1. Differentiability

Recall the definition of derivative from one variable calculus

Definition 1.1. *We say that* $f : \mathbf{R} \to \mathbf{R}$ *is* differentiable *at a point* $a \in \mathbf{R}$ *if the quantity*

$$f'(a) := \lim_{h\to 0} \frac{f(a+h) - f(a)}{h}$$

exists. We then call $f'(a)$ *the* derivative *of* f *at* a.

One way to transfer this definition to higher dimensions is via 'directional' derivatives.

Definition 1.2. *The* directional derivative *of a function* $F : \mathbf{R}^n \to \mathbf{R}^m$ *at a point* $\mathbf{a} \in \mathbf{R}^n$ *in the direction* $\mathbf{v} \in \mathbf{R}^n$ *is the quantity (if it exists)*

$$D_{\mathbf{v}} F(\mathbf{a}) := \lim_{t\to 0} \frac{F(\mathbf{a} + t\mathbf{v}) - F(\mathbf{a})}{t}$$

When $\mathbf{v} = \mathbf{e}_j$ *is a standard basis vector, we write* $\frac{\partial F}{\partial x_j}(\mathbf{a}) := D_{\mathbf{e}_j} F(a)$ *and call this quantity the* partial derivative *of* F *with respect to* x_j.

Another way of stating this definition is that $D_{\mathbf{v}} F(\mathbf{a}) = h'(0)$ where $h : \mathbf{R} \to \mathbf{R}^m$ is the composite function

$$h(t) := F(\mathbf{a} + t\mathbf{v})$$

obtained by restricting F to the line through $\mathbf{a}$ in the direction $\mathbf{v}$. This way of formulating directional derivatives is quite useful when you actually have to compute one!

A shortcoming of directional derivatives is that they don't always do a very good job of controlling the behavior of F near a given point $\mathbf{a}$ (see Section 3.1 e.g. 2 in Shifrin for a good illustration of this). One needs a little bit more restrictive notion of derivative in order to guarantee this sort of control.

Definition 1.3. *We say that a function* $F : \mathbf{R}^n \to \mathbf{R}^m$ *is* differentiable *at a point* $\mathbf{a} \in \mathbf{R}^n$ *if there exists a linear transformation* $T : \mathbf{R}^n \to \mathbf{R}^m$ *such that*

$$\lim_{\mathbf{h}\to \mathbf{0}} \frac{F(\mathbf{a} + \mathbf{h}) - F(\mathbf{a}) - T\,\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0}. \tag{1}$$

If such a T *exists, then we call it the* derivative *of* F *at* $\mathbf{a}$ *write* $DF(\mathbf{a}) := T$.

So under this definition, the derivative $DF(\mathbf{a})$ of F at $\mathbf{a}$ is not a number but rather a linear transformation. This is not so strange if you remember any linear transformation $T : \mathbf{R}^n \to \mathbf{R}^m$ has a standard matrix $A \in \mathcal{M}_{m\times n}$, so you can think of the derivative of F at $\mathbf{a}$ more concretely as a matrix, i.e. as a collection of mn numbers that describe the way all the different components of $F = (F_1, \ldots, F_m)$ are changing in all the different directions one can approach $\mathbf{a}$. I'm sort of doing that already when I suppress parentheses in $T(\mathbf{h})$ and write $T\mathbf{h}$ instead.

In particular, if $f : \mathbf{R} \to \mathbf{R}$ is just a scalar function of a single variable, then the number $f'(a)$ above is just the lone entry in the 1×1 matrix for the linear transformation $T : \mathbf{R} \to \mathbf{R}$ given by $T(h) = f'(a)h$.

Note that Equation (1) can be written in several slightly different but equivalent ways. For instance, one could take the magnitude of the numerator and write instead (I'll use $DF(\mathbf{a})$ in place of T now).

$$\lim_{\mathbf{h}\to \mathbf{0}} \frac{\|F(\mathbf{a} + \mathbf{h}) - F(\mathbf{a}) - DF(\mathbf{a})\mathbf{h}\|}{\|\mathbf{h}\|} = \mathbf{0}.$$

Or one could set $\mathbf{x} := \mathbf{a} + \mathbf{h}$ and rewrite the limit as

$$\lim_{\mathbf{x}\to\mathbf{a}} \frac{F(\mathbf{x}) - F(\mathbf{a}) - DF(\mathbf{a})(\mathbf{x}-\mathbf{a})}{\|\mathbf{x}-\mathbf{a}\|}.$$

Another very useful way to restate (1) is to say that

$$F(\mathbf{a}+\mathbf{h}) = F(\mathbf{a}) + DF(\mathbf{a})\mathbf{h} + E(\mathbf{h}), \tag{2}$$

where the 'error term' $E(\mathbf{h})$ satisfies $\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|E(\mathbf{h})\|}{\|\mathbf{h}\|} = 0$.

