

**Advanced Numerical Analysis**  
**Prof. Sachin Patwardhan**  
**Department of Chemical Engineering**  
**Indian Institute of Technology - Bombay**

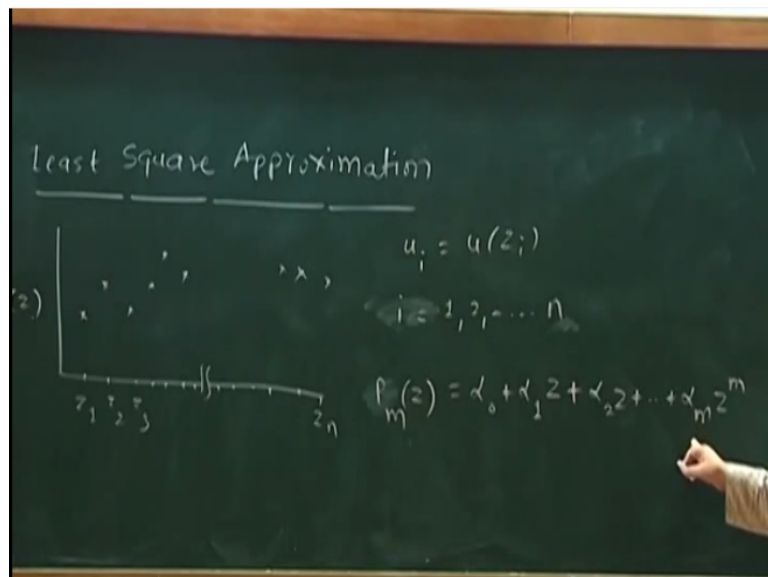
**Lecture - 17**

**Least Square Approximations, Necessary and Sufficient Conditions for Unconstrained Optimization**

So now we want to look at the problem of approximating a function using least square approximation or optimization. So now this is as I said in my last lecture this is different from what we have looked till now. We have been looking at interpolation and interpolation you wanted the approximate polynomial ultimately to pass through every point.

Now the difference is that I want a lower order polynomial, which may not pass through every point but which is the best fit in some sets okay. So let me first formulate the problem and then we will move on to necessary and sufficient conditions for optimality and then we will move back to solution. So before we need to do some extra work in between before we actually solve the problem.

**(Refer Slide Time: 01:28)**



So the problem at hand is now why least squares? Why not something else? And we have been using least squares probably in your under graduate program in exploited methods or fitting curves and fitting lines and so on. Why we use least squares? Why do not we use some of absolute error and why do not we use 2-norms has some special.

**“Professor - student conversation starts.”** After teaching you about norms you are telling me only 2 norms define distance. No. So 2-norm also defines angle I mean not 2-norm defines angle the definition angle comes free with the 2-norm right so you buy 2-norm you get angle free okay **“Professor - student conversation ends.”** That is the advantage, you do not get orthogonality, you do not get all those definitions when you use 1-norm, infinite norm.

One could formulate a problem instead of least square fitting, fitting a function in the absolute norm or fitting a function in the infinite norm one can very much do that, but 2-norm has something special. Now we are going to see why it is special but before that we have to do some work. So let me again restate the problem. My problem was that I have these points here.

They need not be equal space, there are some points here where I have this function  $u$  and I know values of this function at different points. So this is a function defined over some domain you can have it 0 to 1 does not matter or it can be from some  $A$  to  $B$ . I know values of this function at different locations. So this is the dependent variable  $u_i = u$  at independent variable  $z_i$  where  $i$  goes from 1, 2, up to  $n$  okay,  $i$  goes from 1 to  $n$ .

Now the main difference is that I want to fit now a polynomial, which is not of  $n$ th order or  $n-1$ th order. I want to fit a polynomial, which is of a lower order okay. So typical problem is that I want to fit a polynomial say  $P_m(z)$ . I want to fit a polynomial  $P_m(z)$ , which is  $\alpha_0 + \alpha_1 z + \alpha_2 z^2$  up to  $\alpha_m z^m$  to power  $m$  okay. I want to fit a polynomial of this form okay.

Now I cannot say earlier in interpolation, we said that the polynomial value or the polynomial should pass through every point I cannot say that now okay.

**(Refer Slide Time: 05:05)**

$$\begin{aligned}
 e_i &= u_i - P_m(z_i) \\
 &= u(z_i) - P_m(z_i) \\
 i &= 1, 2, \dots, n
 \end{aligned}$$
  

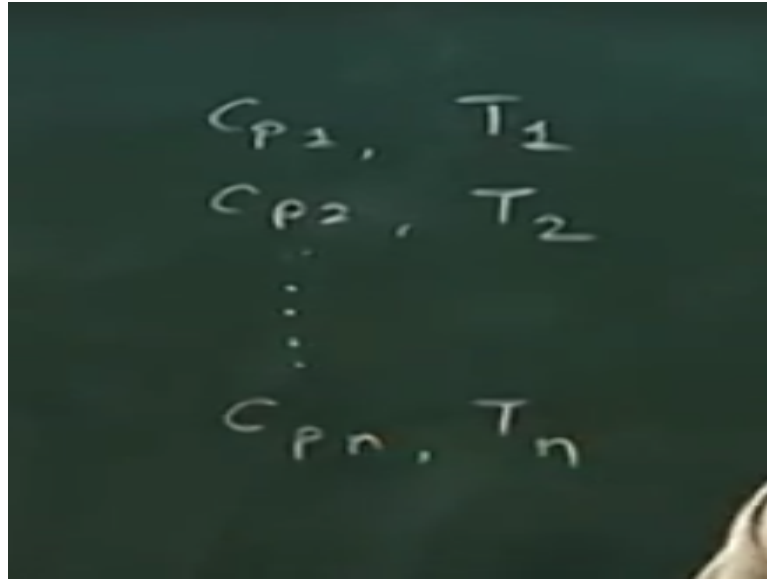
$$\begin{aligned}
 &\dots + \alpha_m z^m \\
 C_p &= A + BT + CT^2 \\
 &= A + BT + CT^2 + DT^3
 \end{aligned}$$

What I am going to say now is that this error I am going to define error, which is  $u_i$  that is  $u_i - P_m(z_i)$  okay which is nothing but  $u_i - P_m(z_i)$  okay. This error has to be small in some sense, but this error is not just at 1 point, you have error defined at  $i$  goes from 1, 2 up to  $n$  okay. Now you have vector of errors not 1 error right. You have  $n$  points, this  $n$  could be large number of points,  $n$  could be 100,  $n$  could be 1000 okay.

