Advanced Numerical Analysis

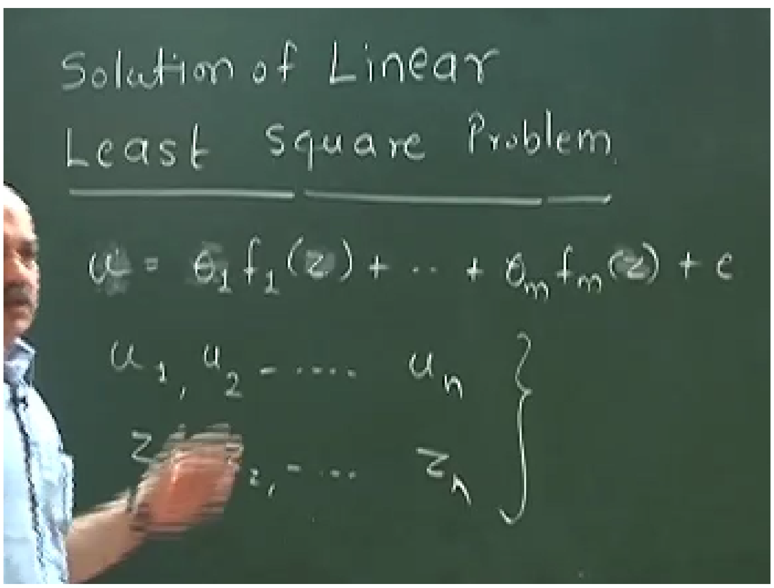
Prof. Sachin Patwardhan
Department of Chemical Engineering
Indian Institute of Technology- Bombay

Lecture - 19

Linear Least Square Estimation and Geometric Interpretation of the Least Square Solution

Okay so in our last lecture we looked at optimization multivariable optimization unconstrained optimization. So, we derived conditions for optimality necessary and sufficient condition followed by the next lecture we looked at application to solving linear in parameter least square problem. Okay, what is nice about linear in parameter of the least square problem is that the solution optimal solution for the optimal value of parameters least square estimates can be computed analytically. So, just to have a recap.

(Refer Slide Time: 01:03)

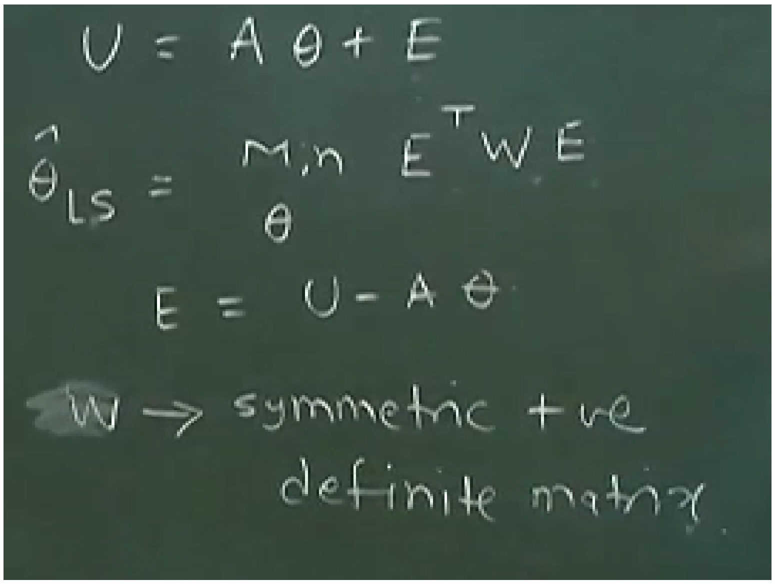


So, this is just solution of linear least squares okay we have collected data and we have some model. This model could be in general I said could be y=theta1 f1x theta m+error in general you have a model of this form where f1, f2, fm are some known functions simplest one we looked at was polynomials. But it did not be polynomials it could be any functions which are known function and then you are doing a function approximation.

If it is polynomials it is polynomial approximation. We have collected data we said this is u not y and one minute we did not use x we use z here. So, let me correct that, so z is the independent variable it could be anything it depends upon. In this case it need not be space it could be any independent variable. We looked at many examples from chemical engineering where we could use this method for and then we have this data collected which is u1, u2+un at points z1 z2 zn,

And using this data we wrote number of linear equations.

(Refer Slide Time: 02:48)



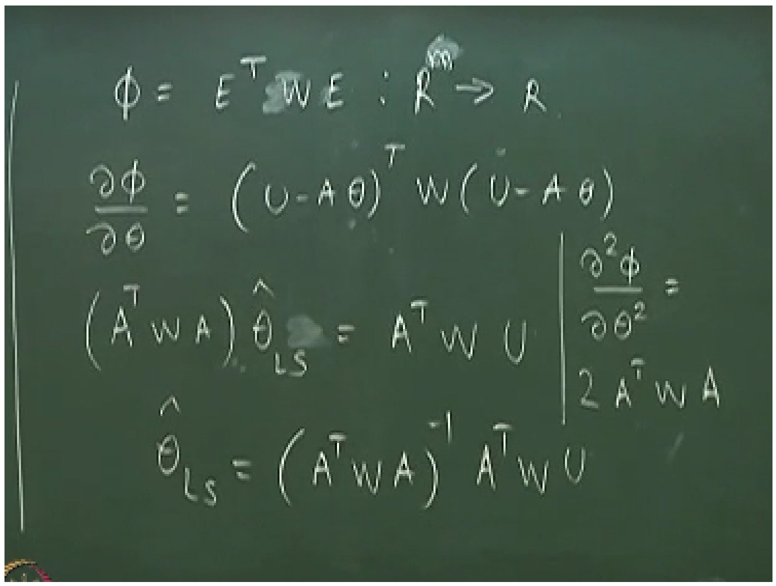
And finally, we put it in a matrix form U=A theta+E where E is the modeling error and then you know we found out theta least square=minimum or minimize with respect to theta E transpose. Now I am going to make a little small modification here as compared to the previous development. I am going to say here E transpose WE okay where E=U-A theta I want to solve this problem where W is a positive definite matrix is a symmetric positive definite matrix.

