Differentiability

Alex Nita

Abstract

In this section we try to develop the basics of differentiability of vector-valued functions $f = (f_1, \ldots, f_m) : \mathbb{R}^n \to \mathbb{R}^m$. It turns out that, as with continuity, it is enough to know how to differentiate the component functions $f_i : \mathbb{R}^n \to \mathbb{R}$, i.e. how to differentiate *real-valued* functions of several variables. Of course, with more dimensions come more ways to differentiate. We can differentiate in different directions as well as in some overall sense, and these are related in a specific way, namely by a matrix of partial derivatives. The necessary hypotheses are spelled out in great detail, and you should study these carefully. They are important. Also, the notation and use of matrices, though standard in math, differ somewhat from the textbook. These notes are a supplement, of course, but I believe they give a unified view of differentiation, and add some important features that are missing from the book. Namely, we see that we may take directional derivatives of vector-valued, and not just real-valued, functions, and additionally we give a slick presentation of the chain rule in terms of total derivatives, which has the side benefit of explaining why, componentwise, in terms of partial derivatives, the chain rule is a sum–because that's how matrix products work! The theorems are stated first, and their proofs are postponed to the Appendix at the end, because they are technical and would probably clutter the notes with unnecessary detail–but for those curious to see why the theorems are true, they are there at the back.

1 The Total Derivative

The thing we want to do now is to locally appproximate a complicated function $f : \mathbb{R}^n \to \mathbb{R}^m$ by a much simpler *linear* function. This is the idea behind the total derivative of f at a point $\mathbf{a}$ in $\mathbb{R}^n$. Formally, we say that f is **differentiable at a point a** $\in \mathbb{R}^n$ if there exists an $m \times n$ matrix (m and n here depend on the domain and range of f!)

$$Df(\mathbf{a}) = \begin{pmatrix} m_{11} & \cdots & m_{1n} \\ \vdots & \ddots & \vdots \\ m_{m1} & \cdots & m_{mn} \end{pmatrix} \qquad \text{(a real } m \times n \text{ matrix)} \tag{1.1}$$

called the **total derivative** (or the **Jacobi matrix**), which satisfies the following limit condition:

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{|f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) - Df(\mathbf{a})\mathbf{h}|}{|\mathbf{h}|} = 0 \tag{1.2}$$

Equivalently, f must locally be approximated by a linear function, that is

$$f(\mathbf{a}+\mathbf{h}) \approx Df(\mathbf{a})\mathbf{h} + f(\mathbf{a}) \tag{1.3}$$

where the **error** in the approximation

$$\begin{aligned} E(\mathbf{h}) &= f(\mathbf{a}+\mathbf{h}) - \big(\text{linear approximation of the value of } f \text{ at } \mathbf{x} = \mathbf{a}+\mathbf{h}\big) \\ &= f(\mathbf{a}+\mathbf{h}) - \Big(Df(\mathbf{a})\mathbf{h} + f(\mathbf{a})\Big) \end{aligned}$$

satisfies

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{|E(\mathbf{h})|}{|\mathbf{h}|} = 0 \tag{1.4}$$

This last statement, (1.4), is obviously the same statement as (1.2).

Remark 1.1 *The expression $Df(\mathbf{a})\mathbf{h}$ denotes matrix multiplication. Here, $\mathbf{h} = (h_1, \dots, h_n)$ is a vector in $\mathbb{R}^n$ thought of as an $n \times 1$ column vector:*

$$Df(\mathbf{a})\mathbf{h} = \begin{pmatrix} m_{11} & \cdots & m_{1n} \\ \vdots & \ddots & \vdots \\ m_{m1} & \cdots & m_{mn} \end{pmatrix} \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix} = \begin{pmatrix} m_{11}h_1 + \cdots + m_{1n}h_n \\ \vdots \\ m_{m1}h_1 + \cdots + m_{mn}h_1 \end{pmatrix} \qquad \blacksquare$$

Remark 1.2 *Just as in the single variable case, where $h = \Delta x = x - a$, so, too, here:*

$$\mathbf{h} = \Delta\mathbf{x} = \mathbf{x} - \mathbf{a} = \begin{pmatrix} x_1 - a_1 \\ \vdots \\ x_n - a_n \end{pmatrix}$$

Thus, the idea is that going a distance of $\mathbf{h}$ away from $\mathbf{a}$, to the point $\mathbf{x}$, we may approximate the value of $f(\mathbf{x}) = f(\mathbf{a} + \mathbf{h})$ by the "tangent plane" $Df(\mathbf{a})\Delta\mathbf{x}$. Referring to the notes on Points, Vectors and Matrices, this approximation, $Df(\mathbf{a})\Delta\mathbf{x}$, is literally *the tangent plane approximation in the real-valued case of $m = 1$, for then*

$$\begin{aligned} Df(\mathbf{a})\Delta\mathbf{x} &= \begin{pmatrix} m_1 & \dots & m_n \end{pmatrix} \begin{pmatrix} x_1 - a_1 \\ \vdots \\ x_n - a_n \end{pmatrix} \\ &= m_1(x_1 - a_1) + \cdots + m_n(x_n - a_n) \\ &= m_1x_1 + m_2x_2 + \cdots + m_nx_n + d \end{aligned}$$

where $d = -m_1a_1 - \cdots - m_na_n$, which is precisely the expression for an n-dimensional plane. For if this equals some number, k, then letting $d' = d - k$ we get the equation of the plane, $m_1x_1 + m_2x_2 + \cdots + m_nx_n + d' = 0$. In the general case, where m may be greater than 1, we have some higher-dimensional analog of the plane. ■

Remark 1.3 *Thus to say that a function f is differentiable at a point is equivalent to saying that f has a total derivative there. We shall see that f may be* partially *differentiable, and to have directional derivatives in all directions, yet not be differentiable. We will explain this further below.* ■

2 The Directional Derivative

Suppose $\mathbf{v}$ is a 'vector' in $\mathbb{R}^n$ and $\mathbf{a}$ is a 'point' in $\mathbb{R}^n$, and let $T : \mathbb{R} \to \mathbb{R}^n$ be the translation-by-$t\mathbf{v}$ function

$$T(t) = \mathbf{x} + t\mathbf{v}$$

We say that $f : \mathbb{R}^n \to \mathbb{R}^m$ has a **directional derivative** at $\mathbf{a}$ in the direction of $\mathbf{v}$ if the composition $f \circ T : \mathbb{R} \to \mathbb{R}^m$, $(f \circ T)(t) = f(\mathbf{a} + t\mathbf{v})$, is "differentiable at 0", in the sense that the limit

$$\begin{aligned} f_{\mathbf{v}}(\mathbf{a}) \ \text{ or } \ D_{\mathbf{v}}f(\mathbf{a}) &= \left.\frac{d}{dt}\right|_{t=0} f(\mathbf{a} + t\mathbf{v}) \\ &= \lim_{t \to 0} \frac{f(\mathbf{a} + t\mathbf{v}) - f(\mathbf{a})}{t} \end{aligned} \tag{2.1}$$

exists in $\mathbb{R}^m$.