I have large number of data points, so I know the function value at large number of points but I want to fit a polynomial of order 2 or order 3 okay. Classic example from chemical engineering, I would give here is  $C_p$  as a function of temperature we fit  $C_p$  as some  $A + BT + CT^2$  square okay or sometimes we fit  $A + BT + CT^2 + DT^3$  cube.

There is no unique way of choosing depending upon what range of  $C_p$  values and what range of temperature you are considering okay. What the choice of polynomial would differ okay.

**(Refer Slide Time: 07:02)**



You may have 100 values of  $C_p$  okay so I may have you know  $C_{p1}$  at temperature 1,  $C_{p2}$  at temperature 2 and so on so  $C_{pn}$  at temperature  $n$ . So these are different measurements points, I know  $C_p$  at large number of temperatures okay. So if I actually try to fit into this, I will have problem okay. The problem is that none of this you know there are more number of equations than the unknowns okay.

So now we have to do something about this okay. "Professor - student conversation starts." Here  $i$  goes from 1 to  $n$  not 1 to  $m$ . The number of coefficients here are  $m+1$  but  $i$  goes from 1 to  $n$  okay. "Professor - student conversation ends." So now the trouble is that this when you write this equation, it is not exact okay that is actually one more term missing here. So let us concentrate right now on the quadratic equation, cubic equation will worry about later. Look at this quadratic equation.

What I have to say that  $C_p = \text{this} + \text{an error}$ . This is not exact representation. This is an approximation. This  $e$  is the approximation error okay. Why am I allowed to fit a polynomial? Because (()) (08:55) theorem applies. Continuous function can be approximated by a polynomial function that is why I am fitting a polynomial okay, but this is not exact so there is an error here.

**(Refer Slide Time: 09:15)**

$$N_{eq} = n \quad \text{Variables} = n+3$$

$$C_{p1} = A + BT_1 + CT_1^2 + e_1$$

$$C_{p2} = A + BT_2 + CT_2^2 + e_2$$

$$C_{p3} = A + BT_3 + CT_3^2 + e_3$$

$$\vdots$$

$$C_{pn} = A + BT_n + CT_n^2 + e_n$$

So actually if I take this data points and I start writing  $C_{p1} = A + BT_1 + CT_1^2 + e_1$  right. So there is some, this is not exact. This will give you most of the variation but this is not exactly equal to this there is some error here. So  $C_{p2} = A + BT_2 + CT_2^2 + e_2$ ,  $C_{p3} = A + BT_3 + CT_3^2 + e_3$  and so on. I can write these equations up to  $C_{pn} = A + BT_n + CT_n^2 + e_n$ . So these are my equations now. Can I solve these equations? Is there a problem here? How many variables and how many equations?

How many equations I have right now? I have  $n$  equations. How many variables I have? So  $N_{equations} = \text{number of equation} = n = \text{number of data points}$ . How many variables?  $n+3$  right so  $\text{variables} = n+3$ . What are the  $n$  variables?  $e_1, e_2$  okay and plus  $A, B$  and  $C$  so there are 3 unknowns  $A, B, C$  plus there are unknowns these are the errors, which are unknown okay. All these errors are unknown.

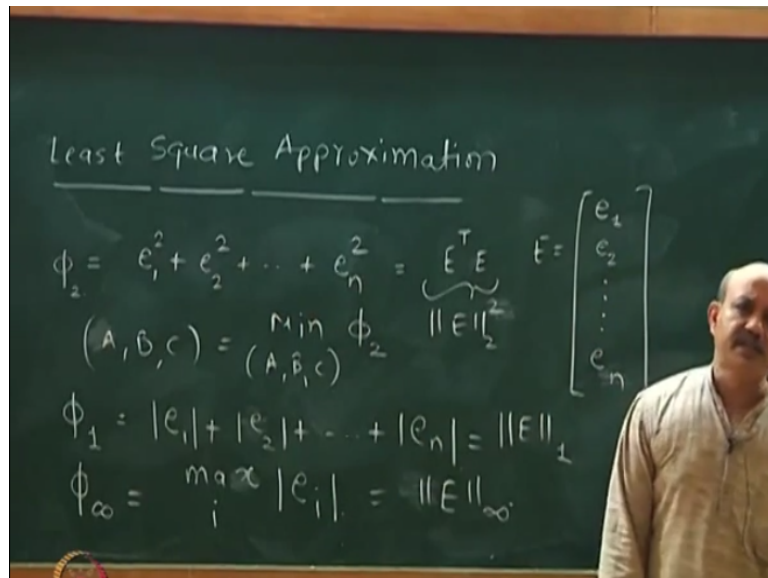
Now the trouble is how do I solve this? There are infinite solutions to this problem why? If I choose  $A, B, C$  in one particular way okay if I choose  $A, B, C$  in one particular way if I fix  $A, B, C$  by some means okay. Then I will get 1 value for  $e_1, e_2, e_3$  because once I specify  $A, B, C$  I will have  $n$  equations in unknowns I can solve them okay. The trouble is how do I fix  $A, B$  and  $C$ ? Differentiate why?

