

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

Lecture - 18

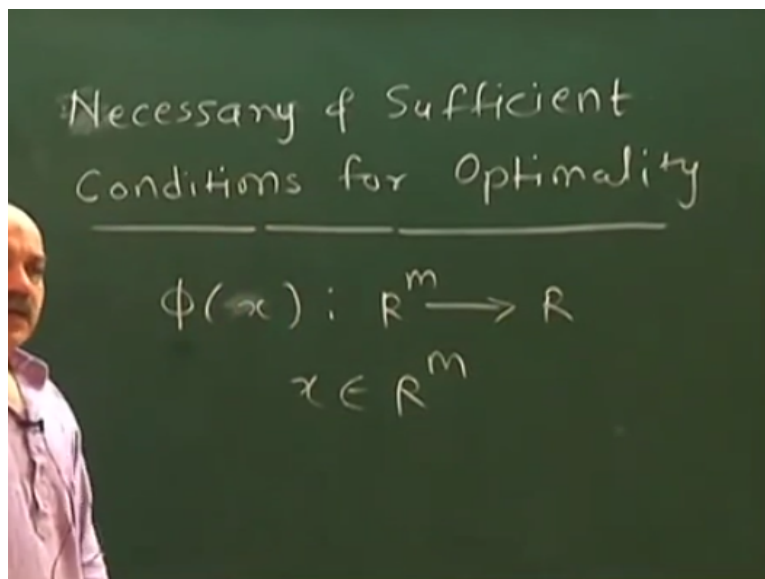
Least Square Approximations: Necessary and Sufficient Conditions for Unconstrained Optimization Least Square Approximations (contd.)

Criteria for deciding whether a point is maximum or minimum. How do you qualify a point to be maximum or minimum? The problem at hand was minimizing a scalar function from m dimensions to one-dimension \mathbb{R}^m to \mathbb{R} and we wanted to qualify which point in the state space qualifies to be an optimum point.

So we said the point at which the first derivative the gradient goes to 0, that point is the stationary point to be very precise. It could be a maximum, it could be a minimum, it could be a saddle point. We do not know okay. So to come up with you know further qualification, we have to look at the second derivative the Hessian matrix and if Hessian matrix is positive definite or positive semi-definite it uses way to qualify the point to be a minimum or a maximum.

So it should be strictly positive definite if it is to be a minimum, it should be strictly negative definite if it is to be a maximum okay. So let me just summarize what we looked at.

(Refer Slide Time: 01:51)



So we are looking at necessary and sufficient conditions for optimality. We have this function $\phi: \mathbb{R}^m \rightarrow \mathbb{R}$ and we said that this ϕ function is twice differentiable okay at any point. Right now this x belongs to \mathbb{R}^m m dimensional space okay.

(Refer Slide Time: 02:28)

Stationary Pt.
 $x = \bar{x}$
 $\nabla_x \phi(\bar{x}) = \begin{bmatrix} \frac{\partial \phi}{\partial x_1} \\ \frac{\partial \phi}{\partial x_2} \\ \vdots \\ \frac{\partial \phi}{\partial x_n} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$
 Computed at $x = \bar{x}$

And then we said that if a point $x = \bar{x}$, if this point $x = \bar{x}$ we want to see whether this is a stationary point. Then gradient of ϕ with respect to x at $x = \bar{x}$. So this is nothing but $\frac{d\phi}{dx_1} \frac{d\phi}{dx_2}$. This gradient vector should be $= 0$ and the way this necessary condition was proved by considering the fact that if you take it non-zero, it contradicts the fact that \bar{x} is the local minimum.

Well we are talking only about local minimum. This is the condition to be satisfied at a local minimum because all the arguments regarding a point to be local minimum or a local maximum. That is a stationary point were made using Taylor series expansion. Taylor series expansion holds only locally; it does not hold everywhere globally. So this necessary condition is a local condition, just remember that.

So first thing is given an objective function, first thing is to compute its gradient and set it $= 0$. If you can find such a point, then that qualifies to be a stationary point. A stationary point where you have gradient $= 0$. We further qualify so this \bar{x} is the stationary point and whether the stationary point further qualifies to be optimum.

That is maximum or minimum that will depend upon this matrix so by the way this, this, this derivative is computed at $x = \bar{x}$. This is computed at $x = \bar{x}$ okay.

(Refer Slide Time: 04:55)

$$\phi(\bar{x} + \Delta x) \approx \phi(\bar{x}) + \nabla_x \phi(\bar{x})^T \Delta x + \frac{1}{2} (\Delta x)^T \nabla_x^2 \phi(\bar{x}) \Delta x$$
$$\phi(\bar{x} + \Delta x) - \phi(\bar{x}) \approx \frac{1}{2} (\Delta x)^T \underbrace{\nabla_x^2 \phi(\bar{x})}_{\text{+ve definite}} \Delta x$$

Now ϕ of $\bar{x} + \Delta x$ we wrote this as you know ϕ of $\bar{x} + \text{grad } \phi$ at \bar{x} transpose $\Delta x + 1/2$. If we approximate this locally using Taylor series, it is Δx transpose $\Delta^2 \phi$ and now we have said that this is a stationary point. So this is 0 right and then locally we want to look at this difference, which is governed by locally in a small neighborhood of \bar{x} .

