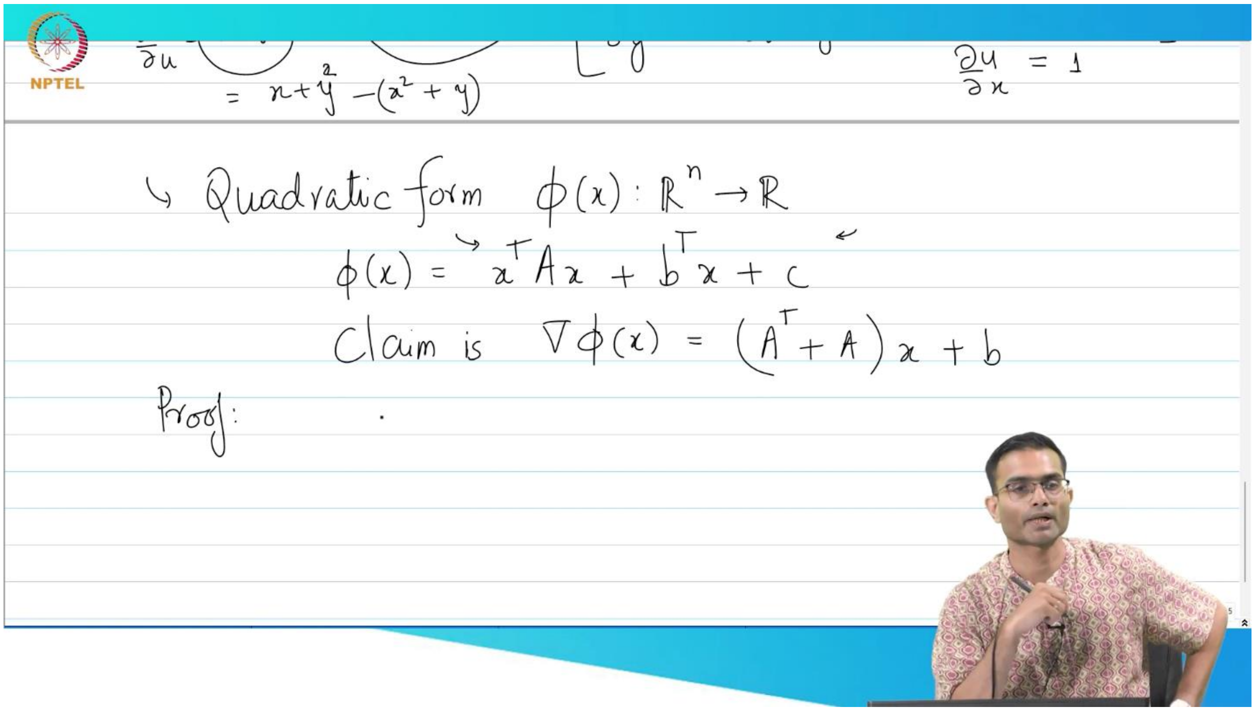
Course Name: Optimization Theory and Algorithms
Professor Name: Dr. Uday K. Khankhoje
Department Name: Electrical Engineering
Institute Name: Indian Institute of Technology Madras
Week - 02
Lecture - 13

Gradient of Quadratic Form and Product Rule

A common symbol for it is $\phi(x)$, which is a function from $\mathbb{R}^n$ to $\mathbb{R}$. A quadratic form is nothing but a generalization of a quadratic equation in one dimension to a quadratic equation in n dimensions; we call it a quadratic form. How do we define it? We have seen this in a previous section. Do I need to define what a, b, and c are, or is it self-explanatory from here? It is self-explanatory, right? c has to be a scalar, b has to be a vector, and A has to be a matrix. Right now, I am not making any special qualification on A; A can be any square matrix.



