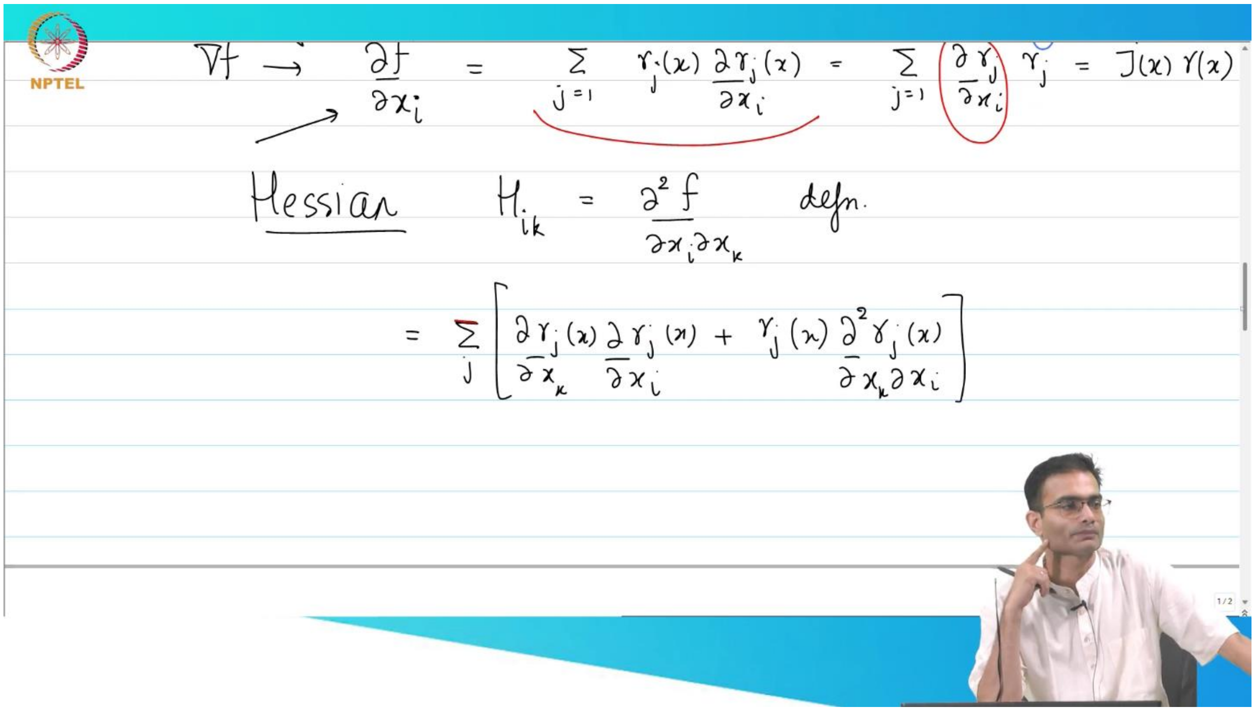
Course Name: Optimization Theory and Algorithms
Professor Name: Dr. Uday K. Khankhoje
Department Name: Electrical Engineering
Institute Name: Indian Institute of Technology Madras
Week - 08
Lecture - 57

Linear least squares - Part 2

Ok, once the gradient is done, what else do I need? Since we have previously studied Newton's method, what would I also need? Look at the Hessian, why not? Right, it looks... does not look very difficult. Now, Hessian again, I have I am keeping j for the way the index of summation. So, I am going to use i and k so that I do not get confused. This is the definition of the Hessian. So, my starting point again is this red expression over here, I have to take two derivatives, right, or I can take this guy and then do what? Take one derivative with respect to x_k, and then I will be done. Ok, so let us do that, that will be simpler.



$$\nabla f \longrightarrow \frac{\partial f}{\partial x_i} = \underbrace{\sum_{j=1} r_j(x) \frac{\partial r_j(x)}{\partial x_i}}_{} = \sum_{j=1} \boxed{\frac{\partial r_j}{\partial x_i}} r_j = J(x)\, r(x)$$

Hessian $\quad H_{ik} = \frac{\partial^2 f}{\partial x_i \partial x_k}$ defn.

