2 The Derivative

2.2 Differentiability in One Variable

We seek to view the derivative in a way that generalizes to multiple dimensions. From our studies of functions $f:\mathbb{R}\to\mathbb{R}$ and $f:\mathbb{R}^2\to\mathbb{R}$, we have the idea of a tangent line/plane approximation. That is, as we get closer and closer to the point of tangency, our tangent line (resp. plane) and curve (resp. surface) get closer and closer. We derive a more mathematically precise way to describe this.

Let $f:\mathbb{R}\to\mathbb{R}$ be a function, and say a is the point at which we want to construct our tangent line $l(x)=mx+b$. Our first obvious constraint on l is that it must intersect f at the point of tangency, meaning $l(a)=f(a)$. This forces $ma+b=f(a)$, and so $b=f(a)-ma$. Plugging this back into our equation for the line, we get $l(x)=f(a)+m(x-a)$. Thus, the family of all potential tangent lines reduces from all lines to only the lines passing through $(a,f(a))$. Great! But how do we pin down the correct slope m?

[2.2.1] Example

To see what goes wrong with a bad guess, let's isolate a specific example: $f(x)=x^2$ at $a=0$. Since $f(0)=0$, our candidate tangent lines take the form $l(x)=mx$. So...is $l(x)=5x$ a tangent line? Obviously not—we want our tangent line to fit snugly with f near $x=0$.

Try slopes like $2x$, x, or $0.0001x$. At some scale, these lines might *look* tangent to f at $x=0$, but if we zoom in close enough, they all inevitably stab through the parabola rather than hugging it. While the absolute gap between the curve and the line vanishes as $x\to 0$, it shrinks at the *same rate* that x itself is shrinking.

Returning to our general $f(x)$ and $l(x)$, we see that merely requiring $l(x)\to f(x)$ as $x\to a$ isn't enough; *any* intersecting line satisfies that! For $l(x)$ to fit snugly with $f(x)$ as the true tangent line, we want $l(x)$ to get closer to $f(x)$ at a rate *faster* than $x\to a$. If the gap between $f(x)$ and $l(x)$ shrinks at the same rate as $x\to a$, then the relative error between them remains constant, meaning the line intersects the curve at a sharp angle. To fit our notion of "tangency," the error must be a smaller order of magnitude than the distance we are zooming in.

[2.2.2] Definition (*Tangent Line*)

This behavior is perfectly captured by little-o notation. We demand that the error $f(x)-l(x)$ is $o(|x-a|)$. In other words, as $x\to a$, the ratio of the error to the distance $|x-a|$ must completely vanish:

$$\lim_{x\to a}\frac{f(x)-l(x)}{|x-a|}=0.$$

[2.2.3] Theorem (*Equivalence of Little-o and Limit Definitions*)

Suppose $f:\mathbb{R}\to\mathbb{R}$ is a function, and let $l(x)=f(a)+m(x-a)$ be a generic line passing through $(a,f(a))$. The error $f(x)-l(x)$ is $o(|x-a|)$ if and only if $m=f'(a)$.

Proof. By definition, the error being $o(|x-a|)$ implies that

$$\lim_{x\to a} \frac{f(x)-(f(a)+m(x-a))}{x-a} = 0.$$

Splitting the fraction, we get

$$\lim_{x\to a} \left(\frac{f(x)-f(a)}{x-a} - \frac{m(x-a)}{x-a} \right) = 0,$$

which evaluates to

$$\lim_{x\to a} \left(\frac{f(x)-f(a)}{x-a} \right) - m = 0.$$

Rearranging to solve for m yields

$$m = \lim_{x\to a} \frac{f(x)-f(a)}{x-a},$$

which is exactly $f'(a)$. By simply enforcing that a generic line intersects $f(a)$ and that its error is $o(|x-a|)$, the only valid slope remaining is exactly the standard limit definition of the derivative. ❦

2 The Derivative

2.2 Bachmann-Landau Notation

[2.2.1] DEFINITION ($o(1), o(\|h\|), \mathcal{O}(\|h\|)$ *Mappings*)

Suppose $\varphi : B(\mathbf{0}_n, \varepsilon) \to \mathbb{R}^m$ is a map from a ball about $\mathbf{0}_n \in \mathbb{R}^n$ to $\mathbb{R}^m$.