$= x + y^2 - (x^2 + y)$ $\qquad \frac{\partial u}{\partial x} = 1$

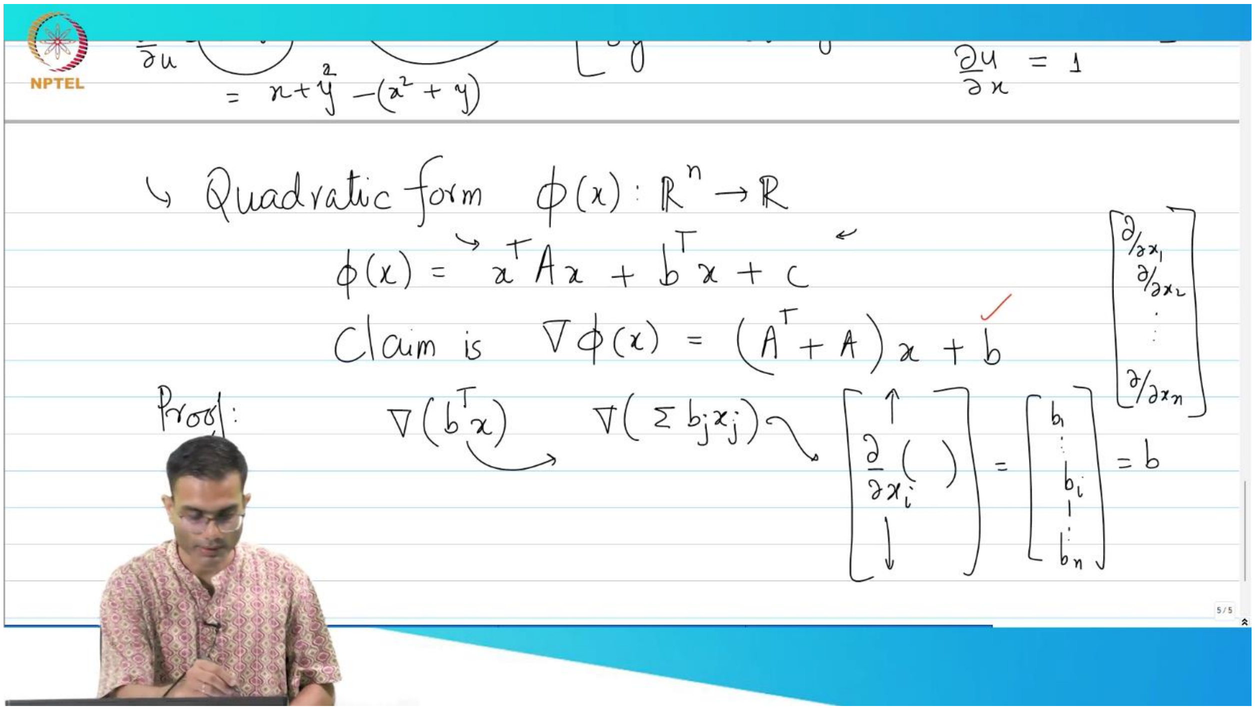
↳ Quadratic form $\phi(x) : \mathbb{R}^n \rightarrow \mathbb{R}$

$$\phi(x) = x^T A x + b^T x + c$$

Claim is $\nabla \phi(x) = (A^T + A)x + b$

Proof:

Now, this is your first opportunity to put your knowledge of gradients into practice. The claim is that, and this is something that we will use throughout the course, the gradient of this expression is given by $A^T A x + b$. Those of you who have done machine learning would be very familiar with this, but we are not going to assume that. Let us prove it. This is one proof that everyone should become very comfortable and familiar with because, as I mentioned, we will use it throughout the semester.



