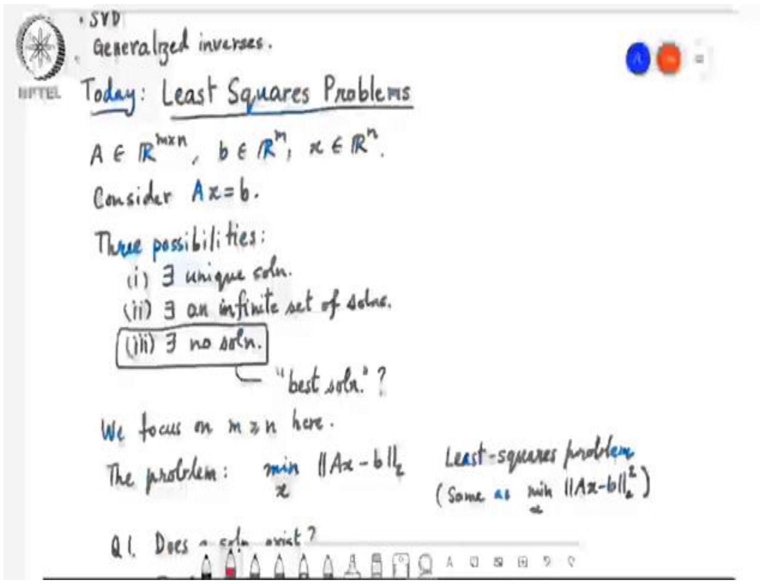
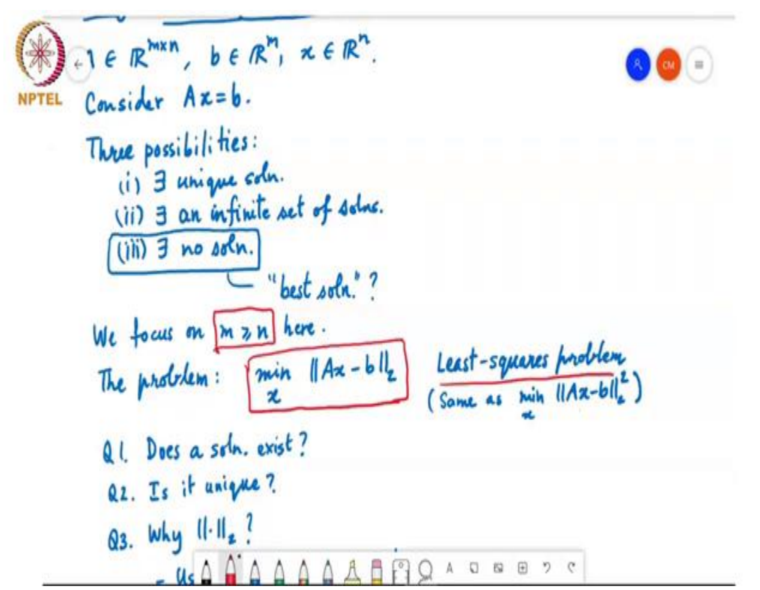
Matrix Theory
Professor Chandra R Murthy
Department of Electrical Communication Engineering
Indian Institute of Science Bangalore
Least squares

(Refer Slide Time: 0:15)





So, the last time we looked at the singular value decomposition and generalized inverses. Today, we will connect these ideas of generalized inverses to least squares problems. So, what are these least squares problems? So, if you consider a matrix A of size m by n and B, a vector B of size of length m and a vector x of length n, then consider the problem a Ax equals b.

So, if you want to solve this problem, this is basically m equations in n unknowns. And there are three possibilities regardless of what m and n are. Either there can exist a unique solution, or there could exist an infinitely large set of solutions, or there could be no solution, there is no nothing in between.

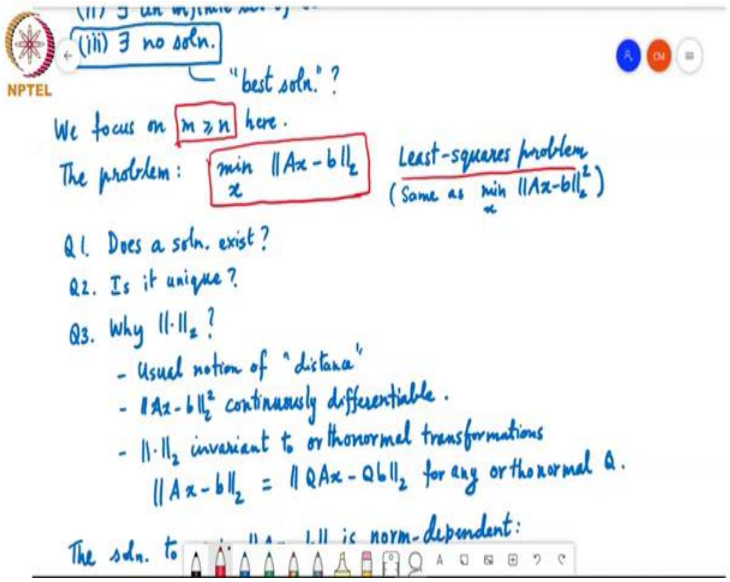
So, for example, you can never have two solutions, exactly two solutions. So, in the case where there is no solution, it is natural to ask what would be the best possible solution you can find. And, so the, so we will, in this part of this course, we will focus on the case where m is greater than or equal to n.