This difference is governed by this Δx transpose; this term governs the local behavior in the neighborhood of \bar{x} okay. Now if \bar{x} is a minimum what should happen? If I move away from the minimum what should happen? Value should increase so this difference should always be positive. If \bar{x} is a minimum any direction, I try to move away from \bar{x} I should have this difference to be positive.

That will be possible only when this matrix is positive definite. If this matrix is positive definite what is the definition of positive definiteness? x transpose Ax is always > 0 for any non-zero x . Anyway, any direction I try to move okay this will always be positive. That is positive definite matrix. If this is always positive irrespective of whichever direction, I try to move away from \bar{x} then \bar{x} is a minimum okay.

Just whatever you can visualize in 2 dimensions or 3 dimensions same thing holds in n dimensions. If you try to move away from that point, you just think of being in a valley any which way if it is the lowest point in the valley any which way you try to move you will only

increase height. You will not go further below okay if you are at a minimum. Then that is the simple logic okay.

How do you mathematically express this? Using positive definite matrices okay. Mathematical quantification of simple fact that you know that when you are in a valley at a lowest point in the valley, any direction you try to move okay height will increase, it cannot further decrease. Same thing is being said here. Just remember that simple thing and core that with this mathematical analysis.

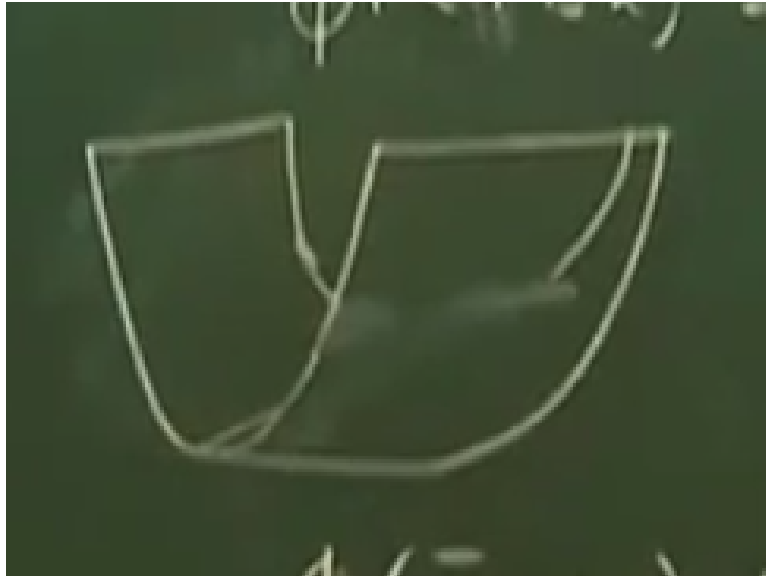
Then you will understand it better. Now other way round, now if \bar{x} is a maximum, you are at a peak what should happen? Anyway which you move now height will decrease, anyway which you move so what should happen is if I move make any movement Δx from \bar{x} this different should be negative okay. When will that happen for any Δx that is very important for any Δx . That will happen if this is negative definite.

If this Hessian matrix is negative definite, I am sure that any which way I move okay this is going to be negative so that means \bar{x} is a local peak. Of course, this local condition should also be satisfied at the global condition; however, this analysis does not tell us how to reach the global conditions. It just tells us given a point if the derivative is 0 how do we qualify it? Minimum or a maximum locally okay?

Now this condition can be checked for an objective function which is twice differentiable of course. You cannot check if the objective function is not differentiable or twice differentiable so that is very, very critical. So this is how what if your matrix is neither positive definite nor negative definite. Then the point is neither the minimum nor a maximum.

You know in terms of a hill, it could be you know something like this, like a step where gradient with respect is becoming 0 or saddle point and what you know in saddle point. So a multi-dimensional extension of a saddle point.

(Refer Slide Time: 10:28)



You know let us say you have a valley but a valley which is something like this. Well it is difficult to draw, I am not good at. Let us say there is no one unique point. There is no one unique point where you have you know you have multiple points where the gradient goes to 0 so you may have minimum respect to one variable not a minimum respect to other variable. So you may have a saddle condition where there is not a unique point.

So a saddle point is qualified by looking at property of this local matrix. The nice thing is that definiteness of this particular matrix will allow us to find out whether particular point is you know maximum, minimum or a saddle point. Okay now let us come back to our problem of fitting C_p versus temperature okay. I am going to generalize this set of equations and then going to solve it.

(Refer Slide Time: 11:47)

$$\begin{bmatrix} C_{p1} \\ C_{p2} \\ \vdots \\ C_{pn} \end{bmatrix} = \begin{bmatrix} 1 & T_1 & T_1^2 \\ 1 & T_2 & T_2^2 \\ \vdots & \vdots & \vdots \\ 1 & T_n & T_n^2 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \\ \theta \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

So now let me derive this problem which we had at hand. We had the C_p values, C_{p1} , C_{p2} , C_{pn} . I had this n values correct. I had n values. I had collected the C_p values at different temperatures say T_1 , T_2 , T_3 , T_n large number of right. I am just rewriting those equations in a matrix form. Earlier, I had just put them one below each other as single equations. I am now just rewriting them as one matrix equation.