Sigma of all the errors not a great idea, is it a norm? Sigma of errors is it a norm? So it might happen that all the errors might sum up to 0 so that means is that the solution? Why only square? Why not absolute? So I could formulate my problem. Now I need to generate 3 more

equations. To complete the problem, I have  $n+3$  variables, I have  $n$  equations. To make the problem exact, I need to generate 3 more equations.

What are these equations? Okay so now let us look at how do I formulate the problem? I can formulate this problem in multiple ways. I can formulate this problem, I can formulate an index.

**(Refer Slide Time: 13:00)**



So I can say an index phi which is  $e_1^2 + e_2^2 + \dots + e_n^2$  let me call this 2 index  $e_n$  square okay, which I can write this as  $E^T E$  okay where this  $E$  is a vector of  $e_1, e_2, \dots, e_n$  okay. So I can propose that find  $A, B, C$  such that you minimize this, I minimize this phi 2 with respect to  $A, B, C$ . Some of the square of errors is as small as possible okay. So I try to get a polynomial such that sum of the square of errors.

See this is individual error in each equation. I am just squaring it and summing it up okay. I want to find out that value which gives me minimum sum of the square of errors okay. So this is one way of formulating the problem. I still do not know whether doing this is going to give me 3 additional equations. I have to generate them so there is another way of formulating the problem.

This is not the only way. I could say that you know I formulate this phi 1, I will say that you know  $|e_1| + |e_2| + \dots + |e_n|$  somebody might say that why this sum of the square of errors? Why not sum of the absolute errors? Nobody stops you from doing that. So this will be nothing but

1-norm of  $E$ . What is this? 2-norm of  $E$  square right. If  $E$  is the vector, it is 2-norm of  $E$  square, this is 1-norm of  $E$  vector okay.

So I want to minimize instead of minimizing this  $\phi_2$ , I could post the problem as minimize  $\phi_1$  right and somebody else might say well I do not believe in this, I would like to minimize  $\phi_\infty$ , which is  $\max_i$ , minimize the maximum error, minimize the maximum deviation okay. So instead of minimizing  $\phi_2$ , I could choose to minimize  $\phi_1$ , I could choose to minimize  $\phi_\infty$ .

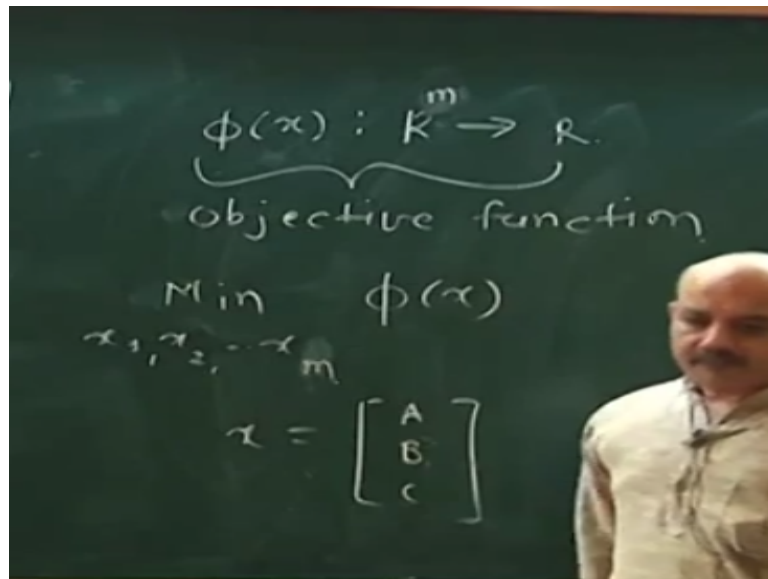
This is infinite norm of error vector. This is nothing but anyone of them is fine okay. The question is how do I solve this problem? Now first of all, you should notice that this is an optimization problem, you want to minimize sum of the square of errors with respect to these parameters okay. This is something different and what you have done in your undergraduate optimization or minimization.

In undergraduate course, you normally study most of you may be some of you have done advanced things but most of you study maximizing or minimizing a function of 1 variable. Here you have a function of 3 variables okay. So we need to generalize what you have studied in undergraduate. How do I come up with minimization of a function, which is multi-dimensional okay which is multi-dimensional?

So what is this function? I will generalize this. I generalize this problem here okay. Now let us push this to the background and will come back to this  $C_p$  versus temperature business little later. Let us look at an abstract problem now. My abstract problem is I want to minimize an objective function. Why just look at 3 variables? I will look at general  $n$  variables okay. I have a scalar function.

What is this function  $\phi_1$ ,  $\phi_2$ ,  $\phi_\infty$ ? What are these functions? These are scalar functions. Norm is a scalar function right. You always get but I need not always define an objective, which is coming through norm. There could be other ways in some other problem. So in general I am worried about minimizing a function.