$\frac{\partial}{\partial u}$

$= x + y^2 - (x^2 + y)$

$\frac{\partial u}{\partial x} = 1$

↳ Quadratic form $\phi(x) : \mathbb{R}^n \to \mathbb{R}$

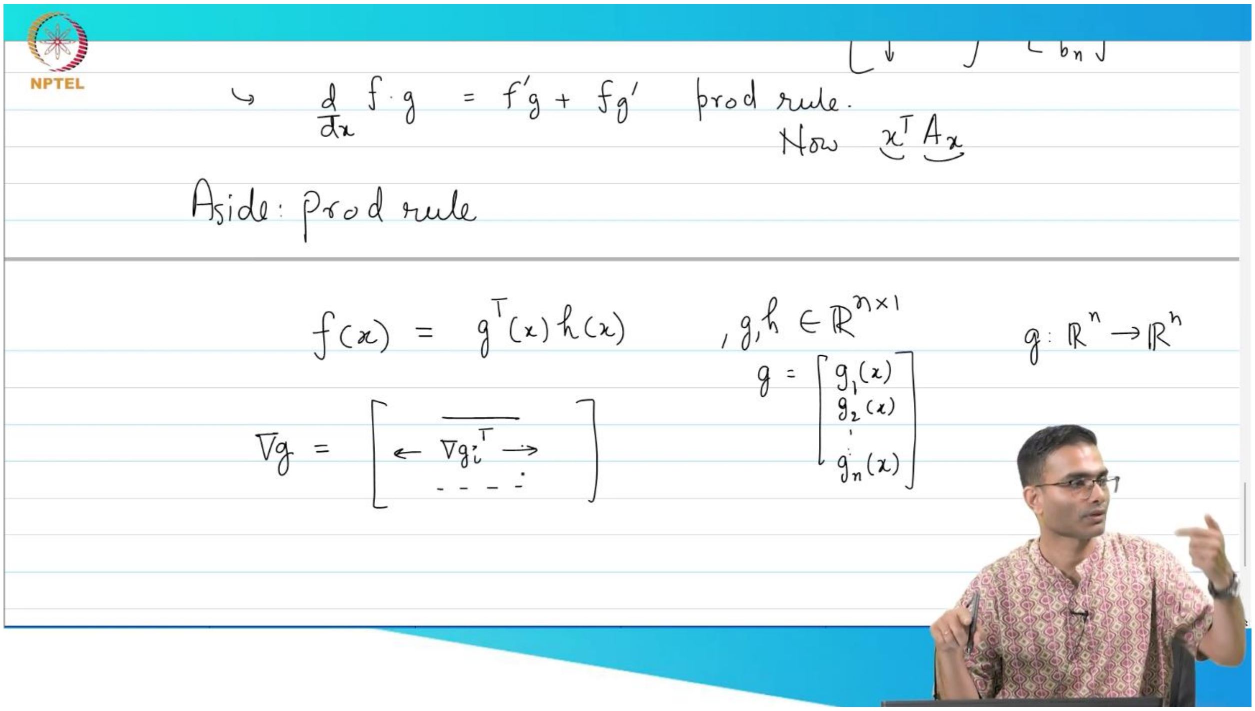
$\phi(x) = x^T A x + b^T x + c$

Claim is $\nabla \phi(x) = (A^T + A) x + b$ ✓

$$\begin{bmatrix} \partial/\partial x_1 \\ \partial/\partial x_2 \\ \vdots \\ \partial/\partial x_n \end{bmatrix}$$

Proof: $\nabla(b^T x) \quad \nabla\left(\sum b_j x_j\right) \quad \begin{bmatrix} \uparrow \\ \frac{\partial}{\partial x_i}(\ ) \\ \downarrow \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_i \\ \vdots \\ b_n \end{bmatrix} = b$

Let us start, as always, with the cycle first, then go to the car. We will not start from the left-hand side of the expression; we will start from the right-hand side, the rightmost term. When the gradient operator hits c, what happens? The constant gets eliminated, right? So, the next term that we need to look at is $b^T x$, and keep in mind the basic definition of ∇f. $\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots\right)$. So, just keep that picture in mind.



↳ $\frac{d}{dx} f \cdot g = f'g + fg'$ prod rule.