In general, you can solve a problem earlier we had looked a special case of this service that is E transpose E where W was identity matrix. Identity matrix is a symmetric positive definite matrix okay this is a special case which we have looked at earlier I am just generalizing this. W can be see for example in some situations when you collect data and fit a model you know that a particular observation is more reliable okay or particular observation is less reliable.

So, you could attach weight positive weight okay small weight if it is less reliable a large weight if it is more reliable. So that when you optimize okay the optimizer will give more importance to those which are accurate measurements include less important to those which are less accurate good means we could actually twist this or sometimes you need to do this because of you know the variables have different values and so on. So, this is in general general formulation.

In which I have some weighted matrix here which a symmetric positive definite matrix is. As you know that this will define 2 norm and so on. Right so take a positive definite matrix it will be defining a 2 norm.

(Refer Slide Time: 05:15)



If I have this phi which is E transpose E E transpose WE then if I use necessary condition for optimality, then it is dou phi/dou theta=U-A theta transpose U-A theta and then we had rules of differentiation of a scalar function. So, this is a scalar function from N to R this is a scalar function so M to R RM to R theta is M dimensional vector this is a function of theta theta is M dimensional vector.

In general, this is a this is a scalar function and we had rules of differentiation of a scalar function. With respect to a vector and using that we came up with a formula which is we will get this equation if you use it A transpose WA theta least square=A transpose WU okay and finally we argued that if columns of A are linearly independent then this matrix is invertible this is a symmetric matrix symmetric positive definite matrix very nice matrix.

And theta least square=, so my least square estimate my least square estimate can be written as A transpose WA inverse. We also said a special case where W=I we look at a special case when W=I A transpose A inverse A transpose is called pseudo inverse of matrix A. Okay remember here A is a non-square matrix, A is a n cross m matrix theta is a m cross 1 vector E is n cross 1 vector okay this is a non-square matrix say non-square matrix.

And then we talked about its inverse or we talked about it pseudo inverse. Pseudo inverse is defined by A transpose A Inverse A transpose. Okay so far so good so we have this derivation one of your classmate was asking me after the last lecture is that she knows about this formation formula using summations you know least square estimate using summation summation summation yeah actually this formula.

And that is not different they are one and the same and deriving that summation formula starting from this formula is part of one of the exercise problems okay. So, the exercise problem which I give you next for least square estimation will have that problem. You will derive that summation formula for least square estimates. Okay so those two things are not different this is a more elegant compact way of expressing the same thing which it is not different.

It is one and the same thing. Okay so the regression formula which you know with multiple summations can be very elegantly expressed through this A Transpose A Inverse A Transpose the same thing exactly identical thing. When you solve the problem, you will realize that it is nothing different the same same problem. Okay what I want to do now is so far so good we have done lot of algebra.

We have found out the condition for optimality then we said the second derivative here what is the second derivative here the second derivative here that is dou2 phi/ dou theta square this is nothing but A transpose WA to A transpose WA that will be second derivative. The second derivative is always symmetric positive definite okay which means the stationery point which you have got through this is a stationery point.

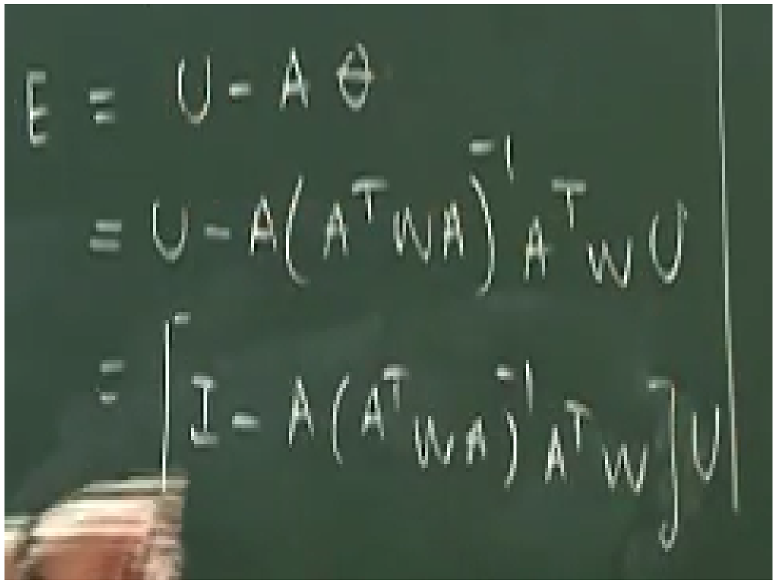
This is a point at which the gradient=0 okay at this point at this point the second derivative is given by this matrix and this particular matrix is symmetric positive definite you have reached the global minimum. Okay so this is fine this is lot of algebra, I want to give some geometric insights into what is really happening okay, how do you relate this to your school geometry okay that is what I want to elaborate next.

So, what was the thing here here you had a non-square matrix right you had 100 data points may be only 3 parameters. We saw the example of CP versus temperature okay, so m was small, and n was large you are fitting some function some polynomial and the number of parameters were much much less than the number of variables. Okay to get in an insight into what is happening let does not work with 100 dimensional spaces.