The first indication that our definition of differentiability will give us sufficient control of F at nearby values of $\mathbf{a}$ is the following.

Theorem 1.4. *If $F : \mathbf{R}^n \to \mathbf{R}^m$ is differentiable at $\mathbf{a}$, then F is continuous at $\mathbf{a}$.*

Proof. From Equation (2) and continuity of linear transformations, I find that

$$\lim_{\mathbf{x}\to\mathbf{a}} F(\mathbf{x}) = \lim_{\mathbf{x}\to\mathbf{a}} F(\mathbf{a}) + DF(\mathbf{a})(\mathbf{x}-\mathbf{a}) + E(\mathbf{x}-\mathbf{a}) = F(\mathbf{a}) + DF(\mathbf{a})\mathbf{0} + \lim_{\mathbf{x}\to\mathbf{a}} E(\mathbf{x}-\mathbf{a}).$$

Moreover, since F is differentiable at $\mathbf{a}$, I can dismiss the last limit as follows.

$$\lim_{\mathbf{x}\to\mathbf{a}} E(\mathbf{x}-\mathbf{a}) = \lim_{\mathbf{x}\to\mathbf{a}} \|\mathbf{x}-\mathbf{a}\| \frac{E(\mathbf{x}-\mathbf{a})}{\|\mathbf{x}-\mathbf{a}\|} = \lim_{\mathbf{x}\to\mathbf{a}} \|\mathbf{x}-\mathbf{a}\| \lim_{\mathbf{x}\to\mathbf{a}} \frac{E(\mathbf{x}-\mathbf{a})}{\|\mathbf{x}-\mathbf{a}\|} = 0 \cdot 0.$$

Note that in the last equality, I use continuity of the magnitude function $\mathbf{x} \to \|\mathbf{x}\|$. At any rate, I conclude that

$$\lim_{\mathbf{x}\to\mathbf{a}} F(\mathbf{x}) = F(\mathbf{a}),$$

i.e. F is continuous at $\mathbf{a}$. □

The next fact about our new notion of derivative $Df(a)$ is that it's not that far from partial and directional derivatives.

Theorem 1.5. *Suppose that $F : \mathbf{R}^n \to \mathbf{R}^m$ is differentiable at a point $\mathbf{a} \in \mathbf{R}^n$. Then the directional derivative of F at $\mathbf{a}$ in direction $\mathbf{v} \in \mathbf{R}^n$ exists and is given by*

$$D_{\mathbf{v}}F(\mathbf{a}) = DF(\mathbf{a})\mathbf{v}. \tag{3}$$

In particular, the standard matrix for the linear transformation $DF(\mathbf{a}) : \mathbf{R}^n \to \mathbf{R}^m$ is given column-wise by

$$\begin{bmatrix} \frac{\partial F}{\partial x_1}(\mathbf{a}) & \ldots & \frac{\partial F}{\partial x_n}(\mathbf{a}) \end{bmatrix} \tag{4}$$

Among other things, this theorem tells us that there is only one candidate for $Df(a)$ and gives us a practical means for finding out what it is (by taking partial derivatives). It does not, however, tell us how to determine whether our candidate is a winner, i.e. whether F is actually differentiable at $\mathbf{a}$. For most purposes, the following condition suffices for that purpose.

Proof. The main thing here is to justify the formula (3) for the directional derivative. This formula implies in particular that

$$\frac{\partial F}{\partial x_j}(\mathbf{a}) = D_{\mathbf{e}_j}F(\mathbf{a}) = DF(\mathbf{a})\mathbf{e}_j.$$

So the expression (4) for the standard matrix of $DF(\mathbf{a})$ proceeds immediately from this and the fact that the jth column of the standard matrix of a linear transformation is obtained by applying the transformation to the standard basis vector $\mathbf{e}_j$.