Remark 2.1 *Notice that for each fixed nonzero t the difference $f(\mathbf{a} + t\mathbf{v}) - f(\mathbf{a})$ in the numerator is a vector in $\mathbb{R}^m$, while $1/t$ is a real number, so the quotient $\frac{f(\mathbf{a}+t\mathbf{v})-f(\mathbf{a})}{t}$ is actually scalar multiplication of the vector $f(\mathbf{a} + t\mathbf{v}) - f(\mathbf{a})$ by $1/t$.*

Another thing to notice is that in the sum $\mathbf{a} + t\mathbf{v}$ we added a point to a vector. Since we have emphasized blurring the lines between points and vectors in $\mathbb{R}^n$, on account of algebraically *they are indistinguishable, at least in $\mathbb{R}^n$, the sum makes sense.* ■

3 The Partial Derivative

Now suppose f is a real-valued function, $f : \mathbb{R}^n \to \mathbb{R}$. The i**th partial derivative** of f at $\mathbf{a}$ is the directional derivative of f at $\mathbf{a}$ in a coordinate direction, that is in the direction of a unit coordinate vector $\mathbf{e}_i = (0, \ldots, i, \ldots, 0)$,

$$\begin{aligned} \left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{x}=\mathbf{a}} \ \text{ or } \ f_{x_i}(\mathbf{a}) \ \text{ or } \ D_i f(\mathbf{a}) &= \left.\frac{d}{dt}\right|_{t=0} f(\mathbf{a} + t\mathbf{e}_i) \\ &= \lim_{t \to 0} \frac{f(\mathbf{a} + t\mathbf{e}_i) - f(\mathbf{a})}{t} \\ &= \lim_{t \to 0} \frac{f(a_1, \ldots, a_{i-1}, a_i + t, a_{i+1}, \ldots, a_n) - f(a_1, \ldots, a_i, \ldots, a_n)}{t} \end{aligned} \tag{3.1}$$

Remark 3.1 *The practical import of this definition will become clear in a minute. For now, notice that the derivative $\left.\frac{d}{dt}\right|_{t=0} f(\mathbf{a} + t\mathbf{e}_i)$ is an ordinary derivative from Calc 1. It's the derivative of the real-valued function of a real variable $f \circ T : \mathbb{R} \to \mathbb{R}$, $(f \circ T)(t) = f(\mathbf{a} + t\mathbf{e}_i)$. This means that all the other coordinates, which we normally treat as variables, since they may vary, are treated here as constants. Thus, if we label the variables $x_1, \ldots, x_i, \ldots, x_n$, all the other x_j for $j \neq i$ are treated as constants in any expression for f. For example, if $f(x, y, z) = xyz + x^2y + z^2$, in the partial derivative $\frac{\partial f}{\partial x}$ with respect to x the 'variables' y and z are treated as constants, so we can do what we normally do when computing a Calc 1 derivative, pull the constants out. Here, for example, we'd have $\frac{\partial f}{\partial x} = yz + 2xy$.* ■

4 The Relationship Between the Total and Directional and Partial Derivatives

Nobody wants to compute an actual limit, though the limit idea is extremely important theoretically. Luckily, we don't have to here. The directional derivative, though defined in terms of a limit, is in fact computable in terms of a matrix product!

Theorem 4.1 *If $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\mathbf{a}$ in $\mathbb{R}^n$, then all of its directional derivatives at $\mathbf{a}$ exist, and for any choice of vector $\mathbf{v}$ in $\mathbb{R}^n$ we have*

$$\boxed{D_{\mathbf{v}} f(\mathbf{x}) = Df(\mathbf{x})\mathbf{v}} \tag{4.1}$$

The left-hand side is a limit, while the right-hand side is a matrix product, with $\mathbf{v}$ treated as a column vector. ∎

Example 4.2 *Suppose $f : \mathbb{R}^3 \to \mathbb{R}^2$ is differentiable at the point $\mathbf{a} = (1, 1, 2)$ and $\mathbf{v} = (-1, 4, 2)$ is a vector in $\mathbb{R}^3$. If we already knew the total derivative of f, say*

$$Df(\mathbf{a}) = \begin{pmatrix} 2 & 3 & 2 \\ 1 & 0 & 5 \end{pmatrix}$$

then computing the directional derivative of f at $\mathbf{a}$ in the direction of $\mathbf{v}$ would be easy, namely

$$f_{\mathbf{v}}(\mathbf{a}) = Df(\mathbf{a})\mathbf{v} = \begin{pmatrix} 2 & 3 & 2 \\ 1 & 0 & 5 \end{pmatrix} \begin{pmatrix} -1 \\ 4 \\ 2 \end{pmatrix} = \begin{pmatrix} 14 \\ 9 \end{pmatrix}$$ ∎

Thus, if f is differentiable at a, the task is to find a way to compute $Df(\mathbf{a})$. For then we can compute all directional derivatives by simple matrix multiplication. Luckily, we can do this, using the previous theorem, provided we know the that the total derivative $Df(\mathbf{a})$ has rows consisting of the total derivatives of the component functions, that is

$$\underbrace{D\begin{pmatrix} f_1(\mathbf{a}) \\ \vdots \\ f_m(\mathbf{a}) \end{pmatrix}}_{Df(\mathbf{a})} = \begin{pmatrix} Df_1(\mathbf{a}) \\ \vdots \\ Df_m(\mathbf{a}) \end{pmatrix}$$

In this case I claim that **the total derivative $Df(\mathbf{a})$ is the matrix of partial derivatives of the component functions f_i of f,**

$$\boxed{Df(\mathbf{a}) = \begin{pmatrix} \left.\dfrac{\partial f_1}{\partial x_1}\right|_{\mathbf{a}} & \cdots & \left.\dfrac{\partial f_1}{\partial x_n}\right|_{\mathbf{a}} \\ \vdots & \ddots & \vdots \\ \left.\dfrac{\partial f_m}{\partial x_1}\right|_{\mathbf{a}} & \cdots & \left.\dfrac{\partial f_m}{\partial x_n}\right|_{\mathbf{a}} \end{pmatrix}} \tag{4.2}$$

To see this, take a real-valued function first, $f : \mathbb{R}^n \to \mathbb{R}$ (for any vector-valued function as above is made up of its m real-valued component functions) and look at the directional derivative in the ith coordinate direction $\mathbf{e}_i = (0, \ldots, 1, \ldots, 0)$ (it has a 1 in the ith slot and 0 everywhere else). Letting

$$Df(\mathbf{a}) = \begin{pmatrix} m_1 & m_2 & \cdots & m_n \end{pmatrix}$$

be the $1 \times n$ matrix defining the total derivative of f, and noting that the ith partial derivative is the directional derivative in the ith coordinate direction, we have

$$\begin{aligned} \left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{a}} = D_{\mathbf{e}_i} f(\mathbf{a}) = Df(\mathbf{a})\mathbf{e}_i &= \begin{pmatrix} m_1 & m_2 & \cdots & m_n \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \\ &= m_1 \cdot 0 + m_2 \cdot 0 + \cdots m_i \cdot 1 + \cdots m_n \cdot 0 \\ &= m_i \end{aligned}$$

This is true for each $i = 1, \ldots, n$, so

$$Df(\mathbf{a}) = \begin{pmatrix} m_1 & m_2 & \cdots & m_n \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{pmatrix}$$

Now consider a vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$, $f = (f_1, \ldots, f_m)$. Then, each of its component functions f_i is a real-valued function, and the above result applies separately to each, from which we get our result,

$$Df(\mathbf{a}) = D \begin{pmatrix} f_1(\mathbf{a}) \\ \vdots \\ f_m(\mathbf{a}) \end{pmatrix} = \begin{pmatrix} Df_1(\mathbf{a}) \\ \vdots \\ Df_m(\mathbf{a}) \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

The only questionable thing about this is the legality of the second equality, where we "pulled the D inside the column vector." It was, in fact, legal, and moreover our ability to do this gives us another useful way to decide the differentiability of a function. Let us state and prove this result carefully.