Now $x^T A x$

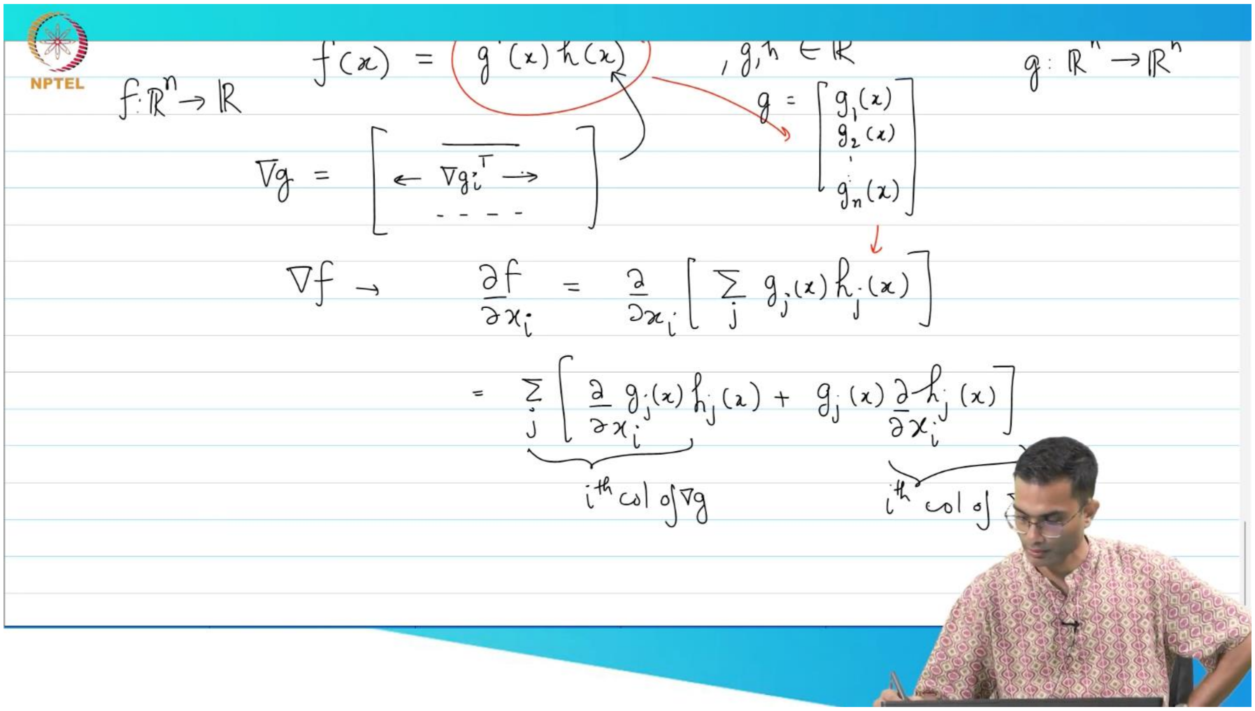
Aside: prod rule

$f(x) = g^T(x) h(x)$, $g, h \in \mathbb{R}^{n \times 1}$ $\quad g : \mathbb{R}^n \to \mathbb{R}^n$

$g = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_n(x) \end{bmatrix}$

$\nabla g = \begin{bmatrix} \leftarrow \nabla g_i^T \rightarrow \\ - - - - \end{bmatrix}$

Now, let us try to find out what the gradient of $b^T x$ is. Can I expand it in terms of the individual components? What would that be? It will be $\sum b_i x_i$. Actually, let me make that b_j. Now, when I want to take the gradient of this, I need to find out $\frac{d}{dx_i}$ of this expression. When I take the derivative with respect to x_i, does anything survive? Which term? Only b_i; everything else goes to 0. So, this is going to be equal to b_i. If I now do this over all i going from 1 to n, what will I get? I will get the entire vector b. It does not get easier than this. I just took a summation, took the derivative with respect to each variable, and I am left with b.



