John Nachbar
Washington University
March 24, 2018

Multivariate Differentiation[1]

1 Preliminaries.

I assume that you are already familiar with standard concepts and results from univariate calculus; in particular, the Mean Value Theorem appears in two of the proofs here.

To avoid notational complication, I take the domain of functions to be all of $\mathbb{R}^N$. Everything generalizes immediately to functions whose domain is an open subset of $\mathbb{R}^N$. One can also generalize this machinery to "nice" non-open sets, such as $\mathbb{R}^N_+$, but I will not provide a formal development.[2]

In my notation, a point in $\mathbb{R}^N$, which I also refer to as a vector (*vector* and *point* mean exactly the same thing for me), is written

$$x = (x_1, \dots, x_N) \underset{\text{def}}{=} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}.$$

Thus, a vector in $\mathbb{R}^N$ *always* corresponds to an $N \times 1$ (column) matrix. This ensures that the matrix multiplication below makes sense (the matrices conform).

If $f : \mathbb{R}^N \to \mathbb{R}^M$ then f can be written in terms of M *coordinate functions* $f_m : \mathbb{R}^N \to \mathbb{R}$,

$$f(x) \underset{\text{def}}{=} (f_1(x), \dots, f_M(x)).$$

Again, $f(x)$, being a point in $\mathbb{R}^M$, can also be written as an $M \times 1$ matrix. If $M = 1$ then f is *real-valued.*

2 Partial Derivatives and the Jacobian.

Let e^n be the unit vector for coordinate n: $e^n = (0, \dots, 0, 1, 0, \dots, 0)$, with the 1 appearing in coordinate n. For any $\gamma \in \mathbb{R}$, $x^* + \gamma e^n$ is identical to x^* except for coordinate n, which changes from x^*_n to $x^*_n + \gamma$.

[2] Recall that if $a, b \in \mathbb{R}^N$, then $a \geq b$ means $a_n \geq b_n$ for every n. $a > b$ means $a \geq b$ and $a_n > b_n$ for at least one n. $a \gg b$ means $a_n > b_n$ for every n. $\mathbb{R}^N_+ = \{x \in \mathbb{R}^N : x \geq 0\}$. $\mathbb{R}^N_{++} = \{x \in \mathbb{R}^N : x \gg 0\}$.

Given a function $f : \mathbb{R}^N \to \mathbb{R}^M$ and a point $x^* \in \mathbb{R}^N$, if the limit

$$\lim_{\gamma \to 0} \frac{f_m(x^* + \gamma e^n) - f_m(x^*)}{\gamma}$$

exists, then it is called the *partial derivative* of f, evaluated at x^*, for coordinate function f_m with respect to variable x_n; I usually denote this partial derivative as $D_n f_m(x^*)$.[3] Standard alternative notation for the partial derivative is

$$\frac{\partial f_m}{\partial x_n}(x^*).$$

I tend to use $D_n f_m(x^*)$ instead of $\partial f_m(x^*)/\partial x_n$ because I find $D_n f_m(x^*)$ more legible. I reserve the notation,

$$\frac{df_m}{dx}(x^*)$$

for situations in which $N = 1$.

If you can take derivatives in the one-dimensional case, you can just as easily take partial derivatives in the multivariate case. When taking the partial derivative with respect to x_n, just treat the other variables like constants.

Example 1. $f : \mathbb{R}^2_+ \to \mathbb{R}$ is defined by $f(x_1, x_2) = \sqrt{x_1 + 3x_2}$. Then at the point $x^* = (1, 1)$,

$$\begin{aligned} D_1 f(x^*) &= \frac{1}{2} \frac{1}{\sqrt{x_1^* + 3x_2^*}} = \frac{1}{4}, \\ D_2 f(x^*) &= \frac{1}{2} \frac{1}{\sqrt{x_1^* + 3x_2^*}} 3 = \frac{3}{4}. \end{aligned}$$

□

The $M \times N$ matrix of partial derivatives is called the *Jacobian* of f at x^*, denoted $Jf(x^*)$.

$$Jf(x^*) \underset{\text{def}}{=} \begin{bmatrix} D_1 f_1(x^*) & \dots & D_N f_1(x^*) \\ \vdots & \ddots & \vdots \\ D_1 f_M(x^*) & \dots & D_N f_M(x^*) \end{bmatrix}.$$

Remark 1. Different authors use the word "Jacobian" in different ways. For example, in Rudin (1976), the Jacobian refers to the determinant of $Jf(x^*)$ when $N = M$. Here, however, Jacobian refers to the matrix of partial derivatives, even if $N \neq M$. □

[3] Recall from the Continuity notes that this limit notation means that for any sequence (γ_t) in $\mathbb{R}$ such that $\gamma_t \neq 0$ for all t and $\gamma_t \to 0$, the quotient

$$\frac{f_m(x^* + \gamma_t e^n) - f_m(x^*)}{\gamma_t}$$

converges to $D_n f_m(x^*)$.

Example 2. In the previous example,

$$Jf(x^*) = \left[\begin{array}{cc} 1/4 & 3/4 \end{array}\right].$$

□

If the partial derivatives $D_n f_m(x^*)$ are defined for all x^* in the domain of f then one can define a function $D_n f_m : \mathbb{R}^N \to \mathbb{R}$.