Theorem 4.3 *A function $f : \mathbb{R}^n \to \mathbb{R}^m$, $f = (f_1, \ldots, f_n)$, is differentiable at $\mathbf{a}$ if and only if each of its component functions $f_i : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\mathbf{a}$. In that case, we have*

$$Df(\mathbf{a}) = D \begin{pmatrix} f_1(\mathbf{a}) \\ \vdots \\ f_m(\mathbf{a}) \end{pmatrix} = \begin{pmatrix} Df_1(\mathbf{a}) \\ \vdots \\ Df_m(\mathbf{a}) \end{pmatrix} \tag{4.3}$$

That is, to compute $Df(\mathbf{a})$ we can just compute the $1 \times n$ derivative matrices of the f_i first, which we know are of the form $Df_i(\mathbf{a}) = \begin{pmatrix} \left.\frac{\partial f_i}{\partial x_1}\right|_{\mathbf{a}} & \cdots & \left.\frac{\partial f_i}{\partial x_n}\right|_{\mathbf{a}} \end{pmatrix}$, and enter the result into the ith row of the larger matrix. ■

Now that we know how to compute $Df(\mathbf{a})$ if we know that $Df(\mathbf{a})$ exists, we have to answer the question, **"How do we determine the existence of $Df(\mathbf{a})$?"**. Well, we have seen that it boils down do determining the existence of the m separate total derivatives of the component functions $Df_i(\mathbf{a})$. The remaining question, therefore, is, **"How do we determine the existence of the m separate total derivatives $Df_i(\mathbf{a})$ of the component functions f_i?"** The naïve answer is, "Well, just compute the partials $\frac{\partial f_i}{\partial x_j}$ at $\mathbf{a}$ of each f_i and put them in a matrix," unfortunately, is not entirely correct. It *would* be if we knew that the partials were also *continuous* on a neighborhood of the point $\mathbf{a}$, but not otherwise. Here is an **example of why the existence of the partials $\frac{\partial f_i}{\partial x_j}$ at a alone is not enough to conclude the existence of $Df(\mathbf{a})$ (we must also have their continuity)**:

Example 4.4 *Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by*

$$f(x,y) = \begin{cases} \dfrac{x^2 y}{x^4 + y^2}, & \text{if } (x,y) \neq (0,0) \\ 0, & \text{if } (x,y) = (0,0) \end{cases}$$

First, notice that all its directional derivatives exist at the origin, for if $\mathbf{v} = (h,k)$ is any vector in $\mathbb{R}^2$, then the directional derivative $D_{\mathbf{v}} f(\mathbf{0})$ is computable directly:

$$D_{\mathbf{v}} f(\mathbf{0}) = \lim_{t\to 0} \frac{f(\mathbf{0} + t\mathbf{v}) - f(\mathbf{0})}{t} = \lim_{t\to 0} \frac{(th)^2(tk) - 0}{(th)^4 + (tk)^2} \cdot \frac{1}{t} = \lim_{t\to 0} \frac{\cancel{t^3} h^2 k}{\cancel{t^3}(t^2h^4 + k^2)} = \begin{cases} \dfrac{h^2}{k} & \text{if } k \neq 0 \\ 0 & \text{if } k = 0 \end{cases}$$

In particular, choosing $\mathbf{v} = \mathbf{e}_1 = (1,0)$ and $\mathbf{v} = \mathbf{e}_2 = (0,1)$ shows that it has partial derivatives $\frac{\partial f}{\partial x}|_{(0,0)} = \frac{\partial f}{\partial y}|_{(0,0)} = 0$ at the origin. Outside the origin it is easily seen to be partially differentiable, and its partial derivatives exist everywhere on $\mathbb{R}^2$, and are given by

$$\frac{\partial f}{\partial x} = \begin{cases} \dfrac{(x^4+y^2)2xy - 4x^5y}{(x^4+y^2)^2}, & \text{if } (x,y) \neq (0,0) \\ (0,0), & \text{if } (x,y) = (0,0) \end{cases}$$
$$\frac{\partial f}{\partial y} = \begin{cases} \dfrac{(x^4+y^2)x^2 - 2x^2y^2}{(x^4+y^2)^2}, & \text{if } (x,y) \neq (0,0) \\ (0,0), & \text{if } (x,y) = (0,0) \end{cases}$$

Thus, f is partially differentiable everywhere in $\mathbb{R}^2$. However, f is not differentiable at the origin, in fact it is not even continuous there. For notice that on the parabola $y = x^2$ the function is constant with value $1/2$:

$$f(h,h^2) = \frac{h^4}{2h^4} = \frac{1}{2}$$

so that arbitrarily close to the origin there are points for which $f(x,y) = 1/2$, while $f(0,0) = 0$. On the other hand, along any straight line $y = mx$ the function satisfies

$$f(x,mx) = \frac{mx^3}{x^2(x^2+m^2)} = \frac{mx}{x^2+m^2}$$

so f approaches 0 along straight lines. By one of your homework problems, however, all differentiable functions must be continuous, so we conclude that f is not differentiable at the origin. (We prove that differentiability implies continuity below!)

The problem here, of course, is that the partials $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ are not continuous at the origin. For example, $\frac{\partial f}{\partial x}$ approaches 0 along the parabola $y = x^2$ while it diverges to $-\infty$ along the line $y = x$. (Check this!) ■

Remark 4.5 *The problem point* $(0,0)$ *isn't special. We could make any point a problem point, for example* $(1,5)$*, by translating the above example function by* $(1,5)$*, i.e. by considering* $f(x,y) = \frac{(x-1)^2(y-5)}{(x-1)^4+(y-5)^2}$ *when* $(x,y) \neq (1,5)$ *and* $f(0,0) = (0,0)$*.* ■

OK, so now we know that the mere existence of the partials $\frac{\partial f_i}{\partial x_j}\big|_{\mathbf{a}}$ of $f = (f_1, \ldots, f_m)$ isn't enough to ensure the existence of $Df(\mathbf{a})$. What we need is the *continuity of the partials* $\frac{\partial f_i}{\partial x_j}$ *on a neighborhood of* $\mathbf{a}$.

Theorem 4.6 *Let* $f : \mathbb{R}^n \to \mathbb{R}^m$*. If all the partial derivatives* $\frac{\partial f_i}{\partial x_j}\big|_{\mathbf{a}}$ *of* f *exist and are continuous at* $\mathbf{a}$*, then* f *is differentiable at* $\mathbf{a}$*.* ■

Conclusion: If we know that the component functions of $f = (f_1, \ldots, f_m)$ are each *continuously* differentiable on a neighborhood of our point a in $\mathbb{R}^n$, then we know that the f_i, and therefore f itself, are differentiable, and the total derivative $Df(\mathbf{a})$ is in fact the $m \times n$ matrix of partial derivatives $\frac{\partial f_i}{\partial x_j}\big|_{\mathbf{a}}$!