$f: \mathbb{R}^n \to \mathbb{R}$

$f'(x) = g'(x) h(x)$, $g, h \in \mathbb{R}$

$g: \mathbb{R}^n \to \mathbb{R}^n$

$$g = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_n(x) \end{bmatrix}$$

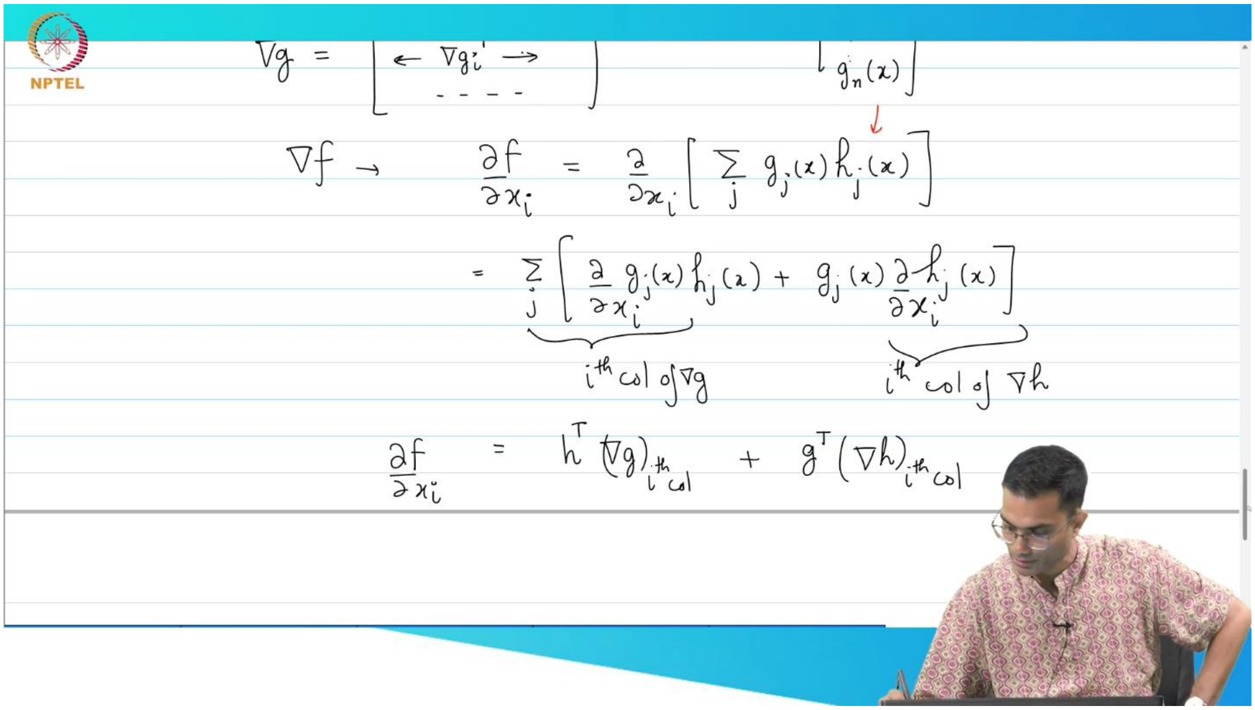
$$\nabla g = \begin{bmatrix} \leftarrow \overline{\nabla g_i^T} \rightarrow \\ - - - - \end{bmatrix}$$

$$\nabla f \rightarrow \quad \frac{\partial f}{\partial x_i} = \frac{\partial}{\partial x_i}\left[\sum_j g_j(x) h_j(x)\right]$$

$$= \sum_j \left[\underbrace{\frac{\partial}{\partial x_i} g_j(x)}_{i^{th} \text{ col of } \nabla g} h_j(x) + g_j(x) \underbrace{\frac{\partial}{\partial x_i} h_j(x)}_{i^{th} \text{ col of}}\right]$$

So, the first part is over here; this is fine, right? Now, any difficulty here? Quite straightforward, right? Okay, now we need to look at $x^T A x$. Let us just build some intuition. When you did simple one-dimensional calculus, when you encountered something like $f \cdot g$ and you wanted to do $\frac{d}{dx}$ of this, what did you do? The product rule, right? You could say $f'g + fg'$, right? That was our product rule.