To prove (3), I must show that

$$\lim_{t\to 0} \frac{F(\mathbf{a}+t\mathbf{v})-F(\mathbf{a})}{t} = DF(\mathbf{a})\mathbf{v}$$

If $\mathbf{v}=\mathbf{0}$, the two sides are clearly equal. Otherwise, I can use equation (2) to rewrite the difference quotient on the left side as follows

$$\frac{F(\mathbf{a}+t\mathbf{v})-F(\mathbf{a})}{t} = \frac{t\,DF(\mathbf{a})\mathbf{v}+E(t\mathbf{v})}{t} = DF(\mathbf{a})\mathbf{v} + \frac{E(t\mathbf{v})}{t}.$$

So from here it suffices to show that $\lim_{t\to 0}\frac{E(t\mathbf{v})}{t} = 0$. To this end, let $\epsilon > 0$ be given. Differentiability of F at $\mathbf{a}$ guarantees that there exists $\tilde{\delta} > 0$ such that $\|\mathbf{h}\| < \tilde{\delta}$ implies that $\frac{\|E(\mathbf{h})\|}{\|\mathbf{h}\|} < \frac{\epsilon}{\|\mathbf{v}\|}$. I therefore choose $\delta = \frac{\tilde{\delta}}{\|\mathbf{v}\|}$. If $|t| < \delta$, then $\|t\mathbf{v}\| < \tilde{\delta}$. Hence

$$\left\|\frac{E(t\mathbf{v})}{t}\right\| = \|\mathbf{v}\| \left\|\frac{E(t\mathbf{v})}{t\mathbf{v}}\right\| < \|\mathbf{v}\| \frac{\epsilon}{\|\mathbf{v}\|} = \epsilon.$$

Hence $\lim_{t\to 0}\frac{E(t\mathbf{v})}{t} = 0$. I conclude that $D_{\mathbf{v}}F(\mathbf{a}) = DF(\mathbf{a})\mathbf{v}$. □

Definition 1.6. *A function $F : \mathbf{R}^n \to \mathbf{R}^m$ is said to be* continuously differentiable *at $\mathbf{a} \in \mathbf{R}^n$, if all partial dervatives $\frac{\partial F}{\partial x_j}$ exist near and at $\mathbf{a}$, and each is continuous at $\mathbf{a}$. If F is continuously differentiable at each point in its domain, then we say simply that 'F is continuously differentiable' (or 'C^1' for short).*

Theorem 1.7. *If $F : \mathbf{R}^n \to \mathbf{R}^m$ is C^1, then F is differentiable at every point $\mathbf{a}$ in its domain.*

The following preliminary result reduces the proof of the theorem to the special case where $F = f : \mathbf{R}^n \to \mathbf{R}$ is scalar-valued.

Lemma 1.8. *Let $F : \mathbf{R}^n \to \mathbf{R}^m$ be a vector-valued function with component functions $F_1, \dots, F_m : \mathbf{R}^n \to \mathbf{R}$. Then F is differentiable at $\mathbf{a} \in \mathbf{R}^n$ if and only if each component function F_j is differentiable at $\mathbf{a}$. In this case, the standard matrix for $DF(\mathbf{a})$ has jth row equal to the standard matrix for $DF_j(\mathbf{a})$ (note that this is a $1\times n$ matrix—i.e. a row vector).*

Proof. Exercise: follows from the definition of differentiable and the fact (Proposition 6.7 in my glossary) finding the limit of a vector-valued function reduces to finding the limits of each of its component functions. □

To restate the lemma a bit less formally, F is differentiable at exactly those points where all its components are differentiable, and at each of these points the components of the derivative of F are equal to the derivatives of the components of F.

I will also need to use the following signal fact from one variable calculus

Theorem 1.9 (Mean Value Theorem)**.** *If $(a,b) \subset \mathbf{R}$ is open and $f : (a,b) \to \mathbf{R}$ is differentiable on (a,b) then for any two distinct points $x, y \in (a,b)$, there exists a point c between x and y such that*

$$\frac{f(x)-f(y)}{x-y} = f'(c).$$

Now back to the program:

Proof of Theorem 1.7. I will give the proof in the special case where $F = f : \mathbf{R}^2 \to \mathbf{R}$ is scalar-valued and depends on only *two* variables. The proof for scalar-valued functions of $n > 2$ variables is similar and left as an exercise.