So, you have more equations than you have unknowns, the case m less than n is something that I cover, in some sense in the course on compressed sensing. So, m is greater than or equal to n here, and we will focus on this problem, minimize with respect to x, the l2 norm of Ax minus b, this is called the least squares problem because the, this norm square is the same.

So, if you take the, so, minimizing the l2 norm of Ax minus b is the same as minimizing the square of the l2 norm because squaring is a monotonic function. So, whatever x minimizes this will also be the x that minimizes this cost function. And this itself is the sum of the squares of all the individual entries in Ax minus b. And so that is why it is called the least squares problem. We are finding the solution that achieves the least square error in this Euclidean norm sense between Ax and b.

So, there are several questions that one can ask one is, does this have a solution? Can you solve this problem? Second, is that solution unique? Or are there other solutions, which are equally good? And the third point, of course, is why are we choosing to use the l2 norm here? So the third question is the easiest to answer and it is just going to be a, it is just going to be a completely hand waving answer.

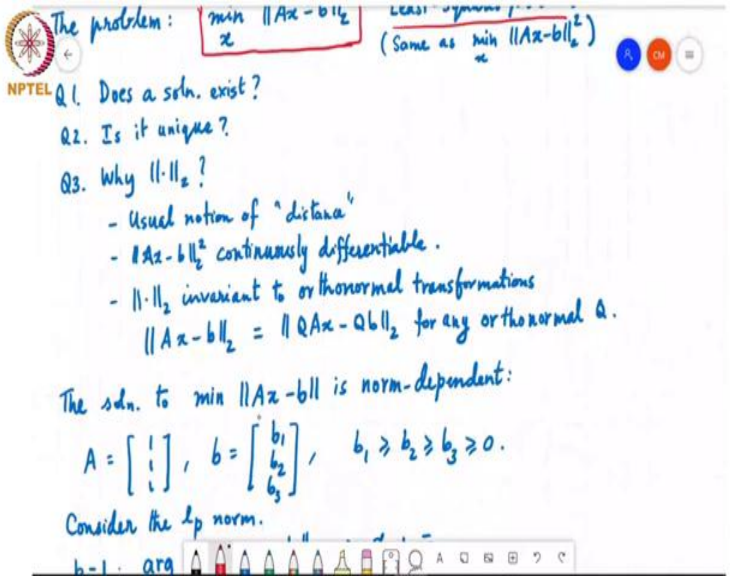
(Refer Slide Time: 3:09)



The answer is that l2 norm is kind of the most logical or easiest to understand notion of distances, the Euclidean distances, the one we are most familiar with. But more importantly, from an analytical point of view, Ax minus b square is a continuously differentiable function of x. So for example, you can find its derivative and so on.

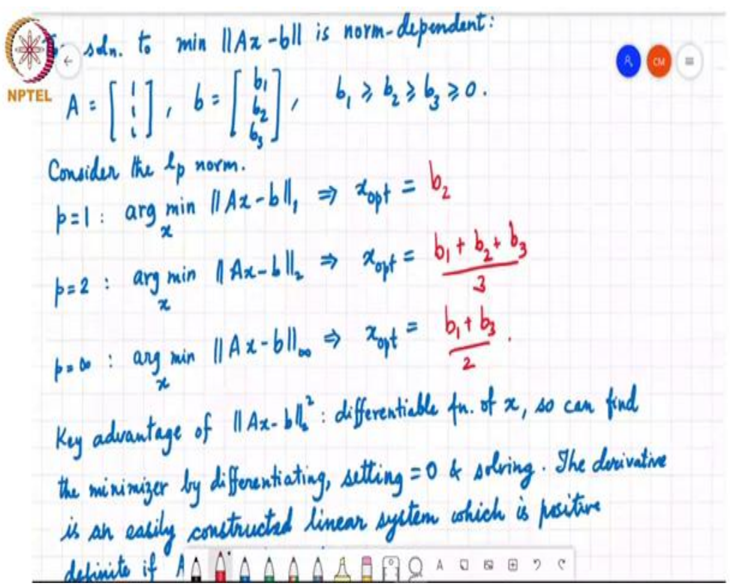
And so it is amenable to optimization much more so than other cost functions you could imagine. And thirdly, this l2 norm is invariant to orthonormal transformations. So, Ax minus b l2 is the same as Q times Ax minus Qb l2 for any orthonormal Q, we will see that this is actually, you know, very useful for, very useful property, for simplifying this problem that we want to solve.