Example 3. If, as above, $f : \mathbb{R}^2_+ \to \mathbb{R}$ is defined by $f(x_1, x_2) = \sqrt{x_1 + 3x_2}$ then the functions $D_n f$ are defined by

$$\begin{aligned} D_1 f(x) &= \frac{1}{2\sqrt{x_1 + 3x_2}}, \\ D_2 f(x) &= \frac{3}{2\sqrt{x_1 + 3x_2}}. \end{aligned}$$

□

And one can ask whether the $D_n f_m$ are continuous and one can compute partial derivatives of the $D_n f_m$, which would be second order partials of f. And so on.

3 Directional Derivatives.

Informally, imagine that you are standing on the side of a hill and considering walking in some compass direction. Taking the partial derivatives is like measuring the slope of the hill in just two directions, due north (0 degrees on a compass) and due east (90 degrees). But you might be interested in measuring the slope in some other direction, and it is not hard to imagine a hill that is so irregular that knowing the slope when walking north or east doesn't give any information about the slope when walking, say, north-east (45 degrees).

A *direction* in $\mathbb{R}^N$ is a vector $v \in \mathbb{R}^N$ such that $v \neq 0$. Given a function $f : \mathbb{R}^N \to \mathbb{R}^M$, if

$$\lim_{\gamma \to 0} \frac{f_m(x^* + \gamma v) - f_m(x^*)}{\gamma},$$

exists then it is called the *directional derivative* of f, evaluated at the point x^*, for coordinate function f_m in the direction v. I denote this directional derivative as $D_v f_m(x^*)$. If $M > 1$, then it is also convenient to write,

$$D_v f(x^*) \underset{\text{def}}{=} \begin{bmatrix} D_v f_1(x^*) \\ \vdots \\ D_v f_M(x^*) \end{bmatrix}.$$

Many texts assume that $\|v\| = 1$, but I do not.

As before, let e^n be the unit vector for coordinate n. Then e^n is a direction and one can verify from the definitions that, for any m,

$$D_{e^n} f_m(x^*) = D_n f_m(x^*).$$

That is, a partial derivative is a special case of a directional derivative.

4 The Derivative.

Consider the following example.

Example 4. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x_1, x_2) = 3x_1 + x_1x_2$ and let $x^* = (1, 1)$ and let $v = (1, 1)$. Then, applying the definition of directional derivative,

$$\begin{aligned} D_v f(x^*) &= \lim_{\gamma \to 0} \frac{(3(x_1^* + \gamma) + (x_1^* + \gamma)(x_2^* + \gamma)) - (3x_1^* + x_1^*x_2^*)}{\gamma} \\ &= \lim_{\gamma \to 0} \frac{(4 + 5\gamma + \gamma^2) - 4}{\gamma} \\ &= \lim_{\gamma \to 0} 5 + \gamma \\ &= 5. \end{aligned}$$

Note also that

$$Jf(x^*) = \left[\begin{array}{cc} 4 & 1 \end{array} \right].$$

Therefore, $Jf(x^*)v = 5$. That is, $D_v f(x^*) = Jf(x^*)v$. □

In fact, it will follow from results below that for the function in Example 4, for any x^* and $v \neq 0$, $D_v f(x^*) = Jf(x^*)v$. That is, one can think of the Jacobian as a machine for computing directional derivatives. Putting the same point slightly differently, once one has computed the partial derivatives for this function, one can easily compute all the (other) directional derivatives.

Unfortunately, there are some real-valued functions for which $D_v f(x^*) \neq Jf(x^*)v$, even though both $D_v f(x^*)$ and $Jf(x^*)$ are well defined. Consider the following.

Example 5. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x) = \begin{cases} \frac{x_1^3}{x_1^2 + x_2^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

The graph of f looks like a wrinkled sheet of paper. By direct application of the definition of partial derivative, one can compute that

$$Jf(0, 0) = \left[\begin{array}{cc} 1 & 0 \end{array} \right].$$

Let $v = (1, 1)$. Then $Jf(0, 0)v = 1$. On the other hand, one can compute that $D_v f(0, 0) = 1/2$. So $Jf(0, 0)v \neq D_v f(0, 0)$. Thus, for this function, the partial derivatives, which give the slope of f in the direction of each axis, do not provide accurate information about the slope of f along the 45 degree line between the two axes. □

In the preceding example, $Jf(x^*)v \neq D_v f(x^*)$ at $x^* = (0, 0)$ because f is not differentiable at x^* in the following sense. Recall first that in the univariate case, $Df(x^*)$ is the derivative of f at x^* iff

$$Df(x^*) = \lim_{w \to 0} \frac{f(x^* + w) - f(x^*)}{w}.$$

This holds iff,

$$\lim_{w \to 0} \left| \frac{f(x^* + w) - f(x^*)}{w} - Df(x^*) \right| = 0,$$

iff,

$$\lim_{w \to 0} \left| \frac{f(x^* + w) - f(x^*) - Df(x^*)w}{w} \right| = 0,$$

iff,

$$\lim_{w \to 0} \frac{|f(x^* + w) - f(x^*) - Df(x^*)w|}{|w|} = 0.$$

This last characterization generalizes easily to the multivariate case.