Theorem 1.5 tells me that there is only one candidate $\left[\frac{\partial f}{\partial x_1}(\mathbf{a}) \quad \frac{\partial f}{\partial x_2}(\mathbf{a})\right]$ for the standard matrix for $Df(\mathbf{a})$. So assuming that f is C^1 at some point $\mathbf{a} = (a_1, a_2)$, I must show that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) - \left(\frac{\partial f}{\partial x_1}(\mathbf{a})h_1 + \frac{\partial f}{\partial x_2}(\mathbf{a})h_2\right)}{\|\mathbf{h}\|} = \mathbf{0}. \tag{5}$$

Being C^1 at $\mathbf{a}$ means that f is at least defined near and at $\mathbf{a}$ (why?). That is, there exists $r > 0$ such that $f(\mathbf{x})$ is defined for all $\mathbf{x} \in B_r(\mathbf{a})$. Moreover, given a point $\mathbf{a} = (a_1, a_2) \in \mathbf{R}^2$, and a displacement $\mathbf{h} = (h_1, h_2) \in \mathbf{R}^2$ with $\|\mathbf{h}\| < r$, I may rewrite the expression inside the limit in equation (5) as follows.

$$\begin{aligned}
&\frac{f(a_1+h_1, a_2+h_2) - f(a_1,a_2) - \frac{\partial f}{\partial x_1}(a_1,a_2)h_1 - \frac{\partial f}{\partial x_2}(a_1,a_2)h_2}{\|\mathbf{h}\|} \\
&\qquad = \frac{f(a_1+h_1, a_2+h_2) - f(a_1, a_2+h_2) - \frac{\partial f}{\partial x_1}(a_1,a_2)h_1}{\|\mathbf{h}\|} \\
&\qquad\qquad + \frac{f(a_1, a_2+h_2) - f(a_1,a_2) - \frac{\partial f}{\partial x_2}(a_1,a_2)h_2}{\|\mathbf{h}\|}.
\end{aligned}$$

So to establish (5) it suffices to show that each of the last two expressions have limit $\mathbf{0}$ as $\mathbf{h} \to \mathbf{0}$. I will show this for the first (i.e. second last) expression only, the argument for the other expression being similar.

Given $\epsilon > 0$, continuity of partial derivatives tells me that there exists $\delta > 0$ such that $\|\mathbf{x} - \mathbf{a}\| < \delta$ implies that

$$\left\| \frac{\partial f}{\partial x_1}(\mathbf{x}) - \frac{\partial f}{\partial x_1}(\mathbf{a}) \right\| < \epsilon.$$

Moreover, if I think of f as a function of only the first variable x_1, then the one variable mean value theorem tells me that

$$\frac{f(a_1+h_1, a_2+h_2) - f(a_1, a_2+h_2) - \frac{\partial f}{\partial x_1}(a_1,a_2)h_1}{\|\mathbf{h}\|} = \frac{\left(\frac{\partial f}{\partial x_1}(a_1+\tilde{h}_1, a_2+h_2) - \frac{\partial f}{\partial x_1}(a_1,a_2)\right) h_1}{\|\mathbf{h}\|}$$

for some number $\tilde{h}_1$ between 0 and h_1. In particular $\left\|(\tilde{h}_1, h_2)\right\| \leq \|\mathbf{h}\|$. So if $\|\mathbf{h}\| < \delta$, I can estimate as follows

$$\begin{aligned}
&\frac{\left\|f(a_1+h_1, a_2+h_2) - f(a_1, a_2+h_2) - \frac{\partial f}{\partial x_1}(a_1,a_2)h_1\right\|}{\|\mathbf{h}\|} \\
&\qquad = \left\| \frac{\partial f}{\partial x_1}(a_1+\tilde{h}_1, a_2+h_2) - \frac{\partial f}{\partial x_1}(a_1,a_2) \right\| \frac{|h_1|}{\|\mathbf{h}\|} < \epsilon \cdot 1 = \epsilon.
\end{aligned}$$

This proves that the left side converges to 0 as $\mathbf{h} \to \mathbf{0}$, which is what I intended to show. □