We cannot visualize in 100 dimensional spaces right but in 3 dimensions I can visualize. Okay so I am going to create a dummy problem from this which is very, very simple 3 equations in 2 unknowns okay 3 equations in 2 unknowns. It is a representative problem for this equation see what is the what is the main thing number of equations is more than number of unknowns and then you are not able to satisfy all the equations simultaneously.

If you are able to satisfy all the equations simultaneously okay, you know error would be 0 but error is not 0. You are finding out least square solution okay what is the error? How do you compute the error here? what is the error vector. If you substitute this theta you will get the error vector* this equation.

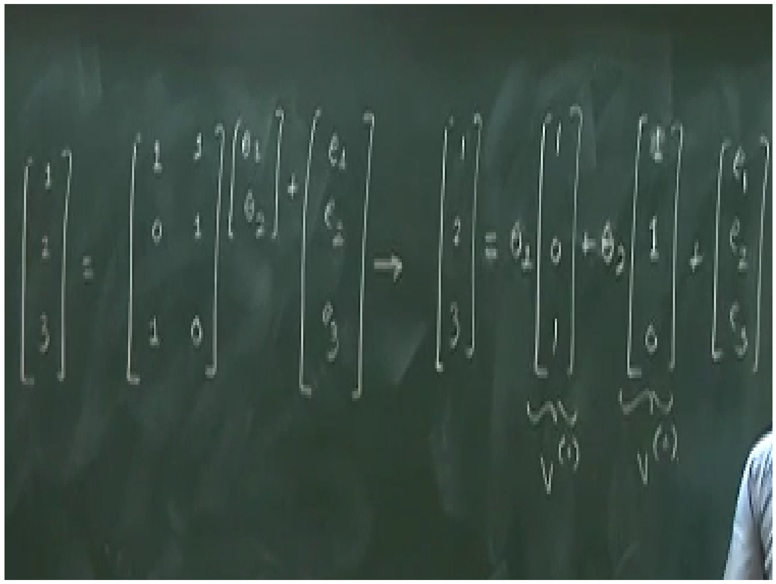
(Refer Slide Time: 11:39)



Okay so if I substitute here if I substitute E=U-A times A transpose WA inverse A transpose WU. Okay so I can write this is I-A A transpose WA inverse A Transpose W Times U unless this matrix is 0 you will not get exact satisfaction. So, these equations are such that you cannot satisfy all of them simultaneously you are trying to find out a least square trying to find out of this square.

And then there is always going to be an error vector such that probably no equation is satisfied. Okay so the curve which you get here will not probably pass through any of the points quite likely okay it only captures the tendency in the data. Not going through every point in the data because okay let us try to get some insights by taking a simple dummy problem. So, I am going to take a simple problem in which I have this equation.

(Refer Slide Time: 13:02)

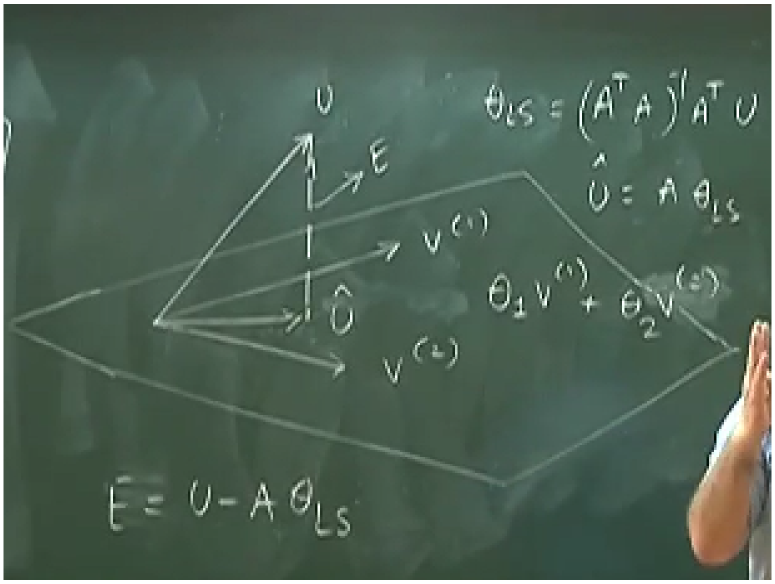


To be satisfied I want to solve this equation 123= I hope that this vector cannot be defined by this linear combination of these 2. I have not I have just created this problem right I do not know whether. So, when can you solve this exactly just think about the geometric inside of it about it when can you solve this when can you actually exactly solve this. How can you write this equation is there another way of writing the same equation?

Okay I will move to here, so I can write this as 123=theta 1 101+theta 2 110+ e1 e2 right I have taken very simple problem 3 equations 2 unknowns. 2 unknowns in the terms of actually there are 5 unknowns, actually there are 5 unknowns e1 e2 e3 theta1 theta 2 there are 5 unknowns but well far as the parameters are considered there are 2 unknowns theta 1 and theta 2 2 parameter unknowns. Okay if I write it like this do get some more insight.

“Professor - student conversion starts” (()) (15:09) yeah so theta 1 times this vector let us call this as v1 and let us call this vector as v2. When can you solve this exactly when will the error be=0. Can you say something about a span do not forget the last quiz all possible linear combinations of v1 and v2 what will it give you (()) (15:42).

(Refer Slide Time: 15:43)



Okay let us say this is my v1 and this is my v2 let us say this is my v1 and this is my v2. Okay what will it span what will it span v1 and v2 all possible linear combinations of v1 and v2 what will span. It will be a plane passing through the origin right when will the error be = 0 (()) (16:20) say it louder this is when this vector on the left-hand side lies in the plane okay if this vector left hand side vector exactly lies in this plane which is span of these 2 vectors **"Professor-Student conversion ends"**.