**(Refer Slide Time: 18:19)**



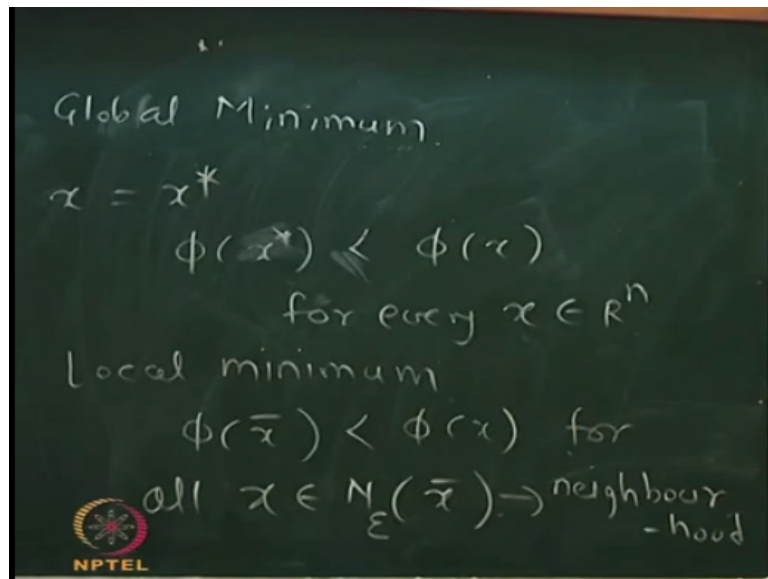
Let us say  $\phi(x)$  from  $\mathbb{R}^m$  to  $\mathbb{R}$ .  $\phi(x)$  is the function from  $\mathbb{R}^m$  to  $\mathbb{R}$  from  $n$  dimension to 1 dimension. So this is my objective function, some objective of a vector  $x$  so  $x$  is the vector which is  $n$  dimensional. In this case,  $x$  is nothing but an error vector. I am now generalizing; I am abstracting the problem. I no longer want to just work with modeling errors. I want to just go and generalize this and say in general actually this is a special case of a problem in which I have a scalar function mapping from  $n$  dimension to 1 dimension.

I want to minimize and I want to find out minimize or maximize, minimize with respect to  $x_1, x_2$ , okay let us take here  $\mathbb{R}^m$  to  $\mathbb{R}$ . I want to minimize this objective function  $\phi(x)$  with respect to  $x_1, x_2, x_m$  okay. There is 1 problem, now do not confuse this  $x$  with this  $n$  okay. If I said that earlier there was a small error. What is  $x$  here? When I am generalizing from here to here what is  $x$ ? Is the unknown parameters. What is  $x$ ?  $A, B, C$  okay.

So in this case  $m$  would be 3 in this particular case  $m$  would be 3. I want to minimize this function  $\phi(x)$  okay with respect to  $x_1, x_2, x_3, x_m$ . In this case in this particular  $C_p$  problem,  $x$  would be nothing but  $A, B$  and  $C$ . I want to minimize some function  $\phi(x)$  okay be it 1-norm, infinite norm, 2-norm whatever. I want to minimize some function for  $x$  with respect to  $x$  where in this particular case  $x$  happens to be  $A, B, C$  okay.

**(Refer Slide Time: 20:54)**





How to solve this problem? I need some background definitions here. So well why I am just talking about minimization problem? Because if you have a maximization problem, if you want to maximize phi of x, it is same as minimizing -phi of x. So I can just talk about minimization problem. So it includes maximization. We do not have to worry separately about maximization problem okay.

Now first concept that we need to know is the global minimum. A global minimum is a point say  $x=x^*$ ,  $x=x^*$  is called as a global minimum if  $\phi(x^*) < \phi(x)$ , you call this  $x^*$  to be a global minimum actually equality is not there. So you call it to be a global minimum okay.

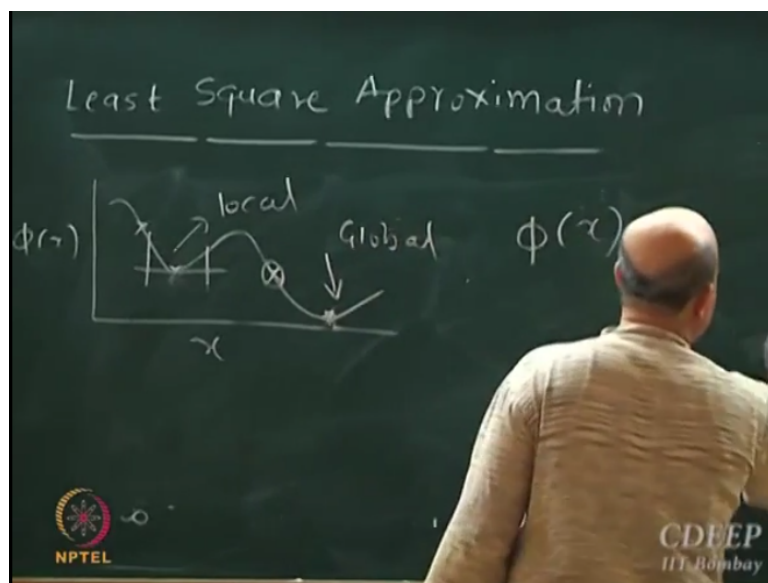
If you take any other value so what is the global minimum for A, B, C? There is some special value of A, B, C for which this particular objective function will assume the minimum value. There is no other value of A, B, C for which you will get a smaller sum of the square of errors okay. If that exist, if that you can reach then that is called as a global minimum okay. Now when you do optimization, you may not be able to reach a global minimum.

It may happen that you start doing a search and then you go to a local minimum okay. Imagine that you know you are standing on a mountain and then there are multiple valleys okay. You might reach a valley, which is not you know which is not the global minimum okay which is just a local minimum and then so we also have to worry about local minimum. So in local minimum, we do not talk about for every x in  $\mathbb{R}^n$  okay.

We talk about some neighborhood of a particular point okay. So in some neighborhood so some epsilon neighborhood. We have defined neighborhood earlier; we can think of a ball okay some ball around  $\bar{x}$  okay. If you can show that  $\phi(\bar{x}) \leq \phi(x)$  for any other  $x$  in that neighborhood, then  $\bar{x}$  will be called as a local minimum. Here  $x^*$  will be called global minimum.

Because anywhere you go in the space, you cannot get a value of  $x$  for which  $\phi$  will be smaller okay. The smallest value of  $\phi$  is obtained at a global minimum and a local minimum is like local minima okay so you know 1 dimension it is easier to draw.