Definition 1. *$f : \mathbb{R}^N \to \mathbb{R}^M$ is* differentiable *at x^* if and only if there is an $M \times N$ matrix $Df(x^*)$, called the* derivative *of f at x^*, such that,*

$$\lim_{w \to 0} \frac{\|f(x^* + w) - f(x^*) - Df(x^*)w\|}{\|w\|} = 0.$$

f is differentiable *iff it is differentiable at every point in $\mathbb{R}^N$.*[4]

It is not hard to show that the derivative, if it exists, is unique. It is also almost immediate that if f is differentiable then it is continuous. One can show that the sum of differentiable functions is differentiable, as is the product or quotient of real-valued differentiable functions (provided the function in the denominator does not take the value 0), as is the inner product of vector-valued differentiable functions. The following theorem, called the Chain Rule, shows that compositions of differentiable functions are differentiable.

Theorem 1 (Chain Rule)**.** *Let $g : \mathbb{R}^N \to \mathbb{R}^M$, let $f : \mathbb{R}^M \to \mathbb{R}^L$, and define $h : \mathbb{R}^N \to \mathbb{R}^L$ by $h(x) = f(g(x))$. If g is differentiable at x^* and f is differentiable at $y^* = g(x^*)$ then h is differentiable at x^* and*

$$Dh(x^*) = Df(y^*)Dg(x^*).$$

[4] Strictly speaking, the formal definition of the derivative is more abstract than this, and what I am calling the derivative is actually the matrix representation of the derivative. Now that I've said this, you can forget it for the purposes of these notes.

Proof. The proof is just a marshalling of definitions, but it is tedious and notationally dense, so I have relegated it to Section 6. Example 7 provides an explicit example of the Chain Rule. ■

Theorem 4 below shows that if f is differentiable in the above sense then $Df(x^*) = Jf(x^*)$. Theorem 3 shows that $D_v f(x^*) = Df(x^*)v$, which then implies $D_v f(x^*) = Jf(x^*)v$. A corollary is that the function in Example 5 is not differentiable. I discuss this issue further after Theorem 4.

As a first step, I record that f is differentiable iff its constituent coordinate functions are differentiable. This enables me to simplify arguments by restricting attention to the case $M = 1$.

Theorem 2. *Let $f : \mathbb{R}^N \to \mathbb{R}^M$. f is differentiable at x^* iff for every m, f_m is differentiable at x^*, in which case,*

$$Df(x^*) = \begin{bmatrix} Df_1(x^*) \\ \vdots \\ Df_M(x^*) \end{bmatrix}.$$

Proof. This follows from the fact that Euclidean convergence in $\mathbb{R}^M$ is equivalent to convergence in each coordinate (see the section on pointwise convergence in the notes on $\mathbb{R}^N$ Completeness and Compactness). ■

Theorem 3. *If $f : \mathbb{R}^N \to \mathbb{R}^M$ is differentiable at x^* then for any m and any $v \in \mathbb{R}^N$, $v \neq 0$, the directional derivative $D_v f_m(x^*)$ exists and*

$$D_v f_m(x^*) = Df_m(x^*)v,$$

hence

$$D_v f(x^*) = Df(x^*)v.$$

Proof. By Theorem 2, it suffices to consider the case $M = 1$. Fix $v \neq 0$ and let $w = \gamma v$. Then the definition of derivative requires that,

$$\lim_{\gamma \to 0} \frac{\|f(x^* + \gamma v) - f(x^*) - Df(x^*)(\gamma v)\|}{\|\gamma v\|} = 0,$$

which holds (see also the discussion immediately preceding Definition 1) iff,

$$\lim_{\gamma \to 0} \frac{f(x^* + \gamma v) - f(x^*)}{\gamma} = Df(x^*)v.$$

But the left-hand side is simply $D_v f(x^*)$. ■

An almost immediate corollary is that $Df(x^*)$, if it exists, equals $Jf(x^*)$.

Theorem 4. *Let $f : \mathbb{R}^N \to \mathbb{R}^M$. If f is differentiable at $x^* \in \mathbb{R}^N$ then all partial derivatives exist at x^* and $Df(x^*) = Jf(x^*)$.*

Proof. If f is differentiable then by Theorem 3, all partial derivatives exist and for any m and any n

$$\frac{\partial f_m}{\partial x_n}(x^*) = D_{e^n} f_m(x^*) = Df_m(x^*)e^n,$$

but, by Theorem 2, the last term is the (m, n) element of $Df(x^*)$. ■

If $N = 1$ then f_m is univariate, $Df_m(x^*)$ is the ordinary univariate derivative, and the expression in Theorem 2 is equivalent to $Df(x^*) = Jf(x^*)$. Thus, if $N = 1$, existence of the partial derivatives is sufficient as well as necessary for differentiability of f.