Now, what do we have? $x^T A x$ again looks like a product rule, right? The product of two functions. However, we have not yet written the generalization of the product rule. So, let us do that, and then this will be a piece of cake. There is one grungy way of doing it. What is that grungy way? Before we go into the product rule, if you did not know the product rule, what would you do? You would expand this big thing, right? You would write the matrix A multiplied by x; you will get n rows with this thing, then you will multiply x^T with it, right? Then you will take the derivative. It will get you the correct answer; obviously, it will just take a little bit more time.



$$\nabla g = \begin{bmatrix} \leftarrow \nabla g_i' \rightarrow \\ - - - - \end{bmatrix} \qquad \begin{bmatrix} g_n(x) \end{bmatrix}$$

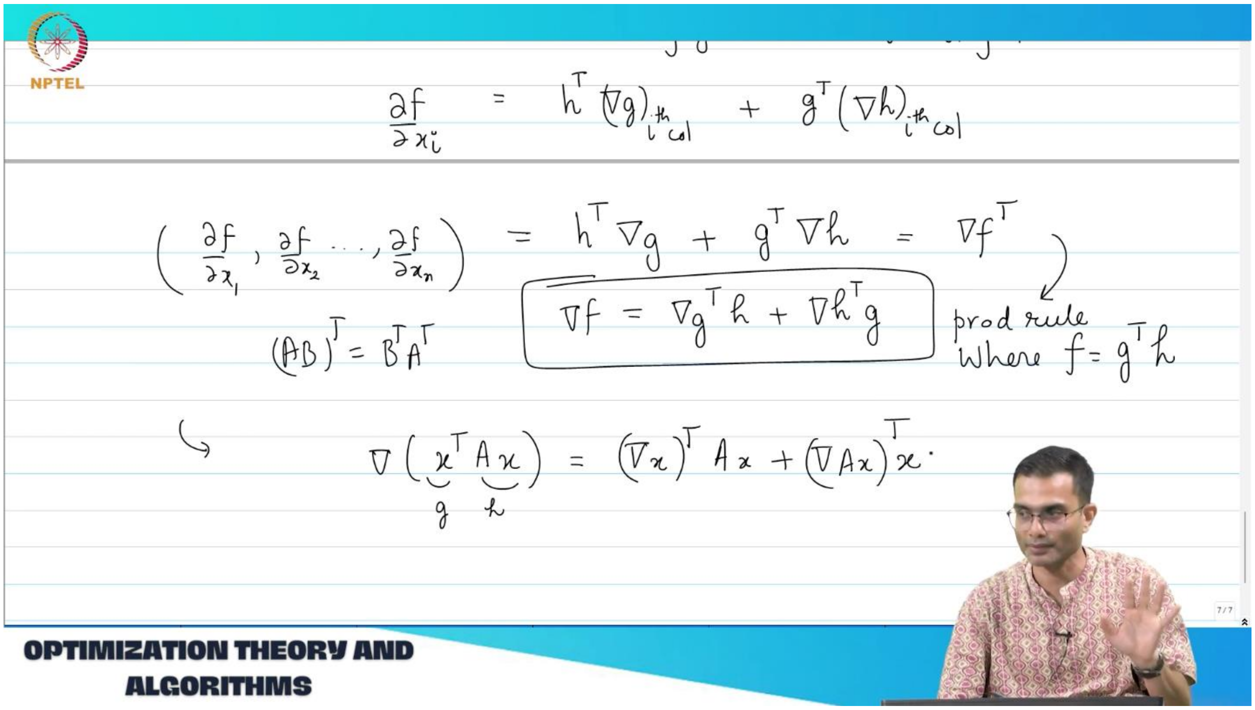
$$\nabla f \rightarrow \quad \frac{\partial f}{\partial x_i} = \frac{\partial}{\partial x_i}\left[\sum_j g_j(x) h_j(x)\right]$$

$$= \sum_j \left[\underbrace{\frac{\partial}{\partial x_i} g_j(x)}_{i^{th}\text{ col of }\nabla g} h_j(x) + g_j(x) \underbrace{\frac{\partial}{\partial x_i} h_j(x)}_{i^{th}\text{ col of }\nabla h}\right]$$

$$\frac{\partial f}{\partial x_i} = h^T (\nabla g)_{i^{th}\text{ col}} + g^T (\nabla h)_{i^{th}\text{ col}}$$

Now we could do it that way, but I want to show you the product rule because later on in the course, you may have places where you have multiple products of functions happening. So, if we know the product rule, you can quickly solve it without this messy kind of math. So, let us take an aside to the product rule.

