CHAIN RULE: A UNIFIED VIEW VIA MATRICES

Matrices. A matrix is an array of numbers. For instance,

$$A = \begin{bmatrix} -1 & 0 & 2 \\ e & 1/2 & -\pi/4 \end{bmatrix}, \quad B = \begin{bmatrix} 3 \\ -4 \\ 0 \end{bmatrix}, \quad C = [5 \ -2 \ 1],$$

are matrices. $A$ has two rows and one three columns, and is therefore called a $2 \times 3$ matrix. $B$ is a $3 \times 1$ matrix and $C$ is a $1 \times 3$ matrix. ($B$ is also called a column matrix and $C$ is called a row matrix.)

The numbers in a matrix are called its entries (or components). The entry in the $i^{th}$ row and $j^{th}$ column is called the $(i, j)$-entry. If a matrix is denoted by $A$, then its $(i, j)$-entry is often denoted by $a_{ij}$. For instance, the $(2, 3)$-entry of $A$ is $a_{23} = -\pi/4$.

Matrix multiplication. Two matrices $A$ and $B$ can be multiplied if the number of columns of $A$ equals the number of rows of $B$. That is, if $A$ is an $m \times k$ matrix and $B$ is an $\ell \times n$ matrix, then $AB$ is defined only if $k = \ell$ and $AB$ is then an $m \times n$ matrix (i.e., it has the same number of rows as $A$ and the same number of columns as $B$).

Here’s how $AB$ is defined. If $A_{i \to}$ denotes the $i^{th}$ row of $A$ and $B_{j \to}$ is the $j^{th}$ column of $B$, then the $(i, j)$-entry of $AB$ is the dot product $A_{i \to} \cdot B_{j \to}$. For example, the $(1, 1)$-entry of $P = AB$, where $A$ and $B$ are as above, is

$$p_{11} = A_{1 \to} \cdot B_{1 \to} = \begin{bmatrix} -1 & 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ -4 \\ 0 \end{bmatrix} = (-1) \cdot 3 + 0 \cdot (-4) + 2 \cdot 0 = -3.$$

The whole matrix $P = AB$ looks like this:

$$P = \begin{bmatrix} -3 \\ 3e - 2 \end{bmatrix}.$$

Note that $AC$ is not defined, but $BC$ and $CB$ are defined: $BS$ is a $3 \times 3$ matrix, while $CB$ is a $1 \times 1$ matrix (i.e., a number): $CB = 23$.

Derivatives. Now let $f: \mathbb{R}^n \to \mathbb{R}^m$ be a map. We will mostly deal with the cases $n = 2, 3$ and $m = 1, 2$. For instance, assume $f: \mathbb{R}^3 \to \mathbb{R}^2$. This means that for every $x = (x_1, x_2, x_3)$ in the domain of $f$, $f(x)$ is a pair of real numbers $(f_1(x), f_2(x))$. We call $f_1$ and $f_2$ the component functions (or just components) of $f$.

As defined in class for maps $f: \mathbb{R}^2 \to \mathbb{R}$, we will call $f$ differentiable at a point $p$ if $f$ can be approximated by a linear map near $p$. We will not dwell on the intricacies of this definition. Instead, here’s a fact we’ll use: if $f$ is differentiable at some point $p$, then its derivative, denoted by $Df(p)$ is the matrix whose entries are the partial derivatives of $f$ at $p$. For instance, if $f: \mathbb{R}^3 \to \mathbb{R}^2$ is as above, with components $f_1, f_2$, then, denoting the partial derivative with respect to the variable $x_i$ by $\partial x_i$, we have

$$Df(p) = \begin{bmatrix} \partial_{x_1} f_1(p) & \partial_{x_2} f_1(p) & \partial_{x_3} f_1(p) \\ \partial_{x_1} f_2(p) & \partial_{x_2} f_2(p) & \partial_{x_3} f_2(p) \end{bmatrix}.$$
Note that the first row of $Df(p)$ consists of the partials of $f_1$ and the second row consists of the partials of $f_2$.

