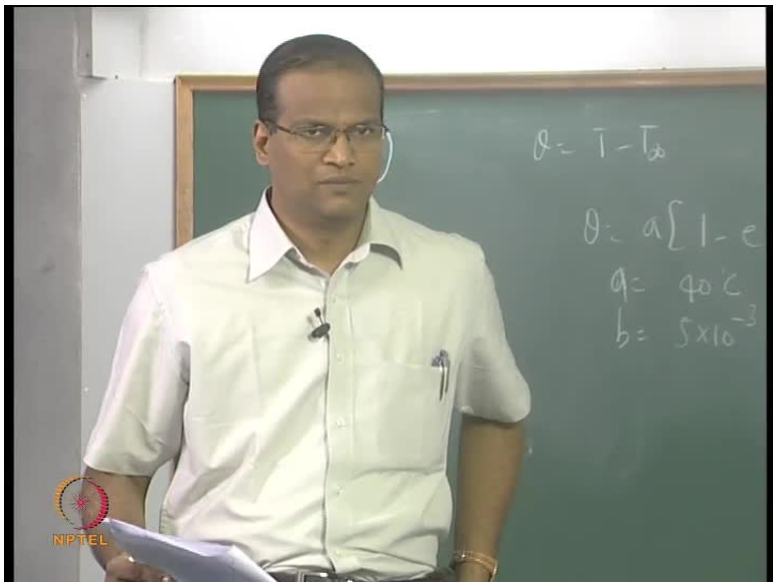
That is what? That is concisely brought out by the matrix. When you do this, it is – you are minimizing the least squares, least square deviation. That is why I taught the linear least squares with matrix algebra yesterday, so that you do not get confused. The e i gets merged into this, because the e i's are submerged in this. Are you getting the point? e i is a variation of the… e i is the difference between Y i minus Y fit for each of the points. So, what we do is e i whole square sigma i equal to 1 to n, what we minimize. So, if you want to minimize that, the equation is this. Since people may get a doubt, sir, how do I believe this? Yesterday, though it was not required for this course, I taught you how to do linear least square with matrix algebra; without that also, we solved. Are you able to get the connection now? Now, it may look nebulous and vague; do not worry, we will solve a problem; everything will become clear.

Now, what are the cardinal steps involved in this? Basically, the non-linear least squares is an iterative procedure; it is not going to be easy. So, in the exam, maximum I will ask one iteration. If you really want to do non-linear regression, you have to write a computer program; go to matlab or you have to use a non-linear regression tool. As is the case with any non-linear regression procedure; as is the case with any iterative technique, successful solutions… For example, if you want to get good estimates of a and b, it is critically dependent on an initial case. If the initial case is way off, then the solution may converge slowly, the solution may oscillate, the solution may also diverge. So, the Gauss-Newton algorithm is not the universal penertia; it is not a cure all; it cannot solve all the optimization problems. But, it is a strategy, which is worth.

Suppose, you go through the Gauss-Newton algorithm and you get stuck in between, there are some powerful tools, which are available for you to correct; that we will see

after I give the example, after we do the example. The powerful algorithms are basically the steepest descent method, the Levenberg-Marquardt algorithm, conjugate gradient; there are so many other techniques, which are available. Please remember that, we are not using the second order terms here (Refer Slide Time: 25:21). So, there is a charge that any scheme which works only with this, stops with the linear expansion… the first order terms to the Taylor series as a chance that it can go this way, that way. Are you getting the point? So, it does not mean that, we have a solution, which will have an algorithm, which will solve all optimization problems. If this were to be true and simulated annealing, genetic algorithm, ant colony optimization; none of these techniques would have developed; there are so many problems; where, you have got non-linear regression, you have got several parameters, and the Gauss-Newton algorithm – all these algorithms will get chocked. You will struggle; they will get stuck in what is called a local optimum; they cannot come out.