If $N > 1$ things are different. For simplicity of notation, suppose $M = 1$. The fundamental issue is then the following. Suppose that for each $v \neq 0$ the directional derivative $D_v f(x^*)$ exists. Then for each $v \neq 0$ there is a $1 \times N$ matrix, call it A_v, such that $D_v f(x^*) = A_v v$ (indeed, for $N > 1$, there are typically infinitely many matrices A_v that satisfy this expression). The problem is that this allows different matrices A_v for different v. Differentiability requires using the same matrix for all v. This is precisely what goes wrong in Example 5: all the directional derivatives exist but there is no single $1 \times N$ matrix A such that for all $v \neq 0$, $D_v f(x^*) = Av$. This raises the question of what functions *are* differentiable.

Theorem 5 below gives a sufficient condition for existence of $Df(x^*)$. To state the theorem, I first need to define what it means for a multivariate function to be continuously differentiable. If f is differentiable then one can consider the function Df that gives the derivative for every point in the domain of f. By Theorem 4, if f is differentiable then for every $x^* \in \mathbb{R}^N$ the partial derivatives all exist and $Df(x^*)$ is the matrix of partial derivatives at x^*. Therefore, if Df exists, define Df to be continuous iff the partial derivatives of f are continuous.[5] Say that f is *continuously differentiable*, written f is $\mathcal{C}^1$, iff f is differentiable and Df is continuous.

Theorem 5. *$f : \mathbb{R}^N \to \mathbb{R}^M$ is $\mathcal{C}^1$ iff $D_n f_m$ exists and is continuous for every n and m.*

Proof. Somewhat messy in terms of notation and therefore relegated to Section 6. The hard step in the proof is showing that if $D_n f_m$ exists and is continuous for every n and m then Df exists. ■

[5] This definition of continuity for Df can be shown to be consistent with defining continuity of Df using the norm on $M \times N$ matrices given by $\|Df(x^*)\|_\mu = \sup_{w \in S} \|Df(x^*)w\|$, where S is the unit sphere, centered on the origin, in $\mathbb{R}^N$. This norm is also used in the proof, which appears in Section 6, of the Chain Rule (Theorem 1).

Example 6. In Example 5, with $f(x_1, x_2) = x_1^3/(x_1^2 + x_2^2)$ for $(x_1, x_2) \neq (0, 0)$, $f(0, 0) = 0$, the partial derivatives are not continuous at 0.

$$Jf(x) = \begin{cases} \begin{bmatrix} 1 & 0 \end{bmatrix} & \text{if } x = 0, \\ \\ \begin{bmatrix} \frac{x_1^2(x_1^2+3x_2^2)}{(x_1^2+x_2^2)^2} & \frac{-2x_1^3 x_2}{(x_1^2+x_2^2)^2} \end{bmatrix} & \text{otherwise.} \end{cases}$$

One can check that for any $t \neq 0$, $Jf(t, t) = \begin{bmatrix} 1 & -1/2 \end{bmatrix}$. So neither of the partial derivatives is continuous at 0. The partial derivatives are, however, continuous at every $x \neq 0$ and so f is $\mathcal{C}^1$ except at $x = 0$. □

In practice, I either assume (for abstract functions) that the functions are $\mathcal{C}^1$ or choose explicit functional forms that are $\mathcal{C}^1$.

Example 7. $g : \mathbb{R} \to \mathbb{R}^2$, $g(x) = (x, x^2)$, $f : \mathbb{R}^2 \to \mathbb{R}$, $f(y_1, y_2) = y_1^2 + y_2$, hence

$$Jg(x) = \begin{bmatrix} 1 \\ 2x \end{bmatrix}$$

and

$$Jf(x) = \begin{bmatrix} 2y_1 & 1 \end{bmatrix}.$$

The partials are continuous (they are linear) and hence g and f are $\mathcal{C}^1$ and $Dg(x) = Jg(x)$, $Df(x) = Jf(x)$.

To illustrated the Chain Rule (Theorem 1), let $h : \mathbb{R} \to \mathbb{R}$, $h = f \circ g$. Then $h(x) = f(g(x)) = 2x^2$ and $Dh(x) = 4x$. On the other hand, by the Chain Rule, since $g(x) = (x, x^2)$,

$$\begin{aligned} Dh(x) &= Df(x, x^2)Dg(x) \\ &= \begin{bmatrix} 2x & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2x \end{bmatrix} \\ &= 4x. \end{aligned}$$

□

If f is differentiable then it is possible to define the derivative of Df at x, written $D^2 f(x)$, but I will not do so explicitly. If this derivative exists, it is the second derivative of f. If Df is differentiable then one can consider the function $D^2 f$ that gives the second derivative for every point in the domain of f. One can define continuity of $D^2 f$, although I will not do so. Say that f is twice continuously differentiable, or f is $\mathcal{C}^2$, iff the function $D^2 f$ is well defined and continuous. And so on. An extension of Theorem 5 says that f is $\mathcal{C}^r$ iff its rth order partial derivatives all exist and are continuous. If $D^r f$ exists for every positive integer r, say that f is $\mathcal{C}^\infty$ or *smooth*.