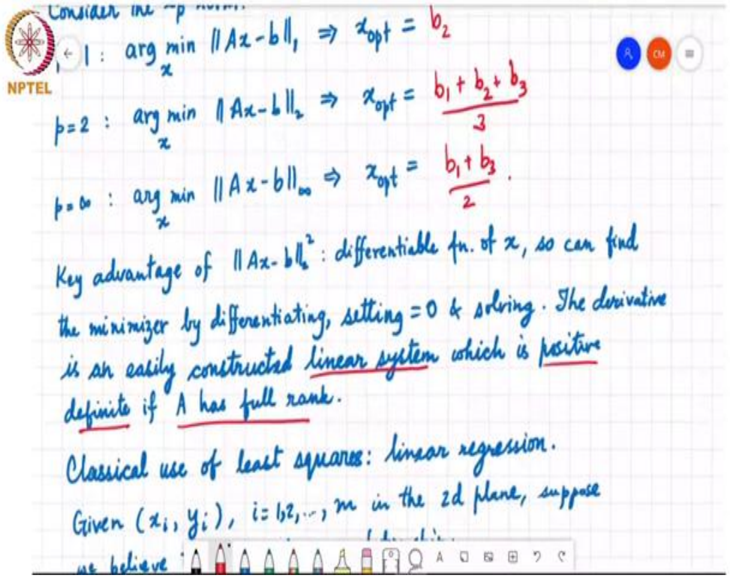
(Refer Slide Time: 4:06)



So, before I actually go about solving this problem, I want to make one small point that the solution to minimizing a norm of Ax minus b is indeed dependent on the norm. So, take a very simple example, where A is a 3 by 1 vector of all ones, and B has 3 components b1, b2 and b3, where b1 is greater than or equal to b2 is greater than or equal to b3, which is greater than or equal to 0. So, they are positive numbers with this ordering.

(Refer Slide Time: 4:35)





Now, let us look at the lp norm. When I take p equals 1, what is the, so x here is a scalar. Because this is just a number, this times this vector, you want it to be as close to be as possible. So if I am looking at the x that minimizes the l1 norm of Ax minus b, so what is that x? What is the optimal x that minimizes this?

What you can see is that it will, when I do Ax minus b, I will get x minus b1, mod of x minus b1 plus mod of x minus b2 plus mod of x minus b3. If you look at this for a few minutes, given that b1 is greater than or equal to b2 is greater than or equal to b3, which is greater than or equal to 0, you can reason out that the x that minimizes this should be equal to b2, because if I use any other x, then all three terms will be greater than 0. But the first and last term will add up to a constant value if the x is between b1 and b3. If x is beyond b1 and b3, it will actually be even higher.

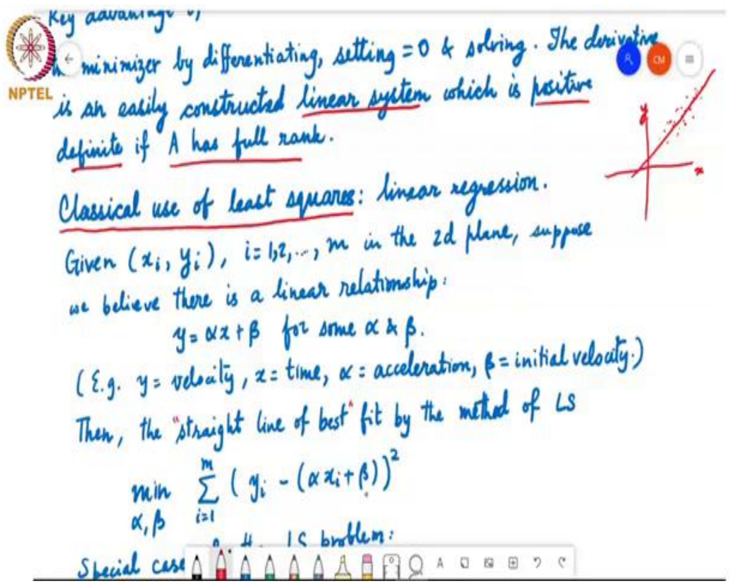
If it is, for example, if x is bigger than b1 or less than b3, the total cost will be even higher. But if it is between b1 and b3, mod x minus b1 plus mod of x minus b3 will be a constant, and this will be greater than mod of x minus b2 will be greater than 0 unless x is equal to b2. So the solution is x optimum is equal to b2.

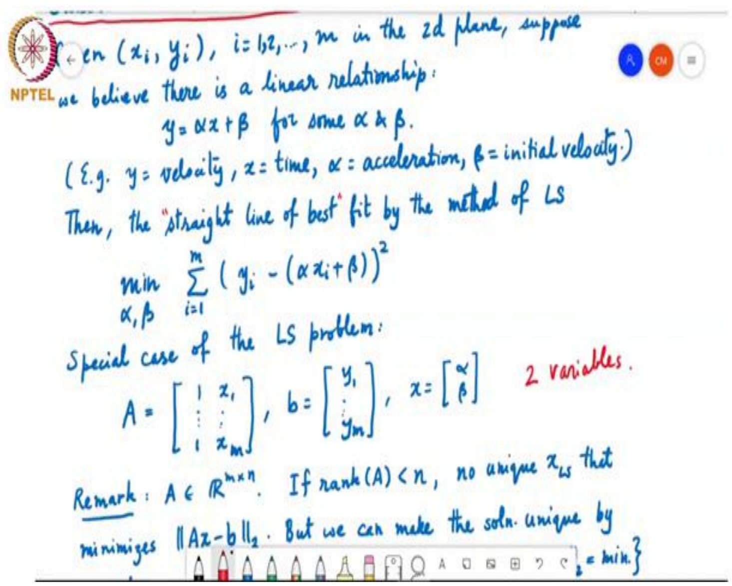
Similarly, if I take the two norm, this is x minus b1 square plus x minus b2 square plus x minus b3 squared. If you just differentiate that with respect to x and set it equal to 0 and solve, you can easily show that the solution is b1 plus b2 plus b3 over 3, it is just the average, it minimizes the mean squared error between these two.