Okay all possible linear combinations any vector in this two-dimensional plane can be generated by a linear combination. If it does not what it means when it is vector outside okay let us picturize this as with this vector. So, this is my u vector this is my u vector this is u. Okay, so this is here this is my u vector, so u does not lie in this plane, it does not lie in this plane, so I need to find out what do I need to find out a least square approximation.

See because ultimately, I am finding out an approximation which is theta 1 times this vector+thetha 2 times this vector. That approximation is going to lie where the approximation will lie somewhere here right a vector this vector will be theta 1 times v1+theta 2 times v2. This vector theta 1 times v1+theta 2 times v2 cannot leave this plane right it is linear combination of these 2 vectors it cannot leave this plane.

Okay in your school geometry you have studied this problem point which is closest point which is closest to the vector in this plane how do you get it drop a perpendicular. Okay I just want to show that what we have done till now by so called least square is nothing but drop a perpendicular. Okay the point which you will hit the point which you will hit here if you drop a perpendicular what is the best approximation in the least square sense.

I will show that this point is nothing but the point which is the least square approximation of this vector in this plane okay, I want to find out best approximation of this vector in this plane in the least square sense. From the school geometry I know just draw a perpendicular okay this is the point which is closest to this plane okay point which is closest to this plane. Okay so this point and what I actually get by solving theta least square.

So, what should be the theta that gives you this theta least square should be A Transpose A Inverse A Transpose U this will give you the theta and this this particular vector would be you know this approximation. Let us call this U cap will be A times theta LS this will be U cap okay U cap is my approximation. What is this vector error vector this is if this is U cap this U cap is A times theta LS okay my equation is error=U-A times theta LS my equation is U-?

Okay if I just complete this you know law of vector additions this is nothing but error vector. In least square approximation the error vector is perpendicular to the plane this you know from your school geometry right. I am just generalizing that result in any inner product space what is very, very elegant is that same result can hold in any inner product space in any Hilbert space. You can have the same result.

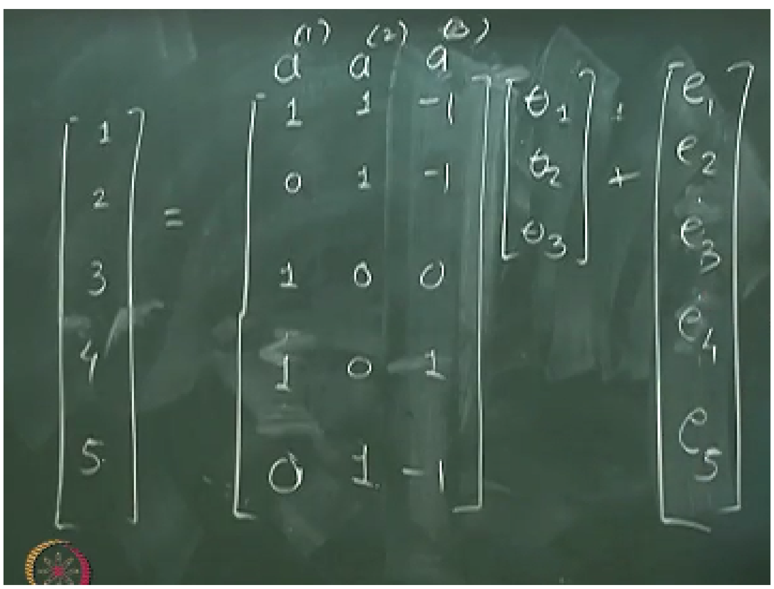
Which you know from school geometry finding minimum distance of a point from a plane. Okay so this problem this problem see you cannot visualize when suppose this suppose here let us go back here suppose you had 10 equations in 2 unknowns. Okay I cannot visualize in a 10-dimensional plane but linear combination of 2 vectors in 10 dimensions what is it like? it will look like it will look like a plane like this something like this.

If I were a creature in 10 dimension I could be you know it would be possible for me to visualize a 9-dimensional plane, but we cannot so, but it will look something like this. And then what you are doing you are just finding out the point in the plane which is at a minimum distance from this point which is lying outside the plane. Okay geometrically what you are doing is what is called as projections okay I am projecting this vector onto this plane.

Projections you probably would have done when you have done your engineering drawing right. So, projections is something which you know from your engineering or right from your school. Okay so even though you cannot visualize 10-dimensional thing conceptually or it is not going to be different. I mean if you were able to visualize that it would look almost the same. So, I just want to show that if I just proceed through this geometric idea I will get the same thing.

Okay that is my next that is my next task. Okay let me do a derivation with only 2 vectors okay in my notes there is a derivation with only 1 vector. 1 vector means distance of a point from a line. I will start with that and then I generalize it to distance of a point from a subspace actually in general if there were 3 if there were 3 vectors okay see if this problem.

(Refer Slide Time: 23:19)