That will help me solve the problem very easily. So now this is $1 \ T_1 \ T_1 \text{ square}$ $1 \ T_2 \ T_2 \text{ square}$. Just have a look at those earlier equations. I am just revisiting them, writing them in a different form, everyone with me on this. Same equations how many unknowns? $n+3$ and number of equations right now we have are only 3 okay. So first thing that we need to do is to define an objective function.

Well objective function should be twice differentiable right and then we should you know minimize that and see whether the minimum is the positive definite matrix or negative definite matrix. Then we will be able to qualify whether we have reached a minimum or not okay. So now I am going to call this as θ , I am going to call this vector as capital E , x is the vector which is 3 dimensional in this case.

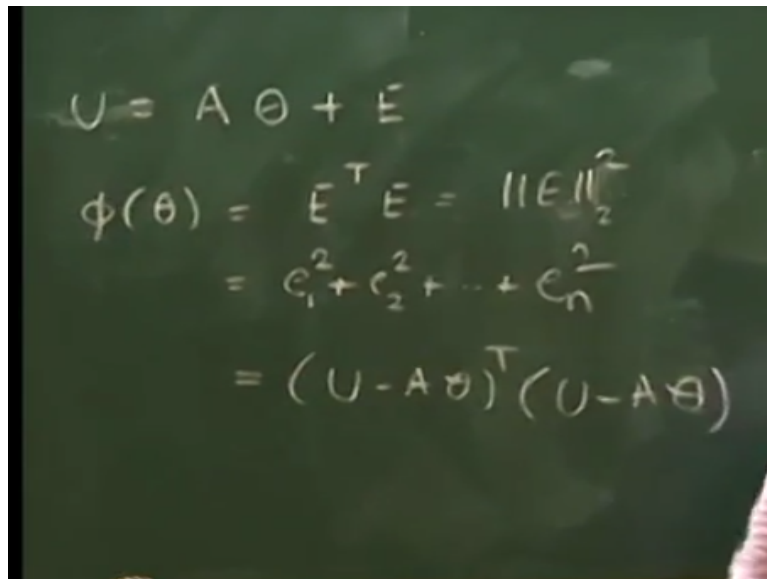
In general, θ will be a vector which is m dimensional. Suppose you have cubic equation which you are fitting will be 4 okay and so on. So if you fit a higher order equation, you will get higher order polynomial and I am going to call this matrix as matrix A let me see whether I am consistent with the notation, slight change in the notation. I am going to call this as θ .

I am calling it θ because θ is typically if you see literature on parameter estimation, a parameter vector is called θ okay that is why I am calling it θ , nothing particular about this. Also in the notes, I have developed everything with notation as θ so when you read the notes it will be easier for you to follow okay. So this particular vector I am going to call as capital U .

This particular vector I am going to call as capital U this is A , this is θ this is E . My large number of linear equations, are these linear equations? Why but there is $T_1 \text{ square}$ $T_2 \text{ square}$. They are known, I have taken measurements of temperature T_1 , T_2 , T_3 , T_n are known. I can

compute this so these will be just columns of a matrix with some numbers. So this is the linear matrix equation okay.

(Refer Slide Time: 15:35)


$$\begin{aligned}U &= A\theta + E \\ \phi(\theta) &= E^T E = \|E\|_2^2 \\ &= e_1^2 + e_2^2 + \dots + e_n^2 \\ &= (U - A\theta)^T (U - A\theta)\end{aligned}$$

So my equation reduces to $U=A\theta+E$, E is the vector of errors. Okay now I need to define an objective function and then this objective function should be differentiable, I should take its first derivative and set it equal to 0, take its derivative with respect to A , B , C , okay. Now I am just moving from this specific problem to a general problem where θ in general could be m dimensional okay.

This A will be n cross m matrix, see here m is 3 so this is n cross 3 matrix, it is a tall matrix so this n could be 100 and you have only 3 parameters to be estimated okay. I am going to define a ϕ of θ which is $E^T E$ which is nothing but $e_1^2 + e_2^2 + \dots + e_n^2$. Sum of the square of errors what are these? What is e_1 to e_n ? Errors in modeling, we are developing a correlation model of C_p versus temperature okay.

It is not exact fit; it is an approximate correlation. This e_1 to e_n are errors okay. I want to develop a model that minimizes sum of the square of errors, simple idea, which probably you have heard or done in your undergraduate studies for single parameter probably. We fit a line, we very, very usually fit a line. I do not know whether we do in undergraduate. We are taught to find out the best or least square fit.

But we fit a line by many times in experiments, we fit a line by you know just observing what is the trend visual. We do not try to do all this business, but then if you ask Excel or

MATLAB they will do this for you and give you okay. So what is this? This is $\|U - A \theta\|^2$ right. What is E ? $U - A \theta$ okay. Now how do I find the optimum? This is a scalar function.