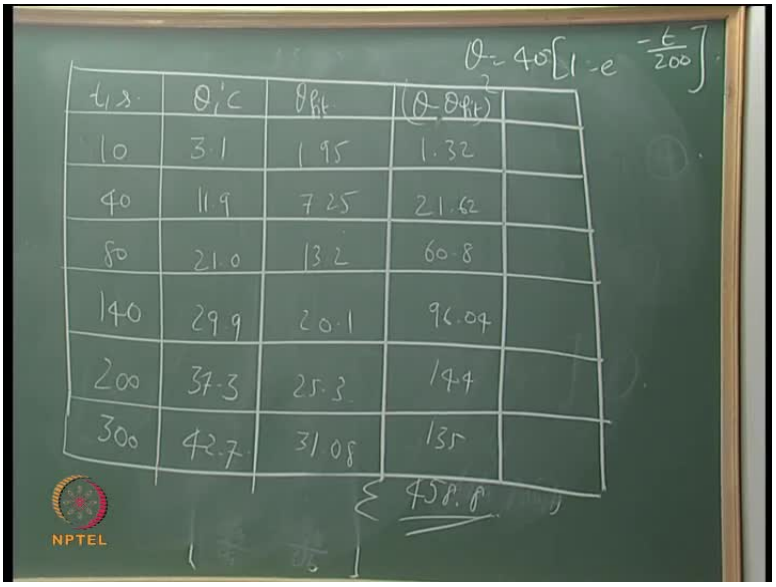
(Refer Slide Time: 26:31)



Let us solve a problem now. Things should become clear. Let us take the case of heating of a first order system. So, this is example number? Problem number? We will keep this (Refer Slide Time: 26:18). Problem 19? Please take down. Consider the heating of a first order system… Consider the heating of a first order system… Consider the heating of a first order system, whose temperature excess theta… whose temperature excess theta is known to vary as follows; whose temperature excess theta is known to vary as follows; I mean varying the form, which is given on the black board; where, a and b are constants

and t is the time in seconds; theta is in degree centigrade and t is in seconds. Using the data given below with an initial guess of a equal to 40; using an initial guess of a equal to 40; a is equal to 40 degree centigrade and b is 5 into 10 to the power of minus 3 second inverse. Perform one iteration of the GNA; perform one iteration of the Gauss-Newton algorithm. Compare the residuals before and after the iteration.

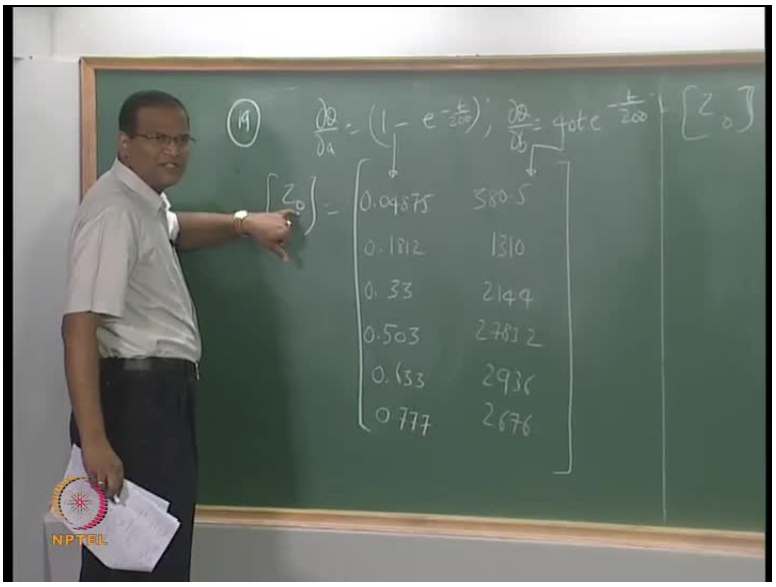
(Refer Slide Time: 28:57)



I will give you the data now. Second quiz; one question surely on; Gauss-Newton. So, do it carefully now. That should carry at least 20 marks out of 40 marks. I go to one extra column; just leave it. Is it correct? So, it is very nice. To start with, this is very nice. You can substitute the value of t and get the theta fit; and then, calculate the sigma. But, that is not end of the story; that is the beginning of the story. You will not be in a position to fill up this column until you do the sigma x i, y i, sigma x square and all these for the linear least square. Here to start with, you assume some values a and b and find out how this is. If the theta fit is excellent, then straight away you go home. That is rarely the case. So, you first evaluate this and then take dou theta by dou a, dou theta by dou b; get the Jacobian matrix, get the forcing vector; you will get still a 2 by 2 system only, because a and b; you can either invert or solve the simultaneous equations. I except you have to complete within half an hour. We have 18.4 minutes; you should be able to proceed. Start. If the residuals after iteration one spill over to the next class, it is all right. But, at least, you should able to get a and b within 20 minutes. Let them do. You can give a break of 5 minutes. I will start filling out. Correct? Is it ok? Deepak, you got it? It

is pretty high. So, it is not the correct value of a; you are not using the correct values of a and b; which is all right. So, we get a high value of the residue.