Example 4.7 *Let* $f : \mathbb{R}^3 \to \mathbb{R}^2$ *be given by* $f(x,y,z) = (x^2+y-z,\ e^{xy}\sin z + xz)$*. Then* $f_1(x,y,z) = x^2+y-z$ *and* $f_2(x,y,z) = e^{xy}\sin z + xz$ *are each clearly continuously differentiable in each partial derivative (for example,* $\frac{\partial f_1}{\partial x} = 2x$ *is continuous on all of* $\mathbb{R}^n$*). Therefore,* f *is differentiable and, say at* $\mathbf{a} = (1,1,2)$*, we have*

$$
\begin{aligned}
Df(1,1,2) &= \begin{pmatrix} \frac{\partial f_1}{\partial x}\Big|_{(1,1,2)} & \frac{\partial f_1}{\partial y}\Big|_{(1,1,2)} & \frac{\partial f_1}{\partial z}\Big|_{(1,1,2)} \\ \frac{\partial f_2}{\partial x}\Big|_{(1,1,2)} & \frac{\partial f_2}{\partial y}\Big|_{(1,1,2)} & \frac{\partial f_2}{\partial z}\Big|_{(1,1,2)} \end{pmatrix} \\
&= \begin{pmatrix} 2x\big|_{(1,1,2)} & 1\big|_{(1,1,2)} & -1\big|_{(1,1,2)} \\ ye^{xy}\sin z + z\big|_{(1,1,2)} & xe^{xy}\sin z\big|_{(1,1,2)} & e^{xy}\cos z + x\big|_{(1,1,2)} \end{pmatrix} \\
&= \begin{pmatrix} 2 & 1 & -1 \\ e\sin 2 + 2 & e\sin 2 & e\cos 2 + 1 \end{pmatrix}
\end{aligned}
$$

Moreover, if $\mathbf{v} = (-1,4,2)$ *is a vector in* $\mathbb{R}^3$*, then we can compute the directional derivative of* f *at* $(1,1,2)$ *in the direction of* $\mathbf{v}$ *by simple matrix multiplication:*

$$
\begin{aligned}
D_{(-1,4,2)}f(1,1,2) &= Df(1,1,2)\begin{pmatrix} -1 \\ 4 \\ 2 \end{pmatrix} \\
&= \begin{pmatrix} 2 & 1 & -1 \\ e\sin 2 + 2 & e\sin 2 & e\cos 2 + 1 \end{pmatrix}\begin{pmatrix} -1 \\ 4 \\ 2 \end{pmatrix} \\
&= \begin{pmatrix} 0 \\ 3e\sin 2 + 2e\cos 2 \end{pmatrix}
\end{aligned}
$$

■

5 Further Properties of the Total and Partial Derivative

Theorem 5.1 (Chain Rule I) *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^p$ be functions such that $g \circ f$ is defined (i.e. the image of f is contained in the domain of g). If f differentiable at $\mathbf{a} \in \mathbb{R}^n$ and g is differentiable at $\mathbf{b} = f(\mathbf{a})$, then their composite $g \circ f : \mathbb{R}^n \to \mathbb{R}^p$ is differentiable at $\mathbf{a}$ and their derivative is a matrix product, namely the product of their two respective total derivatives,*

$$\boxed{D(g \circ f)(\mathbf{a}) = Dg\big(f(\mathbf{a})\big) \cdot Df(\mathbf{a})} \tag{5.1}$$

The components of the matrix $D(g \circ f)(\mathbf{a})$ in (6.7) may explicitly be given by the formulas:

$$\frac{\partial (g \circ f)_i}{\partial x_j}\bigg|_{\mathbf{a}} = \frac{\partial g_i}{\partial y_1}\bigg|_{\mathbf{b}} \frac{\partial f_1}{\partial x_j}\bigg|_{\mathbf{a}} + \cdots + \frac{\partial g_i}{\partial y_m}\bigg|_{\mathbf{b}} \frac{\partial f_m}{\partial x_j}\bigg|_{\mathbf{a}} \tag{5.2}$$

or, if we let $z_i := g_i(y_1, \ldots, y_m)$ and $y_k := f_k(x_1, \ldots, x_n)$,

$$\boxed{\frac{\partial z_i}{\partial x_j} = \frac{\partial z_i}{\partial y_1}\frac{\partial y_1}{\partial x_j} + \cdots + \frac{\partial z_i}{\partial y_m}\frac{\partial y_m}{\partial x_j}} \tag{5.3}$$

■

Theorem 5.2 (Clairaut: Equality of Mixed Partial Derivatives) *If $f : \mathbb{R}^n \to \mathbb{R}^m$ has twice continuously differentiable partial derivatives or equivalently if for all $1 \leq i, j \leq n$ the partial derivatives $\frac{\partial^2 f}{\partial x_i \partial x_j}$ and $\frac{\partial^2 f}{\partial x_j \partial x_i}$ exist on a neighborhood of a point $\mathbf{a}$ and are continuous at $\mathbf{a}$, then*

$$\boxed{\frac{\partial^2 f}{\partial x_i \partial x_j}\bigg|_{\mathbf{a}} = \frac{\partial^2 f}{\partial x_i \partial x_j}\bigg|_{\mathbf{a}}} \tag{5.4}$$

for all $1 \leq i, j \leq n$. This is also frequently denoted $f_{x_i x_j}(\mathbf{a}) = f_{x_j x_i}(\mathbf{a})$. ■

Remark 5.3 *Failure of continuity at $\mathbf{a}$ may lead to inequality of the mixed partials at $\mathbf{a}$. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by*

$$f(x,y) = \begin{cases} \dfrac{x^3 y - xy^3}{x^2 + y^2}, & \text{if } (x,y) \neq (0,0) \\ 0, & \text{if } (x,y) = (0,0) \end{cases}$$

Then,

$$\begin{aligned} \frac{\partial f}{\partial x} &= \begin{cases} \dfrac{(x^2+y^2)(3x^2y - y^3) - 2x(x^3y - xy^3)}{(x^2+y^2)^4}, & \text{if } (x,y) \neq (0,0) \\ 0, & \text{if } (x,y) = (0,0) \end{cases} \\ &= \begin{cases} \dfrac{3x^4y - x^2y^3 + 3x^2y^3 - y^5 - 2x^4y + 2x^2y^3}{(x^2+y^2)^4}, & \text{if } (x,y) \neq (0,0) \\ 0, & \text{if } (x,y) = (0,0) \end{cases} \\ &= \begin{cases} \dfrac{x^4y + 4x^2y^3 - y^5}{(x^2+y^2)^4}, & \text{if } (x,y) \neq (0,0) \\ 0, & \text{if } (x,y) = (0,0) \end{cases} \end{aligned}$$

and

$$\frac{\partial f}{\partial y} = \begin{cases} \dfrac{(x^2+y^2)(x^3-3xy^2)-2y(x^3y-xy^3)}{(x^2+y^2)^4}, & \text{if } (x,y)\neq(0,0) \\ 0, & \text{if } (x,y)=(0,0) \end{cases}$$

$$= \begin{cases} \dfrac{x^5-3x^3y^2+x^3y^2-3xy^4}{(x^2+y^2)^4}, & \text{if } (x,y)\neq(0,0) \\ 0, & \text{if } (x,y)=(0,0) \end{cases}$$

$$= \begin{cases} \dfrac{x^5-4x^3y^2-xy^4}{(x^2+y^2)^4}, & \text{if } (x,y)\neq(0,0) \\ 0, & \text{if } (x,y)=(0,0) \end{cases}$$