- φ is said to be **smaller than order 1** ("little-o of 1", $o(1)$) if, for any $c > 0$, there is some $H > 0$ such that $\|\varphi(h)\| \leq c$ for every $h \in \mathbb{R}^n : \|h\| < H$.
- φ is said to be **smaller than order** h ("little-o of h", $o(\|h\|)$) if, for every $c > 0$, there is some $H > 0$ such that $\|\varphi(h)\| \leq c\|h\|$ for every $h \in \mathbb{R}^n : \|h\| < H$.
- φ is said to be **of order** h ("big-o of h", $\mathcal{O}(\|h\|)$) if, for a specific $c > 0$, there is some $H > 0$ such that $\|\varphi(h)\| \leq c\|h\|$ for every $h \in \mathbb{R}^n : \|h\| < H$.

Fix some domain ball $B(\mathbf{0}_n, \varepsilon)$ and some codomain $\mathbb{R}^m$. The set of all $o(1)$ mappings with respect to this domain and range is denoted by $o(1)$, and analogous statements follow for $o(\|h\|)$ and $\mathcal{O}(\|h\|)$.

[2.2.2] REMARK (*Intuition Behind Mappings*)

Saying something is an $o(1)$, $o(\|h\|)$, $\mathcal{O}(\|h\|)$ mapping is a statement of the function being bounded above for small enough inputs (inputs near the origin).

- An $o(1)$ mapping is such that the norm of output vectors near $\mathbf{0}_n$ get arbitrarily small relative to any constant. That is, $o(1) \to 0$ as $h \to \mathbf{0}_n$.
- An $o(\|h\|)$ mapping is such that the norm of output vectors near $\mathbf{0}_n$ get arbitrarily small relative to any linear function. That is, $o(\|h\|)/\|h\| \to 0$ as $h \to \mathbf{0}_n$.
- A $\mathcal{O}(\|h\|)$ mapping is such that the norm of output vectors near $\mathbf{0}_n$ get arbitrarily small relative to a specific linear function. That is, $o(\|h\|)/\|h\|$ is bounded above by some $C > 0$ as $h \to \mathbf{0}_n$.

Suppose $\varphi : \mathbb{R} \to \mathbb{R}$. If φ is $o(1)$, then $\varphi(0) \to 0$. If φ is $o(\|h\|)$, then for **any** constant $C > 0$, there is some interval $[-h, h]$ in which φ is bounded in between the lines $y = \pm Cx$. If φ is $\mathcal{O}(\|h\|)$, then there exists **some** constant $C > 0$ for which there is some interval $[-h, h]$ in which φ is bounded in between the lines $y = \pm Cx$.

Note that since an $o(\|h\|)$ mapping is less than **any** $c\|h\|$ as $h \to \mathbf{0}_n$, it is less than **some** $c\|h\|$ as $h \to \mathbf{0}_n$, and so it is an $\mathcal{O}(\|h\|)$ mapping. Since there is some $c > 0$ such that $\mathcal{O}(\|h\|) \leq c\|h\|$ as $h \to \mathbf{0}_n$, it follows that $\mathcal{O}(\|h\|) \to 0$, and so a $\mathcal{O}(\|h\|)$ mapping is an $o(1)$ mapping as well.

[2.2.3] THEOREM (*Classification of Mappings*)

$o(\|h\|) \subseteq \mathcal{O}(\|h\|) \subseteq o(1)$.

[2.2.4] THEOREM (*Dominance Principle for Landau Spaces*)

Suppose $\varphi = o(1)$, and suppose $\|\psi(h)\| \leq \|\varphi(h)\|$ for small enough h (in terms of norm). Then $\psi = o(1)$. Similar statements follow for $o(\|h\|)$ and $\mathcal{O}(\|h\|)$.