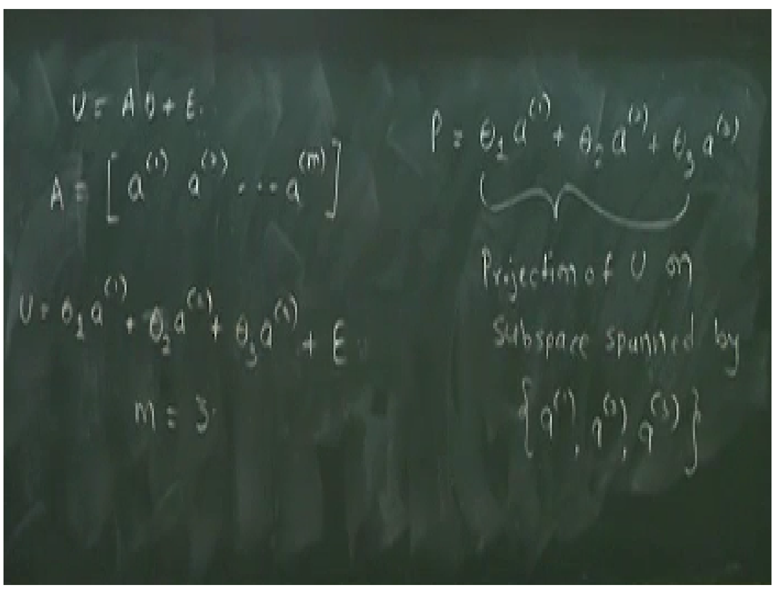
If I just modify this problem and then here left-hand side is let us say this is not coming from some physical problem. I am just arbitrarily creating some set of equations. Let us say I have this problem okay now here here what can you say the least square solution will be lying in the subspace spanned by this column vector this column vector and this column vector. The least square solution will lie in the subspace spanned by this column vector.

This column vector and this column vector. Okay the component which is outside the subspace is given by this error okay we are able to split actually geometrically speaking we are able to split a vector into 2 components. 1 lying in the subspace and 1 orthogonal to the subspace. Okay we are able to split the vector theory the least square, what is the subspace? Subspace we find by linear combination of columns okay linear combination of column will give me subspace.

This e1, e2 to e5 will give me the component which is outside the subspace together they form this whole vector. Okay together they form this whole vector. Now I want to generalize this and then just show you that what we have derived but I am going to take a case where W=I Okay weighted matrix=I. I am not going to complicate life by so that will give me some handle to okay let us write.

So, this is let us call this okay in my notes I am calling this vector as a1 this vector as a2 and this vector as a3.

(Refer Slide Time: 25:33)



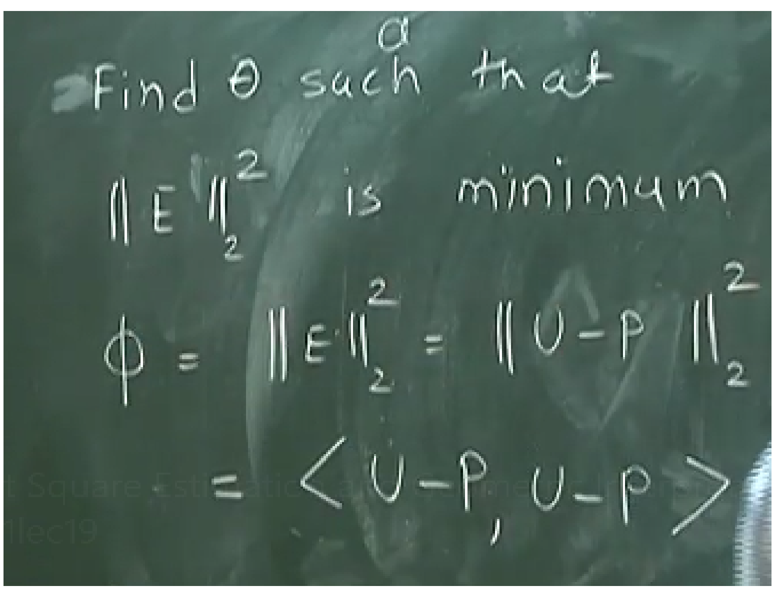
Okay so in general in general I can write my model as U=capital U this is my original model is U=A theta+error and I am writing my A matrix as a1, a2 and am, there are m columns these are column vectors, there are M columns in this case, here there are 3 columns, here there are m columns. Okay so I can write this equation as U=theta1 a1+theta2 a2. I can write this vector equation.

Okay now to simplify life let us take a case where there are only 3 vectors. Okay so generalization to m is not so difficult so I will just take theta 1+theta3 a3+E this is the case when we have taken m=3, there are 3 parameters theta1 theta2 and theta 3. I want to find out least square estimates of this okay using okay I am going to call this vector see this vector belongs to the subspace which vector this vector p.

I am going o call this as a is this as a projection vector projection vector theta 1 a1+theta 2 a2+theta 3 a3, this is the projection vector this is the projection projection of U on okay this is projection of say it is linear combination of a1 a2 a3. So, this p vector has to lie in the subspace spanned by a1 a2 a3 right it has to lie in the subspace spanned by a1 a2 a3.

So, this is a projection I am going to call this as projection of u onto subspace spanned by what is it that I want to minimize, how do I find out this, I find this out by minimizing square of distance right.

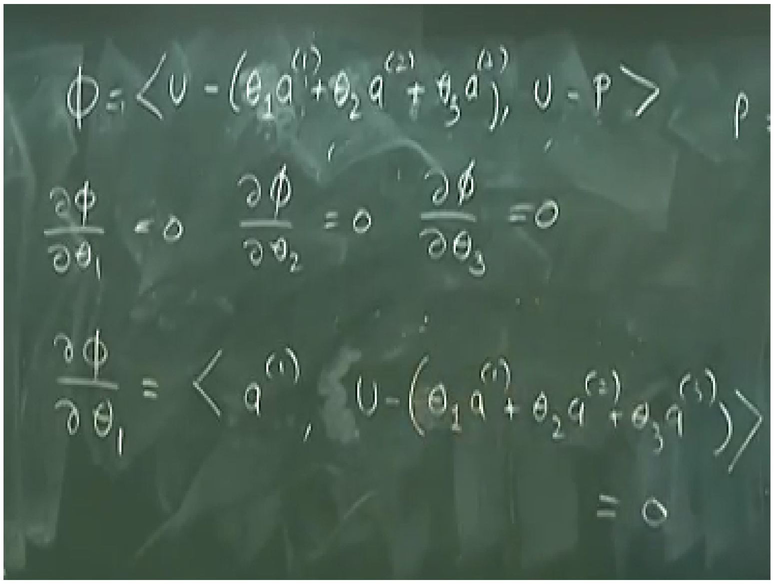
(Refer Slide Time: 28:43)



I want to find out, so my problem is find theta such that error 2 norm is minimum right, 2 norm of error is minimum right. Okay let us start doing this, so what does it mean so I want to minimize a phi which is 2 norm error square which is p-u or u-p right but what is this is a 2 norm, we are working on an inner order space so 2 norm square is related to the inner product how? so this is u-p, u-p right inner product of vector u-p, u-p.