Therefore, for $a, b \neq 0$ *we have*

$$\left.\frac{\partial f}{\partial x}\right|_{(0,b)} = \frac{-5b^5}{b^4} = -b \qquad \left.\frac{\partial f}{\partial y}\right|_{(a,0)} = \frac{a^5}{a^4} = a$$

and consequently

$$\left.\frac{\partial^2 f}{\partial y\partial x}\right|_{(0,0)} = \lim_{t\to 0}\frac{\left.\frac{\partial f}{\partial x}\right|_{(0,t)} - \left.\frac{\partial f}{\partial x}\right|_{(0,0)}}{t} = \lim_{t\to 0}\frac{-t-0}{t} = -1$$

$$\left.\frac{\partial^2 f}{\partial x\partial y}\right|_{(0,0)} = \lim_{t\to 0}\frac{\left.\frac{\partial f}{\partial y}\right|_{(t,0)} - \left.\frac{\partial f}{\partial y}\right|_{(0,0)}}{t} = \lim_{t\to 0}\frac{t-0}{t} = 1$$

and so $\left.\frac{\partial^2 f}{\partial y\partial x}\right|_{(0,0)} \neq \left.\frac{\partial^2 f}{\partial x\partial y}\right|_{(0,0)}$. *The problem, of course, is the discontinuity of the second derivatives at* $(0,0)$*:*

$$\frac{\partial^2 f}{\partial x\partial y} = \begin{cases} \frac{(x^2+y^2)^2(5x^4-12x^2y^2-y^4)-2(x^2y^2)2x(x^5-4x^3y^2-xy^4)}{(x^2+y^2)^4}, & \text{if } (x,y)\neq(0,0) \\ 0, & \text{if } (x,y)=(0,0) \end{cases}$$

$$\frac{\partial^2 f}{\partial y\partial x} = \begin{cases} \frac{(x^2+y^2)^2(x^4+12x^2y^2-5y^4)-2(x^2y^2)2y(x^4y+4x^2y^3-y^5)}{(x^2+y^2)^4}, & \text{if } (x,y)\neq(0,0) \\ 0, & \text{if } (x,y)=(0,0) \end{cases}$$

For example, along the line $x = y$ *we have* $\frac{\partial^2 f}{\partial x\partial y} = 2(1-x)$*, so it approaches a value of* 2*, while along the line* $x = 0$ *it stays constant at* 1*, as noted above.* ■

6 Appendix: Proofs of the Theorems

Theorem 6.1 *If $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at* $\mathbf{a}$ *in* $\mathbb{R}^n$*, then all of its directional derivatives at* $\mathbf{a}$ *exist, and for any choice of vector* $\mathbf{v}$ *in* $\mathbb{R}^n$ *we have*

$$D_{\mathbf{v}}f(\mathbf{x}) = Df(\mathbf{x})\mathbf{v} \tag{6.1}$$

The left-hand side is a limit, while the right-hand side is a matrix product, with $\mathbf{v}$ *treated as a column vector.*

Proof: Since f is differentiable at $\mathbf{a}$, fix $\mathbf{v}$ and consider $\mathbf{h} = t\mathbf{v}$ for some sufficiently small $t \in \mathbb{R}$. Applying the linear approximation (1.3) and the linearity of the derivative $Df(\mathbf{a})$ (i.e. $Df(\mathbf{a})(a\mathbf{x} + b\mathbf{y}) = aDf(\mathbf{a})\mathbf{x} + bDf(\mathbf{a})\mathbf{y}$) we get

$$\begin{aligned} f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x}) - tDf(\mathbf{x})\mathbf{v} &= f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x}) - Df(\mathbf{x})(t\mathbf{v}) \\ &= f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h} \\ &= E(\mathbf{h}) \\ &= E(t\mathbf{v}) \end{aligned} \tag{6.2}$$

and applying the limit (1.4)

$$\lim_{t\to 0} \frac{|E(t\mathbf{v})|}{|t|} = \lim_{t\to 0} \frac{|E(t\mathbf{v})|}{|t||\mathbf{v}|} \cdot |\mathbf{v}| = \lim_{t\to 0} \frac{|E(t\mathbf{v})|}{|t\mathbf{v}|} \cdot |\mathbf{v}| = \lim_{\mathbf{h}\to\mathbf{0}} \frac{|E(\mathbf{h})|}{|\mathbf{h}|} \cdot |\mathbf{v}| = 0 \cdot |\mathbf{v}| = 0$$

By (6.2) this means

$$\lim_{t\to 0} \frac{|f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x}) - tDf(\mathbf{x})(\mathbf{v})|}{t} = 0$$

and hence

$$\begin{aligned} \lim_{t\to 0} \frac{f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x})}{t} - Df(\mathbf{x})(\mathbf{v}) &= \lim_{t\to 0} \frac{f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x})}{t} - \lim_{t\to 0} \frac{tDf(\mathbf{x})(\mathbf{v})}{t} \\ &= \lim_{t\to 0} \frac{f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x}) - tDf(\mathbf{x})(\mathbf{v})}{t} \\ &= \mathbf{0} \end{aligned}$$

i.e.

$$D_{\mathbf{v}}f(\mathbf{x}) = \lim_{t\to 0} \frac{f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x})}{t} = Df(\mathbf{x})(\mathbf{v})$$

■

Theorem 6.2 *A function $f : \mathbb{R}^n \to \mathbb{R}^m$, $f = (f_1, \ldots, f_n)$, is differentiable at* $\mathbf{a}$ *if and only if each of its component functions $f_i : \mathbb{R}^n \to \mathbb{R}$ is differentiable at* $\mathbf{a}$*. In that case, we have*

$$Df(\mathbf{a}) = D\begin{pmatrix} f_1(\mathbf{a}) \\ \vdots \\ f_m(\mathbf{a}) \end{pmatrix} = \begin{pmatrix} Df_1(\mathbf{a}) \\ \vdots \\ Df_m(\mathbf{a}) \end{pmatrix} \tag{6.3}$$

That is, to compute $Df(\mathbf{a})$ we can just compute the $1 \times n$ derivative matrices of the f_i first, which we know are of the form $Df_i(\mathbf{a}) = \left(\frac{\partial f_i}{\partial x_1}\big|_{\mathbf{a}} \quad \cdots \quad \frac{\partial f_i}{\partial x_n}\big|_{\mathbf{a}} \right)$, and enter the result into the ith row of the larger matrix.