Remark 2. One sometimes runs across notation of the following form. Suppose that $f : \mathbb{R}^2 \to \mathbb{R}$. Then some authors write,

$$\begin{aligned} \frac{df}{dx_1} &= \frac{\partial f}{\partial x_1}\frac{dx_1}{dx_1} + \frac{\partial f}{\partial x_2}\frac{dx_2}{dx_1} \\ &= \frac{\partial f}{\partial x_1} + \frac{\partial f}{\partial x_2}\frac{dx_2}{dx_1}, \end{aligned}$$

where the second equality follows from $dx_1/dx_1 = 1$. Some authors refer to the above expression as the *total derivative* (other authors, including Rudin (1976), call Df the total derivative). The usual interpretation is that the total change in f from a change in x_1 is the sum of the direct change (the first term) and an indirect change, through x_2 (the second term).

One can justify the total derivative expression as follows. Suppose that x_2 can be written as a differentiable function, call it ψ, of x_1: $x_2 = \psi(x_1)$; the total derivative implicitly assumes existence of the function ψ; I am being explicit. Then we are interested in $f(x_1, \psi(x_1))$. This is not quite the correct form for the Chain Rule. Define the auxiliary $g : \mathbb{R} \to \mathbb{R}^2$ by $g(x_1) = (x_1, \psi(x_1))$. Then we are interested in the composite function $h : \mathbb{R} \to \mathbb{R}$ given by $h(x_1) = f(g(x_1)) = f(x_1, \psi(x_1))$. By the Chain Rule, at a point $x^* = (x_1^*, x_2^*)$,

$$\begin{aligned} Dh(x_1^*) &= Df(x^*)Dg(x_1^*) \\ &= \begin{bmatrix} \frac{\partial f}{\partial x_1}(x^*) & \frac{\partial f}{\partial x_2}(x^*) \end{bmatrix} \begin{bmatrix} 1 \\ \frac{d\psi}{dx_1}(x_1^*) \end{bmatrix} \\ &= \frac{\partial f}{\partial x_1}(x^*) + \frac{\partial f}{\partial x_2}(x^*)\frac{d\psi}{dx_1}(x_1^*). \end{aligned}$$

This is the total derivative given above, provided you interpret df/dx_1 to mean $Dh = dh/dx_1$ and interpret dx_2/dx_1 to mean $d\psi/dx_1$.

The total derivative thus "works;" it is an implication of the Chain Rule. As notation, however, I find it problematic and encourage you to avoid it. It uses the same notation, f, for two different mathematical objects, the original function f and the composite function h. And it uses the same notation, x_2, both for a variable and for the function ψ, disguising the fact that one must assume the existence of ψ. □

5 Real-Valued Functions.

5.1 The tangent plane.

If $f : \mathbb{R}^N \to \mathbb{R}$ is differentiable at x^* then one can define the *tangent plane* to be the graph of the function,

$$B(x) = Df(x^*)[x - x^*] + f(x^*).$$

If $N = 1$ then the tangent plane is a line that (a) has a slope equal to that of the function f at x^* and (b) touches the graph of f at $(x^*, f(x^*))$. For $N > 1$, the tangent plane is an N dimensional flat surface.

The tangent plane can be thought of as the line or plane that best approximates the graph of f near the point $(x^*, f(x^*))$.

5.2 The gradient.

Consider $f : \mathbb{R}^N \to \mathbb{R}$. Then

$$Df(x^*) = \begin{bmatrix} D_1 f(x^*) & \cdots & D_N f(x^*) \end{bmatrix},$$

which is a row matrix. Its transpose is a column matrix, which can also be interpreted as a vector in $\mathbb{R}^N$. This vector is called the *gradient* of f at x^*, written $\nabla f(x^*)$:

$$\nabla f(x^*) \underset{\text{def}}{=} Df(x^*)' = \begin{bmatrix} D_1 f(x^*) \\ \vdots \\ D_N f(x^*) \end{bmatrix}.$$

The gradient has the following important interpretation. Let θ be the angle between $\nabla f(x^*)$ and $v \neq 0$. Then

$$\begin{aligned} Df(x^*)v &= \nabla f(x^*) \cdot v \\ &= \cos(\theta) \|\nabla f(x^*)\| \|v\|. \end{aligned}$$

(On the cosine formula for inner product, see the notes on Vector Spaces and Norms.)

Assuming that $\nabla f(x^*) \neq 0$, this implies that f is increasing the fastest when $\cos(\theta) = 1$. (If $\nabla f(x^*) = 0$ then f is not increasing in any direction.) But $\cos(\theta) = 1$ when $\theta = 0$ and if $\theta = 0$ then it must be that the v for which f is increasing the fastest are the v that are positively collinear with $\nabla f(x^*)$: the gradient points in the direction of fastest increase of f.

5.3 The Hessian.

Let $f : \mathbb{R}^N \to \mathbb{R}$ and suppose that all second-order partial derivatives exist. Then it is common practice to arrange these partial derivatives into an $N \times N$ matrix called the *Hessian* of f at x^*, which I denote $Hf(x^*)$:

$$Hf(x^*) \underset{\text{def}}{=} \begin{bmatrix} D^2_{11} f(x^*) & \cdots & D^2_{1N} f(x^*) \\ \vdots & & \vdots \\ D^2_{N1} f(x^*) & \cdots & D^2_{NN} f(x^*) \end{bmatrix},$$

where

$$D^2_{ij} f(x^*) \underset{\text{def}}{=} D_i(D_j f)(x^*) \underset{\text{def}}{=} \frac{\partial^2 f}{\partial x_i \partial x_j}(x^*).$$

(The term Hessian is sometimes also used for the determinant of this matrix.)