(Refer Slide Time: 33:16)



Now, we have to work out the Jacobian matrix. Correct? Is it ok? Fill out the values using appropriate value of time; 10, 40, 80, 140, 200. So, you do that; you will get this. Please correct me if I make some mistake. Alok, you are doing? You can just check for 2 or 3 values. All of you should get this. Swaroop? Hope everybody has a Calci? Abhinandan, it is fine? If the Jacobean matrix has mistakes, then the whole thing comes. It is not difficult, but it is painful. So, you have to do it patiently; that is all. Heavy calculations are involved. Unfortunately, this course is like that. Now, the other one. So, what I will do is; already we spent so much time on this. So, what I will do in my calculator is; I will take 0.04875; I will subtract 1; I will subtract 1 multiplied by 40; and then, multiply by this. I believe it will speed up; there may be other ways for you to do that. Other way is 1 minus is e to the power of minus t… I do not have to do that e to the power minus t by 200 again; whatever. So, the first value is 380.5…

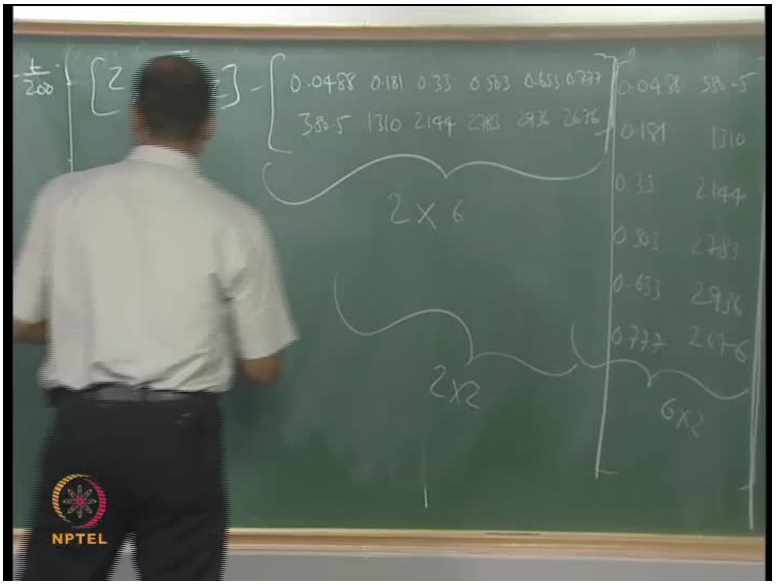
Student: (( ))

I do not understand what you are saying.

Student: Theta minus theta fit will not be incorporated in that value on the (( )).

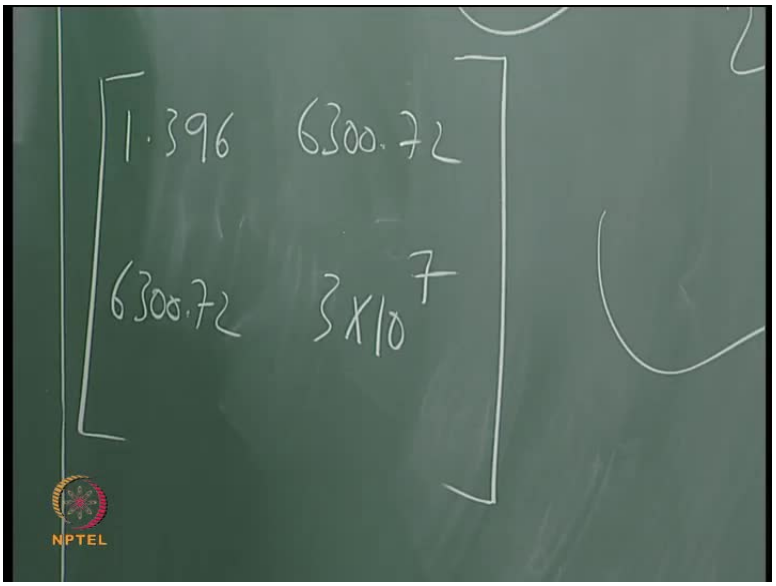
Theta minus theta fit is actually driving the whole thing. We are minimizing that error. But, the theta minus theta fit is the approximate measure; it is a performance metric; it is only the measure, because we are taking theta minus theta fit individually and sigma of that we are trying to minimize. But, what is the algorithm for that is this. This is indicative. This gives you the health of the fit. Shall I erase the main one? Everybody has this.