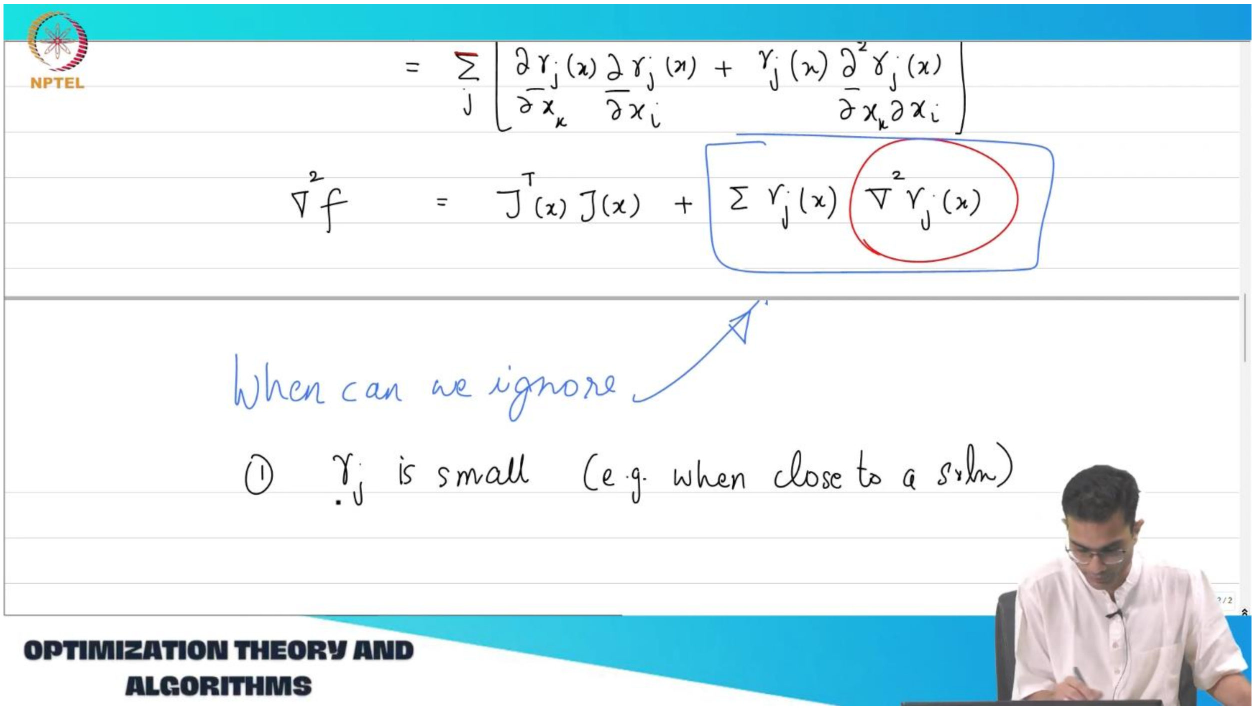
$$= \sum_j \left[\frac{\partial r_j(x)}{\partial x_k} \frac{\partial r_j(x)}{\partial x_i} + r_j(x) \frac{\partial^2 r_j(x)}{\partial x_k \partial x_i} \right]$$

So, this is going to be equal to. So, two terms over here, let me underline it once again, this is what I am looking at. I need to take the derivative with respect to x_k. Ok, so this summation, the summation over j, will remain. Ok, so

$$\frac{dr_j(x)}{dx_k}\frac{dr_j(x)}{dx_i}$$

The first term is as it is and then the second term becomes $r_j(x)$. The first term again looks like something that can be simplified.

What does it remind? I mean, if you got the gradient, this should be just as easy. The first term on the Hessian, what should it, does it look like the product of two matrices? Yeah, what? The Jacobian, that is clear that they are the product of two Jacobian matrices. The only thing that we have to get right is whether the order, whether it is transpose or not, right. So, no, nothing very great over here. I am going to give you the final answer.



$$= \sum_j \left[\frac{\partial r_j(x)}{\partial x_k} \frac{\partial r_j(x)}{\partial x_i} + r_j(x) \frac{\partial^2 r_j(x)}{\partial x_k \partial x_i} \right]$$

$$\nabla^2 f = J^T(x)\, J(x) + \boxed{\sum r_j(x) \left(\nabla^2 r_j(x)\right)}$$

When can we ignore

① r_j is small (e.g. when close to a s.ln)