[2.2.5] Theorem (*Landau Spaces are Vector Spaces*)

Fix and domain-ball $B(\mathbf{0}_n, \varepsilon)$ and any codomain $\mathbb{R}^m$. Then $o(1)$ is a vector space, and $\mathcal{O}(\|h\|)$ and $o(\|h\|)$ are subspaces.

Proof. Define $+ : o(1) \times o(1) \to o(1)$ by traditional function addition. Let $f, g \in o(1)$. Fix $c > 0$. There exists some $H > 0$ such that

$$\|f(h)\| \leq c/2 \qquad \|g(h)\| \leq c/2$$

for every $h \in \mathbb{R}^n : \|h\| < H$. Then note

$$\|(f+g)(h)\| \leq \|f(h)\| + \|g(h)\| = c,$$

meaning $f + g \in o(1)$. Commutativity, associativity, additive identity, and additive inverse axioms follow immediately.

Fix $k \in \mathbb{R}$, and let $f \in o(1)$. With our aforementioned fixed $c > 0$, there is some $H' > 0$ such that

$$\|f(h)\| \leq c/|k|$$

for every $h \in \mathbb{R}^n : \|h\| < H$. Then, of course,

$$\|kf(h)\| = |k|\|f(h)\| \leq c,$$

meaning $kf \in o(1)$. Multiplicative identity and inverse axioms quickly follow. Subspace axioms are trivial to verify. ❦

We provide an example of a class of functions that encompass $o(1)$, $o(\|h\|)$, and $\mathcal{O}(\|h\|)$ functions.

[2.2.6] Example

Define

$$\varphi_k : \mathbb{R}^n \to \mathbb{R} \qquad \varphi_k(x) = \|x\|^k$$

for all $k \in \mathbb{R}$. We show that φ_k is $o(1)$ for every $k > 0$, φ_k is $\mathcal{O}(\|h\|)$ for every $k \geq 1$, and φ_k is $o(\|h\|)$ for every $k > 1$.

First, we show φ_k is $o(1)$ for $k > 0$. Fix $c > 0$, then let $H = c^{1/k}$. Then

$$\|\varphi_k(h)\| = \|h\|^k < (c^{1/k})^k = c$$

for every $h \in \mathbb{R}^n : \|h\| < H$, and so φ_k is $o(1)$.

Now let $k > 1$. For any $c > 0$, we may let $H = c^{1/(k-1)}$ and observe that

$$\|\varphi_k(h)\| = \|h\|^k = \|h\|\|h\|^{k-1} \leq c\|h\|$$

for every $h \in \mathbb{R}^n : \|H\| < H$, and so φ_k is $o(\|h\|)$.

If $k = 1$, then H is undefined, so we tackle it a different way. Note that if we let $c > 1$, it is true that

$$\|\varphi_k(h)\| = \|x\| \leq c\|x\|.$$

Thus, for $k \geq 1$, $\varphi_k \in \mathcal{O}(\|h\|)$ (by stitching together our strategy for $k > 1$ and $k = 1$, we restrict $c > 1$, meaning it is no longer $o(\|h\|)$).

[2.2.7] Theorem (*Finite-Dimensional Linear Maps are Continuous*)

Suppose $T : V \to W$ is a linear map between two finite-dimensional vector spaces. Then T is continuous.

Proof. Let $(v_k) \subseteq V$ be a sequence of vectors such that $v_k \to v \in V$. For T to be continuous, we need that the image sequence $T(v_k) \subseteq W$ is such that $T(v_k) \to T(v) \in W$.

Equip V, W with norms $\|\cdot\|_V, \|\cdot\|_W$. Choose a basis $\mathcal{B} = \{e_1, \ldots, e_n\} \subseteq V$. Express each v_k in the sequence as

$$v_k = a_{k,1}e_1 + \cdots + a_{k,n}e_n,$$

and express

$$v = a_1 e_1 + \cdots + a_n e_n.$$

Since (v_k) is a convergent sequence, we know $\|v_k - v\|_V \to 0$ as $k \to \infty$. Because V is finite-dimensional, convergence in norm implies coordinate-wise convergence. Thus, for each coordinate $i \in \{1, \ldots, n\}$, we have $|a_{k,i} - a_i| \to 0$.