(Refer Slide Time: 38:19)



Now, Z naught into Z… Why is it (Refer Slide Time: 38:30) Z naught? We are starting with the 0-th. Now, you have written down Z transport T Z. So, this is basically a… So, we will get a 2 by 2 matrix.

(Refer Slide Time: 40:46)



Can you tell me the values? Good; 1.396; not 4000.

Student: 6300.72

6000?

Student: 6300.72

Is it? 6300 is correct?

Student: Yes, sir. 30026.2 dot…

No, put as engineering and tell me (( ))

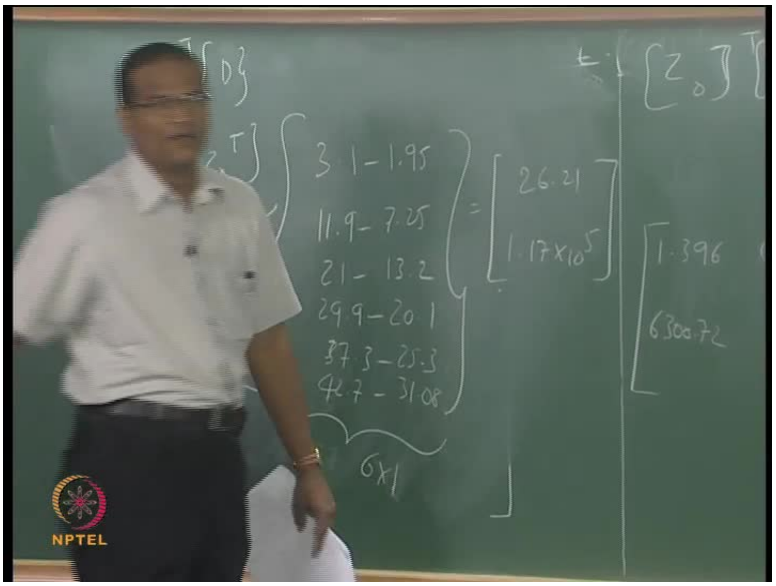
Student: 3 into 10 power 7.

3 into 10 power 7?

Student: Yes, sir.

I got 2.5 into… I think it is good that I left one column; I can put theta minus theta fit new after we get the new values of a and b. Let us keep that. Let us keep this (Refer Slide Time: 41:56). By chance, I am have not erased it so far. Now, we can get the values of… What about the forcing vector?

(Refer Slide Time: 42:20)



What about the forcing vector? What did you get? Into y minus (Refer Slide Time: 42:56)… This 37.3 or… It is 6 by 2 or what is it? 2 by 6; 6 by 1; 2 by 1. So, you will get a column vector. So, what is this?

Student: 26.21

Good; 26.?

Student: 21; 1276

Engineering.

Student: 1.17 into 10 to the power 5.

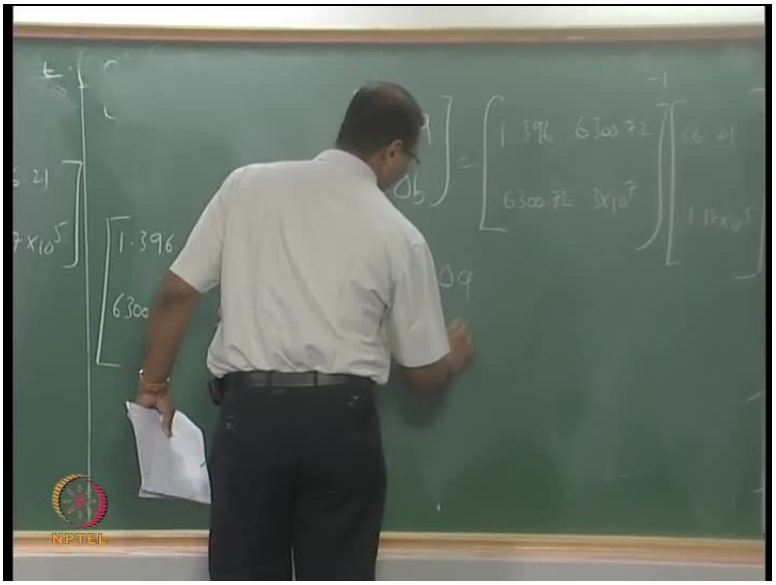
Good. Not bad; in 13 minutes, you have got all the values; go ahead; invert it. Get delta a, delta b; delta a, delta b?