Is this a scalar function? Why? This is simply norm E^2 square right sum of the square of elements norm E^2 square. Norm is the scalar function, it is a positive scalar function, it is always be positive. Well an objective function need not always be positive but in this case it turns out to be strictly positive the smallest value E transpose E can take 0.

That means all the errors are 0 okay whether that is possible for this particular problem or not is a different story but a smallest value it can take 0 in this case okay.

(Refer Slide Time: 19:07)

$$\frac{\partial \phi}{\partial \theta} = \begin{bmatrix} \frac{\partial \phi}{\partial \theta_1} \\ \vdots \\ \frac{\partial \phi}{\partial \theta_m} \end{bmatrix} = \bar{0}$$

$$f = x^T B y$$

$$\frac{\partial f}{\partial x} = B y \quad \frac{\partial f}{\partial y} = B^T x$$

So what I want to find out is I want to find out $\text{d}\phi/\text{d}\theta$ that should be equal to so this should be $\text{d}\phi/\text{d}\theta_1$ to $\text{d}\phi/\text{d}\theta_m$, this should be equal to 0 vector okay. Now I need to know something about how do I differentiate? See this is a function of a vector which maps into a scalar. How do I differentiate this? How do I differentiate with respect to a vector?

We know how to differentiate with respect to a scalar value right. So how do I differentiate with respect to vector? I am just going to state the rules; you can derive them. They are not so difficult to derive. We need to be little bit patient and then do your algebra correctly. So in terms of notation I am just digressing a little bit because I want to differentiate this E transpose E with respect to θ said that equal to 0 okay.

I will get m additional equations. See these are already n equations. I need m more equations, m here happens to be 3. I need m more equations. Those equations will be obtained when I do this but before that I should know how to differentiate a scalar function which is function of a vector? So let us say I have this function f which is $x^T B x$, B is the matrix, x and y are vectors.

I just want to differentiate f now, f is a scalar function you can see that, f is a scalar function okay. How do I differentiate? So $\frac{d}{dx}$ of f differentiating this scalar function with respect to the first argument x okay. That is by $\frac{df}{dx} = B^T x$ okay.

(Refer Slide Time: 21:47)

$$\frac{\partial}{\partial x} (x^T B x)$$

$$= 2 B x$$

if B is symmetric.

And a straightforward corollary of this is that $\frac{d}{dx} x^T B x = 2 B x$ if B is symmetric. If B is symmetric matrix, then it looks B symmetric matrix okay then it is well if it is not a symmetric matrix then what will it be? $B + B^T$. If it is not a symmetric matrix, it will be $B + B^T x$ okay. So we can use these rules to differentiate this function with respect to x and come up with a solution.

So how do I do that? So can you differentiate? Can you tell me what is the solution? Just differentiate just use these rules. Can you expand this?

(Refer Slide Time: 22:58)

$$\frac{\partial \phi}{\partial \theta} = \begin{bmatrix} \frac{\partial \phi}{\partial \theta_1} \\ \vdots \\ \frac{\partial \phi}{\partial \theta_m} \end{bmatrix} = \vec{0} \quad \left| \begin{array}{l} (A^T A)^T = \\ A^T A \end{array} \right.$$

$$\phi = U^T U - U^T A \theta - \theta^T A^T U + \theta^T (A^T A) \theta$$

$$\frac{\partial \phi}{\partial \theta} = -2 A^T U + 2 (A^T A) \theta$$

See this will be $\phi = U^T U - U^T A \theta - \theta^T A^T U + \theta^T (A^T A) \theta$. I have just expanded this. I have just expanded 4 terms. U is the vector, $A \theta$ is the vector right and then I am just expanding this simple matrix multiplication okay. Now what is U ? U is the vector of known values, it is a constant vector. So what is $\frac{\partial \phi}{\partial \theta}$?

This part will be 0. What about this guy? Apply the rules correctly, will be $A^T U$. So if I differentiate this $\frac{\partial \phi}{\partial \theta}$ I will get $-2 A^T U + 2 (A^T A) \theta$. I do not know whether you all agree with me. Why I am writing 2 times here? This $A^T A$ is a special matrix. What kind of matrix that is? Why 1 by 1? It will be m cross m know. In this case, it will be a 3 cross 3 matrix.

Can I say that $A^T A$ is always a symmetric matrix? You take transpose of $A^T A$ you will get back $A^T A$ okay. Positive definite matrix naturally appears here, you do not have to make any effort okay. It just plops out from the equations. So this is a positive definite matrix. In fact, this turn out to be positive definite matrix. This is a symmetric matrix. Will go on to show that this is positive definite also.

So this particular matrix is symmetric matrix, so I can write $2 A^T A$ okay that is because $A^T A$ just check this, $(A^T A)^T = A^T A$. So whatever is this matrix just remembering that this particular matrix whatever is this matrix you know how so tall it is, 1000 equations and 3 variables and it might be filled with all kinds of strange numbers.

This matrix that is $A^T A$ will always be symmetric matrix. In fact, we will show that if these columns are linearly independent then that particular matrix will be positive definite matrix okay. It will be a positive definite matrix so now the task is done. If this is 0, if I said this=0 what do I get? I said this=0 then what is the theta? How do you compute theta? What is this matrix? What is size of this matrix? It is m by m okay so if m is 3, this will be 3 cross 3 okay.