**(Refer Slide Time: 24:05)**



So in a simple 1-dimensional case, this would be you know if you have a function  $\phi$  of  $x$  versus  $x$  so this is my global minimum but this is a local minimum okay. So this point for this function is a global minimum whereas this is a local minimum and this is global minimum. So in some neighborhood, this is the minimum. Take any point in this neighborhood okay you cannot get value of  $\phi <$  this.

But that is not a case if you look at much larger domain okay. So there might be some other value where you get much lower. The conditions that we are going to derive now okay are pertaining to the local minimum. We cannot actually derive general conditions for global minimum and in general optimization problem what we find is typically a local minimum.

If it happens to be global minimum in some cases great, I mean no you have achieved what you wanted to do but in many situations you do optimization by numerical search and then

the solution depends upon your initial guess. If you are far away from the global minimum okay you may not reach, there. So when you do trial error approach you know arrive at the solution you might be starting with a guess here and the solution might hit this minimum.

And you know your numerical method will declare that you have reached a minimum and if you happen to give an initial guess somewhere here, may be you will reach this point okay. Numerical methods you cannot predict you might even go here and get into this, so hard to predict what will happen but likely that if you are here you might reach this global minimum. The conditions that we are going to derive are for local minimum.

What are the necessary conditions for a point to be a local minimum? What is the sufficient condition for a point to be a local minimum? That is what I am going to derive now okay. Now to qualify a point to be a local or a global minimum, I need some more definitions. See what was the key thing when you talked about minimum of a single variable function? First was that a derivative at this point of  $\phi$  with respect to  $x$  become 0.

There is no change locally okay. So the derivative is tangent, which is parallel to  $x$  axis okay. That is the key thing about so the derivative is 0 that was the first thing. How do I extend this to the multi-dimensional case? Now I have a variable, now I have objective function, which is function of  $m$  unknowns, not just 1 unknown. So I should derive an equivalent condition. Second thing is so well done I have to make an assumption that this  $\phi$  is differentiable okay.

One norm and infinite norm, there is a trouble, they are not differentiable functions. Two-norm is a differentiable function,  $X^T X$  sum of the squares differentiable function okay. Second thing is what qualified this point to be a minimum and this point to be a maximum? Second derivative. So second derivative when it was positive it was minimum. Second derivative when it was negative it was a maximum.

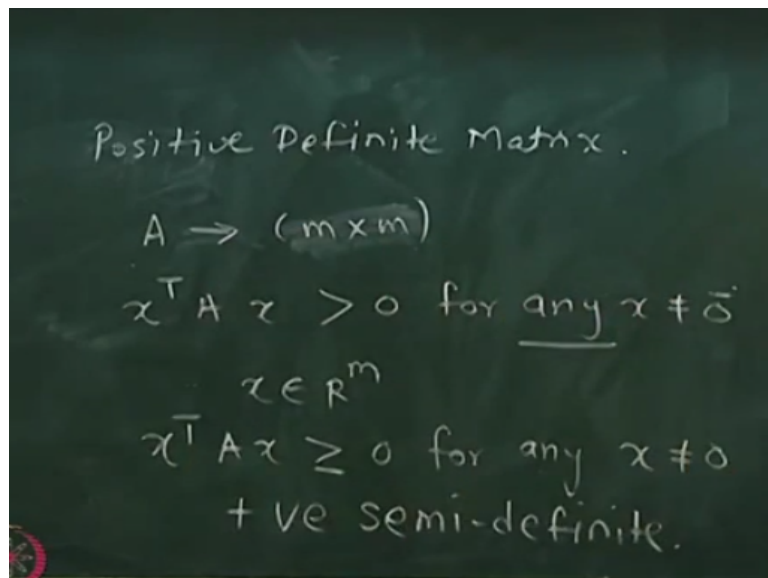
Now in this particular case, if  $\phi(x)$  is a differentiable or if it is twice differentiable function then the first derivative of this with respect to  $x$  is going to be a vector and the second derivative of this with respect to  $x$  is going to be a matrix. So I need some more additional structure, which will help me qualify a point to be maximum or minimum okay. Now this is where you need to define matrices, which are certain special properties.

So what is this special property? Definiteness. I now need to define matrices, which are positive definite or negative definite okay. So now before I proceed and define a positive definite matrix and a negative definite matrix and an indefinite matrix. If you understand this geometric connections of positive definiteness, indefiniteness and then you know it will make much more sense later when you use these concepts okay.

Otherwise many times you know in a course on linear algebra, they are introduced not by connecting it to geometry, they are just introduced by saying a positive definite matrix is 1 which has all Eigen values positive but why? Why do I need this animal which has all positive Eigen values? Okay it is not clear to us. This is where it will become clear. Why do I need such funny matrix?

Well it is not really funny. It happens to be very nice matrix. It helps us in many, many ways and it is going to help us throughout the course in many ways okay.

**(Refer Slide Time: 29:35)**



Positive Definite Matrix.

$$A \rightarrow (m \times m)$$
$$x^T A x > 0 \text{ for any } x \neq 0$$
$$x \in \mathbb{R}^m$$
$$x^T A x \geq 0 \text{ for any } x \neq 0$$

+ve semi-definite.

So we have this first definition is positive definite matrix. So when is a matrix positive definite? Right now I am just digressing from the main theme, I am just going into little bit of linear algebra may appear disconnected. So if I have a matrix A, this A is a m cross m matrix, we are talking about real valued matrices right now. So if  $x^T A x > 0$  for any  $x \neq 0$  okay.