This is simply

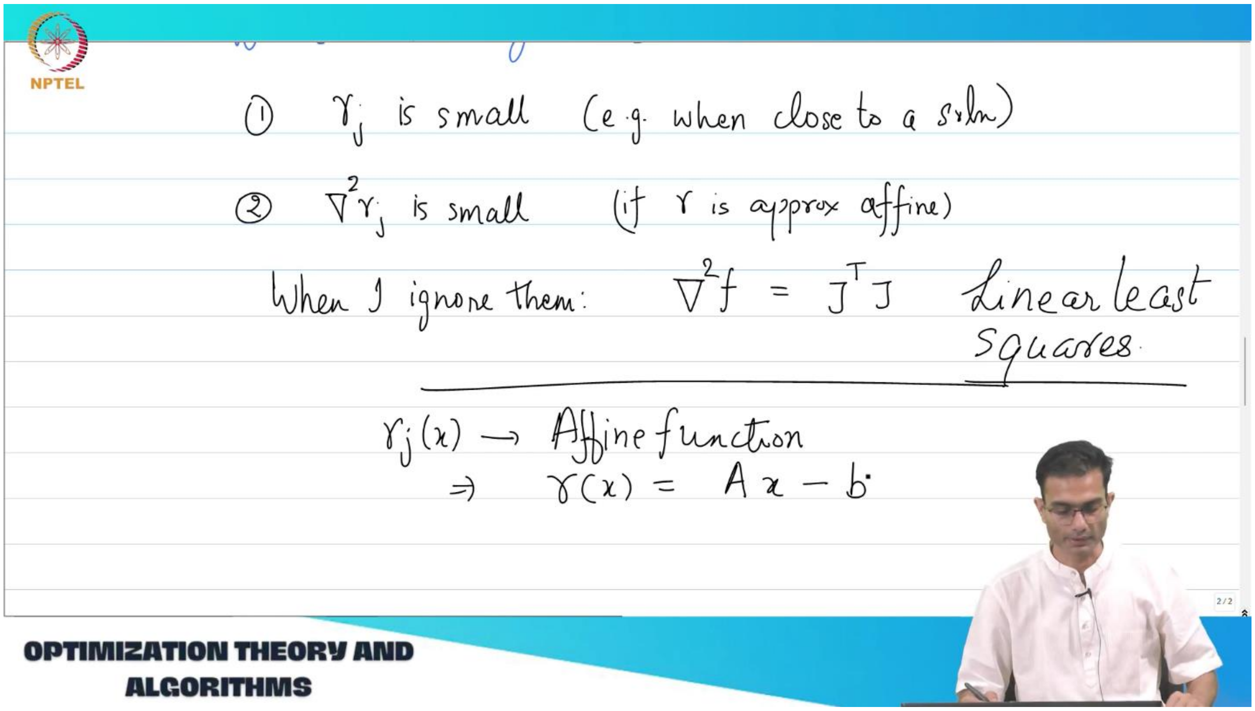
$$J^T(x) \times J(x)$$

Oops, this is the first term. The second term is just going to be, I am just going to write it like this. Ok, so $r_j(x)$ and the second term, what would you call it? This is the Hessian of $r_j(x)$, right. So, we have got our gradient, we have got our Hessian.

Hessian surprisingly did not end up looking very, very complicated. You have two terms over here, right. So, here is where the literature of least squares problems kind of splits into two different approaches. Ok, and why it splits into two different approaches is simply because one is easier to do than the other. Why again? Because of the presence of a second order term over here. Ok.

So, if I were to ask you if you look at the second, this second expression over here, the Hessian has two terms, one is the product of Jacobians, the second is a Hessian over here multiplied by this. If I wanted to ignore the blue term, ok, so there are two terms, this is engineering. So, we are trying to see under what situations can the second term be ignored, that is what I am asking. When can we ignore this blue term? There basically there seem to be two options.

First option is what? What would you say? It is a product of two terms, either the first term is low or the second term is low, there is no other possibility, right. $r_j(x)$ is small, that is one possibility. If $r_j(x)$ is small, again in plain English, what would I say? I am somewhat near the solution. If I am near the solution, the residual is going to be low. So, if my... if I think of this as an iterative process and I am arriving close to my solution, $r_j(x)$ is going to be small. Does that mean that, does not necessarily say anything about the Jacobian, Jacobian is going to be whatever, right?



① r_j is small (e.g. when close to a soln)

② $\nabla^2 r_j$ is small (if r is approx affine)

When I ignore them: $\nabla^2 f = J^T J$ Linear least squares.