(Refer Slide Time: 27:26)

The image shows a chalkboard with the following handwritten equations and dimensions:

$$U = A\theta + E$$

$$\phi(\theta) = E^T E = \|E\|_2^2$$

$$\underbrace{(A^T A)}_{m \times m} \underbrace{\theta}_{m \times 1} = \underbrace{A^T}_{(m \times n)} \underbrace{U}_{(n \times 1)}$$

The result of the matrix multiplication on the right is labeled as $m \times 1$.

So setting gradient=0 gives us this equation $A^T A \theta = A^T U$ okay. So this is m cross m . What is dimension of this matrix? A^T is so this is m cross n cross n cross 1 so this is m cross 1 so finally you are getting an equation which is 3 cross 3 matrix and how many equations you wanted? n equations or you wanted 3 equations.

In this particular case, we wanted 3 equations. We have those 3 equations now. This is m equations okay. How many m unknowns? This is m cross 1, m unknowns. I solve this problem now if columns of A are linearly independent just think about what I am saying and I am going to prove it. If columns of A are linearly independent, then this matrix is invertible not only that this matrix will always be positive definite okay.

Before we proceed further and I show that why it is positive definite. What is the second derivative of this? $A^T A$ right.

(Refer Slide Time: 29:15)

$$\frac{\partial^2 \phi}{\partial \theta^2} = 2 A^T A$$

$$\frac{\partial \phi}{\partial \theta} = -2 A^T U + 2 (A^T A) \theta$$

So the second derivative $\frac{\partial^2 \phi}{\partial \theta^2} = 2 A^T A$. So now depending upon whether this matrix is positive definite okay. If this turns out to be positive definite matrix, then we are done then we have reached the minimum okay. In fact, this is the linear equation you can show that this minimum is not just local. This is the global minimum for linear set of equations, varies the global minimum okay.

Now my task is to show so that $A^T A$ is a positive definite. Symmetric matrix is very straight forward, positive definite matrix okay.

(Refer Slide Time: 30:01)

$A \rightarrow$ columns are linearly independent.

$$Ax = \vec{0} \iff x = \vec{0}$$

$$x^T (A^T A) x > 0 \text{ if } x \neq 0$$

$$(Ax)^T Ax = \|Ax\|_2^2 > 0 \text{ if } x \neq 0$$

So let us assume that columns of A are linearly independent okay. What does it mean? When will you get $Ax=0$? When you will get this? If columns are linearly independent which vector x will give you 0 vector? Only 0 vector. If columns are linearly independent, this will be 0 is

same as saying $x=0$. If columns are linearly independent only where you can get $x=0$ is $x=0$ okay.

Now how do you check? What is the definition of positive definiteness? Positive definiteness is that $x^T A x$ should be >0 if $x \neq 0$ right okay. I am just going to write this as $Ax^T x$. Am I correct? I am just clubbing x and A x and A , $Ax^T x$ okay. So what is $Ax^T x$? Norm x^2 square. When will this be 0?

So there are 2 situations, if A has columns which are linearly dependent there might be an x which will give you 0 but we have said that columns of A are linearly independent so if you have an x which is non-zero this vector is going to be non-zero and this is square of the norm, square of the norm is always a positive number so whatever x value whatever x vector you give me $A^T A$ is going to be a symmetric positive definite very nice matrix okay.

Matrix which we keep studying in linear algebra here it just pops out naturally as part of the development. So $A^T A$ is a symmetric positive definite matrix. I am showing here that for any x you know you will always get a positive number if $x \neq 0$. So this will be >0 if $x \neq 0$ so this satisfies the definition of positive definiteness okay. There are algebraic conditions like Eigen value should be positive when all that.

Let us not worry about it right now, we will visit that little later but here using the basic definition of positive definiteness I am showing that $A^T A$ is always positive okay which means not only that have reached the stationary point I have reached the minimum. It is a linear model, there is only one minimum. I have reached the minimum okay. If I have to do all this using 1 norm of E or infinite norm of E is not possible.

Because 1 norm of e is not differentiable, I cannot use this nice theory okay. Infinite norm is not differentiable I cannot use all this nice theory. So there are problems if I use some other norms, 2 norms very, very nice, you know you can differentiate and get this result. Now I am going to relate all this to projections. I will show that this is nothing but projecting a vector onto a subspace, geometric interpretation what is happening here.