For example, let
\[
f(x, y, z) = (x^2 + y^2 + \cos z, e^{xyz}).
\]
Then
\[
Df(x, y, z) = \begin{bmatrix} 2x & 2y & -\sin z \\ ye^{xyz} & xe^{xyz} & yze^{xyz} \end{bmatrix}.
\]
If $f$ is real-valued, e.g., $f : \mathbb{R}^2 \to \mathbb{R}$, then $Df$ is a row matrix (since $f$ has only one component):
\[
Df(x, y) = \left[ \frac{\partial x f}{\partial y} \right].
\]

If you already know about the gradient $\nabla f$ of $f$, observe that $Df$ is just the transpose of $\nabla f$ (to transpose a matrix, just switch its rows and columns). We will always think of the gradient of a scalar-valued function as a column vector and its derivative as a row vector (later we'll see why).

**The Chain Rule.** Now we'll get to see why it was worth learning a little about matrices and thinking of derivatives as matrices.

Recall from Calculus I that for two differentiable functions $f, g$ whose composition $(g \circ f)(x) = g(f(x))$ is defined, we have $(g \circ f)'(x) = g'(f(x))f'(x)$.

Suppose now that $f$ and $g$ are two differentiable functions of several variables, such that the composition $g \circ f$ is defined. Then as it turns out $g \circ f$ is also differentiable and $D(g \circ f)$ is just the matrix product of their derivatives. More precisely, we have:

**Theorem (Chain rule).** If $f : \mathbb{R}^n \to \mathbb{R}^k$ and $g : \mathbb{R}^k \to \mathbb{R}^m$ are both differentiable, then $g \circ f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable, and for every $x \in \mathbb{R}^n$,
\[
D(g \circ f)(x) = Dg(f(x)) \, Df(x).
\]

Note that apart from denoting the derivative by $Df$ instead of $f'$, the form of the chain rule is exactly the same as for functions of one variable. We just need to multiply matrices instead of real numbers! Observe also that we need to evaluate the derivative of $g$ at the point $f(x)$.

All the “cases” of the chain rule discussed in Stewart follow immediately from this general result.

**Case 1.** Suppose we are given a function $z = g(x, y)$ (i.e., $g : \mathbb{R}^2 \to \mathbb{R}$) where both $x$ and $y$ depend on a variable $t$, i.e., $x = x(t)$ and $y = y(t)$. Then $g(x(t), y(t))$ is a function of $t$; let’s call it $h(t)$. How do we set things up in order to use the chain rule? We can think of $(x(t), y(t))$ as a vector-valued function $f(t)$, i.e., $f(t) = (x(t), y(t))$, so that $h(t) = g(f(t))$. Then
\[
Dh(t) = Dg(f(t))Df(t) = \begin{bmatrix} g_x(f(t)) & g_y(f(t)) \end{bmatrix} \begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} = g_x(f(t))x'(t) + g_y(f(t))y'(t).
\]
Observe that $Dh(t)$ is just a number, as it should be, since $h$ takes a real number $t$ to a real number $g(x(t), y(t))$.

**Case 2.** Now $f : \mathbb{R}^2 \to \mathbb{R}^2$ and $g : \mathbb{R}^2 \to \mathbb{R}$. Let’s say that $z = g(x, y)$ and $f(s, t) = (x, y)$, so that $x = x(s, t)$ and $y = y(s, t)$. Then
\[
Dg(x, y) = \begin{bmatrix} g_x & g_y \end{bmatrix}
\]
\[
Df(s, t) = \begin{bmatrix} x_s & x_t \\ y_s & y_t \end{bmatrix}
\]
hence the derivative of $h = g \circ f$ equals

$$Dh(s,t) = Dg(x,y)Df(s,t) = \begin{bmatrix} g_x & g_y \\ x_s & x_t \end{bmatrix} = \begin{bmatrix} g_x x_s + g_y y_s & g_x x_t + g_y y_t \end{bmatrix}.$$  