Now my model is this U=theta.

(Refer Slide Time: 30:08)

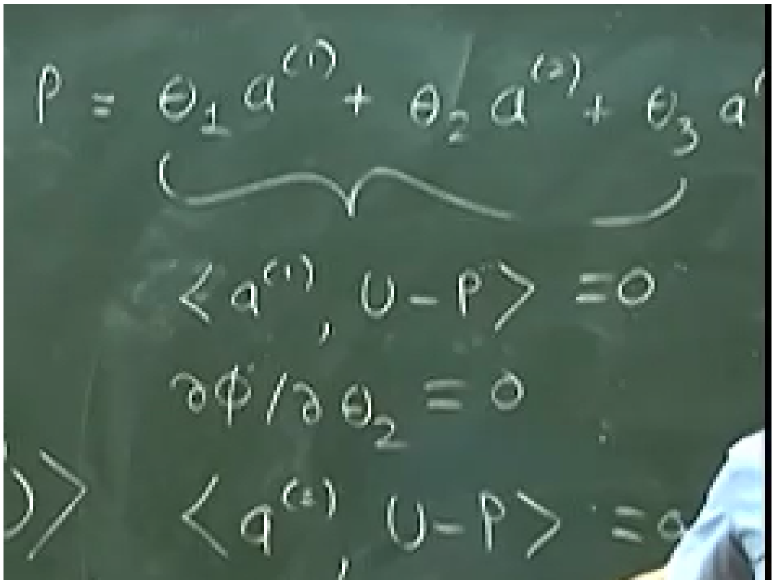


So, I can write phi to be inner product of u-theta 1 a1+theta2 a2+theta3 a3*u-p, p is again this vector, I can write this vector again here right, what is p? theta1a1+theta2 a2+theta3 a3, I can just put this vector in place of p it will be a longer expression. Okay now how do you find out the minimum what is the necessary condition can you do this see my necessary condition.

See my necessary condition for optimality is dou phi/dou theta1=0 dou phi/dou theta2=0 and dou phi/dou theta 3=0 right these are my 3 conditions is everyone with me on this, these are my 3 conditions. Okay if I differentiate this phi with respect to theta okay what will I get okay just look at it you are differentiating only once okay if I differentiate only to this first vector what will remain only a will remain.

Okay you can check this, but I am just going to write the final result that dou phi / dou theta1 this is nothing, but you will get here a1, okay if I am differentiating this I am skipping in between steps you can actually expand you can expand this entire inner product. Okay you will get many terms; okay you have to patiently find out the inner product of each element by element you know the rules of expanding an inner product.

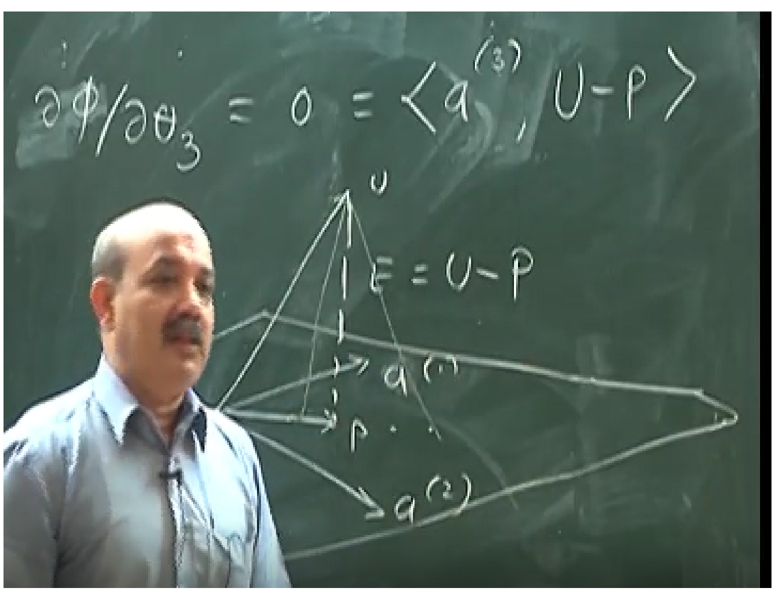
(Refer Slide Time: 33:20)



So, what I will get is a1 inner product with this=0 so actually what I am getting here is that a1 inner product u-p=0 because this is projection vector, this is my projection vector okay so what I am getting is a1 inner product u-p=0, what is this? u-p error vector, so what it says is that error vector is perpendicular to a1 direction same thing If you do with theta 2 okay setting dou phi/dou theta2=0 you will get a2 u-p=0 okay you will get this equation.

And well the third equation you will get is.

(Refer Slide Time: 34:09)



The third equation that you will get is dou phi/dou theta 2=0, so that is=a3 u-p this is=0. So, what is the meaning of this? what is the geometric meaning of this? that this error vector is perpendicular to a1, a2 and a3. Okay if error vector is perpendicular to a1 a2 a3 will be perpendicular to linear combination of a1 a2 and a3 yeah what is linear combination of a1 a2 and a3? it will be in s apan of a1 a2 3 and it will be in a subspace of a1 a2 a3.