Now consider the images

$$T(v_k) = a_{k,1}T(e_1) + \cdots + a_{k,n}T(e_n),$$
$$T(v) = a_1 T(e_1) + \cdots + a_n T(e_n).$$

Let $A = \max\{\|T(e_1)\|_W, \ldots, \|T(e_n)\|_W\}$. Then, using the linearity of T and the triangle inequality,

$$\|T(v_k) - T(v)\|_W = \|T(v_k - v)\|_W = \|\sum_{i=1}^{n}(a_{k,i} - a_i)T(e_i)\|_W \leq \sum_{i=1}^{n}|a_{k,i} - a_i|\|T(e_i)\|_W \leq A\sum_{i=1}^{n}|a_{k,i} - a_i|.$$

As $k \to \infty$, each $|a_{k,i} - a_i| \to 0$, which means the right-hand sum goes to 0. Therefore, $\|T(v_k) - T(v)\|_W \to 0$, establishing that $T(v_k) \to T(v)$. ❦

[2.2.8] Theorem (*Linear Maps in Landau Notation*)

Suppose $T : \mathbb{R}^n \to \mathbb{R}^m$ is linear. Then T is an $\mathcal{O}(\|h\|)$ mapping, and T is an $o(\|h\|)$ mapping if and only if $T \equiv 0$.

Proof. Define the unit ball $B = \{v \in \mathbb{R}^n : \|v\| \leq 1\}$. Then note B is closed and bounded, thus compact. All linear maps are continuous, and so the image of the ball $T(B)$ is compact as well, meaning $T(B)$ is bounded. Thus there is some $M > 0$ such that $T(B) \leq M$.

Let $v \in V$, which we can write as $v = \|v\|v_0$ for some $v_0 \in B$ (with $\|v_0\| = 1$). Note then that

$$\|Tv\| = \|T(\|v\|v_0)\| = \|v\|\|Tv_0\| \leq M\|v\|.$$

Thus T is a $\mathcal{O}(\|h\|)$ mapping.

Note the zero linear map is trivially $o(\|h\|)$, so suppose $T \not\equiv 0$—we show $T \neq o(\|h\|)$. Thus there is some $v_0 \in V$ such that $T(v_0) \neq 0$. Choose $c = \|T(v_0)\|/2$. Let $v = \lambda v_0$ be any scalar multiple of v_0. Then note

$$\|Tv\| = |\lambda|\|T(v_0)\| = 2c|\lambda| = 2c\|v\|.$$

Thus $\|Tv\| > c\|v\|$, meaning $T \neq o(\|h\|)$, completing the proof. ❦

[2.2.9] Theorem (*Product of Landau Functions*)

Suppose

$$f, g, fg : B(\mathbf{0}_n, \varepsilon) \to \mathbb{R}$$

are mappings such that $f = o(1)$ and $g = \mathcal{O}(\|h\|)$. Then $fg = o(\|h\|)$.

Proof. Fix $c > 0$. Choose $d > 0$ be such that there exists an $H > 0$ such that

$$\|f(h)\| \le c/d \quad \|g(h)\| \le d\|h\|$$

whenever $h \in \mathbb{R}^n : \|h\| < H$. Then, of course,

$$\|fg(h)\| = \|f(h)g(h)\| \le c\|h\|$$

for every $h \in \mathbb{R}^n : \|h\| < H$ for arbitrary $c > 0$. Thus $fg = o(\|h\|)$. ❦

Note this theorem applies if $f = \mathcal{O}(\|h\|)$, or $f = o(\|h\|)$, or $g = o(\|h\|)$.

[2.2.10] Theorem (*Composition of Landau Functions*)

$$\begin{aligned} o(o(1)) &= o(1), \\ \mathcal{O}(\mathcal{O}(\|h\|)) &= \mathcal{O}(\|h\|), \\ o(\mathcal{O}(\|h\|)) &= o(\|h\|), \\ \mathcal{O}(o(\|h\|)) &= o(\|h\|). \end{aligned}$$

For Landau functions that tend to 0, composition order is transitive.