Proof: This follows from the inequalities

$$|a_i| \le |\mathbf{a}| \le \sqrt{n} \max_{1\le i\le n} |a_i|$$

for all i, since if f is differentiable at $\mathbf{a}$, then the limit (1.2) exists, so the first inequality above implies that the limit of zero exists in each of the coordinates, and so for each of the coordinate functions. Indeed, by that limit we must have that $Df_i(\mathbf{a})$ is the ith component function of $Df(\mathbf{a})$. Conversely, if the component functions are differentiable at $\mathbf{a}$, then multiplying the limit (1.2) for f_i by $\sqrt{n}$ and using the second inequality above we have that the limit (1.2) for f holds as well (just choose the f_i with maximum absolute value), and moreover we must have that $Df_i(\mathbf{x})$ are the coordinate linear functionals of $Df(\mathbf{a})$ by the first inequality. ■

Theorem 6.3 *Let $f : \mathbb{R}^n \to \mathbb{R}^m$. If all the partial derivatives $\frac{\partial f_i}{\partial x_j}\big|_{\mathbf{a}}$ of f exist and are continuous at $\mathbf{a}$, then f is differentiable at $\mathbf{a}$.*

Proof: By Proposition 4.3 it is enough to prove this for the component functions f_i of f. Indeed, let f_i be a component function of f, and suppose it's partial derivatives all exist and are continuous in a neighborhood of $\mathbf{a}$. Then, since $\frac{\partial f_i}{\partial x_j}\big|_{\mathbf{a}}$ moves only in the jth coordinate direction, we need only $\mathbf{h}_j = (0, \ldots, h_j, \ldots, 0)$ in those directions. By the definition of continuity of $\frac{\partial f_i}{\partial x_j}\big|_{\mathbf{a}}$, for any $\varepsilon > 0$ we choose there is a $\delta > 0$ such that if $|\mathbf{h}| = |h_j| < \delta$ then

$$\frac{\left|\frac{\partial f_i}{\partial x_j}\Big|_{\mathbf{a}+\mathbf{h}} - \frac{\partial f_i}{\partial x_j}\Big|_{\mathbf{a}}\right|}{|h_j|} < \frac{\varepsilon}{n}$$

Let $\mathbf{h}$ be a point in $\mathbb{R}^n$, so that $\mathbf{h} = \mathbf{h}_1 + \cdots + \mathbf{h}_n$ using our notation above. By the Mean Value Theorem from Calc 1, the continuity of f and the existence of the jth partial implies the existence of a point $\mathbf{a} + \mathbf{h}_j + t_j\mathbf{e}_j$ between $\mathbf{a} + \mathbf{h}_j$ and $\mathbf{a} + \mathbf{h}_j + \mathbf{e}_j$ such that

$$f(\mathbf{a} + \mathbf{h}_j) - f(\mathbf{a}) = \left(\frac{\partial f_i}{\partial x_j}\Big|_{\mathbf{x}+\mathbf{h}_i+t_i\mathbf{e}_i}\right) h_j \tag{6.4}$$

(Note: in the jth coordinate, keeping all other coordinates fixed, f_j is a real-valued function of a single variable, so this works. Recall the MVT: If f is continuous on $[a, b]$ and differentiable on (a, b) then there is a point c between a and b such that $f(b) - f(a) = f'(c)(b - a)$!) As a consequence, we have

$$\begin{aligned}
\left|f_i(\mathbf{a}+\mathbf{h}) - f_i(\mathbf{a}) - \left(\frac{\partial f_i}{\partial x_1}\Big|_{\mathbf{a}} \quad \cdots \quad \frac{\partial f_i}{\partial x_j}\Big|_{\mathbf{a}}\right)\mathbf{h}\right| &= \left|f_i(\mathbf{a}+\mathbf{h}) - f_i(\mathbf{a}) - \sum_{j=1}^{n}\left(\frac{\partial f_i}{\partial x_j}\Big|_{\mathbf{a}}\right)h_j\right| \\
&= \left|\sum_{j=1}^{n}\left(\frac{\partial f_i}{\partial x_j}\Big|_{\mathbf{a}+\mathbf{h}+t_j\mathbf{e}_i}\right)h_j - \sum_{j=1}^{n}\left(\frac{\partial f_i}{\partial x_j}\Big|_{\mathbf{a}}\right)h_i\right| \\
&\le \sum_{j=1}^{n}\left|\frac{\partial f_i}{\partial x_j}\Big|_{\mathbf{a}+\mathbf{h}+t_j\mathbf{e}_i} - \frac{\partial f_i}{\partial x_j}\Big|_{\mathbf{a}}\right||h_j| \\
&< \sum_{j=1}^{n}\frac{|h_j|\varepsilon}{n} \\
&\le |\mathbf{h}|\varepsilon
\end{aligned}$$

where the first inequality is from factoring out $|h_j|$ and then using the triangle inequality, the second is by application of (6.4) for each j, and the third by observing that $|h_1| + \cdots + |h_n| \le$

$\sqrt{h_1^2 + \cdots + h_n^2} + \cdots + \sqrt{h_1^2 + \cdots + h_n^2} = n|\mathbf{h}|$. Dividing the above inequality through by $|\mathbf{h}|$ gives our desired inequality,

$$\frac{\left| f_i(\mathbf{a}+\mathbf{h}) - f_i(\mathbf{a}) - \left(\left.\frac{\partial f_i}{\partial x_1}\right|_{\mathbf{a}} \quad \cdots \quad \left.\frac{\partial f_i}{\partial x_j}\right|_{\mathbf{a}} \right) \mathbf{h} \right|}{|\mathbf{h}|} < \varepsilon$$

We have thus demonstrated the limit

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\left| f_i(\mathbf{a}+\mathbf{h}) - f_i(\mathbf{a}) - \left(\left.\frac{\partial f_i}{\partial x_1}\right|_{\mathbf{a}} \quad \cdots \quad \left.\frac{\partial f_i}{\partial x_j}\right|_{\mathbf{a}} \right) \mathbf{h} \right|}{|\mathbf{h}|} = 0$$

which is the definition of differentiability, and moreover, in the course of the proof, we have also shown that $Df(\mathbf{a}) = \left(\left.\frac{\partial f_i}{\partial x_1}\right|_{\mathbf{a}} \quad \cdots \quad \left.\frac{\partial f_i}{\partial x_j}\right|_{\mathbf{a}} \right)$ as well! ■

Proposition 6.4 (Hadamard) *Let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}^m$. Then, for any $\mathbf{x}_0 \in U$ the following are equivalent:*

(1) *f is differentiable at $\mathbf{x}_0$.*

(2) *There exists a map $\varphi_{\mathbf{x}_0} : U \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, continuous at $\mathbf{x}_0$, such that for all $\mathbf{x} \in U$ we have*

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \varphi_{\mathbf{x}_0}(\mathbf{x})(\mathbf{x} - \mathbf{x}_0) \tag{6.5}$$