Student: 6.41 and 2.55 delta minus a.

Is it? Let us work it out. So, we got this; we got this. Shall I erase (Refer Slide Time: 44:45) this? Please pay attention; if you go to the library; if you look at 100 books on statistics, 99 books will not have non-linear regression, if you think, then learn it own, there are very few books, which will give you this; very few books, which will have worked out example and all these. So, please pay attention in the class. So, these

problems we have spent time to work out and this thing; get values of a and b new. I have asked my students to do one more iteration and all that. So, please pay attention. This is not available in most books.

(Refer Slide Time: 45:41)

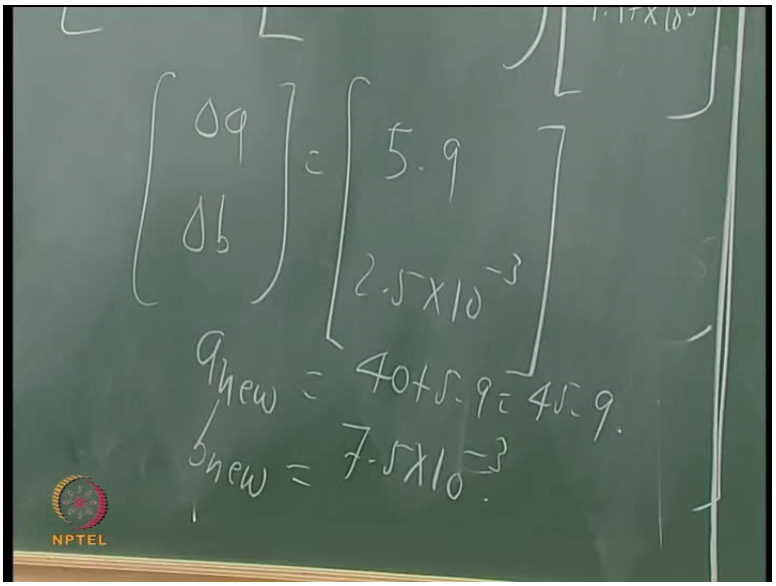


Now, we can get the value of delta a, delta b… This is inverse. Take the determinant; exchange these two; put minus; get the inverse; cross multiply; get the values of delta a, delta b. Abhishek, you got it? No calci? Pavan, got it? You are lost?

Student: I did not get my calci.

I think by the way, the fun is lost. Please bring you calci. Sowmya, you got it? Now, the other values are ok? You got the Z transpose T Z – all those things? Venugopal, you got it? Now, what is delta a, delta b?

(Refer Slide Time: 47:10)



Student: 22…

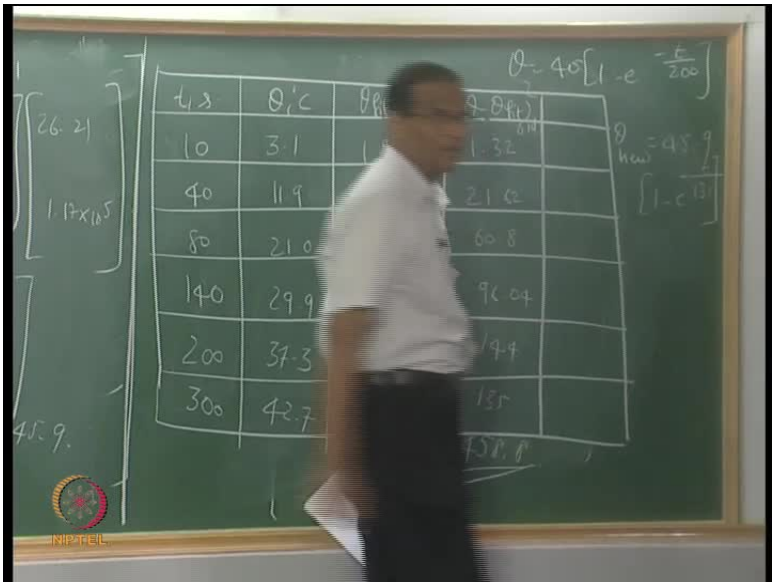
22?