2. Differentiating composite functions

Theorem 2.1 (Chain Rule)**.** *Suppose that $G : \mathbf{R}^n \to \mathbf{R}^m$ is differentiable at $\mathbf{a} \in \mathbf{R}^n$ and $F : \mathbf{R}^m \to \mathbf{R}^\ell$ is differentiable at $G(\mathbf{a})$. Then the composition $F \circ G : \mathbf{R}^n \to \mathbf{R}^\ell$ is differentiable at $\mathbf{a}$ and*

$$D(F \circ G)(\mathbf{a}) = DF(G(\mathbf{a})) \circ DG(\mathbf{a}).$$

Proof. The composition $DF(G(\mathbf{a})) \circ DG(\mathbf{a})$ of two linear maps is linear, so it suffices for me to show that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{F(G(\mathbf{a}+\mathbf{h})) - F(G(\mathbf{a})) - DF(G(\mathbf{a}))DG(\mathbf{a})\mathbf{h}}{\|\mathbf{h}\|} = 0.$$

Differentiability of F at $G(\mathbf{a})$ implies that

$$F(G(\mathbf{a}+\mathbf{h})) - F(G(\mathbf{a})) = DF(G(\mathbf{a}))(G(\mathbf{a}+\mathbf{h}) - G(\mathbf{a})) + E_F(G(\mathbf{a}+\mathbf{h}) - G(\mathbf{a}))$$

where $\lim_{\mathbf{v}\to\mathbf{0}} \frac{\|E_F(\mathbf{v})\|}{\|\mathbf{v}\|} = 0$. Hence the limit above can be rewritten

$$\begin{aligned}
&\lim_{\mathbf{h}\to\mathbf{0}} \frac{DF(G(\mathbf{a}))(G(\mathbf{a}+\mathbf{h}) - G(\mathbf{a}) - DG(\mathbf{a})\mathbf{h}))}{\|\mathbf{h}\|} + \lim_{\mathbf{h}\to\mathbf{0}} \frac{E_F(G(\mathbf{a}+\mathbf{h}) - G(\mathbf{a})))}{\|\mathbf{h}\|} \\
= \;& DF(G(\mathbf{a}))\left(\lim_{\mathbf{h}\to\mathbf{0}} \frac{G(\mathbf{a}+\mathbf{h}) - G(\mathbf{a}) - DG(\mathbf{a})\mathbf{h})}{\|\mathbf{h}\|}\right) + \lim_{\mathbf{h}\to\mathbf{0}} \frac{E_F(G(\mathbf{a}+\mathbf{h}) - G(\mathbf{a})))}{\|\mathbf{h}\|} \\
= \;& DF(G(\mathbf{a}))\mathbf{0} + \lim_{\mathbf{h}\to\mathbf{0}} \frac{E_F(G(\mathbf{a}+\mathbf{h}) - G(\mathbf{a})))}{\|\mathbf{h}\|} \\
= \;& \lim_{\mathbf{h}\to\mathbf{0}} \frac{E_F(G(\mathbf{a}+\mathbf{h}) - G(\mathbf{a})))}{\|\mathbf{h}\|}.
\end{aligned}$$

The first equality holds because $DF(G(\mathbf{a}))$ is linear and therefore continuous. The second equality follows from the definition of differentiability.

For the remaining limit, I use the fact that

$$G(\mathbf{a}+\mathbf{h}) - G(\mathbf{a}) = DG(\mathbf{a})\mathbf{h} + E_G(\mathbf{h}).$$

where $\lim_{\mathbf{h}\to\mathbf{0}} \frac{E_G(\mathbf{h})}{\|\mathbf{h}\|} = 0$. In particular, there exists $\delta_1 > 0$ such that $0 < \|\mathbf{h}\| < \delta_1$ implies $\frac{\|E_G\mathbf{h}\|}{\|\mathbf{h}\|} < 1$. Hence when $\|\mathbf{h}\| < \delta_1$, I can employ the triangle and Cauchy-Schwarz inequality to estimate

$$\|G(\mathbf{a}+\mathbf{h}) - G(\mathbf{a})\| \le \|DG(\mathbf{a})\| \, \|\mathbf{h}\| + \|E_G(\mathbf{h})\| \le (\|DG(\mathbf{a})\| + 1) \, \|\mathbf{h}\| \, .$$