Proof. Trivialized by the definition of Landau functions. ❦

2.3 Revision of the Single-Variable Derivative

[2.3.1] Definition (*Univariate Differentiability*)

A function $f : \mathbb{R} \to \mathbb{R}$ is said to be **differentiable** at $x = a$ if

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

exists.

This is merely a construction, however. In trying to abstract a construction to a more general setting, we seek to determine invariants we seek to keep as we move to more general settings. Let's try to figure out what makes the derivative special.

[2.3.2] Remark (*Derivative as Best Linear Approximation*)

Let's try to consider what it means for the graph of a function $f : \mathbb{R} \to \mathbb{R}$ to have a tangent line of slope t at $(a, f(a))$. Let x be a point close to a. To measure how much the value of our tangent line deviates from the value of f at x, we can define a function

$$g(x) := f(x) - [f(a) + t(x - a)],$$

where the term in brackets is the tangent line evaluated at x. But graphically, this is rather crude: it's a rotated line not necessarily aligned with a nice axis on the coordinate plane. We move away from these "global coordinates" that depend on x to a coordinate system centered at $(a, f(a))$.

Let's define $h := x - a$, the horizontal displacement between our chosen point and the point of tangency. We may now write g in terms of h as such:

$$g(h) := f(a+h) - [f(a) + th].$$

This captures the exact same deviation of our tangent line from the function.

Why is this perspective useful? Because it normalizes our problem. If our candidate slope t is indeed the true tangent slope of f at a, then near a, the function f behaves like a line of slope t. By subtracting the line $f(a) + th$ to create $g(h)$, we are subtracting out the entirety of this linear behavior. A curve of slope t minus a line of slope t leaves a remainder curve g with a slope of exactly 0.

Thus, the messy problem of checking if f has a tangent of slope t out at $x = a$ is entirely reduced to a much tidier problem: checking if g is perfectly horizontal at the origin.

To say that g is horizontal at the origin geometrically means that for any arbitrarily small slope $c > 0$, the graph of g will eventually fall entirely within the wedge bounded by the lines $y = ch$ and $y = -ch$ if we zoom in close enough to the origin. Analytically, this means $|g(h)| \leq c|h|$ for small enough h, which is exactly the definition of $g(h)$ being $o(h)$.

We can now show that $t = f'(a)$. Note that

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

where the numerator is exactly of the form $g(h)$, and this limit expression also agrees that $g(h)$ is $o(h)$. Thus the derivative $f'(a)$ is exactly the value of t for which our tangent line is *actually* tangent to f at $x = a$.

We have to generalize this idea to multiple dimensions. Note we can observe the derivative not just as a constant $f'(a)$, but the linear map $T_a : \mathbb{R} \to \mathbb{R}$ that sends $h \mapsto f'(a)h$.

[2.3.3] Definition (*Multivariate Differentiability*)

Let $A \subseteq \mathbb{R}^n$, and let $f : A \to \mathbb{R}^m$ be a mapping. f is said to be **differentiable** at some interior point $a \in \mathrm{Int}(A)$ if there exists some $T_a : \mathbb{R}^n \to \mathbb{R}^m$ satisfying the condition that

$$f(a+h) - f(a) - T_a(h) = o(h).$$

Alternatively, f is differentiable at a if

$$f(a+h) = f(a) + T_a(h) + o(h),$$

where $o(h)$ is some $o(h)$ mapping.

T_a is said to be the **total derivative** of f at a, written Df_a or $(Df)_a$. When f is differentiable at a, the matrix $[Df_a]$ expressed in the standard basis is said to be the **Jacobian** of f at a.

[2.3.4] Remark (*Differentiation at Interior Points*)

Recall that, for a univariate function $f : [a,b] \to \mathbb{R}$, f' is not well defined at a, b. This is because $a, b \notin \mathrm{Int}([a,b])$, an idea expressed in our definition of multivariate differentiability.