And if I take the L infinity norm, I am looking at mod of, I am looking at the smallest among mod of x minus b1 plus and mod of x minus b2 and mod of x minus b3. And this, you can see, you will have to reflect on it for a minute. But you can show that this is minimized by choosing x optimum equals b1 plus b3 over 2.

So, this is just to illustrate that the solution to the problem does depend on which norm you are considering. But the key advantage of x minus b square is that it is a differentiable function of x. So we can find minimizers by differentiating setting equal to 0 and solving and this derivative it will turn out and I show it to you in a minute is an easily constructed linear system. And this linear system is positive definite if A has full rank. So, that is the property that we will use.

(Refer Slide Time: 7:55)



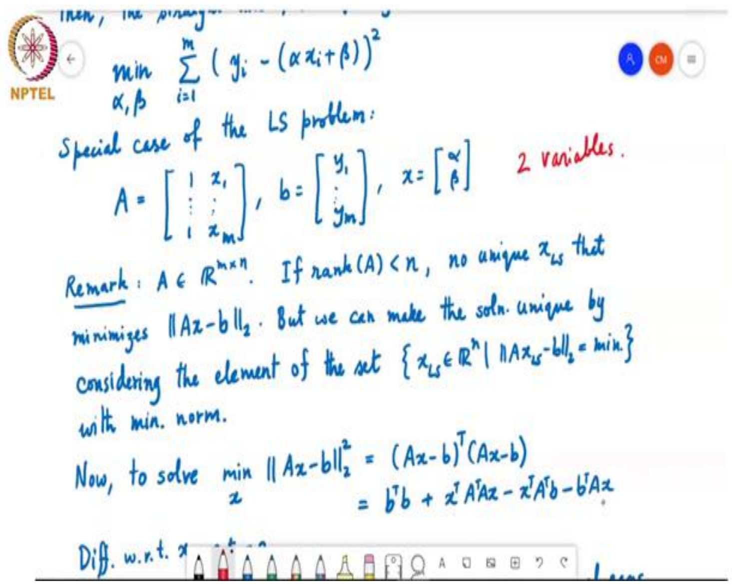


So, before I solve this least squares problem, I just want to mention that these least squares problem is when the origin of it is in linear regression. So, here, you are given points in the two dimensional plane, x1, y1, x2, y2, etc. So, these points could be some points I do not mean to draw them on a straight line, they could be all over here. And what we believe is that there is a linear relationship between x and y and so we want to find a line that fits these points, which explains this relationship between x and y and so we believe that there is a linear relationship of the form x equal to alpha y plus beta for some alpha and beta.

So, for instance, if x is the acceleration, so if alpha is the acceleration, x is the time and y is the velocity, the velocity at time x is going to be alpha times x, plus the initial velocity, which is the velocity when x is equal to 0. And so that is the linear relationship. So, we go out there and we see some vehicle moving and we record the time and the velocity at that time. And then we want to infer this relationship between the velocity and time and determine the parameters alpha and beta.

So, the straight line of best fit the so called straight line of best fit between x and y can be obtained by this method of least squares where you ask what is the alpha and beta that minimizes the mean squared error between yi and alpha xi plus beta. So, this in turn is actually a special case of this general least squares problem we just put down, where I just said, A to be the matrix with all ones as its first column, x1 to xm as its second column. B is the vector containing y1 to ym as its entries. And x is a two dimensional vector with alpha and beta as its two entries. This just has two variables.

(Refer Slide Time: 10:33)