I am going to write $f(x)$. Now, the thing to be careful about when we generalize the product rule is that in scalars, it did not matter how I wrote it. I wrote $f \cdot g$ or $g \cdot f$; it did not matter. Now, I have to take care of which is a transpose and which is a row vector, which is a column vector. So, I am going to write $f(x)$ as $g^T \cdot h(x)$, and obviously, g and h are all n-dimensional vectors to make it clear. Let me say $n \times 1$. So, these are n-dimensional vectors. Now, let us, in particular, look at g. So, g is going to be $g_1(x), g_2(x), \ldots, g_n(x)$; this is the meaning of g, right?



NPTEL

$$\frac{\partial f}{\partial x_i} = h^T (\nabla g)_{i^{th}\,col} + g^T (\nabla h)_{i^{th}\,col}$$

$$\left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \dots, \frac{\partial f}{\partial x_n} \right) = h^T \nabla g + g^T \nabla h = \nabla f^T$$

$$(AB)^T = B^T A^T$$

$$\boxed{\nabla f = \nabla g^T h + \nabla h^T g}$$

prod rule
Where $f = g^T h$

$$\nabla (\underbrace{x^T}_{g} \underbrace{Ax}_{h}) = (\nabla x)^T A x + (\nabla A x)^T x .$$

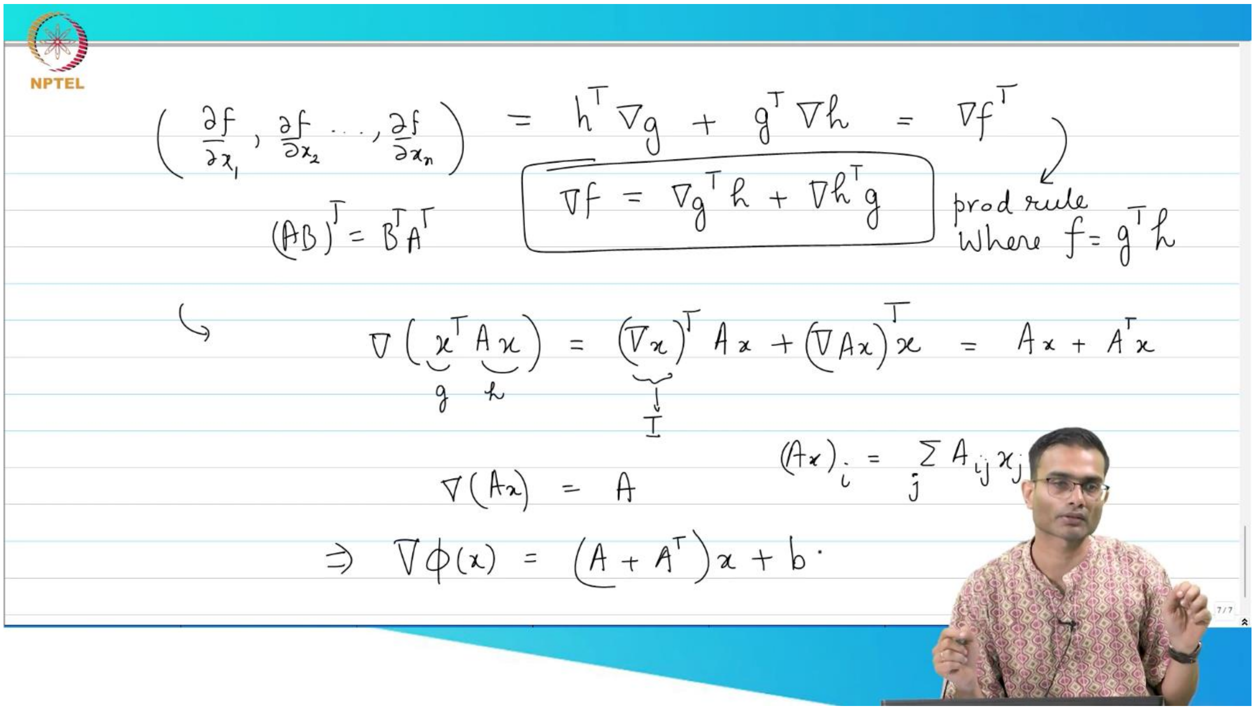
7 / 7

OPTIMIZATION THEORY AND ALGORITHMS

Now, when I say ∇g, I have a function. Now, x is not a scalar; x is also a vector. Obviously, I mean that is what we are dealing with. So, what type of object is ∇g? It is a Jacobian. You know why? Because g can be thought of as $\mathbb{R}^n$ to $\mathbb{R}^n$, right? It takes x_n as input, which is n-dimensional, and gives n-dimensional g out, g_1 to g_n. So, when I ask you what is ∇g? It should be a matrix, which is the Jacobian. We already know what this looks like.

Every row of this matrix is the partial derivative of that component. So, if I write it as row, I mean this is just writing it again. Each row is ∇g_i. So, $\nabla g_1, \nabla g_2$ all the way up to ∇g_n. I am just writing out the definition of the Jacobian, which we saw a short while ago.

So far, so good. Now, what am I doing? I am multiplying g^T and h, and I want to calculate, the whole point of the product rule is to calculate this ∇f for me. ∇f, f is $\mathbb{R}^n$ to $\mathbb{R}$, f is a scalar function. So, ∇f is going to be a column vector $\left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n}\right)$. If I want to find out ∇f, the thing I should find out is $\frac{\partial f}{\partial x_i}$. If I can get this, I am done.