If any of these conditions holds, moreover, then

$$Df(\mathbf{x}_0) = \varphi_{\mathbf{x}_0} \tag{6.6}$$

Proof: (1) ⇒ (2): Suppose f is differentiable at $\mathbf{x}_0$, then there is an ϵ-function $\epsilon : \mathbb{R}^n \to \mathbb{R}^m$ satisfying $f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + Df(\mathbf{x})\mathbf{h} + \epsilon(\mathbf{h})$, where $\lim_{\mathbf{h}\to\mathbf{0}} \frac{|\epsilon(\mathbf{h})|_2}{|\mathbf{h}|_2} = 0$, for all $\mathbf{h}$ with $\mathbf{x}_0 + \mathbf{h} \in U$. Define $\varphi_{\mathbf{x}_0}$ by

$$\varphi_{\mathbf{x}_0}(\mathbf{x}) = \begin{cases} Df(\mathbf{x}_0) + \dfrac{1}{|\mathbf{x}-\mathbf{x}_0|_2^2}\epsilon(\mathbf{x}-\mathbf{x}_0)\cdot(\mathbf{x}-\mathbf{x}_0)^T, & \text{if } \mathbf{x} \in U \backslash \{\mathbf{x}_0\} \\ Df(\mathbf{x}_0), & \text{if } \mathbf{x} = \mathbf{x}_0 \end{cases}$$

where the product $\epsilon(\mathbf{x}-\mathbf{x}_0)\cdot(\mathbf{x}-\mathbf{x}_0)^T$ is a matrix product, producing an $m \times n$ matrix, associated with a linear map in $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. This map is therefore linear, being the sum of two linear functions. Applying $\mathbf{y} = \mathbf{x} - \mathbf{x}_0$ we get that $f(\mathbf{x}) = f(\mathbf{x}_0) + \varphi_{\mathbf{x}_0}(\mathbf{x})(\mathbf{x} - \mathbf{x}_0)$. To see that $\varphi_{\mathbf{x}_0}$ is continuous at $\mathbf{x}_0$, note a fact of linear algebra: the operator norm satisfies

$$|\epsilon(\mathbf{h})\mathbf{h}^T| = \sqrt{\textstyle\sum_{ij}(\epsilon(\mathbf{h})\mathbf{h}^T)_{ij}} = \sqrt{\textstyle\sum_{ij}\epsilon(\mathbf{h})_i h_j} = |\epsilon(\mathbf{h})|_2|\mathbf{h}|_2$$

so $\lim_{\mathbf{h}\to\mathbf{0}} \frac{|\epsilon(\mathbf{h})\mathbf{h}^T|}{|\mathbf{h}|_2^2} = \lim_{\mathbf{h}\to\mathbf{0}} \frac{|\epsilon(\mathbf{h})|_2}{|\mathbf{h}|_2} = 0$, so that $\lim_{\mathbf{h}\to\mathbf{0}} \varphi_{\mathbf{x}_0}(\mathbf{x}_0 + \mathbf{h}) = Df(\mathbf{x}_0)$.

(2) ⇒ (1): Conversely, suppose there is a $\varphi_{\mathbf{x}_0} : U \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, continuous at $\mathbf{x}_0$, such that for all $\mathbf{x} \in U$ equation (6.6) holds. Then, by continuity we have that for all $\epsilon > 0$ there is a $\delta > 0$ such that $|\mathbf{h}|_2 = |\mathbf{x}_0 + \mathbf{h} - \mathbf{x}_0|_2 < \delta$ implies $|\varphi_{\mathbf{x}_0}(\mathbf{x}_0 + \mathbf{h}) - \varphi_{\mathbf{x}_0}(\mathbf{x}_0)| < \epsilon$. Since $\mathbb{R}^n$ and $\mathbb{R}^m$ are finite-dimensional, it is an easy matter to show that any $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ is continuous,

and therefore bounded. Consequently we may use again that fact from linear algebra cited above, and along with (6.5) we have

$$\begin{aligned}\frac{|f(\mathbf{x}_0+\mathbf{h})-f(\mathbf{x}_0)-\varphi_{\mathbf{x}_0}(\mathbf{x}_0)(\mathbf{h})|_2}{|\mathbf{h}|_2} &\overset{(6.5)}{=} \frac{|\varphi_{\mathbf{x}_0}(\mathbf{x}_0+\mathbf{h})(\mathbf{h})-\varphi_{\mathbf{x}_0}(\mathbf{x}_0)(\mathbf{h})|_2}{|\mathbf{h}|_2} \\ &\leq \frac{|\varphi_{\mathbf{x}_0}(\mathbf{x}_0+\mathbf{h})-\varphi_{\mathbf{x}_0}(\mathbf{x}_0)||(\mathbf{h})|_2}{|\mathbf{h}|_2} \\ &< \epsilon\end{aligned}$$

i.e. $\lim_{\mathbf{h}\to\mathbf{0}} \frac{|f(\mathbf{x}_0+\mathbf{h})-f(\mathbf{x}_0)-\varphi_{\mathbf{x}_0}(\mathbf{x}_0)(\mathbf{h})|_2}{|\mathbf{h}|_2} = 0$, and f is differentiable at $\mathbf{x}_0$. ■

Theorem 6.5 (Chain Rule I) *Let $f:\mathbb{R}^n\to\mathbb{R}^m$ and $g:\mathbb{R}^m\to\mathbb{R}^p$ be functions such that $g\circ f$ is defined (i.e. the image of f is contained in the domain of g). If f differentiable at $\mathbf{a}\in\mathbb{R}^n$ and g is differentiable at $\mathbf{b}=f(\mathbf{a})$, then their composite $g\circ f:\mathbb{R}^n\to\mathbb{R}^p$ is differentiable at $\mathbf{a}$ and their derivative is a matrix product, namely the product of their two respective total derivatives,*

$$D(g\circ f)(\mathbf{a}) = Dg\big(f(\mathbf{a})\big)\cdot Df(\mathbf{a}) \tag{6.7}$$

The components of the matrix $D(g\circ f)(\mathbf{a})$ in (6.7) may explicitly be given by the formulas:

$$\frac{\partial(g\circ f)_i}{\partial x_j}\bigg|_{\mathbf{a}} = \frac{\partial g_i}{\partial y_1}\bigg|_{\mathbf{b}}\frac{\partial f_1}{\partial x_j}\bigg|_{\mathbf{a}} + \cdots + \frac{\partial g_i}{\partial y_m}\bigg|_{\mathbf{b}}\frac{\partial f_m}{\partial x_j}\bigg|_{\mathbf{a}} \tag{6.8}$$

or, if we let $z_i := g_i(y_1,\ldots,y_m)$ and $y_k := f_k(x_1,\ldots,x_n)$,

$$\frac{\partial z_i}{\partial x_j} = \frac{\partial z_i}{\partial y_1}\frac{\partial y_1}{\partial x_j} + \cdots + \frac{\partial z_i}{\partial y_m}\frac{\partial y_m}{\partial x_j} \tag{6.9}$$