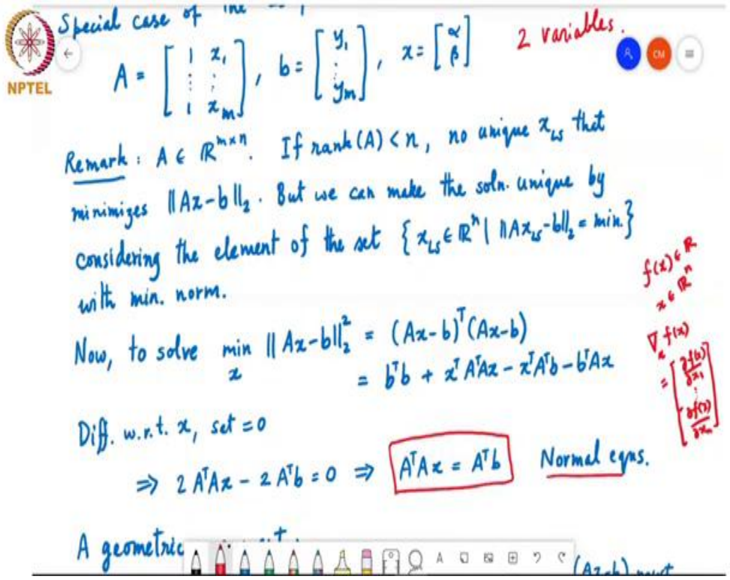


So, another small remark is that, if I look at this problem Ax minus b l2, if the rank of A is less than the number of variables in x, then this, the solution to this problem is not unique, but we can make up the solution unique by considering among all the solutions, that is all the xLS for which this norm of Ax LS minus b l2 is the minimum among all such vectors, which is the one with the minimum norm. And this is something I will talk about in just a few minutes.

But, so there is a way to find a unique solution, but it is unique in the sense that it is the least length least square solution. Now, coming back to our problem, our problem was to minimize the norm of Ax minus b square with respect to x. Of course, this would be the same as minimizing norm of Ax minus b cube Ax minus b to the 4, or in fact, any monotonic function of the norm of Ax minus b.

But square turns out to be very convenient for us. So that is what I will consider here. This is nothing but Ax minus b transpose times Ax minus b. And as usual, if we expand it out, we get b transpose b plus x transpose A transpose Ax minus x transpose A transpose b minus b transpose Ax.

(Refer Slide Time: 11:57)

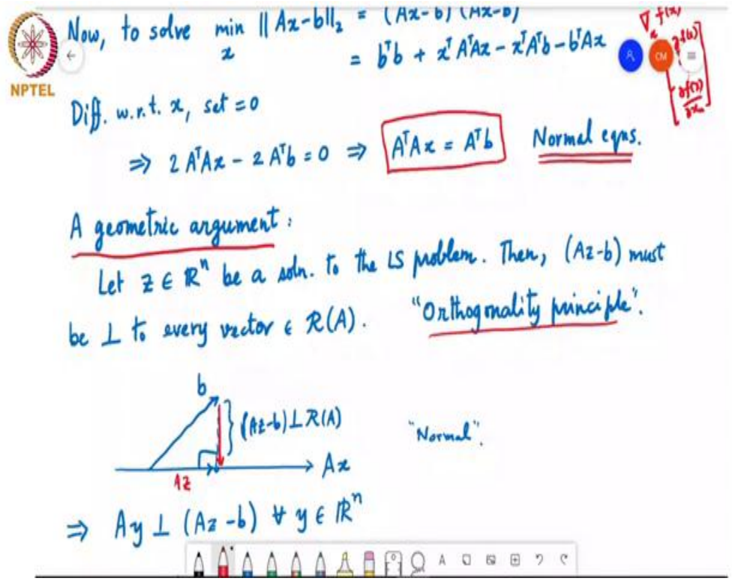


So, now what we do is we differentiate with respect to x. And what I am writing here is actually the vector derivative of this, vector derivatives work in a very similar way as scalar derivatives. But unfortunately, I do not have the time to teach you a module on how to find vector derivatives.

But I will just mention that, if I have f of x is a scalar function of a vector x, then we know that the gradient with respect to x, f of x is equal to the vector containing the partial derivatives. So, this is how you differentiate with respect to a vector. And if you apply this idea, you can show, it is actually very simple, it is very elementary to show these things. But the derivative of x transpose A transpose Ax is 2 times A transpose Ax. And the derivative of the sum of these two terms is the 2 times A transpose B.

So, if we set the derivative equal to 0, we get that A transpose Ax equals A transpose B. So, basically, we need to solve for x that satisfies this problem. So this is a, this is again a linear system of equations. And in fact, they are called normal equations.

(Refer Slide Time: 13:36)



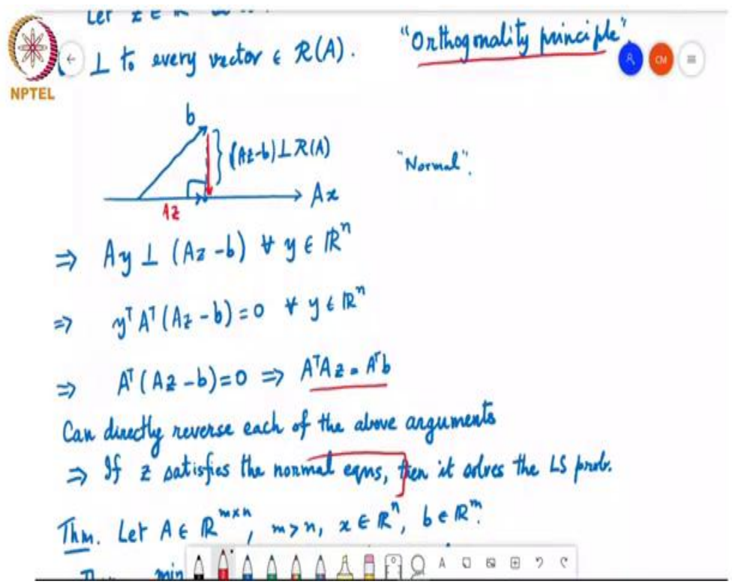
So, in order to see why these are called normal equations, we will go through a small geometric argument, oops. So, what I will show is that if Z is a solution to the least squares problem, then the error vector Az minus b must be perpendicular to every vector which belongs to the range space of A.