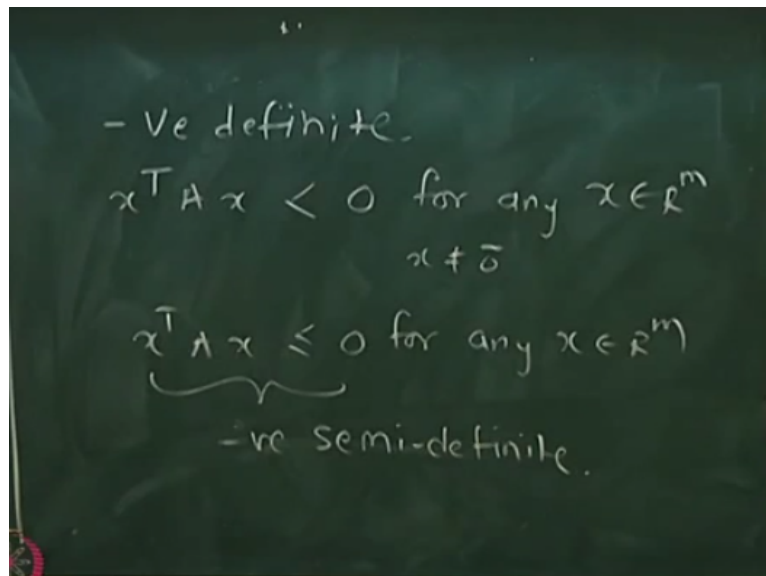
And  $x$  belongs to m dimensional space. This should be m cross m because we are talking about  $\phi(x)$  where  $x$  is m dimensional so m cross m, very, very important most important

word here is for any  $x$  for any  $x$  even if you find 1  $x$  for which this is  $=0$  and or  $<0$ , the matrix is not positive definite. So this is the fundamental definition of positive definiteness okay. Give me any vector in a space okay.

$x^T A x$  is  $>0$  when  $x$  is  $\neq 0$  when will this be  $=0$ ? Only when  $x=0$  okay. Remember that. Well there is one more matrix that you probably have studied in your undergraduate is positive semi-definite matrix. When do you call positive semi-definite matrix? So if this condition becomes  $x^T A x$  is  $\geq 0$  for any  $x$  so there are some vectors  $x$  for which this will be  $= 0$  this will happen when  $A$  matrix is singular, think about it.

When  $A$  matrix is singular, its columns are linearly dependent, null space is not  $0$  and you will get some well this basic definition translates to Eigen values being positive, will look at it a little later, but in this case if you just change from this strict inequality to you know this  $\geq 0$  it becomes positive semi-definite,  $A$  will be positive semi-definite okay. What about negative definite matrix? So this is positive semi-definite okay.

**(Refer Slide Time: 32:34)**



So negative definite is  $x^T A x$  is strictly  $<0$  for any  $x$  that belongs to  $\mathbb{R}^m$  okay and  $x$  is  $\neq 0$  so non-zero vector and any non-zero vector you give me,  $x^T A x$  will be  $<0$  and it is called as a negative definite matrix okay and a fourth one is of course negative semi-definite so this is  $x^T A x \leq 0$  for any  $x$  belonging to  $\mathbb{R}^m$ . So this is negative semi-definiteness.

This is negative semi-definite okay so this is just the background work that I need to proceed further okay. Well the necessary condition for optimality is something that I would like to quickly derive in the class. Though every step I am not going to write because it is there in the notes. I will just give you outline of the proof, how it is done. So the necessary condition is given by this theorem.

**(Refer Slide Time: 34:00)**

Necessary Condition for Optimality

$\phi(x) \rightarrow$  twice differentiable

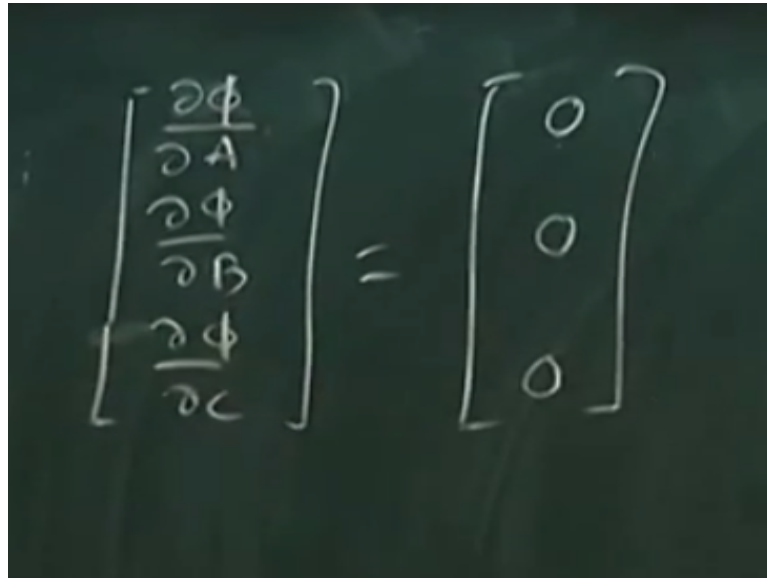
$x = \bar{x} \quad \nabla_x \phi = \begin{bmatrix} \frac{\partial \phi}{\partial x_1} \\ \frac{\partial \phi}{\partial x_2} \\ \vdots \\ \frac{\partial \phi}{\partial x_n} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

So necessary condition for optimality well first of all we assume that phi x is twice differentiable that is what we assume first okay. Then only we can proceed. So if you want to know where this appears in the notes. I do not know how many of you are carrying notes, it is on page 80. It is in the appendix section 8.2. Now to do arguments about you know local optimality I am going to use Taylor series approximation okay.

Taylor series approximation is the bulwark. You know is the one of main tools that we use to prove many, many things. So for a point the statement of the theorem, former statement is given here. I am just stating the main result. So that is if  $x = \bar{x}$  is to qualify as a minimum or maximum or optimum. We do not know what it is. To be precise it a stationary point okay it could be a minimum or a maximum or it could be neither of them.