[2.3.5] Example (*Differentiability for Surfaces*)

Let's apply this idea to some $f : A \to \mathbb{R}$, where $A \subseteq \mathbb{R}^2$. We say f is differentiable at some interior point of $(x,y) \in \mathrm{Int}(A)$ if

$$f(x+h, y+k) - f(x,y) - T_a(h,k) = o(h).$$

Obviously, since T_a is a linear map, we note $T_a(h,k) = ah + bk$ for some choice of values $a, b \in \mathbb{R}$. Our statement of differentiability, rearranged, says that

$$f(x+h, y+k) = f(x,y) + T_a(h,k) + o(h) \approx f(x,y) + T_a(h,k) = f(x,y) + ah + bk.$$

The right-hand side is the equation of a plane in local coordinates (h,k), leading us to believe that f is differentiable at (x,y) if, locally, f behaves like a plane. This is exactly our intuition of what is a "tangent plane" for a surface that is differentiable at some point.

[2.3.6] Theorem (*Uniqueness of Derivative*)

Let $f : A \to \mathbb{R}^m$, where $A \subseteq \mathbb{R}^n$, be differentiable at $a \in A$. Then the derivative of f at a is unique.

> **Proof.** Suppose there exist maps $T_a, \tilde{T}_a : \mathbb{R}^n \to \mathbb{R}^m$ such that $f(a+h) - f(a) - T_a(h)$ and $f(a+h) - f(a) + \tilde{T}_a(h)$ are $o(h)$ mappings. Subtracting the two maps yields $(\tilde{T}_a - T_a)$, which is also $o(h)$ by vector space properties of $o(h)$ mappings. But note that the only linear map that is $o(h)$ is the zero mapping, implying $\tilde{T}_a \equiv T_a$. ❦

[2.3.7] Theorem (*Differentiability Implies Continuity*)

Suppose $f : A \to \mathbb{R}^m$, where $A \subseteq \mathbb{R}^m$, is differentiable at $a \in A$. Then f is continuous at a.

> **Proof.** Note that $f(a+h) - f(a) = T_a(h) + o(h) = o(1) + o(1) = o(1)$ as $h \to 0$, meaning f is continuous at a. ❦

2.4 Basic Results and Chain Rule

[2.4.1] Theorem (*Basic Differentiation Rules*)

(1) Let $f : A \to \mathbb{R}^m$, where $A \subseteq \mathbb{R}^m$, be such that $f(a) = v_0$ for every $a \in A$, where $v_0 \in \mathbb{R}^m$ is fixed. Then $(Df)_a \equiv 0$.

(2) Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be linear. Then $(DT)_a \equiv T$.

(3) The derivative of a function is additive and homogenous (i.e., linear).

Proof.

(1) Note that $f(a+h) - f(a) - (Df)_a$, where $(Df)_a$, is equal to $(Df)_a$. For $(Df)_a$ to be linear and $o(h)$, we note that the derivative is forced to be the zero mapping.

(2) Let $a \in \mathbb{R}^n$. Then note $T(a+h) - T(a) - (DT)_a = T(h) - (DT)_a$ must be $o(h)$, but the subtraction of two linear maps is a linear map, and so $(DT)_a = T(h)$ and so $(DT)_a \equiv T$.

(3) Suppose $f : A \to \mathbb{R}^m$, where $A \subseteq \mathbb{R}^n$, is differentiable at $a \in \mathrm{Int}(A)$. We seek to show that λf is differentiable at a with derivative $(D\lambda f)_a = \lambda (Df)_a$. Note that

$$\lambda f(a+h) - \lambda f(a) - \lambda (Df)_a = \lambda (f(a+h) - f(a) - (Df)_a),$$

a scalar multiple of an $o(h)$ map, which is also $o(h)$ (since $o(h)$ maps form a vector space and are thus closed under scalar multiplication).