Student: 6.5, 5.93

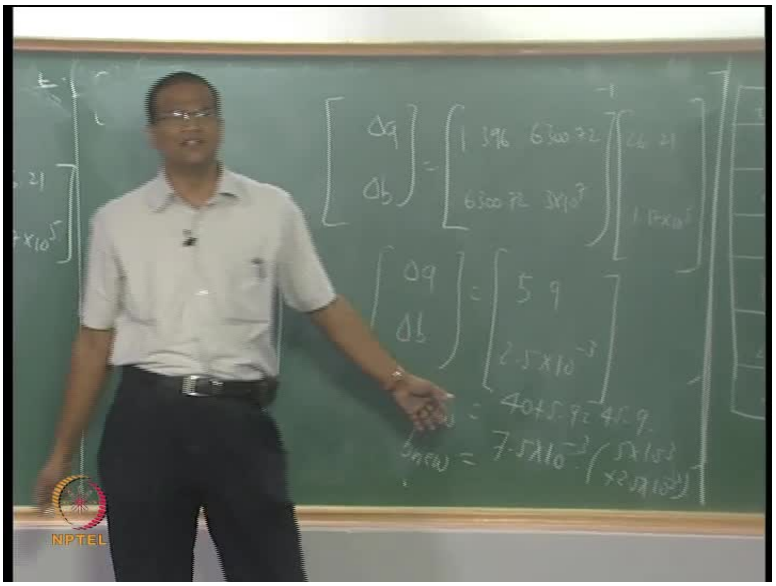
It cannot be written.

Student: 5.9 and 2 into 10 power minus 3; 2 .5 in to 10 power minus 3

(Refer Slide Time: 48:04)



(Refer Slide Time: 48:57)

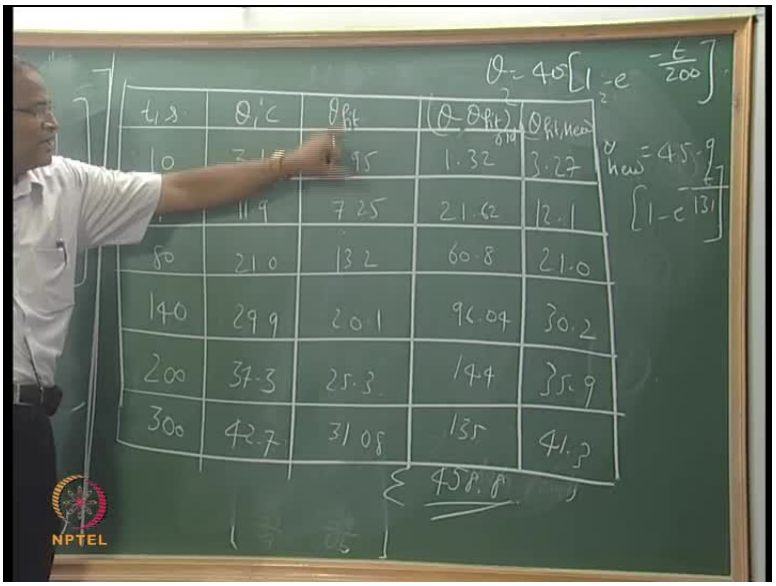


So, a new is equal to 40 plus 5.9 equal to 45.9; b new will be?

Student: 7.5…

Is it? 7.5 into 10 power minus 3. So, this is old (Refer Slide Time: 48:04). What is Z? Can you fill that all? Can you fill the last column? People who are little slow, please look at the steps. And if you are able to understand, you will get a 2 by 2 matrix; you will solve for delta a and delta b. Finally, you will invert this Z transport T Z and then get delta a, delta b. a new will be a old 40 plus delta a; b new will be b old plus… b new is 7.5 is it? Correct, because it is 2.5 plus 5 into 10 to the power of minus 3. You should be able to get this; it is not such a difficult thing that…

(Refer Slide Time: 49:14)



Now, please work this out. Or, at least let us get the theta fit; let us get the… What is theta fit new? First one, 3.3; 3.27.

Student: 12.08, 20.99, 30.15, 35.9, 41.3.

41.?