Okay so this error vector is perpendicular to the span of a1 a2 a3. Okay which means best approximation of u in the span of a1 a2 a3 is obtained just by dropping but you know it is a generalization of result from school geometry to have a perpendicular from a point to a plane. Okay that is the best approximation of that vector in that plane, that is all okay that is all we are generalizing into a general function space or into a general inner product space.

Okay now when is this possible? This is possible only when you have inner product defined that is very, very important. Okay why we why we are so much why we like least square approximation as against to the 1 normal approximation or infinite or approximation because this least square approximation comes attached with you know inner product inner product can be related to the geometry.

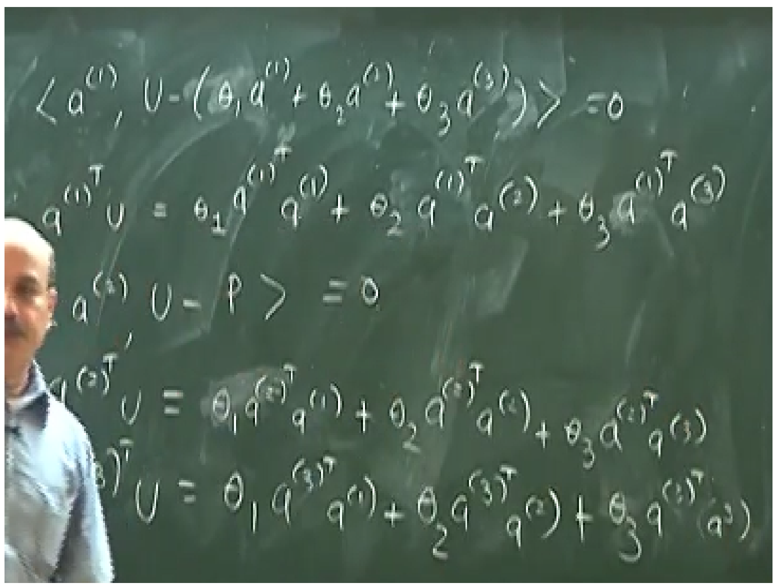
You can talk about perpendicularity you can talk about projections. The same idea of projections which we use in 3 dimensions okay distance of a point from a plane just drop a perpendicular okay same idea is actually being is actually being said here. So, all that we have proved is if I redraw this figure in 2 dimensions if I have this plane which spanned by say v1 and v2. So, this is my a1 and a2 and this is a vector this is my vector U.

Okay this is my U vector then this is the p vector projection vector this is a projection vector Okay this is a projection vector and here what we are getting is this projection vector and this error vector e here is u-p. Okay this is a special vector such that this error see if I take any other vector here this error will not be perpendicular. There are some many ways of approximating you know this in this plane.

I could approximate here I could approximate here somebody will say that this is the approximation. This is one possible way of approximating what is the least square approximation perpendicular okay least square approximation is the perpendicular. Now I need to go back from here and saw I should start with these 3 equations and show that well what you have got actually is nothing but the same formula of A transpose and A inverse.

I need to derive that right, I just showed geometrically that these 3 vectors so if I take a1 a2 and a3 the plane spanned by this and the error is perpendicular to this plane, the plane is spanned by this and so on. So, I need to go back and connect A Transpose A inverse business I will do that. Okay how do I get theta 1, theta2 and theta 3 by solving these 3 equations.

(Refer Slide Time: 38:32)



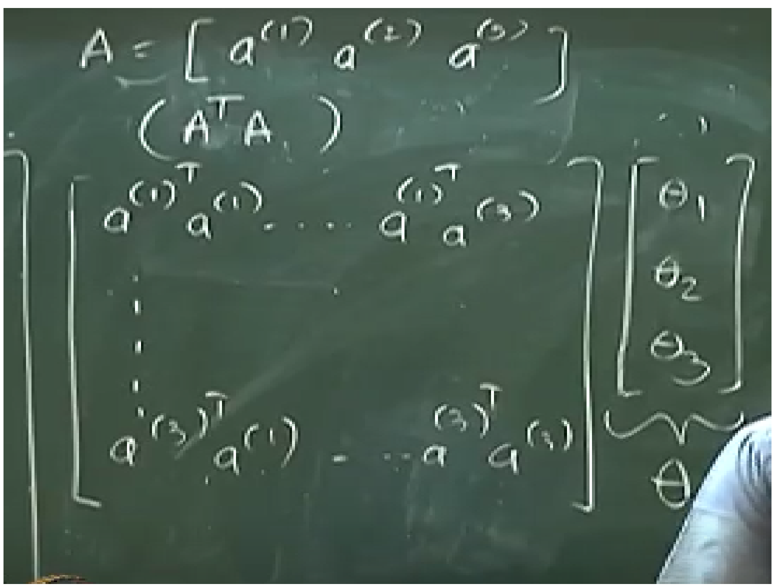
You know a1 inner product u-theta1 a1+theta2 a2+theta3 a3=0 right well in 3 dimensions generally in M dimensions, how do you with real values real values what is the inner product simple inner product A Transpose B you know if there are A and B 2 vectors. So here this will be first equation this first equation I can rewrite as a1 transpose u= I am just skipping 1 step in between.

Okay I am just writing theta1 a1 transpose a1 right a1 with a1 a1 with a2 and a1 with a3 right theta 2 a1 transpose a2+theta 3 a1 transpose a3 right what is the second equation, second equation is a2 u-p=0 right u-p=0 right what will it give you? a2 transpose u=a2 theta 1 a2 transpose, by the way a1 transpose a1 is it a scaler, always a scaler a1 transpose a2 is it a scalar always a scaler right.