This idea is called the orthogonality principle. So, and also the fact that Az minus b must be perpendicular or normal to every vector belonging to R of A is the reason why this set of equations are called normal equations, because they represent this fact of normality between Az minus b and any vector in the range space of A.

So, pictorially what it looks like is this. So, if I take Ax, right now I am for simplicity denoting the space spanned by Ax which is the range space of A by a straight line and b is some other vector, which may not lie in the range space of A, if b lies in the range space of A you know that you can find an x such that Ax exactly equals b and then norm of Ax minus b square will be equal to 0. So you will achieve no, you will, there will be no residual error once you find the least square solution.

But B is outside the range space of A, essentially what we are doing is to project b on to the range space of A. And what is left over is this error. This, this is the A, this is Az. This is a z. And Az minus b is the residual that is this vector and these two vectors, Az and Az minus b. In fact, any vector in the range space of A has to be perpendicular to this Az minus b. Otherwise, you can possibly improve your solution by moving, by picking a different Z.

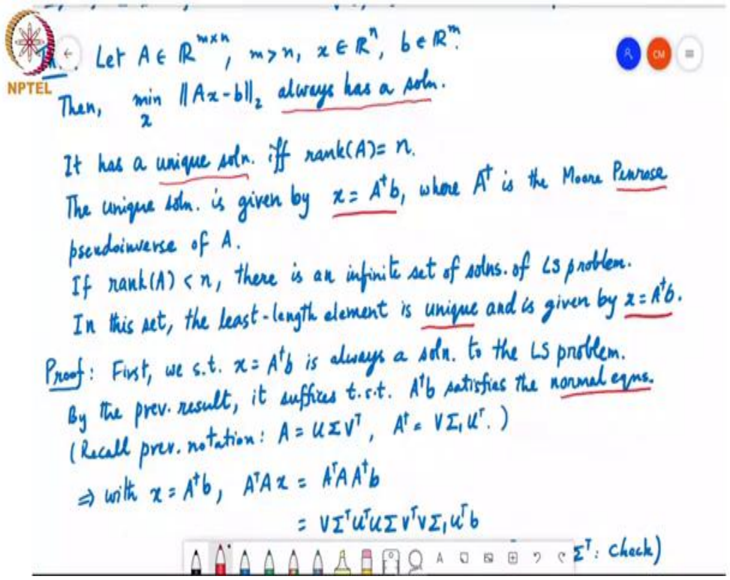
(Refer Slide Time: 15:43)



So that is the basic idea, simple geometric argument. So in other words, what we are saying is Ay should be perpendicular to Az minus b for every y belonging to R to the n, which is the same as saying y transpose A transpose is the transpose of this times this which is nothing but the inner product between Ay and Az minus b must be equal to 0 for every y belonging to R to the n.

Now, if this is true for every y belonging to R to the n, it means that A transpose times Az minus b itself must be equal to the 0 vector, which is the same as saying A transpose Az equals A transpose b, so which actually brings us back to the normal equations. And in fact, all of these arguments are reversible. And that is the reason why these equations, the set of equations, A transpose Az equals A transpose b are called normal equations. So if Z satisfies these normal equations, then it solves the least squares problem.

(Refer Slide Time: 16:51)



So, we have the following theorem. So let A be in R to the m by n and say m is greater than n, I could even choose m greater than or equal to n, and x is in R to the n and b is in R to the m, then the problem of minimizing the l2 norm of Ax minus b with respect to x always has a solution. So there is no case where you cannot solve this problem, you will always be able to solve it.

It has a unique solution, if and only if A is full rank that is m is greater than n here, so it has full column rank. And this unique solution is given by x equal to A dagger b, where A dagger is the Moore Penrose pseudo inverse of A. So I think somebody (())(17:39) up and maybe was asking whether Penrose is the same guy who got the Nobel Prize.

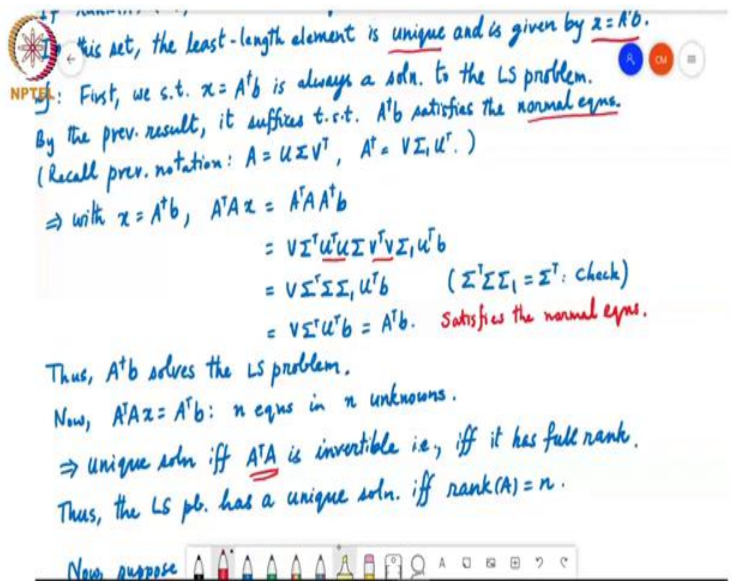
And yes, I checked that. Indeed, Penrose is the guy who got the Nobel Prize in Physics in 2020. So A dagger is the Moore Penrose pseudo inverse of A. And if rank of A is less than n, so A is rank deficient, then there are an infinite set of solutions to the least squares problem. However, in this set, the least length element is unique. And it is still given by A dagger times B. So this A dagger b turns out to be a very beautiful solution.