Note that $Dh(s,t)$ also equals $[h_s, h_t]$, so that

$$h_s = g_x x_s + g_y y_s, \quad \text{and} \quad h_t = g_x x_t + g_y y_t.$$  

**Case 3: the “general version”**. Now we have a function $u = u(x_1, \ldots, x_n)$ and each $x_i$ depends on $t_1, \ldots, t_m$: $x_i = x_i(t_1, \ldots, t_m)$. We set things up like this:

$$f(t_1, \ldots, t_m) = (x_1, \ldots, x_n), \quad \text{and} \quad g(x_1, \ldots, x_n) = u(x_1, \ldots, x_n) = u.$$  

Again, let $h = g \circ f$. Then $h(t_1, \ldots, t_m) = u$ so by the chain rule:

$$Dh(t_1, \ldots, t_m) = Dg(x_1, \ldots, x_n)Df(t_1, \ldots, t_m)$$

$$= \begin{bmatrix} \partial_{t_1} x_1 & \partial_{t_2} x_1 & \cdots & \partial_{t_m} x_1 \\ \partial_{t_1} x_2 & \partial_{t_2} x_2 & \cdots & \partial_{t_m} x_2 \\ \vdots & \vdots & \cdots & \vdots \\ \partial_{t_1} x_n & \partial_{t_2} x_n & \cdots & \partial_{t_m} x_n \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^n u_{x_i} \partial_{t_1} x_i & \sum_{i=1}^n u_{x_i} \partial_{t_2} x_i & \cdots & \sum_{i=1}^n u_{x_i} \partial_{t_m} x_i \end{bmatrix}.$$  

Recall that $h$ is a scalar-valued function, so $Dh$ is just the row-vector

$$Dh = [h_t_1 \ h_t_2 \ \cdots \ h_{t_m}].$$  

Equating the corresponding components, we obtain:

$$h_{t_1} = \sum_{i=1}^n u_{x_i} \partial_{t_1} x_i, \quad \cdots \quad h_{t_m} = \sum_{i=1}^n u_{x_i} \partial_{t_m} x_i.$$  

**Example.** Let $f(x, y) = (x + y^2, e^{x^2+y})$ and $g(u, v) = uv$. Set $h(x, y) = g(f(x, y))$. Let us compute $Dh(0,1)$. Note that $f(0,1) = (1, e)$. Differentiating,

$$Df(x,y) = \begin{bmatrix} 1 & 2y \\ 2xe^{x^2+y} & e^{x^2+y} \end{bmatrix}, \quad \text{so} \quad Df(0,1) = \begin{bmatrix} 1 & 2 \\ 0 & e \end{bmatrix}.$$  

Similarly,

$$Dg(u,v) = [v \ u] \quad \text{so} \quad Dg(1,e) = [e \ 1].$$  

By the Chain Rule,

$$Dh(0,1) = Dg(1,e)Df(0,1) = [e \ 1] \begin{bmatrix} 1 & 2 \\ 0 & e \end{bmatrix} = [e \ 3e].$$  

Recall also that $Dh = [h_x \ h_y]$, which means $h_x(0,1) = e$ and $h_y(0,1) = 3e$. We can check this by observing that

$$h(x, y) = g(x + y^2, e^{x^2+y}) = (x + y^2)e^{x^2+y}.$$  

Hence, by the product rule, $h_x = e^{x^2+y} + (x + y^2)2xe^{x^2+y}$ and $h_y = 2ye^{x^2+y} + (x^2 + y)e^{x^2+y}$, so

$h_{x}(0,1) = e$ and $h_{y}(0,1) = 2e + e = 3e$, which is the same answer as above.