$$\left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \dots, \frac{\partial f}{\partial x_n}\right) = h^T \nabla g + g^T \nabla h = \nabla f^T$$

$$(AB)^T = B^T A^T \qquad \boxed{\nabla f = \nabla g^T h + \nabla h^T g}$$

prod rule
Where $f = g^T h$

$$\nabla(\underbrace{x^T}_{g} \underbrace{Ax}_{h}) = (\underbrace{\nabla x}_{I})^T Ax + (\nabla Ax)^T x = Ax + A^T x$$

$$(Ax)_i = \sum_j A_{ij} x_j$$

$$\nabla(Ax) = A$$

$$\Rightarrow \nabla\phi(x) = (A + A^T)x + b.$$

So, let us write this out. So, $\frac{\partial f}{\partial x_i}$ is going to be the definition of f. So, this is going to be g^T multiplied by h. This is going to be a summation $g_j(x) \cdot h$. I am just writing out using this; that is all I have done. I have n components of g in a row vector, I have n components of h in a row vector; they are multiplied into each other, and we have a scalar. Now, I am taking the partial derivative with respect to x_i.

Here is where we can start applying our product rule. Can I apply a scalar product rule over here? Is $g_j(x)$ a scalar? Yes, it is a function. Let us start applying it. What does this remind you of? $\frac{\partial g_j}{\partial x_i}$ is constant; what is the variable that is roaming around? j, right? If I am changing different j over here, can you think of a compact thing that this thing under the curly brackets is hiding? The gradient g_j has an underlying g and h over here. This will give you ∇g_j and ∇h as the result. So, applying the product rule gives us:

$$\nabla f = \nabla g^T \cdot h + g^T \cdot \nabla h.$$

If I expand this out, I will get $\nabla g^T \cdot h + g^T \cdot \nabla h$ over here. This gives me the gradients of each term, and I will be done.

Let me try to apply it back to the quadratic case.

$$\nabla(x^T Ax) = (A + A^T)x.$$

We can always think of A as being symmetric. The moment we go to $x^T Ax$, which is symmetric, does not matter how you write it out. If A is symmetric, I can always assume that $x^T Ax = x^T A^T x$.

So, this is the result of the product rule over here.

Summary:

- The product rule states that if you have $f(x) = g^T \cdot h$, then $\nabla f = \nabla g^T \cdot h + g^T \cdot \nabla h$.
- The gradient of a quadratic form can be calculated using the properties of Jacobians and the product rule, giving us the final result as $A^T Ax + b$.