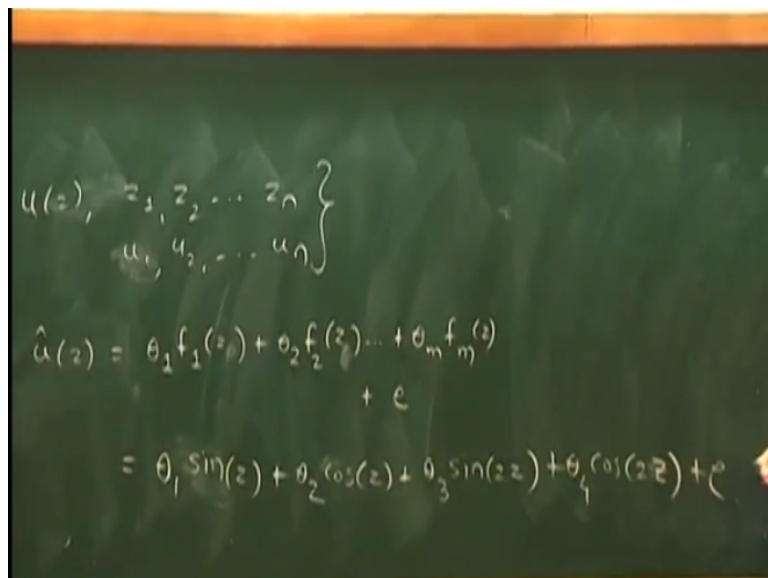
That is what I am going to do next. So why we are able to relate to projections because as I said in an inner product space you know with Hilbert space with the inner product norm gets

free you get a free norm, 2-norm and you get angle, you can talk about orthogonality, you can talk about projections, you can just you know generalize the ideas which you know from 3-dimension school geometry okay.

So this was symmetric positive definite matrix. We have reached the global minimum. This is the solution. Any such problem I can solve using. Do I have to stick to polynomials? I really do not have to okay. I can talk about any function approximation so I am going to take a scenario here I looked at Cp as a function of temperature that was one specific example okay. I do not have to stick to only polynomials.

I could do any function which is linear in parameters. I could actually estimate the parameters by this approach.

(Refer Slide Time: 35:01)



Let us say I am approximating a function say U which is I have this function Uz and I know its values at z1, z2, zn and its values are say u1, u2, un. This is my data set. I do not want to fit a polynomial. Let say I want to fit sin and cos. You are perfectly allowed to do that okay. So I want to fit a function say fz which is or let us call this function by some other notation so I want to fit a function say U cap z okay which is theta 1.

Well this need not be function of only one variable right, it could be function of multiple variables. In this case okay let us take first situation when you have function of only one variable say z1+theta 2 f2 z2. So I am fitting this function approximation of course there will

be an error here. So U cap is my approximation, $\theta_1 * f_1 z$ $\theta_2 * f_2 z$ and so on okay. Polynomial was one type of approximation.

I had chosen this to be 1, I had chosen this to be z , I had chosen the second to be z square, z cube and so on. That is one particular type of approximation. I could have chosen first one to be $\sin z$, second to be $\cos z$, third to be $\sin 2z$, fourth to be $\cos^2 z$ and so on. I could have chosen some other functions. I could have chosen Legendre polynomials. I could have chosen shifted Legendre polynomials.

So it is up to me what should I choose. I could have chosen all kinds of different functions here. I can apply the same theory which is very nice because I know these functions I have chosen these functions okay. I can evaluate them at a particular point okay. So let us take a scenario where you want to approximate this. Let us take instead of being abstract let me put it in a concrete form.

So θ_1 say f_1 is my $\sin z$ $\theta_2 \cos z + \theta_3 \sin 2z + \theta_4 \cos 2z + e$. Let us say this is my approximation function okay. What do I do? I do the same thing which I did earlier.

(Refer Slide Time: 38:32)

$$\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} \sin(z_1) & \cos(z_1) & \sin(2z_1) & \cos(2z_1) & \dots \\ \sin(z_2) & \cos(z_2) & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sin(z_n) & \cos(z_n) & \dots & \dots & \dots \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \vdots \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

I just write u_1, u_2 these are the values which I know at n different points. I can say first one is $\sin z_1 \cos z_1 \sin 2z_1 \cos 2z_1$. Then second row is $\sin z_2$ and finally $\sin z_n \cos z_n$. So I write this huge matrix. How many parameters? 4 parameters, so this is $\theta_1 \theta_2 \theta_3 \theta_4 + I$ have this e_1 to e_n same equation okay. So this f here could be any complex function. Mind you this particular model is linear in parameter okay.

We have formally defined what is the linear function some time back. So we should just check. When you call a function to be linear function? f of $\alpha x + \beta y$ will be α times f of $x + \beta$ times f of y that will exactly hold here. So this is a linear in parameter function, any linear in parameter model you can do this okay. So what do I get here? This is my U vector okay.

This is my A matrix, this is my θ and this is my E . How do I get least square estimate of θ ? $A^T A$.

(Refer Slide Time: 40:46)

$$\hat{\theta}_{LS} = (A^T A)^{-1} A^T U$$

pseudo inverse
of A

So least square estimate of θ is simply $\hat{\theta}_{LS}$ that is what we write, $\hat{\theta}$ why $\hat{\theta}$? It is an estimate of θ . There are different ways of getting θ . If we were to define 1-norm instead of 2-norms we will get a different θ . So I am calling this as $\hat{\theta}_{LS}$ which is obtained through least square of errors okay that is why $\hat{\theta}_{LS}$ is nothing but $A^T A$.

Now we have shown that this is a positive definite symmetric matrix. A positive definite matrix means no Eigen value 0, it must be invertible so I can write inverse $A^T A$. In fact, this matrix is called as pseudoinverse of A . Why pseudo inverse? A is a non-square matrix okay. A is n cross m I cannot invert it by the normal sense of n cross m square matrix. Why it is pseudoinverse?

