The Multivariable Chain Rule

Nikhil Srivastava

February 11, 2015

The chain rule is a simple consequence of the fact that differentiation produces the linear approximation to a function at a point, and that the derivative is the coefficient appearing in this linear approximation.

Let's see this for the single variable case first. It is especially transparent using $o()$ notation, where once again $f(x) = o(g(x))$ means that

$$\lim_{x \to 0} \left| \frac{f(x)}{g(x)} \right| = 0.$$

Suppose we are interested in computing the derivative of $(f \circ g)(x) = f(g(x))$ at x, where f and g are both differentiable functions from $\mathbb{R}$ to $\mathbb{R}$. Since g is differentiable, we have (by the definition of differentiation as a limit):

$$g(x + \Delta x) = g(x) + g'(x)\Delta x + o(\Delta x)$$

for a number $g'(x)$ which we call the derivative of g at x. In words, this says that g is well-approximated by its linear approximation in a neighborhood of x. Similarly, we have

$$f(y + \Delta y) = f(y) + f'(y)\Delta y + o(\Delta y).$$

Letting $y = g(x)$ and $\Delta y = g'(x)\Delta x + o(\Delta x)$, we now find that

$$\begin{aligned}
(f \circ g)(x + \Delta x) &= f(g(x + \Delta x)) \\
&= f\left(g(x) + g'(x)\Delta x + o(\Delta x)\right) \\
&= f(g(x)) + f'(g(x))\left(g'(x)\Delta x + o(\Delta x)\right) + o(\Delta y) \\
&= f(g(x)) + f'(g(x)) \cdot g'(x)\Delta x + o(\Delta x) + o(\Delta y) \quad \text{since } f'(g(x)) \cdot o(\Delta x) = o(\Delta x).
\end{aligned}$$

Thus, we have

$$\lim_{\Delta x \to 0} \frac{(f \circ g)(x + \Delta x) - (f \circ g)(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{f'(g(x)) \cdot g'(x) + o(\Delta x) + o(\Delta y)}{\Delta x} = f'(g(x)) \cdot g'(x),$$

establishing the chain rule.

A very similar thing happens in the multivariable case. Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}^2$ are differentiable. To parallel the notation used in class, let $z = f(x, y)$ and $(x, y) = g(s, t)$. Since both functions are differentiable, they must have linear approximations:

$$f(x + \Delta x, y + \Delta y) = f((x, y) + (\Delta x, \Delta y)) \approx f(x, y) + L_f(\Delta x, \Delta y) \quad (*),$$
$$g(s + \Delta s, t + \Delta t) = g((s, t) + (\Delta s, \Delta t)) \approx g(s, t) + L_g(\Delta s, \Delta t) \quad (**)$$

where $L_f : \mathbb{R}^2 \to R$ and $L_G : \mathbb{R}^2 \to \mathbb{R}^2$ are linear functions, and I have used $\approx$ to indicate equality up to $o(\Delta x)$ terms[1].

But we know that all linear functions are implemented by *matrices* so there must be a 1×2 matrix D_f such that

$$L_f(\Delta x, \Delta y) = D_f \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}.$$

In fact, we know exactly what this matrix is (by comparing coefficients):

$$D_f = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix},$$

so that we have the explicit formula

$$L_f(\Delta x, \Delta y) = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y,$$

which is the same as what is given by the total differential df.

Repeating this process for L_g, we get that for the 2×2 matrix

$$D_g = \begin{bmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \end{bmatrix},$$

the linear approximation of g at (s, t) is given by

$$L_g(\Delta s, \Delta t) = D_g \begin{bmatrix} \Delta s \\ \Delta t \end{bmatrix} = \begin{bmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \end{bmatrix} \begin{bmatrix} \Delta s \\ \Delta t \end{bmatrix} = \begin{bmatrix} \frac{\partial x}{\partial s}\Delta s + \frac{\partial x}{\partial t}\Delta t \\ \frac{\partial y}{\partial s}\Delta s + \frac{\partial y}{\partial t}\Delta t \end{bmatrix}.$$

Now for the punch line: just as in the univariate case, we write:

$$\begin{aligned}
(f \circ g)((s, t) + (\Delta s, \Delta t)) &= f(g((s, t) + (\Delta s, \Delta t))) \\
&\approx f\left(g(s, t) + D_g \begin{bmatrix} \Delta s \\ \Delta t \end{bmatrix}\right) \quad \text{by } (**) \\
&\approx f(g(s, t)) + D_f D_g \begin{bmatrix} \Delta s \\ \Delta t \end{bmatrix} \quad \text{by } (*), \text{ treating } D_g \begin{bmatrix} \Delta s \\ \Delta t \end{bmatrix} \text{ as } \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}
\end{aligned}$$

[1]There is a subtlety about uniform convergence vs pointwise convergence here, but for the purposes of this course you can ignore it, and what is written here is good enough.

On the other hand, since $f \circ g$ is a differentiable function from $\mathbb{R}^2$ to $\mathbb{R}^1$, it must have a linear approximation, which is given by a 2×1 matrix $D_{f \circ g}$:

$$(f \circ g)((s,t) + (\Delta s, \Delta t)) \approx (f \circ g)(s,t) + D_{f \circ g} \begin{bmatrix} \Delta s \\ \Delta t \end{bmatrix},$$

whose entries are by definition given by the partial derivatives of $f \circ g$:

$$D_{f \circ g} = \begin{bmatrix} \frac{\partial (f \circ g)}{\partial s} & \frac{\partial (f \circ g)}{\partial t} \end{bmatrix}.$$

But these two linear approximations must be *the same*, so we have the rather crystalline identity:

$$D_{f \circ g} = D_f D_g,$$

which is the chain rule written in matrix notation. In the single variable case these are 1×1 matrices containing a single entry (the derivative), so we recover the familiar identity:

$$(f \circ g)'(x) = f'(g(x))g'(x).$$

In the two-dimensional case, writing out the entries gives:

$$\begin{bmatrix} \frac{\partial (f \circ g)}{\partial s} & \frac{\partial (f \circ g)}{\partial t} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix} \begin{bmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x}\frac{\partial x}{\partial s} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial s} & \frac{\partial f}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial t} \end{bmatrix},$$

which is identical to the multivariable chain rule that we derived in class using total differentials.

Conceptually, what is going on is that a differentiable function is locally approximated by a linear function (which is just a matrix), so composing two differentiable functions corresponds to composing these linear approximations, which is just matrix multiplication.

The matrix D_f is called the *gradient* of f, and is defined similarly — as a row vector of partial derivatives — for all functions from $\mathbb{R}^n$ to $\mathbb{R}$. The matrix D_g is called the *Jacobian* of g, and is defined similarly for functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. These matrices are a satisfactory generalization of the notion of *derivative* to higher dimensions. Their entries are the partial derivatives, which tell us how individual pairs of variables interact, but to get the whole story you need to look at the entire matrix.