$r_j(x) \to$ Affine function

$\Rightarrow r(x) = Ax - b$

So, $r_j(x)$ is small, this for example, when close to a solution. Ok, very good. Any other possibility? The second term, right. So, when does this happen? When will $\frac{\partial^2 r}{\partial x^2}$ be small? I mean you already know the answer, if you were to take a common sense guess, when is this second term going to be small in magnitude? What would you say? Supposing your model was linear, if your model was linear, then there is no second order term that is identically 0, that is the trivial case. But if it were sort of weakly linear, right?

So, for example, if $r(x)$ is approximately affine or linear, then the second derivative of it is going to basically be either 0 or very close to 0, right. So, these are the two possibilities when we can ignore these terms. Ok, and when I ignore them, what do I get? I get that the Hessian simply becomes

$$J^T J$$

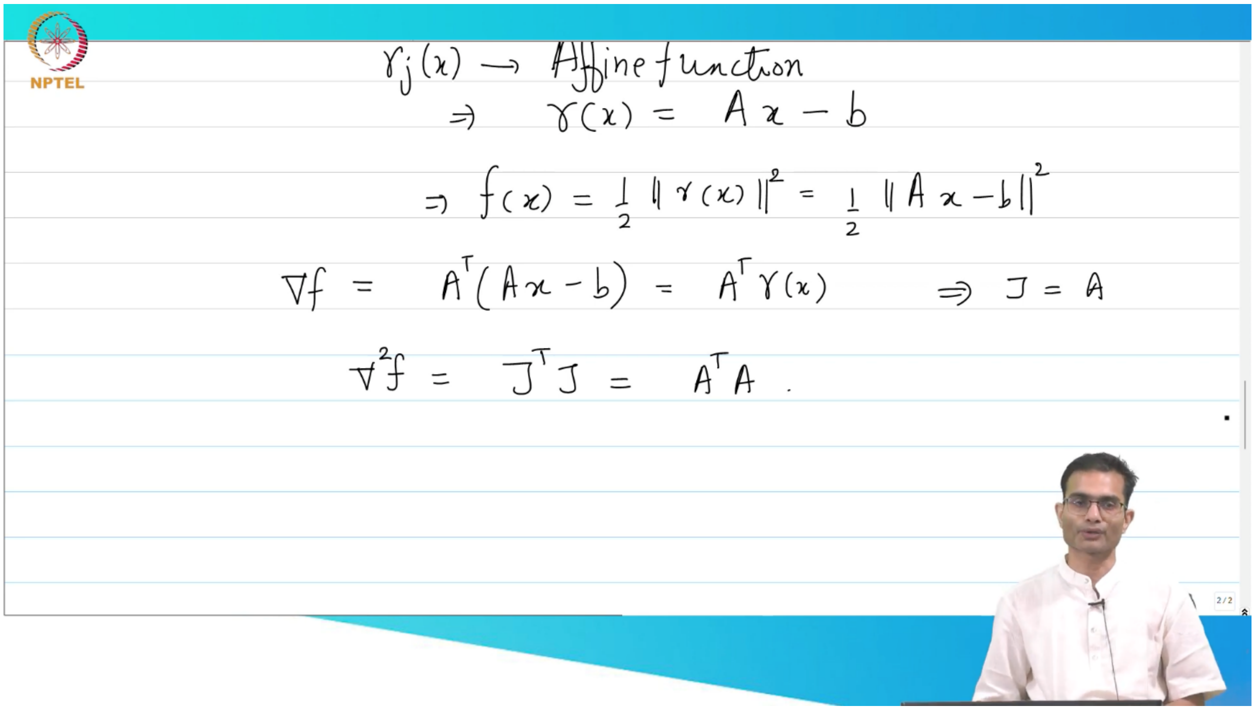
Ok, because the second term is gone, and this is what is called in the literature linear least squares. We will also talk about non-linear least squares when the second term cannot be ignored. Ok.

As you can imagine, there is computationally more work involved in solving a non-linear least squares problem. So, before we go to non-linear least squares, we will spend now some time looking at linear least squares problems because they appear so often and we have already spent some part of the course looking at linear least squares problems. For example, when you solve $Ax = b$, $Ax = b$, I wrote it as

$$\| Ax - b \|$$

for example, that is a linear, the model is linear, and I am trying to minimize it. So, it is a least square minimization problem.