What happens if you post multiplied by A? When you call something to be inverse? You multiply the matrix and matrix inverse you should get identity matrix. Take this matrix post multiplied not pre multiplied post multiplied by A. What will happen? Identity matrix. So that is why this $A^T A^{-1} A^T$ is called as pseudo inverse of A matrix and in MATLAB there is a function called PINV pseudo inverse.

Inv can be used for square matrices, pinv can be used for non-square matrices. You just say pinv doing this in MATLAB is in 2 minutes, formulate A matrix say pinv A times U, you get the theta the square theta okay in fraction of a second you can do this. You should know the theory behind this. That is very important okay. So is this clear? I do not have to have polynomials.

I can have any complex function okay. I will give you an example from chemical engineering and sometimes to begin with you may not have a linear equation but you might be able to come to a linear in parameter equation once you do a transformation say I will give you an example. I want to fit data, I am carrying out some reaction and I want to estimate the kinetics of the reaction okay.

(Refer Slide Time: 43:30)

The image shows a chalkboard with the following handwritten equations:

$$-r_A = k_0 C_A^{-n} e^{-\frac{E}{RT}} + \epsilon$$

$$\log(-r_A) = \log k_0 - \frac{E}{RT} + n \log(C_A)$$

$$= \log k_0 - \left(\frac{E}{R}\right)\left(\frac{1}{T}\right) + n \log(C_A)$$

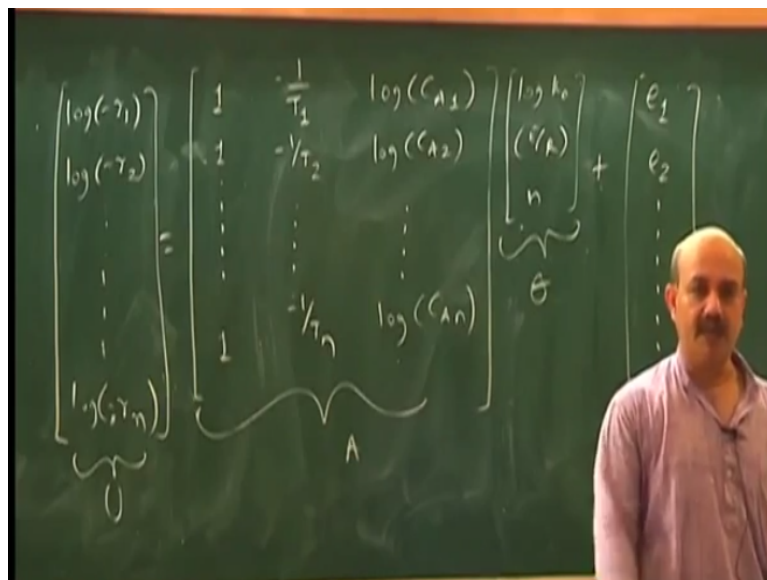
Other faint text visible on the board includes $\hat{\theta}_{LS} = (A^T A)^{-1} A^T y$ and "pseudo".

So I have this $r_A =$ or $-r_A = k C_A$ to power n or let us put it $k_0 e^{-E/RT}$ right or maybe I should write in the order should be $k_0 e^{-E/RT} C_A$ to power n. This is my model okay. Actually I should write this model with an error here. We do not write it because this may not be exact model we are proposing this model. So it is an approximate model so strictly speaking should write an error here say epsilon because do not confuse with E here.

E here is exponential so this is my model and I want to get least square estimates. What are the unknown parameters? k_0 , E , r is known and n . What is the data that you have? C_A and T okay is the data that you have. I could do a log transformation of this model okay. I can say \log of $-r_A = \log k_0 - E/RT + n \log C_A$ right okay. Now I can write this as $\log k_0 - E/R * 1/T + n \log C_A$.

I have collected data for concentration and at different concentrations at different temperature I have data for rate expression. I want to get a least square estimate okay. Is this a linear in parameter model? What are the parameters? $\log k_0$, E/R or E whatever, E if you want to take r on this side you can do that since you know r . So E is unknown parameter and n okay. So here in this matrix what you will get?

(Refer Slide Time: 46:29)



In this matrix, what I will get is 1 1 here then $-1/T_1$ $-1/T_2$ $-1/T_n$ this column, this column is coming because of this variable $1/T$ okay. What will be the third column? $\log C_{A1}$, $\log C_{A2}$ actually $\log I$ should write \ln natural log, $\log k_0$ E/R and n and what should be here? It should be $\log -r_1$ $\log -r_2$ you know all these reaction rates. Just remember $-$ is the notation not so, we are not taking log of a negative number.

So this is the equation that you get this is the U vector which you know because you know the rates at different temperature and concentration okay. You will have estimates of the rates. This matrix is known to you A matrix, temperatures are known to you, concentrations are

known to you so this is known matrix. This is my theta which I want to estimate okay. What is the least square solution?

$A^T A^{-1} A^T U$ just remember this formula very, very important $A^T A^{-1} A^T$ this is called as pseudoinverse of A , $A^T A^{-1} A^T$ is always a square matrix and if columns of A are linearly independent it is always invertible matrix and what you get here from the least square sense is the minimum okay.