Example 8. Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is given by $f(x) = \ln(x_1)\ln(x_2)$. Then

$$Df(x) = \left[\begin{array}{cc} \frac{\ln(x_2)}{x_1} & \frac{\ln(x_1)}{x_2} \end{array} \right],$$

and

$$D^2(x) = \left[\begin{array}{cc} -\frac{\ln(x_2)}{x_1^2} & \frac{1}{x_1 x_2} \\ \frac{1}{x_1 x_2} & -\frac{\ln(x_1)}{x_2^2} \end{array} \right].$$

Thus

$$D^2(1,1) = \left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array} \right].$$

□

In this example, the mixed partials are equal: $D^2_{12}f = D^2_{21}f$. This is true whenever f is $\mathcal{C}^2$, as formalized by Theorem 6. The result goes by a number of different names; in economics it is typically referred to as Young's Theorem.

Theorem 6 (Young's Theorem.)**.** *Let $f : \mathbb{R}^N \to \mathbb{R}$ be $\mathcal{C}^2$. Then the Hessian of f is symmetric: $D^2_{ij}f(x^*) = D^2_{ji}f(x^*)$ for i and j.*

Proof. The proof is in Section 6; see also Remark 3 in that section. ■

In the next example, symmetry fails even though the function is twice differentiable, but not $\mathcal{C}^2$.

Example 9. Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is given by

$$f(x) = \begin{cases} \frac{x_1 x_2 (x_1^2 - x_2^2)}{x_1^2 + x_2^2} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

One can show that f is twice differentiable but it is not $\mathcal{C}^2$ at the origin: neither $D^2_{12}f$ nor $D^2_{21}f$ is continuous at the origin. Direct calculation shows that $D_{12}f(0) = 1 \neq -1 = D_{21}f(0)$. □

If f is twice differentiable then ∇f, which is a function from $\mathbb{R}^N$ to $\mathbb{R}^N$, is differentiable. The derivative of ∇f is (strictly speaking) the transpose of $Hf(x^*)$. By Young's Theorem, if f is $\mathcal{C}^2$, then

$$D(\nabla f)(x^*) = Hf(x^*).$$

One can think of $Hf(x^*)$ as a machine for computing second derivatives along lines in the domain. Explicitly, fix any $x^* \in \mathbb{R}^N$ and any direction $v \in \mathbb{R}^N$. Let $g(\gamma) = x^* + \gamma v$. Let $F : \mathbb{R} \to \mathbb{R}$, $F(\gamma) = f(g(\gamma))$. $D^2F(0)$ is thus the second derivative of f, evaluated at the point x^*, along the line through x^* given by g. One can compute that $D^2F(0)$, which is the second derivative of a univariate function, satisfies

$$D^2F(0) = v'Hf(x^*)v.$$

6 Proofs for Theorems 1, 5, and 6.

Proof of Theorem 1 (Chain Rule). To ease notation at least somewhat, let $A = Df(y^*)$, $B = Dg(x^*)$. I need to show that

$$\lim_{w \to 0} \frac{\|h(x^* + w) - h(x^*) - ABw\|}{\|w\|} = 0.$$

For any w, define

$$\zeta(w) = g(x^* + w) - g(x^*) - Bw,$$

and, for $w \neq 0$,

$$\eta(w) = \frac{\|\zeta(w)\|}{\|w\|}.$$

By definition of B, $\lim_{w \to 0} \eta(w) = 0$.

For any w, let

$$v(w) = g(x^* + w) - g(x^*).$$

Note that $\lim_{w \to 0} v(w) = 0$. Note also that

$$\zeta(w) = v(w) - Bw.$$

Define

$$\phi(w) = f(y^* + v(w)) - f(y^*) - Av(w),$$

and, provided $v(w) \neq 0$,

$$\varepsilon(w) = \frac{\|\phi(w)\|}{\|v(w)\|}.$$

(If $\varepsilon(w) = 0$ then $g(x^* + w) = g(x^*)$. If this holds for every w then g is constant, hence h is constant, hence both $Dg(x^*)$ and $Dh(x^*)$ are matrices of zeroes.) By definition of $A = Df(x^*)$, $\lim_{w \to 0} \varepsilon(w) = 0$.

Now,

$$\|v(w)\| = \|\zeta(w) + Bw\| \leq \|\zeta(w)\| + \|Bw\|.$$

From above, $\|\zeta(w)\| = \eta(w)\|w\|$.

As for $\|Bw\|$, let $S = \{x \in \mathbb{R}^N : \|x\| = 1\}$ be the solid unit sphere centered on the origin and define

$$\|B\|_\mu = \max_{x \in S} \|Bx\|.$$

$\|B\|_\mu$ is well defined since S is compact and $\|Bx\|$ is continuous as a function of x (see the notes on Continuity).