So, I have already looked at it, but what I will do is I am going to put it into the language of that that we have just discussed so far. Ok. And we will see, we will look at for example, this minimum norm solution which I mentioned or how to use the SVD, when can we truncate the SVD, all of these tricks which are found all over numerical analysis, we will have a look at it now. Ok, in the context of linear least squares problems. So, let us look at this now. So, this is an affine function which implies that $r(x)$ is simply something like this



$r_j(x) \rightarrow$ Affine function

$$\Rightarrow \quad r(x) = Ax - b$$

$$\Rightarrow f(x) = \frac{1}{2} \| r(x) \|^2 = \frac{1}{2} \| Ax - b \|^2$$

$$\nabla f = A^T (Ax - b) = A^T r(x) \qquad \Rightarrow J = A$$

$$\nabla^2 f = J^T J = A^T A .$$

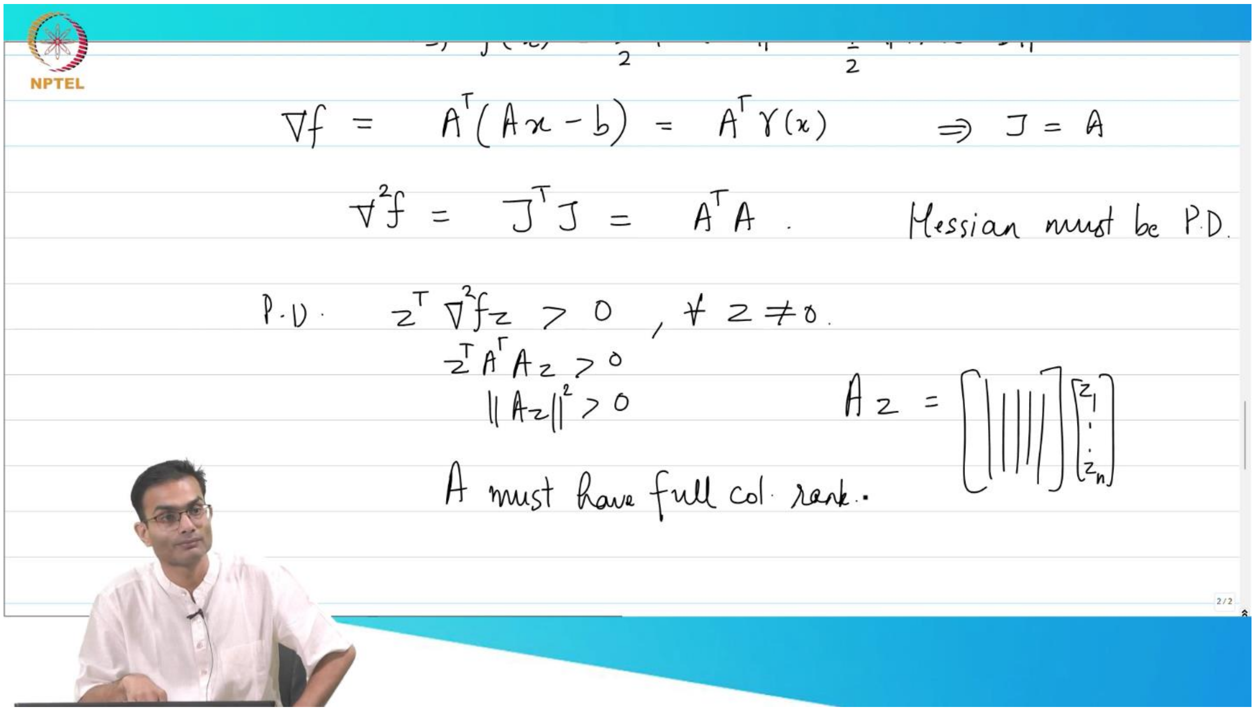
$$r(x) = x - b$$

This is as general a linear expression as I can find, right. So, my objective function I had said was

$$\frac{1}{2} \| R(x) \|^2$$

right.

So, that will continue over here. So, this is just going to be a good kind of a revision for all of you because you are very familiar with all of this now, right. If I take ∇f, I do not need to remember anything of what I did just now. If you look at this expression and calculate ∇f, what will you write? Using what we did in the very beginning of the course, right, taking the gradient of something like this, what should you get? Multivariate chain rule A^T multiplied by you should not forget this. This is how we had from first principles if you remember when we tried to take the derivative, we had done it for the gradient, we had done it for the Hessian, this is ∇f, right. This expression will come so many times, it is useful for you to just sort of keep it in your memory because you know whether you are going to do ML or whatever, this expression will come again and again. Again, it is useful to remember this.



$$\nabla f = A^T(Ax - b) = A^T r(x) \Rightarrow J = A$$

$$\nabla^2 f = J^T J = A^T A .$$ Hessian must be P.D.