It is always the solution to this least squares problem, it is a unique solution if rank of A equals n. But if rank of A is less than n, it is the unique least length solution to this problem. There will be infinitely many solutions for this problem, but the least length one is given by A dagger times b. So the way we show this is first, we will show that A dagger b is always a solution to the least squares problem by using this geometric argument.

And then by, so, in order to show this it suffices to show that A dagger b actually satisfies these normal equations. So, we arrived at the normal equations also by differentiating this norm of Ax minus b square and setting it equal to 0. So, if it solves the normal equations, it does minimize the, solve this optimization problem.

Also, recall that in the previous class, we wrote down these pseudo inverses and we said if A is u sigma v transpose, we can write A dagger to be v sigma 1 u transpose where sigma 1 is the matrix which has its top left r cross r entries as the inverses as the of the r cross r diagonal matrix, which is the top left diagonal matrix in sigma containing the nonzero singular values and everything else being equal to 0. And this is of size m by n, sigma 1 is of size n by n.

(Refer Slide Time: 20:10)

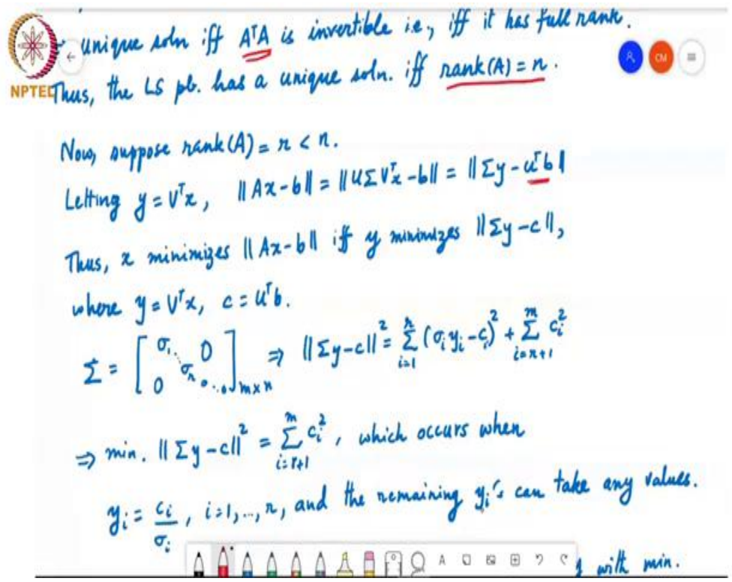


So, now if I let x equal to A dagger b, then and I look at what happens to A transpose Ax, that is the same as A transpose AA dagger b. And if I substitute these two formulas, A transpose is v sigma transpose u transpose A is u sigma v transpose, and A dagger is v sigma 1 u transpose times b, u transpose u is the identity matrix and v transpose v is the identity matrix. So, that is the same as v sigma transpose sigma, sigma 1 u transpose b and sigma 1 is the pseudo inverse of sigma. And so, you can check that sigma transpose sigma sigma 1 is just sigma transpose.

And so, what we have then is v sigma transpose u transpose times b, which is nothing but A transpose b. So, A transpose Ax equals A transpose b when you let x equal to this A dagger times b, so, A dagger b solves the least squares problem, it satisfies the normal equations. Now, if I look at this system, A transpose Ax equals A transpose b, this is a system of linear equations, n linear equations in n unknowns, A is m by n. So, A transpose A is of size n cross n. So, this is n equations and n unknowns.

And we know that this will have a unique solution if and only if the matrix A transpose A is invertible, which is true if and only if rank of A equals n. So, the least square least squares problem has a unique solution if and only if rank of A equals n.

(Refer Slide Time: 22:05)



Now, suppose rank of A is some number r, which is less than n. So if we let y equal to, we define y to be v transpose x, then we can write Ax minus b as, so going forward, I am not going to write the two norm everywhere. But this is all for the l2 norm. So, this is the same as norm of u sigma v transpose times x minus b and v transpose x is y. So, I can write this as sigma y minus u transpose v.