So, I will get another equation theta2*a2 transpose a2+theta 3 a2 transpose a3. Okay I will just write the third equation this follow the same thing. I will get a3 transpose u=theta1 a3 transpose a1+theta2 a3 transpose a2+theta3 a3 transpose a3, is everyone with me on this, I have just rewritten those equations. I started with the geometric I started by saying that well what we have done is nothing but projections.

Okay this error vector is perpendicular to a1 error vector is perpendicular to a2. Okay these are the equations which I have written error vector perpendicular to a1 error vector perpendicular to a2 and in this case error vector perpendicular to a3, How many equations and how many unknowns? 3 equations and 3 unknowns. What are the 3 unknowns? theta1, theta2 and theta 3 So, what do I get here?

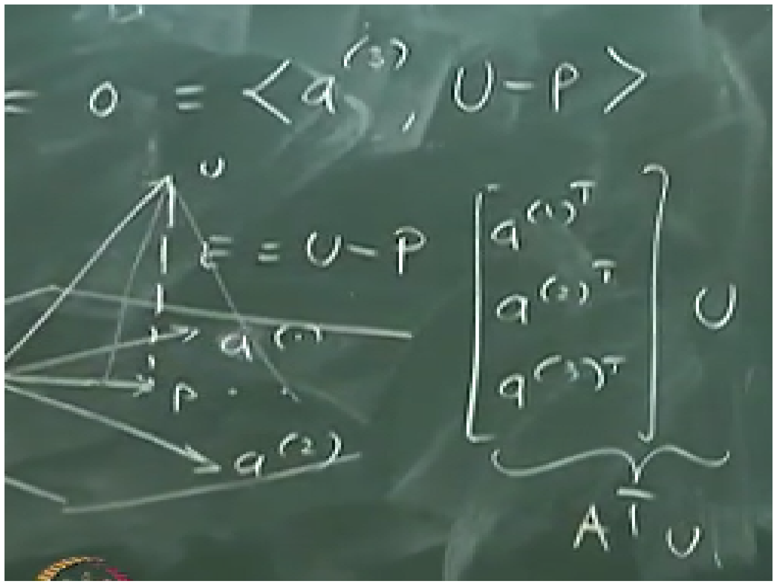
(Refer Slide Time: 42:27)



Right hand side I can write this as a1 transpose a1, a1 transpose a3 and we can just collect it in the vector matrix form this is a3 transpose a1, a3 transpose a3 this *theta1 theta2 theta3 right the right-hand side I can write as 3 cross 3 matrix. Okay just check this matrix is nothing but if I start with A=a1 a2 a3 they are the 3 columns of A matrix. Okay very, very easy to check that this matrix is nothing, but A transpose A multiplication of A Transpose A is this matrix.

Okay the algebra ties in with the geometry very, very nicely okay algebraically we arrived at this condition A transpose a theta=A transpose u. Now just look at this left-hand side what is left hand side, the left-hand side I can write here the LHS.

(Refer Slide Time: 43:58)



I can write as a1 transpose a2 transpose a3 transpose u what is this A transpose u right A transpose u okay that is the result which we got purely by doing you know algebra of necessary conditions of optimality. So, A transpose u=A transpose a times theta least square solution right. So, this is my theta okay I started with geometric argument of projection onto a subspace. I was able to recover the least square formula which is A transpose a times theta=A transpose u.

What is the least square estimate this theta will be least square estimate of course? We call it estimate so we give a hat above it. So, this is estimate of theta this is not there is no way unique way of finding out theta there are multiple ways, but least square estimate is unique. How is it computed A transpose a times theta=, A transpose u, A transpose a if columns of a are linearly independent okay A transpose a is always invertible.

It is always a square invertible matrix symmetric positive-definite square invertible very, very nice matrix. Okay this matrix which we study linear algebra does appear in a very, very practical application. We are trying to fit a curve in some data or fit a function in some data positive definite symmetric invertible matrix. So, this this particular result A transpose a times theta least square=A transpose u.

We could recover this result purely through geometric arguments of projection. We said that basically what is happening here is we are trying to find out that vector in the plane spanned by a1 a2 which is at a closest distance from a vector vigil which is outside this plane. Why we need least squares because this vector is not lying in this plane. What will happen if this vector is lying in this plane error would be 0.

Okay what is the extreme situation thus U is perpendicular. okay so what is the best approximation 0. Okay if U is perpendicular probably you have done wrong modeling you know error should be small not the error the vector is perpendicular. So, this is a very, very nice result, this ties in with actually there is 1 more angle to hold the whole of this thing. So, we derive this result through algebra. We derive this result through geometry.

We can derive the same result through statistics and you will get those summation summation which you are familiar with. I am not going to go with statistical interpretation of the same thing I will upload my notes on statistical interpretation. But if go into statistical interpretation of this result. It will take at least two weeks of you know I have introduced so many concepts but finally finally you will derive the same results.

We will get a different insight from the statistics view point. See I got a different insight into the same result through geometric view point. Okay algebra did not tell me much you know it just said that derivative=0. Here I can relate to my school geometry that is very, very important okay so that is the beauty of this result. You can actually show that the least square estimate obtained algebraically is nothing but projection of a point onto a subspace.

Okay finding out a vector in a subspace which is at a minimum distance from a point outside the subspace. Okay that is the that is the that is where that is why we work with inner product spaces, Hilbert spaces because you know angle comes you know attached with the inner product. you can talk of orthogonality. Okay all these are not possible in other you know banach spaces. So, that is why Hilbert spaces are so special.

Okay so next lecture we will continue on some more algebraic properties of least square. I will show something more and then we will move to variety of fields engineering applications that is functional approximations and also solving partial differential equations or boundary value problems. We will revisit our problems again through this least square approximation.