For any $w \neq 0$, $w/\|w\| \in S$ and therefore $\|Bw\| = \|B(w/\|w\|)(\|w\|)\| \leq \|B\|_\mu \|w\|$. Therefore,

$$\|v(w)\| \leq [\eta(w) + \|B\|_\mu] \, \|w\|. \tag{1}$$

Also, noting that $f(y^*+v(w))-f(y^*) = \phi(w)+Av(w)$ and recalling that $v(w)-Bw = \zeta(w)$,

$$\begin{aligned} h(x^* + w) - h(x^*) - ABw &= f(y^* + v(w)) - f(y^*) - ABw \\ &= \phi(w) + Av(w) - ABw \\ &= \phi(w) + A\left[v(w) - Bw\right] \\ &= \phi(w) + A\zeta(w). \end{aligned}$$

Combining this with inequality (1), defining $\|A\|_\mu$ in a manner analogous to $\|B\|_\mu$, and recalling from above that $\|\phi(w)\| = \varepsilon(w)\|v(w)\|$ and $\|\zeta(w)\| = \eta(w)\|w\|$, yields, for $w \neq 0$,

$$\begin{aligned} \frac{\|h(x^* + w) - h(x^*) - ABw\|}{\|w\|} &= \frac{\|\phi(w) + A\zeta(w)\|}{\|w\|} \\ &\leq \frac{\|\phi(w)\|}{\|w\|} + \frac{\|A\zeta(w)\|}{\|w\|} \\ &\leq \frac{\varepsilon(w)\|v(w)\|}{\|w\|} + \frac{\|A\|_\mu\|\zeta(w)\|}{\|w\|} \\ &\leq \frac{\varepsilon(w)\left[\eta(w) + \|B\|_\mu\right]\|w\|}{\|w\|} + \frac{\|A\|_\mu\left[\eta(w)\|w\|\right]}{\|w\|} \\ &= \varepsilon(w)\left[\eta(w) + \|B\|_\mu\right] + \|A\|_\mu\eta(w). \end{aligned}$$

As $w \to 0$, the right-hand side goes to 0, since $\eta(w) \to 0$ and $\varepsilon(w) \to 0$. ■

Proof of Theorem 5. ⇒. By Theorem 4, if Df is differentiable then, for any $x^* \in \mathbb{R}^N$, $Jf(x^*)$ exists and $Df(x^*) = Jf(x^*)$. Therefore, by the definition of continuity of Df at x^*, the partial derivatives that constitute $Jf(x^*)$ are continuous at x^*.

⇐. By Theorem 2, it suffices, for showing that f is differentiable, to focus on the case $M = 1$. Continuity of Df then follows from the definition and the fact that, for any $x^* \in \mathbb{R}^N$, $Df(x^*) = Jf(x^*)$.

It remains to prove that f is differentiable. Fix x^*. I will show that

$$Jf(x^*) = \begin{bmatrix} D_1 f(x^*) & \cdots & D_N f(x^*) \end{bmatrix}$$

satisfies the definition of derivative. Consider any sequence (w_t) in $\mathbb{R}^N$ such that for all t, $w_t \neq 0$ and $\|w_t\|_{\max} < 1/t$.

The notation below gets somewhat messy but the basic idea of the argument is to decompose the derivative ratio,

$$\frac{|f(x^* + w_t) - f(x^*) - Jf(x^*)w_t|}{\|w_t\|},$$

into a sum of N expressions, each involving the change of only a single coordinate, then apply the Mean Value Theorem to transform these N expressions into differences in partial derivatives, and finally appeal to continuity of the partial derivatives to argue that the overall expression must converge to zero, which implies the result.

Explicitly, the expression $f(x^* + w_t) - f(x^*)$ can be decomposed as the sum of N terms, each reflecting a change in a single coordinate, starting with

$$f(x_1^* + w_{t1}, \ldots, x_N^* + w_{tN}) - f(x_1^*, x_2^* + w_{t2}, \ldots, x_N^* + w_{tN})$$

then

$$f(x_1^*, x_2^* + w_{t2}, \ldots, x_N^* + w_{tN}) - f(x_1^*, x_2^*, x_3^* + w_{t3}, \ldots, x_N^* + w_{tN})$$

and ending with,

$$f(x_1^*, \ldots, x_{N-1}^*, x_N^* + w_{tN}) - f(x^*).$$

By the Mean Value Theorem, for each of these univariate changes, there is a θ_n in the interval $N_{|w_{tn}|}(x_n^*)$ (I allow $w_{tn} < 0$) such such that the change in the value of f equals the derivative, evaluated with θ_n in coordinate n, times the change in x_n, namely w_{tn}. For example, for $n = 1$, there is a θ_1 such that

$$f(x^* + w_t) - f(x_1^*, x_2^* + w_{t2}, \ldots, x_N^* + w_{tN}) = Df(\theta_1, x_2^* + w_{t2}, \ldots, x_N^* + w_{tN})w_{t1}.$$

To simplify notation, let

$$\begin{aligned} c_t^1 &= (\theta_1, x_2^* + w_{t2}, \ldots, x_N^* + w_{tN}), \\ c_t^2 &= (x_1^*, \theta_2, x_3^* + w_{t3}, \ldots, x_N^* + w_{tN}), \end{aligned}$$

down through

$$c_t^N = (x_1^*, \ldots, x_{N-1}^*, \theta_N).$$

Thus, for example,

$$f(x^* + w_t) - f(x_1^*, x_2^* + w_{t2}, \ldots, x_N^* + w_{tN}) = Df(c_t^1)w_{t1}.$$

Note that each c_t^n is in the $1/t$ ball (or cube) around x^* in the max metric, and thus the c_t^n all converge to x^*.