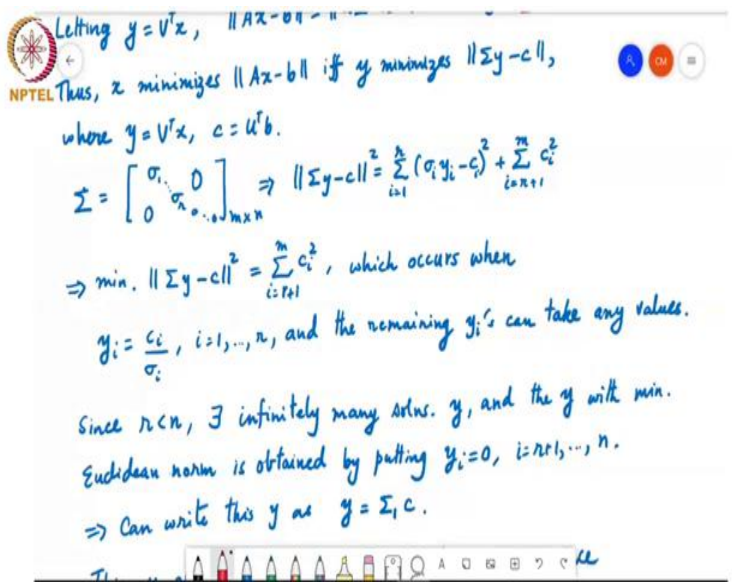
So, here I am using this property that u is a unitary matrix. So, multiplying this whole thing by u transpose will not change this norm. But if I multiply by u (())(22:50). So, what I was saying is that if the rank of A is less than n, then we can write norm of Ax minus b as norm of sigma y minus u transpose b. And here I am using the fact that u is a unitary matrix. And that is why and so multiplying this thing by u transpose does not change the value of the norm.

So, what this means is that x will minimize the norm of Ax minus b, if and only if y minimizes the norm of sigma y minus c, this is a unitary transformation, it is one to one. So, this x will minimize norm of Ax minus b, if and only if y minimizes sigma y minus c, where c is this matrix, this vector u transpose b.

But this is a beautiful thing, simple systems. So, it is very easy to see what is going on here. If I write sigma to be sigma 1 to sigma r and 0s everywhere else, so these are the top left r cross r block and it is the diagonal matrix. Then if I expand sigma y minus c square, that is going to be sigma i equal to 1 to r sigma i yi minus ci square plus from r plus 1 to n sigma is 0, so it will just be ci square.

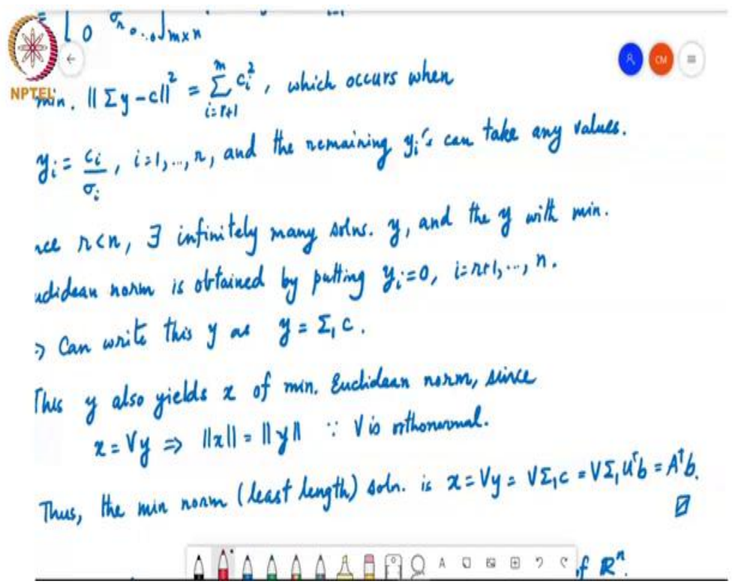
So, now if I look at what happens as I choose different possible values for yi, I can see that by choosing yi equal to ci over sigma i, for the first r values of yi, I can make these terms 0, but y does not touch these terms. So the minimum value of sigma y minus c square is equal to this second term here sigma i equal to r plus 1 to m ci square. And it occurs when yi is ci over sigma i. And the remaining yi can take any possible, any values we wish.

(Refer Slide Time: 24:46)



And since r is less than n, we can have infinitely many solutions for y. But the one with minimum Euclidean norm is obtained by putting yi equals 0, i equal to r plus 1 to n, and in fact, we can write this y as sigma 1 times c, because the norm of y is just the sum of all these chi square plus yr plus 1 square plus etcetera. And the norm of yr plus 1 square plus etcetera can be minimized or the sum of yr plus 1 plus yr plus yr plus one square plus yr plus 2 square plus etcetera up to plus yn square can be minimized by choosing all of those guys equal to 0.

(Refer Slide Time: 25:31)



And this actually yields the solution of minimum Euclidean norm since the norm of x equals the norm of y. So whatever y minimizes the norm of y is also the x that the corresponding x is actually the x with minimum Euclidean norm. That is because v is an orthonormal matrix. So the minimum norm, least length solution is x equal to Vy and y itself is equal to sigma 1 times c. So and c itself is equal to u transpose times b. So V sigma 1 u transpose is nothing but A dagger. So the least length solution is A dagger times b, which completes the proof.