Given $\epsilon > 0$, I can then choose $\delta_2 > 0$ such that $0 < \|\mathbf{k}\| < \delta_2$ implies that $\frac{\|E_F(\mathbf{k})\|}{\|\mathbf{k}\|} < \frac{\epsilon}{(\|DG(\mathbf{a})\|+1)}$. So if $0 < \|\mathbf{h}\| < \delta := \min\{\delta_1, \frac{\delta_2}{(\|DG(\mathbf{a})\|+1)}\}$, then

$$\|G(\mathbf{a}+\mathbf{h}) - G(\mathbf{a})\| \le (\|DG(\mathbf{a})\| + 1) \, \|\mathbf{h}\| < \delta_2,$$

and therefore

$$\|E_F(G(\mathbf{a}+\mathbf{h}) - G(\mathbf{a})))\| < \frac{\epsilon}{(\|DG(\mathbf{a})\| + 1)} \|G(\mathbf{a}+\mathbf{h}) - G(\mathbf{a})\| \le \frac{\epsilon(\|DG(\mathbf{a})\| + 1) \, \|\mathbf{h}\|}{(\|DG(\mathbf{a})\| + 1)} = \epsilon \, \|\mathbf{h}\| \, .$$

In short, $0 < \|\mathbf{h}\| < \delta$ implies that

$$\frac{E_F(G(\mathbf{a}+\mathbf{h}) - G(\mathbf{a})))}{\|\mathbf{h}\|} < \epsilon.$$

Hence

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{E_F(G(\mathbf{a}+\mathbf{h})-G(\mathbf{a}))}{\|\mathbf{h}\|} = \mathbf{0},$$

which is the thing it remained for me to show. □

3. Equality of mixed partial derivatives

First a cautionary tale.

Example 3.1. *Let* $f(x_1,x_2) = \frac{x_1^2-x_2^2}{x_1^2+x_2^2}$. *Observe that*

$$\lim_{x_1\to 0}\lim_{x_2\to 0} f(x_1,x_2) = \lim_{x_1\to 0}\frac{x_1^2}{x_1^2} = 1.$$

However,

$$\lim_{x_2\to 0}\lim_{x_1\to 0} f(x_1,x_2) = \lim_{x_2\to 0}\frac{-x_2^2}{x_2^2} = -1.$$

The moral? One cannot generally switch the order in which one takes limits and expect to get the same answer.

Definition 3.2. *A function* $F:\mathbf{R}^n\to\mathbf{R}^m$ *is said to be* C^2 *(or* twice continuously differentiable*) if all first and second partial derivatives of* f *exist and are continuous at every point* $a\in\mathbf{R}^n$ *that belongs to the domain of* F.

Now that I've defined C^1 and C^2, you can probably imagine then what C^k means when $k>2$. The following theorem tells us that order is irrelevant when we take second (and higher order) partial derivatives of a 'decent' function of several variables.

Theorem 3.3. *Suppose that* $f:\mathbf{R}^n\to\mathbf{R}$ *is* C^2. *Then for any* $1\le i,j\le n$ *and any* $a\in\mathbf{R}^n$ *in the domain of* f, *one has*

$$\frac{\partial^2 f}{\partial x_i\partial x_j}(a) = \frac{\partial^2 f}{\partial x_j\partial x_i}(a).$$

My proof is quite similar to Shifrin's, but (in my humble opinion) mine ends a little more honestly. In any case, the main thing is to show that one can reverse the order of the two limits involved in taking a second partial derivative.

Proof. To start with, note that since we are only considering derivatives of f with respect to x_i and x_j, we might as well assume that these are the *only* variables on which f depends. That is, it suffices to assume that $n=2$ in the statement of the theorem, fix a point $a=(a_1,a_2)$ in the domain of f and show that

$$\frac{\partial^2 f}{\partial x_1\partial x_2}(a) = \frac{\partial^2 f}{\partial x_2\partial x_1}(a).$$

To this end, I go back to the definition of derivative, applying it to both partial derivatives:

$$\begin{aligned}\frac{\partial^2 f}{\partial x_1 \partial x_2}(a) &= \lim_{h_1\to 0} \frac{\frac{\partial f}{\partial x_2}(a_1+h_1,a_2) - \frac{\partial f}{\partial x_2}(a_1,a_2)}{h_1} \\ &= \lim_{h_1\to 0} \frac{\lim_{h_2\to 0}\left(\frac{f(a_1+h_1,a_2+h_2)-f(a_1+h_1,a_2)}{h_2}\right) - \lim_{h_2\to 0}\left(\frac{f(a_1,a_2+h_2)-f(a_1,a_2)}{h_2}\right)}{h_1} \\ &= \lim_{h_1\to 0}\lim_{h_2\to 0} \frac{f(a_1+h_1,a_2+h_2)-f(a_1+h_1,a_2)-f(a_1,a_2+h_2)+f(a_1,a_2)}{h_1h_2}\end{aligned}$$

Let me (for brevity's sake) call the quantity inside the last limit $Q(h_1, h_2)$.