It follows that

$$\begin{aligned} \frac{|f(x^* + w_t) - f(x^*) - Jf(x^*)w_t|}{\|w_t\|} &= \frac{|(\sum_{n=1}^N D_n f(c_t^n)w_{tn}) - Jf(x^*)w_t|}{\|w_t\|} \\ &= \frac{|\sum_{n=1}^N (D_n f(c_t^n)w_{tn} - D_n f(x^*)w_{tn})|}{\|w_t\|} \\ &\leq \sum_{n=1}^N |D_n f(c_t^n) - D_n f(x^*)| \frac{|w_{tn}|}{\|w_t\|} \\ &\leq \sum_{n=1}^N |D_n f(c_t^n) - D_n f(x^*)|, \end{aligned}$$

where the last inequality comes from the fact that $|w_{tn}| \leq \|w_t\|_{\max} \leq \|w_t\|$. Since $c_t^n \to x^*$ and $D_n f$ is continuous, it follows that the right-hand side converges to zero and hence $Jf(x^*)$ satisfies the definition for $Df(x^*)$. ■

Proof of Theorem 6 (Young's Theorem). For notational simplicity, I assume that $N = 2$; this is without loss of generality, since only two coordinates are involved.

For any $\gamma = (\gamma_1, \gamma_2) \gg 0$, define

$$\Delta(\gamma) = f(x^* + \gamma) - f(x_1^* + \gamma_1, x_2^*) - f(x_1^*, x_2^* + \gamma_2) + f(x_1^*, x_2^*).$$

The underlying idea of the proof is that if f is $\mathcal{C}^2$ then $\Delta(\gamma)/(\gamma_1\gamma_2)$ is a discrete approximation to both $D^2_{21}f(x^*)$ and to $D^2_{12}f(x^*)$, which implies, for γ sufficiently small, that these two derivatives must be close to each other.

For any $\gamma \gg 0$, I claim that there is an x in the rectangle $(x_1^*, x_1^* + \gamma_1) \times (x_2^*, x_2^* + \gamma_2)$ such that

$$\frac{\Delta(\gamma)}{\gamma_1\gamma_2} = D^2_{21}f(x).$$

One can view this is an $N = 2$ version of the Mean Value Theorem. To see this, define $g : \mathbb{R} \to \mathbb{R}$ by

$$g(x_1) = f(x_1, x_2^* + \gamma_1) - f(x_1, x_2^*).$$

g is differentiable since f is. Then,

$$\Delta(\gamma) = g(x_1^* + \gamma_1) - g(x_1^*).$$

By the Mean Value Theorem, there is an $x_1 \in (x_1^*, x_1^* + \gamma_1)$ such that

$$g(x_1^* + \gamma_1) - g(x_1^*) = Dg(x_1)\gamma_1,$$

hence

$$\begin{aligned}\Delta(\gamma) &= Dg(x_1)\gamma_1 \\ &= D_1 f(x_1, x_2^* + \gamma_2)\gamma_1 - D_1 f(x_1, x_2^*)\gamma_1.\end{aligned}$$

Applying the Mean Value Theorem again, this time with respect to the second coordinate, there is an $x_2 \in (x_2^*, x_2^* + \gamma_2)$ such that, setting $x = (x_1, x_2)$,

$$D_1 f(x_1, x_2^* + \gamma_2)\gamma_1 - D_1 f(x_1, x_2^*)\gamma_1 = D^2_{21}f(x)\gamma_1\gamma_2$$

implying $\Delta(\gamma) = D^2_{21}f(x)\gamma_1\gamma_2$, which proves the above claim.

Fix $\varepsilon > 0$. By continuity of $D^2_{21}f$, for γ sufficiently small, and hence x sufficiently close to x^*,

$$\left|D^2_{21}f(x) - D^2_{21}f(x^*)\right| < \varepsilon/2,$$

and hence, by the above claim,

$$\left|\frac{\Delta(\gamma)}{\gamma_1\gamma_2} - D^2_{21}f(x^*)\right| < \varepsilon/2.$$

A similar argument, this time exploiting continuity of $D^2_{12}f$, yields, again for γ sufficently small,

$$\left|\frac{\Delta(\gamma)}{\gamma_1\gamma_2} - D^2_{12}f(x^*)\right| < \varepsilon/2.$$

Hence, by the triangle inequality,

$$\left|D^2_{12}f(x^*) - D^2_{21}f(x^*)\right| < \varepsilon,$$

Since this must hold for all $\varepsilon > 0$, the result follows. ■

Remark 3. A variant of the above proof requires that only one of the second-order cross partials be continuous; hence the requirement that f be $\mathcal{C}^2$ is stronger than necessary. See Rudin (1976). I have chosen to prove a weaker result (using a stronger assumption) because I find this proof more transparent and in applications the weaker result is almost invariably good enough. □

References

Rudin, Walter. 1976. *Principles of Mathematical Analysis.* third ed. New York: McGraw-Hill.