P.D. $z^T \nabla^2 f z > 0 , \forall z \neq 0.$

$$z^T A^T A z > 0$$

$$\|Az\|^2 > 0$$

$$Az = \begin{bmatrix} | & | & | & | & | \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}$$

A must have full col. rank.

Now, $Ax - b$ is also my residual. So, I can write this as

$$A^T r(x)$$

Ok, and if I look at the language of least squares over here, where did it go? Here you go. In a linear least squares problem, this was $J^T r$, not surprisingly, here I have $A^T r$. So, it implies simply that the Jacobian is the same as A.

I am trying to take my well-known linear $Ax - b$ problem and fit it into the least squares language. Ok, and that is fine, and then the second term is

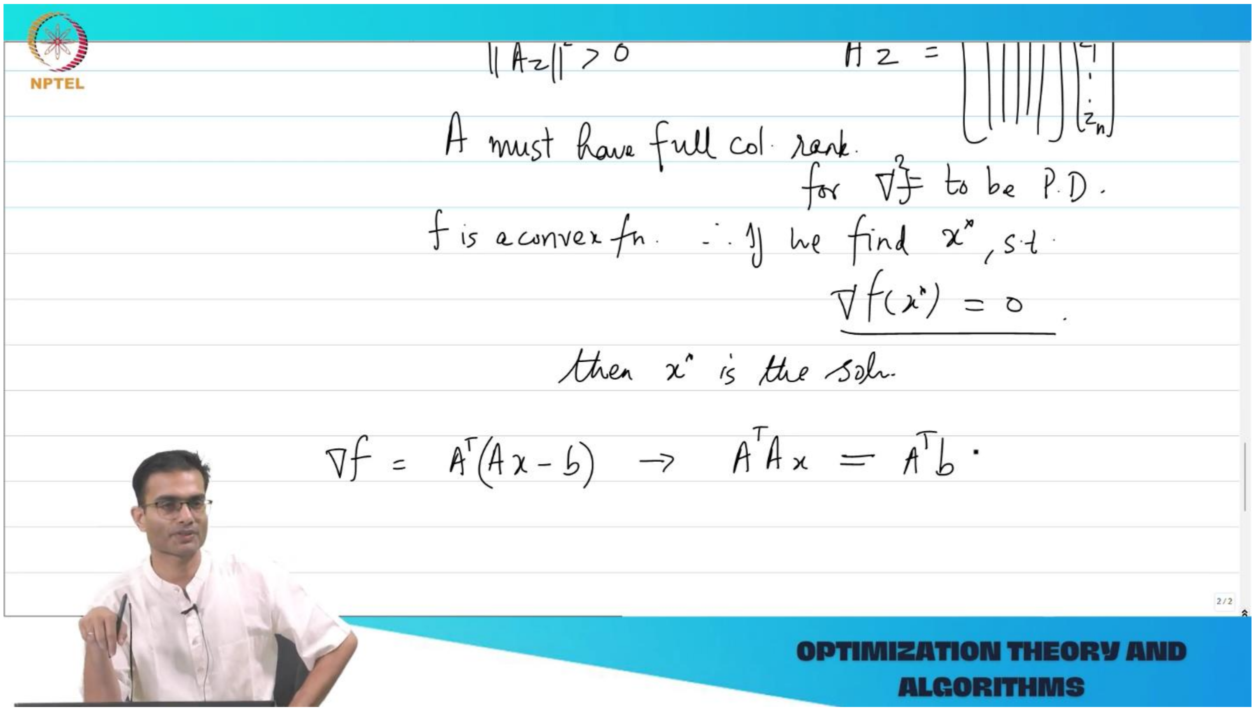
$$\frac{\partial^2 f}{\partial x^2}$$

What will it be? There is r, which is only linear, so the second derivative of r is going to give me 0, right? So, this is $J^T J$, or in other words, $A^T A$. This is what I get, ok. Now, supposing I wanted to do, for example, the Newton method or something like that over here, right.