Unnecessary motivational digression: *Similarly, when the partial derivatives are reversed, one finds:*

$$\frac{\partial^2 f}{\partial x_2 \partial x_1}(a) = \lim_{h_2\to 0}\lim_{h_1\to 0} Q(h_1,h_2)$$

That is, we get the same thing as before, except that the order of the limits is reversed. If we could switch the limits, we'd be home-free. But without justification, we can't. Instead we take a less direct but more justifiable approach that relies on the mean value theorem.

Lemma 3.4. *For each $h = (h_1, h_2) \in \mathbf{R}^2$, there exists $\tilde{h} = (\tilde{h}_1, \tilde{h}_2)$ inside the rectangle determined by h and 0 such that*

$$Q(h) = \frac{\partial^2 f}{\partial x_2 \partial x_1}(a+\tilde{h})$$

Proof. Note (i.e. really—check it!) that we can rewrite

$$Q(h_1,h_2) = \frac{1}{h_2}\frac{g(a_1+h_1)-g(a_1)}{h_1}$$

where $g : \mathbf{R} \to \mathbf{R}$ is given by $g(t) := f(t, a_2+h_2) - f(t, a_2)$. In particular g is a differentiable function of one variable with derivative given by $g'(t) = \frac{\partial f}{\partial x_1}(t, a_2+h_2) - \frac{\partial f}{\partial x_1}(t, a_2)$. So I can apply the mean value theorem, obtaining a number $\tilde{h}_1$ between 0 and h_1 such that

$$Q(h_1,h_2) = \frac{1}{h_2}\left(\frac{g(a_1+h_1)-g(a_1)}{h_1}\right) = \frac{1}{h_2}g'(a_1+\tilde{h}_1) = \frac{1}{h_2}\left(\frac{\partial f}{\partial x_1}(a_1+\tilde{h}_1, a_2+h_2) - \frac{\partial f}{\partial x_1}(a_1+\tilde{h}_1, a_2)\right).$$

Applying the Mean Value Theorem a second time, to this last expression, gives me a number $\tilde{h}_2$ between 0 and h_2 such that

$$Q(h_1,h_2) = \frac{\partial^2 f}{\partial x_2 \partial x_1}(a_1+\tilde{h}_1, a_2+\tilde{h}_2)$$

□

To finish the proof of the theorem, I will use the convenient notation $A \approx_\epsilon B$ to mean that $A, B \in \mathbf{R}$ satisfy $|A - B| < \epsilon$. Note that (by the triangle inequality) we have 'approximate transitivity'—i.e. $A \approx_{\epsilon_1} B$ and $B \approx_{\epsilon_2} C$ implies $A \approx_{\epsilon_1+\epsilon_2} C$.

It will suffice to show that

$$\frac{\partial^2 f}{\partial x_1 \partial x_2}(a) \approx_\epsilon \frac{\partial^2 f}{\partial x_2 \partial x_1}(a)$$

for every $\epsilon > 0$. So let $\epsilon > 0$ be given. By continuity of second partial derivatives, there exists $\delta > 0$ such that $\|h\| < \delta$ implies that

$$\left| \frac{\partial^2 f}{\partial x_2 \partial x_1}(a+h) - \frac{\partial^2 f}{\partial x_2 \partial x_1}(a) \right| < \frac{1}{3}\epsilon.$$

Using the definition of limit twice and then the above lemma, I therefore obtain that when h_1 and then h_2 are small enough,

$$\frac{\partial^2 f}{\partial x_1 \partial x_2}(a) = \lim_{h_1 \to 0} \lim_{h_2 \to 0} Q(h_1, h_2) \approx_{\epsilon/3} \lim_{h_2 \to 0} Q(h_1, h_2) \approx_{\epsilon/3} Q(h_1, h_2) = \frac{\partial^2 f}{\partial x_2 \partial x_1}(a+\tilde{h}) \approx_{\epsilon/3} \frac{\partial^2 f}{\partial x_2 \partial x_1}(a).$$

In short,

$$\frac{\partial^2 f}{\partial x_1 \partial x_2}(a) \approx_{\epsilon} \frac{\partial^2 f}{\partial x_2 \partial x_1}(a),$$

which is what I sought to show. □