It depends upon some more conditions. So the gradient of phi with respect to x that is  $\frac{\partial \phi}{\partial x_1} \frac{\partial \phi}{\partial x_2}$ . This should be =, if phi is the twice differentiable function okay what we can prove is that a necessary condition for optimality is that the first derivative of phi with respect to each of the variables not  $x_n$  we are working with  $x_m$  just remind me. We are working in m dimensional space okay with respect to  $x_m$  should be =0.

**(Refer Slide Time: 36:40)**


$$\begin{bmatrix} \frac{\partial \phi}{\partial A} \\ \frac{\partial \phi}{\partial B} \\ \frac{\partial \phi}{\partial C} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

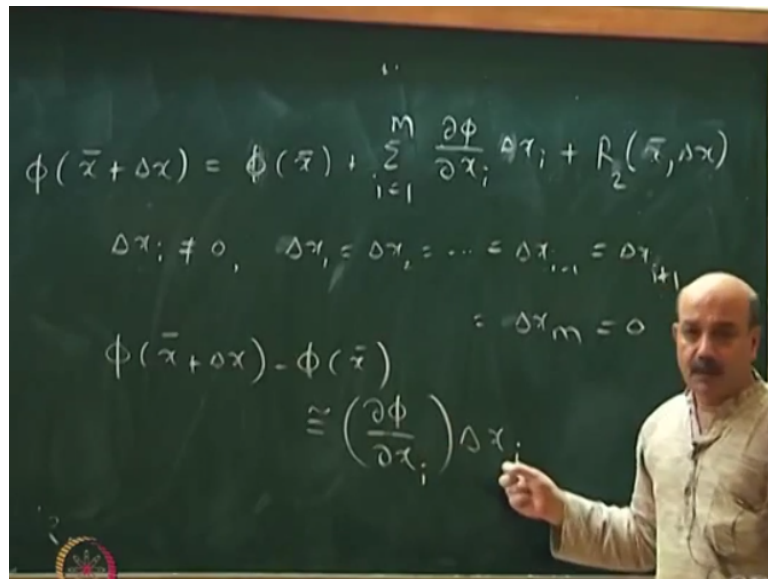
In the specific problem for  $C_p$  okay, it will be what is the specific condition for  $C_p$  value problem? So it will be  $\text{d}\phi/\text{d}A$ , in a specific problem for  $C_p$  it will be  $\text{d}\phi/\text{d}A$ ,  $\text{d}\phi/\text{d}B$  and  $\text{d}\phi/\text{d}C$ . This should be  $=0, 0, 0$  okay. So this particular equation actually is the necessary condition for optimality. Well I started by saying for the  $C_p$  problem we need 3 extra equations.

These are the 3 extra equations,  $\text{d}\phi/\text{d}A=0$ ,  $\text{d}\phi/\text{d}B=0$ ,  $\text{d}\phi/\text{d}C=0$ . In general, if there are  $m$  such parameters with respect to which we want to optimize then we have  $m$  equations here right. We have  $m$  equations coming from the first order derivative to be exactly  $=0$  at the optimum point okay.

Now the proof of this, so the proof of this actually goes by I think contradiction what we do is we assume that you are allowed to vary a particular variable only in 1 direction keeping all the other values constant so I just want to perturb say along  $x_1$  or  $x_2$  okay and then we assume that you know that this condition does not hold okay. We assume that this condition does not hold but the point is a minimum okay.

And what you see is that if you make an assumption that a derivative is not 0 and the point is a minimum will lead to a contradiction okay. These 2 cannot be true okay. So the way it is done is you know you write the full proof you should read here in the notes.

**(Refer Slide Time: 38:53)**



But the way this is done is that I can write  $\phi(\bar{x} + \Delta x)$  so I take a small perturbation let us say  $\bar{x}$  is the minimum let us say  $\bar{x}$  is the local minimum okay. I can write this as  $f(\bar{x})$  using Taylor series expansion okay  $\sum_{i=1}^m \frac{\partial \phi}{\partial x_i} \Delta x_i + R_2(\bar{x}, \Delta x)$  the residual term at  $\bar{x} + \Delta x$ . This is the second order residual term okay. Now if I say that the perturbation is only along 1 direction.

Let us say we will put this as small  $\Delta x$  here  $\Delta x_i$  so let us assume that only  $\Delta x_i$  is not equal to 0 but  $\Delta x_1 = \Delta x_2 = \Delta x_{i-1} = \Delta x_{i+1} = \Delta x_m$  all are 0 but only one of them are not 0. Let us make that assumption okay. All other  $\Delta x$ 's are 0. I am choosing perturb only from  $\bar{x}$  I am choosing to perturb only  $x_i$  variable. You have done this kind of thing have you?

If you have programmed numerical Jacobian, you are keeping all the variables constant just perturbing one value right. We did this in the programming okay the same thing. I am just perturbing one variable at a time okay. So the question is  $\phi(\bar{x} + \Delta x) - \phi(\bar{x})$  okay so this is dominated by  $\frac{\partial \phi}{\partial x_i}$ . So this difference is dominated by this partial derivative times.

If you take  $\Delta x_i$  to be very, very small the second order term here will be insignificant. This difference is dominated by okay. Now tell me what should happen if  $\bar{x}$  is the minimum and if I move away from it? What should happen to objective function? It should increase. So what should happen if this? This should be always  $> 0$ , any moment I make okay. Now let us take the situation that this gradient is not 0 okay.



Let us take the situation this gradient is negative okay. I can choose a  $\Delta x$  which is small negative value such that this will become multiplication will become positive. If multiplication becomes positive, it contradicts the fact that  $x$  bar is the minimum. See if I can make this multiplication positive, it will contradict the fact that this is the minimum. So because if you move away what should happen?