Proof: If f is differentiable at $\mathbf{x}_0$, then by Hadamard's lemma there is a an operator valued function $\varphi: U\to\mathcal{L}(\mathbb{R}^n,\mathbb{R}^m)$, continuous at $\mathbf{x}_0$, such that

$$f(\mathbf{x})-f(\mathbf{x}_0) = \varphi_{\mathbf{x}_0}(\mathbf{x}-\mathbf{x}_0), \quad \text{with} \lim_{\mathbf{x}\to\mathbf{x}_0}\varphi_{\mathbf{x}_0}(\mathbf{x}) = Df(\mathbf{x}_0) \tag{6.10}$$

and similarly since g is differentiable at $f(\mathbf{x}_0)$ there is a $\psi: V\to\mathcal{L}(\mathbb{R}^m,\mathbb{R}^p)$ such that

$$g(\mathbf{y})-g(f(\mathbf{x}_0)) = \psi_{\mathbf{y}_0}(\mathbf{y})(\mathbf{y}-f(\mathbf{x}_0)), \quad \text{with} \lim_{\mathbf{y}\to f(\mathbf{x}_0)}\psi_{\mathbf{y}_0}(\mathbf{y}) = Dg(f(\mathbf{x}_0)) \tag{6.11}$$

Letting $\mathbf{y}=f(\mathbf{x})$ and substituting into (6.11) we get, by (6.10),

$$\begin{aligned}(g\circ f)(\mathbf{x})-(g\circ f)(\mathbf{x}_0) &= \psi_{\mathbf{y}_0}(f(\mathbf{x}))(f(\mathbf{x})-f(\mathbf{x}_0)) \\ &= \psi_{\mathbf{y}_0}(f(\mathbf{x}))\circ\varphi_{\mathbf{x}_0}(x)(\mathbf{x}-\mathbf{x}_0)\end{aligned} \tag{6.12}$$

By the second parts of (6.10) and (6.11) we have $\lim_{\mathbf{x}\to\mathbf{x}_0}\psi_{\mathbf{y}_0}(f(\mathbf{x}))\circ\varphi_{\mathbf{x}_0}(\mathbf{x}) = Dg(f(\mathbf{x}_0))\circ Df(\mathbf{x}_0)$. The linearity of $Dg(f(\mathbf{x}_0))\circ Df(\mathbf{x}_0)$ follows from that of $\psi_{\mathbf{y}_0}(f(\mathbf{x}_0))$ in (6.12), so when we take the limit as $\mathbf{x}\to\mathbf{x}_0$ of (6.12) we get by Hadamard's lemma again that $D(g\circ f)(\mathbf{x}_0) = Dg(f(\mathbf{x}_0))\circ Df(\mathbf{x}_0)$. ■

Theorem 6.6 (Clairaut: Equality of Mixed Partial Derivatives) *If $f : \mathbb{R}^n \to \mathbb{R}^m$ has twice continuously differentiable partial derivatives or equivalently if for all $1 \le i, j \le n$ the partial derivatives $\frac{\partial^2 f}{\partial x_i \partial x_j}$ and $\frac{\partial^2 f}{\partial x_j \partial x_i}$ exist on a neighborhood of a point $\mathbf{a}$ and are continuous at $\mathbf{a}$, then*

$$\left.\frac{\partial^2 f}{\partial x_i \partial x_j}\right|_{\mathbf{a}} = \left.\frac{\partial^2 f}{\partial x_i \partial x_j}\right|_{\mathbf{a}} \tag{6.13}$$

for all $1 \le i, j \le n$. ■

Proof: It will simplify notation a little if we write D_j instead of $\partial/\partial x_j$. In view of Proposition 4.3 it suffices to prove this for all component functions f_k of f. Without loss of generality, we may suppose that $i < j$. Let $r : \mathbb{R}^n \to \mathbb{R}$ be given by

$$r(\mathbf{y}) = \frac{f_k(\mathbf{y}) - f_k(y_1, \ldots, y_{i-1}, x_i, y_{i+1}, \ldots, y_n) - f_k(y_1, \ldots, y_{j-1}, x_j, y_{j+1}, \ldots, y_n) + f_k(\mathbf{x})}{(y_i - x_i)(y_j - x_j)}$$

and define $g : \mathbb{R} \to \mathbb{R}$ by

$$g(t) = f_k(y_1, \ldots, y_{i-1}, t, y_{i+1}, \ldots, y_n) - f_k(y_1, \ldots, y_{i-1}, t, y_{i+1}, \ldots, y_{j-1}, x_j, y_{j+1}, \ldots, y_n)$$

so that

$$r(\mathbf{y}) = \frac{g(y_i) - g(x_i)}{(y_i - x_i)(y_j - x_j)}$$

We will show that both sides of (6.13) are equal to $\lim_{\mathbf{y}\to\mathbf{x}} r(\mathbf{y})$. The denominator of r is the area of the rectangle with vertices (y_i, y_j), (x_i, y_i), (y_j, x_j) and (x_i, x_j) in the i-j plane, while the numerator is the alternating sum of the values of f at these vertices. Note that since the partial derivatives of f (up to order 2), and so those of each component function f_k of f, exist on a neighborhood $N \subseteq U$ of $\mathbf{x}$, we have that g is differentiable on N. By the Mean Value Theorem for $\mathbb{R}$, there is a ξ_i between x_i and y_i such that

$$\begin{aligned} r(\mathbf{y}) &= \frac{g(y_i) - g(x_i)}{(y_i - x_i)(y_j - x_j)} = \frac{g'(\xi_i)(y_i - x_i)}{(y_i - x_i)(y_j - x_j)} = \frac{g'(\xi_i)}{y_j - x_j} \\ &= \frac{D_i f_k(y_1, \ldots, y_{i-1}, \xi_i, y_{i+1}, \ldots, y_n) - D_i f_k(y_1, \ldots, y_{i-1}, \xi_i, y_{i+1}, \ldots, y_{j-1}, x_j, y_{j+1}, \ldots, y_n)}{y_j - x_j} \end{aligned} \tag{6.14}$$

Define $h : \mathbb{R} \to \mathbb{R}$ is defined on a sufficiently small neighborhood of x_j by

$$h(t) = D_j f_k(y_1, \ldots, y_{i-1}, \xi_i, y_{i+1}, \ldots, y_{j-1}, t, y_{j+1}, \ldots, y_n)$$

then, again, h is differentiable and the Mean Value Theorem gives the existence of ξ_j between x_j and y_j. Consequently, from (6.14) we get

$$\begin{aligned} r(\mathbf{y}) = \frac{h(y_j) - h(x_j)}{y_j - x_j} = \frac{h'(\xi_j)(y_j - x_j)}{y_j - x_j} &= h'(\xi_j) \\ &= D_j D_i f(y_1, \ldots, y_{i-1}, \xi_i, y_{i+1}, \ldots, y_{j-1}, \xi_j, y_{j+1}, \ldots, y_n) \end{aligned}$$

Let $\xi = \xi(\mathbf{y}) = (y_1, \ldots, y_{i-1}, \xi_i, y_{i+1}, \ldots, y_{j-1}, \xi_j, y_{j+1}, \ldots, y_n)$, and note that $|\xi - \mathbf{x}|_2 \le |\mathbf{y} - \mathbf{x}|_2$, so the continuity of $D_j D_i f$ at $\mathbf{x}$ implies

$$\lim_{\mathbf{y}\to\mathbf{x}} r(\mathbf{y}) = \lim_{\xi\to\mathbf{x}} D_j D_i f(\xi) = D_j D_i f(\mathbf{x})$$

Reversing the roles of x_i and x_j above shows that, with ξ' probably different from ξ above, that

$$\lim_{\mathbf{y}\to\mathbf{x}} r(\mathbf{y}) = D_i D_j f(\mathbf{x})$$ ■