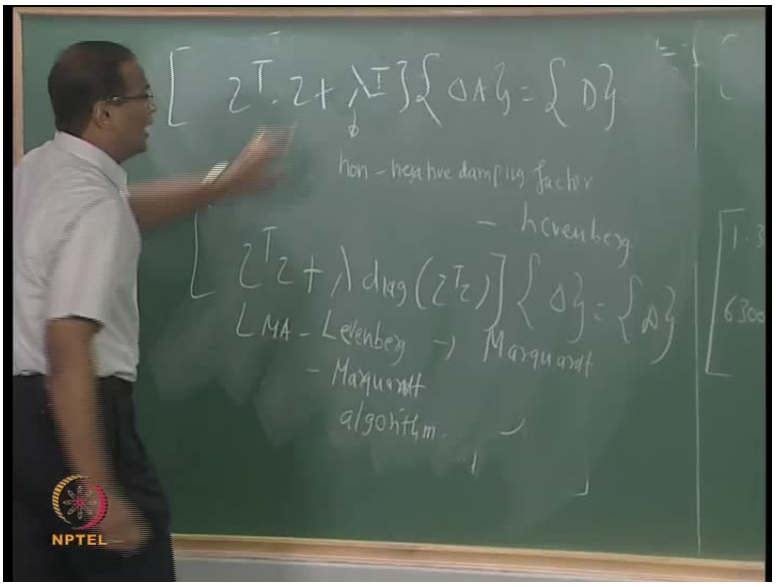
Student: 3

Look at the power of the Gauss-Newton algorithm. When theta was 3.1, the theta fit was 1.95 – first iteration, 0-th iteration. Immediately at first iteration, from 1.95, it became 3.27, which is close to this. From 7.25 it went to 12.1; the exact value is 11.9. From 13 it went to 21; exact value is 21. From 20, it went to 30.2; the exact value is 29.9. From 25.3, it went to 35.9; exact is 37.3; from 31, it went to 41; exact value is 42.7. Just one iteration of the Gauss-Newton algorithm. If you do the next iteration, you will get the correct answer. So, there will be a significant improvement in the residuals, if the problem is well-conditioned and all that. Now, this is fine? Now, can you work out the theta minus theta fit whole square? Please take it as an exercise. I will close the discussion with just one or two points.

What are the difficulties with the Gauss-Newton algorithm? Sometimes it is not possible to calculate dou f by dou a and dou f by dou b. It is very difficult to calculate, because analytically, it will be a very complicated expression. In those cases, you will

numerically evaluate dou f by dou a and dou f by dou b. You can do numerical differentiation – point 1. Point 2 – if the initial case is not chosen properly or if this system is not properly conditioned, the GNA may diverge, may not converge or may converge very slowly. So, additional modifications are done.

(Refer Slide Time: 51:33)



One possibility is to have Z transpose T Z plus lambda I into delta A is equal to D. If you do not have this, that is the Gauss-Newton algorithm. Z transport T Z; you add a lambda into I; where, I is the identity matrix. So, this is the non-negative damping factor. So, this is the contribution by Levenberg. He said that, it will condition the matrix properly and it will lead to better results, if you do this. Additional improvement was; instead of using the… use only the diagonal elements of Z transport T Z. For all the other elements use 0 0 0. This was the modification proposed by Marquardt. Since he borrowed the original idea from Levenberg, there should be a correction to the Gauss-Newton algorithm. But, instead of identity matrix, use the diagonal of Z transport T Z. This is called as the LMA – the Levenberg-Marquardt algorithm. This is one of the most powerful and most widely used in regression. It is a very very powerful optimization tool – Levenberg-Marquardt algorithm. Lambda is a non-negative damping factor.

What is the value of lambda to be chosen? That comes only from experience. Does it mean hand waving? I am not explaining properly? No. So, what you have to do is you have to start with one value of lambda and find out the result. And, take lambda by

alpha, some other factor. Take lambda 1 and lambda 2 and see how it progresses. If the convergence is very fast, then you can cool down lambda and lambda can be reduced. If the convergence is very slow, you can put higher values of lambda. Are you getting the point? So, if lambda is equal to 0, it is the… If lambda is equal to 0, Gauss-Newton algorithm. If the lambda is very high, it approaches the steepest descent or the steepest ascent. Do not get worried; steepest descent and ascent – we will discuss when we go to optimization. So, quietly, from linear least squares, we come to stage, where I have discussed one of the most powerful optimization algorithm. This can be used for any minimization. We used it for least square minimization, but it can be used for any minimization.

Now, if it is a simple non-linear problem, two parameter problem, it should be possible for you to do in the exam or in life. You should be able to write a program or do hand calculations, if it is a few data points and get the new values of a and b from old values of a and b till it approaches convergence.

Thank you.