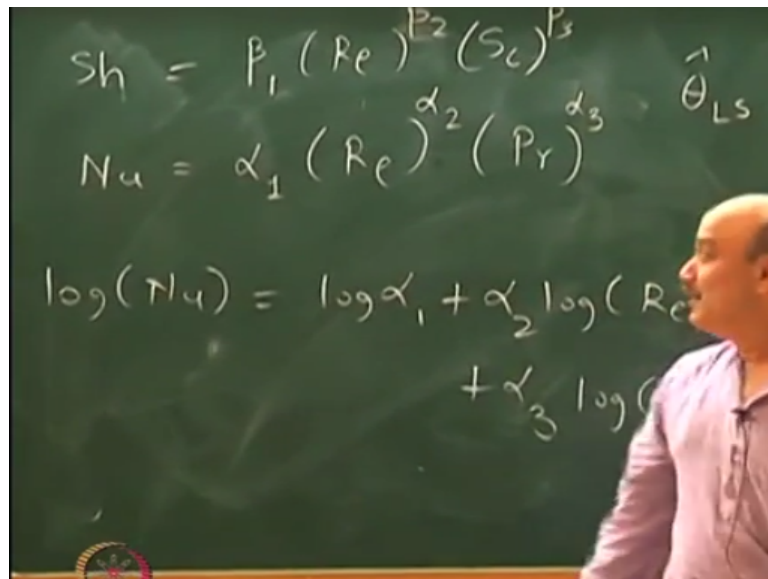
There cannot be any other value of theta which will give you smaller sum of the square of errors. For a linear in parameter model, you cannot get a model which will give you smaller value than this optimum value which you get okay. So this is my least square problem, this least square problem is used in many, many ways.

In the assignment, I will give you other problems from chemical engineering where you do least square fitting to estimate the parameters. Sometimes what happens is that sometimes a model is not transformable to a linear, it cannot be linearized. I will call this as a linearization step. In the linearization step, you can do some transformation and convert originally the model is not a linear in parameter.

E and n and k_0 they are multiplying and they do not appear you know in the fundamental definition will not satisfy here that is $f(\alpha x + \beta y)$ is same as $\alpha f(x) + \beta f(y)$ that will not be satisfied for this function but that will be satisfied this transform function. Now this is not possible always to get this. I will tell you one more model in the chemical engineering where you can do this transformation okay.

And estimate the parameters this is how it is done. One more classic model in chemical engineering is correct me if I am wrong.

(Refer Slide Time: 50:47)



Nusselt number = $\alpha_1 \cdot \text{Reynolds number}^{\alpha_2} \cdot \text{Prandtl number}^{\alpha_3}$ right. This is not a linear in parameter model. What are the known quantities here? I would have data for Nusselt number, Reynolds number for different flows I have collected data for you know Prandtl number, Nusselt number and I want to fit and I want to estimate α_1 α_2 α_3 okay.

This is not a linear in parameter model but simple log transformation so if I take log of Nu it will be $\log \alpha_1 + \alpha_2 \log \text{Reynolds number} + \alpha_3 \log \text{Prandtl number}$. This transform model is linear in parameter okay. I could use least square estimation technique to estimate $\log \alpha_1$ α_2 α_3 . If I get $\log \alpha_1$ I can get α_1 not an issue right. So this is simple transformation.

What is mass transfer correlation? Sherwood number $Sh = \beta_1 \cdot \text{Reynolds number}^{\beta_2} \cdot \text{Schmidt number}^{\beta_3}$ Well for me it is almost 20 years now, so you are fresh so right. This is some point 6 3 and you may have viscosity correction μ/μ_s raised to something fourth parameter. So you may have 4 parameters, but once you do this transformation it is a linear in parameter model.

You can use least square parameter estimation, this column will be different, this column will be different same idea works. You get least square estimates of α_1 , α_2 α_3 α_4 just by taking $A^T A^{-1}$. So polynomial fitting is not the only thing. I am just now polynomial fitting we started with because we wanted to do something with discretizing boundary value problem or partial differential equation.

But least square estimation goes much, much beyond. All these were discovered by Gauss. The famous mathematician well in the realm of mathematics or in the history of mathematics Gauss is called as prince of mathematics and the work he did actually is now major fields okay least square estimation, Gaussian densities, Gaussian quadrature and there are so many things.

I mean I do not know what we would be doing if Gauss had not discovered. Gauss was as child prodigy. He discovered many things at very, very early age and the things that he started actually he started looking at least square estimation because he wanted to fit I think he was looking at a problem of fitting an orbit into the data for obtain from motion of planets around the sun that was the problem that was being discussed at that time.

So what is the best fit? It does not happen to be a circle, actually it is an ellipse and you have to fit because if you look at a data, the data has some errors and then you have to do a best fit to get the correct model. So what he started about 150 or 200 years back is now just spans I mean all these tools are used in image reconstruction. They are used in soft sensing; they are used in just name it.

Least square estimation forms the foundation okay or works by Gauss forms the foundation. So we are here today because of this prince of mathematics Gauss.