

M I C H A E L O R L I T Z K Y

The derivative of a quadratic form

the problem

You want to take the derivative of $f(x) = \langle Ax, x \rangle = x^T Ax$ over the real numbers. You want it to make sense, so that you don't forget it.

notation

Assume that all vectors are column vectors.

derivatives

First, we need to talk about derivatives. The usual definition of $f'(x)$ is,

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

Where $f : \mathbb{R} \rightarrow \mathbb{R}$ is simply a function from one real number to another.

This expression is called the *derivative* of f at x . Intuitively, it represents the rate of change of f at the point x . In particular—starting at x —if you move a distance of h and want to figure out how much f changed, you would multiply the derivative $f'(x)$ by the amount of change, h , and add it to $f(x)$ which is where you started:

$$f(x+h) \approx f(x) + f'(x) \cdot h$$

Note that the last expression, $\gamma(h) = f'(x) \cdot h$ is a linear function which approximates f around the point x .

generalized derivatives

Without getting into too much detail, I'll say that this approach can be generalized to functions $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$. The idea behind a generalized derivative is that we still want a linear approximation to f at a point x , although now we allow x to be a vector and not just a real number.

We want these generalized derivatives to approximate f the same way the usual derivative does. If f' is a generalized derivative for a function f , then we would like,

$$f(x + h) \approx f(x) + f'(x) \cdot h$$

We use the same notation for the generalized derivative as we do for the usual derivative. It shouldn't matter. The usual derivative is a special case, so we'll just think of $f'(x)$ as that thing which satisfies the equation above.

jacobians

You may have seen these generalized derivatives before; they're called *Jacobians*. The Jacobian of f at x (which we'll call $J(f; x)$ below) is a matrix, and the point x is a vector. I'll claim without proof that this works:

$$f(x + h) \approx f(x) + J(f; x) \cdot h$$

where of course h is a vector as well. The vector h still represents the amount by which x changed. The formula for the Jacobian looks something like,

$$J(f; x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_m} \end{bmatrix}$$

where f_1, f_2, \dots are the component functions of f . That is, since the result of $f(x)$ is an n -vector, we can think of f as working component-wise: $f(x) = (f_1(x), f_2(x), \dots)$. In any case, when you write everything out,

$$f(x + h) \approx f(x) + \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_m} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_m \end{bmatrix}$$

None of this is too important if you've never seen a Jacobian before, but it will make the rest easier to remember if you have.

gradients

When $f : \mathbb{R}^m \rightarrow \mathbb{R}$, that is, when $n = 1$, the Jacobian is a matrix with one row:

$$J(f; x) = f'(x) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_m} \right)$$

Since there's only one component function of f , we've dropped the subscript on f_1 for convenience. Again, without proof, I'll claim that this works:

$$f(x + h) \approx f(x) + f'(x) \cdot h = f(x) + \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_m} \right) \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_m \end{bmatrix}$$

The transpose of the row $J(f; x) = f'(x)$ is usually called the gradient of f at x , and is denoted by $\nabla f(x)$. This causes great confusion, since the matrix multiplication above no longer works if we transpose the Jacobian.

From now on, we'll ignore the gradient, and let $J(f; x) = f'(x)$ be the *row vector* that acts as the derivative of $f : \mathbb{R}^m \rightarrow \mathbb{R}$.

derivative of an inner product

The first thing you'll need to remember to differentiate the quadratic form is the derivative of the standard inner product on \mathbb{R}^m . Let $b \in \mathbb{R}^m$ be some other vector; then the inner product itself is defined as,

$$f(x) = \langle x, b \rangle = b^T x = \sum_{k=1}^m b_k \cdot x_k$$

If you have any intuition about these things, it should be clear (and why) that the derivative of this function is simply b^T itself. Since b is a column vector, the sum above is $b^T x$. One way to remember it is that “the derivative of a constant times x is the constant.” Another way is to note that $b^T x$ is a linear function that approximates $f(x)$ —since it *is* $f(x)$ —therefore, as we previously mentioned, it serves as the derivative of f .

If all else fails, compute the Jacobian, and confirm that it equals b^T .