So, for that, what is the requirement on the Hessian? Positive definite, right? So, Hessian must be positive definite. Hessian must be... is this expression positive definite? I see, I hear one, I hear several yeses, anyone disagree? If A is invertible, ok, I tell you that. Ok, so for invertible, that means what should the size of A be? Square. A has to be square if I say invertible, but I said from the common-sense point of view I want more measurements than parameters. That means A is tall, fat, or square? Tall. So, A is definitely not going to be invertible.

So, A is not invertible. Let us take, is this Hessian positive or...? This, I hope someone will disagree. There is... That is important. It is not a given, right? So, only when A is full column rank will this expression be positive definite. Why? So, let us see for positive definite, what is the definition? Anytime you have positive definite, what do you need to do? Just pick a z, z^T on both sides, right?

So, if it is positive definite, $z^T A^T A z$ should be strictly greater than 0. And what is the important qualifier? For all z? For all z? Let us make it even more precise. There is one more thing.



$\|Az\|^2 > 0$ $\qquad$ $Az = \begin{bmatrix} | & | & | & | & | \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}$

A must have full col. rank. for $\nabla^2 f$ to be P.D.

f is a convex fn. $\therefore$ If we find x^*, s.t. $\underline{\nabla f(x^*) = 0}$.

then x^* is the soln.

$\nabla f = A^T(Ax - b) \rightarrow A^T A x = A^T b$.

Not equal to 0. This is the precise statement. Now, this becomes

$$z^T A^T A z > 0$$

And this becomes

$$\| Az \|^2 > 0$$

Now, if A does not have full column rank, for example, if the columns of A are linearly dependent, then there will exist a non-trivial z for which $A \times z = 0$, and so it will be positive semi-definite.

So, this expression is always positive semi-definite; it is not positive definite. On the other hand, if the columns of A are linearly independent, right, so remember when I am talking, when I say this, I am looking at these. These are the columns, and I have a z_1 up to z_n over here. So, $A \times z$ is also a linear combination of the columns of A; that is one picture of $A \times z$.

If the columns of A are linearly independent, that means there is no linear combination of this which is going to give me 0 because they are linearly independent, right? So, that is why A should have... You can say the columns are linearly independent, or a shorter way of saying it is full column rank, right. So, A must have full column rank. Correct. So, that is a good point.

So, this... So, yeah. So, Ruban has asked a good question. It depends on how you chose the data points. If you had a bad strategy, you could have essentially duplicated duplicate points, for example, or you may still have full column rank, but the condition number may be bad; we will come to that analysis later, right. So, choice of data points is important.

I mean, and this issue is there all throughout numerical analysis. For example, have you heard of this Runge phenomenon? You take an evenly spaced number of points and you try to fit a polynomial. As a simple level of experiment, right, take some function $f(x)$ defined from, say, 0 to 1, and take uniformly spaced points and try to fit a polynomial. Start with linear, quadratic, cubic, and so on. As you start increasing the number of points, what happens? If you... This is a well-known thing; there is a nice Wikipedia page also about it.

As you increase the number of points, the function approximation that you get, it goes through the data points very nicely, but in between those data points, you have a huge fluctuation that happens, right? And that is because you chose those points uniformly. On the other hand, if you choose the points non-uniformly, this oscillating behavior goes away completely. Ok. So, and those non-uniform points, often there is, if you study numerical analysis, what are they called? Anyone has an idea? Wants to take a guess? Chebyshev points, for example, they are called Chebyshev points, different from uniformly spaced points.

So, there is a lot of deep analysis that goes into how to choose your measurement points. It seems naively just take uniformly spaced points; it is often not such a good strategy. So, all of that will come into the numerical properties of A. Ok, and we will come to that a little bit further down the line, but A must have full column rank for it to be positive. Assuming this is the case, is f a convex function? The Hessian is positive definite; is f clearly, you can see it is a convex function, right?

So, f is a convex function, ok. Therefore, if we find some point x^* such that $\nabla f(x^*) = 0$, are we done? Then we are done, ok. So, what is the equation that I need to solve now to finish this problem? $\nabla f = 0$. ∇f you already had an expression; where was it? ∇f was here:

$$\nabla f = A^T(Ax - b)$$

Right, so $\nabla f = A^T(Ax - b)$, and this basically becomes... We have seen this kind of equation before, and we have made one comment about this guy, saying that this has a bad condition number because there is an A^TA. If A has condition number κ, A^TA has condition number κ^2, ok.

These equations are also called the normal equations. Do not ask me why they are called that, I do not know what is normal about them, but they are called normal equations, ok.