This should always remain see whichever way I go from the minimum just imagine if you are in a valley okay whichever way you go from the minimum point, you know your height increases, it will never decrease okay. Now if the local gradient let say this is positive, this  $\phi$  is positive then I can choose a  $\Delta x$  which is positive  $\Delta x$  and make this positive, which means that this—this is positive which means  $x$  bar is not a minimum.

No, you have to argue like that. Suppose this is negative then I can choose  $\Delta x$  to be negative so multiplication will become positive, which means this difference will become positive, which means  $x$  bar cannot be minimum you know so you have to argue in a way that well one minute I think the argument I have to repeat. So this if I move away from  $x$  bar, the value should increase okay.

Now if this is negative, I can choose  $\Delta x$  positive okay. If I choose  $\Delta x$  positive, if this is negative, if this positive multiplication is negative, which means I can move in one particular way and reduce this further which means  $x$  bar is not the minimum. I made a wrong argument earlier. Now argue other way if this is positive, I can choose  $\Delta x$  negative. If I choose  $\Delta x$  negative, this will become negative okay.

So I can further reduce by moving away so then  $x$  bar is not a minimum so you can show this that for each variable you can argue like this. So only way this point can be a minimum is if this derivative is 0 okay because if derivative is non-zero you would be able to move a little bit and go further down, which cannot happen if it is a minimum okay. So only way this point can be minimum is this derivative is 0 okay.

So this is the necessary condition, you look at the proof here. To derive the sufficient condition, what we do is will look at the second derivative.

**(Refer Slide Time: 45:11)**

$$\phi(x) \rightarrow \text{twice differentiable}$$

$$x = \bar{x} \quad \nabla_x^2 \phi = \begin{bmatrix} \frac{\partial^2 \phi}{\partial x_1^2} & \frac{\partial^2 \phi}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 \phi}{\partial x_1 \partial x_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \phi}{\partial x_m \partial x_1} & \dots & \dots & \frac{\partial^2 \phi}{\partial x_m^2} \end{bmatrix}$$

The second derivative here is del square phi so del square phi would be the so called Hessian matrix. The Hessian matrix is given by so Hessian in this case will be a m cross m matrix, I keep writing n, it is m we are working in m dimensional space so this is m here okay, yeah so this is m here so we have to look at the Hessian. Why we have to look at the Hessian okay? Let us go back to this equation here.

We said that only way  $\bar{x}$  is the minimum is all derivatives are 0. If all derivatives are 0, this term vanishes right. When I use Taylor series expansion, all derivatives are 0 at  $\bar{x}$  so the first derivative term will vanish okay. Then, we have to look at the second derivative part okay so then I will write this here as okay.

**(Refer Slide Time: 46:20)**

$$\phi(\bar{x} + \Delta x) \cong \phi(\bar{x}) + \frac{1}{2} (\Delta x)^T \underbrace{\left[ \nabla_x^2 \phi \right]}_{\text{Hessian}} \Delta x$$

$\nabla_x^2 \phi(\bar{x}) \rightarrow +ve \text{ definite}$   
 $\bar{x} = \bar{x} \text{ is a minimum}$   
 $\nabla_x^2 \phi(\bar{x}) \rightarrow -ve \text{ definite}$   
 $\bar{x} \text{ is maximum}$

So at the optimum the first derivative is 0, so this is governed by  $\phi(\bar{x} + \Delta x) \approx \phi(\bar{x}) + \Delta x^T \nabla \phi(\bar{x}) + \frac{1}{2} \Delta x^T \nabla^2 \phi(\bar{x}) \Delta x$ . What will govern the local behavior of the function in the neighborhood of the optimum point? See the first derivative is 0 so look at the second derivative. Second derivative is  $\Delta x^T \nabla^2 \phi(\bar{x}) \Delta x$ . This Hessian matrix, which is the Hessian\*okay.

What should happen if you move away from  $\bar{x}$ ? It should increase okay. When will it increase? If this particular matrix is computed at  $x = \bar{x}$ , so we are computing this at  $x = \bar{x}$  okay. The first derivative at  $x = \bar{x}$  is 0. I look at the second derivative to get an idea about what is the local behavior.

So if second derivative this if this Hessian is positive definite or positive semi-definite okay, I can move away from the point but the objective function value will not decrease okay. It will only increase, such a point will be a minimum point. So if  $\Delta x^T \nabla^2 \phi(\bar{x}) \Delta x$  computed at  $\bar{x}$  if this is positive definite or positive semi-definite then  $x = \bar{x}$  is a minimum okay. What is this matrix is indefinite?

What is the meaning of indefiniteness? For some  $x$ , this is positive, for some  $x$  this is negative okay. So in some directions, the function will decrease. In some directions, the function will increase. You know saddle point, the derivative is 0 but in some direction function decreases, in some direction function increases okay. If this matrix is indefinite okay then I cannot say anything about this point, this is not a maximum nor minimum okay.

So if this happens to be positive definite, it is a minimum so this matrix happens to be negative definite it is a maximum okay. If this Hessian matrix is negative semi-definite or negative definite, then it is a maximum otherwise if it is positive definite or positive semi-definite it is a minimum and if a Hessian matrix is neither positive definite nor negative definite, it is an indefinite matrix okay.

So sometimes this multiplication is positive, sometimes this is negative then you know the point is neither a minimum nor a maximum it is a saddle point okay. Now this particular these 2 are sufficient conditions for to qualify a point to be optimum. So first thing is that gradient should be 0, second thing is Hessian should be either positive definite or a minimum or it should be negative definite for a maximum.

And then we can qualify a point to be an optimum point. So this is generalization of the result that you know from one dimension. This now in the next lecture will apply to the specific problem of polynomial approximations okay.