Note that it doesn't matter which order x and b appear in. Over the real numbers, the inner product is symmetric, so $b^T x = x^T b$ both have the same

derivative, b^T .

derivative of a matrix times a vector

Now we'll compute the derivative of $f(x) = Ax$, where A is an $m \times m$ matrix, and $x \in \mathbb{R}^m$. Intuitively, since Ax is linear, we expect its derivative to be A . This turns out to be the case.

Recall that in the discussion of generalized derivatives, we said that we wanted a linear approximation of our function that satisfied,

$$f(x + h) \approx f(x) + f'(x) \cdot h$$

Well, $f(x) = Ax$, so $f(x + h) = A(x + h) = Ax + Ah$. If we set this equal to $f(x) + f'(x) \cdot h$, then,

$$Ax + Ah = f(x) + f'(x) \cdot h$$

Subtracting $f(x) = Ax$ from both sides,

$$Ah = f'(x) \cdot h$$

And dividing by h ,

$$A = f'(x)$$

total derivatives

Let's imagine now that $f(x, y)$ is a function of two variables x and y which returns a real number. Then the derivative of f with respect to x is the partial derivative $\frac{\partial f}{\partial x}$. Recall: the derivative of f with respect to x tells us how fast f changes when x changes a little bit:

$$f(x + h, y) \approx f(x, y) + \frac{\partial f}{\partial x} \cdot h$$

This all holds likewise for y . But what if y isn't independent of x ? Write $f(x, y(x))$ to make explicit the fact that y depends on x .

It makes intuitive sense that if we change x by a little bit (and thus y changes some as well), then the total change in f will be equal to (the change in f due to x changing) plus (the change in f due to y changing).

The change in f due to x is roughly $\frac{\partial f}{\partial x} \cdot h$ as before. Likewise, the change in f due to y is about $\frac{\partial f}{\partial y} \cdot q$, where q is how much y changed. But we can compute q ; how much does y change when x changes by h ? We already know:

$$y(x + h) \approx y(x) + \frac{\partial y}{\partial x} \cdot h$$

Putting this all together, we have,

$$\begin{aligned} f(x + h, y(x + h)) &\approx f(x, y(x)) + \frac{\partial f}{\partial x} \cdot h + \frac{\partial f}{\partial y} \frac{\partial y}{\partial x} \cdot h \\ &= f(x, y(x)) + \left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial x} \right) \cdot h \end{aligned}$$

Remember that the derivative of f was the thing that, when multiplied by the change h , tells us how much f changed? The above formula therefore shows us that,

$$f'(x, y(x)) = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial x}$$

since it satisfies that criteria. Ultimately, f is a function of one variable x , so it makes sense to write f' and speak of “the derivative of f .”

the chain rule

The only other fact we'll use is the chain rule. This should be familiar from calculus,

$$f'(y(x)) = f'(y) \cdot y'(x)$$

The chain rule actually applies to all generalized derivatives, in particular the single-row Jacobian. So it's safe to use the formula above, even when f and y are vector-valued functions. We'll use this below.

the quadratic form

With all that out of the way, this should be easy. Let,

$$f(x) = x^T A x$$

where $x \in \mathbb{R}^m$, and A is an $m \times m$ matrix. We can let $y(x) = Ax$ so that,

$$f(x, y(x)) = x^T \cdot y(x)$$

Using the formula for the total derivative above,

$$f'(x, y(x)) = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} \cdot y'(x)$$

The first term $\frac{\partial f}{\partial x}$ is the derivative of an inner product; if we hold $y(x)$ fixed, then the derivative of $x^T y(x)$ is $y(x)^T$, by the earlier discussion. Remember, it's a row vector.

Likewise, in the second term, $\frac{\partial f}{\partial y}$ is equal to x^T . It's a row vector too.

Furthermore in the second term, $\frac{\partial y}{\partial x}$ is the derivative of a matrix times a vector. We determined above that its derivative is the coefficient matrix A . Substituting,

$$f'(x, y(x)) = y(x)^T + x^T \cdot A$$

Since $y(x) = Ax$, we have $y(x)^T = x^T A^T$. We substitute again,

$$f'(x, y(x)) = x^T A^T + x^T A = x^T (A^T + A)$$

This is a derivative, and therefore, *a row vector*.

If A is symmetric (often the case), then $A^T = A$ and we can simplify,

$$f'(x, y(x)) = 2x^T A$$

copyright sucks