Now suppose $g : B \to \mathbb{R}^m$, where $B \subseteq \mathbb{R}^n$, is also differentiable at our aforementioned $a \in \mathrm{Int}(A) \cap \mathrm{Int}(B)$. Then we seek to show $[D(f+g)]_a = (Df)_a + (Dg)_a$. Note that if $a \in \mathrm{Int}(A) \cap \mathrm{Int}(B)$, then $a \in \mathrm{Int}(A \cap B)$, so $f+g$ is differentiable at a. Then

$$(f+g)(a+h) - (f+g)(a) - (Df)_a - (Dg)_a = [f(a+h) - f(a) - (Df)_a] + [g(a+h) - g(a) - (Dg)_a]$$

is the sum of two $o(h)$ maps, which is also $o(h)$, completing the proof.

[2.4.2] Theorem (*Chain Rule*)

Let $f : A \to \mathbb{R}^m$, where $A \subseteq \mathbb{R}^n$ be a mapping. Let $g : B \to \mathbb{R}^\ell$, where $B \supseteq f(A)$, be a mapping. Then the composition $g \circ f : A \to \mathbb{R}^\ell$ is well defined. If f is differentiable at some $a \in A$, and g is differentiable at $f(a) \in B$, then $g \circ f$ is differentiable, such that

$$D(g \circ f)_a = D(g)_{f(a)} \circ D(f)_a.$$

With a choice of basis, the Jacobian is such that

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

Proof. To avoid maintaining constant terms throughout multiple steps of algebra, we translate f and g such that a is treated as the zero vector in $\mathbb{R}^n$ and $\mathbb{R}^m$, as well as its image under f and g being the zero vector in $\mathbb{R}^m$ and $\mathbb{R}^\ell$ respectively.

Recalling that $f(a+h) = f(a) + D(f)_a[h] + o(h)$, we may eliminate the constant term by subtracting $f(a)$. With this in mind, define

$$\tilde{f}(h) = f(a+h) - f(a) \qquad \tilde{g}(k) = g(b+k) - g(b),$$

where $b = f(a) \in B$. Then note that $\tilde{f}(\mathbf{0}_n) = \mathbf{0}_m$ and $\tilde{g}(\mathbf{0}_m) = \mathbf{0}_\ell$. As a result,

$$\begin{aligned}\tilde{f}(h) &= D(\tilde{f})_{\mathbf{0}_n}[h] + o(h)\\ &= f(a+h) - f(a) = D(f)_a[h] + o(h),\end{aligned}$$

$$\begin{aligned}\tilde{g}(k) &= D(\tilde{g})_{\mathbf{0}_m}[k] + o(k)\\ &= g(b+k) - g(b) = D(g)_b[k] + o(k),\end{aligned}$$

and so $D(\tilde{f})_{\mathbf{0}_n} = D(f)_a$ and $D(\tilde{g})_{\mathbf{0}_m} = D(g)_b$. Now, note that

$$\begin{aligned}\tilde{g}(\tilde{f}(h)) &= \tilde{g}(f(a+h) - f(a))\\ &= g(b + f(a+h) - f(a)) - g(f(a))\\ &= g(f(a+h)) - g(f(a))\end{aligned}$$

as desired, and an expansion analogous to our previous results reveal that $D(g \circ f)_a = D(\tilde{g} \circ \tilde{f})_{\mathbf{0}_n}$.

As a little aside, these results signify that differentiation is essentially translation-invariant. It is best to work in local coordinates where the problem simplifies nicely (our $\tilde{f}, \tilde{g}$ functions work on a translated coordinate plane), as opposed to global coordinates. Global coordinates are great for absolute data, but it may contain irrelevant data that can prove to be a hindrance.

In any case, we want to write $\tilde{g}(\tilde{f}(h))$ in the form $D(\tilde{g} \circ \tilde{f})_{\mathbf{0}_n}[h] + o(h)$. Expanding reveals

$$\begin{aligned}\tilde{g}(\tilde{f}(h)) &= \tilde{g}(D(\tilde{f})_{\mathbf{0}_n}[h] + o(h))\\ &= D(\tilde{g})_{\mathbf{0}_m}\left[D(\tilde{f})_{\mathbf{0}_n}[h] + o(h)\right] + o\left(D(\tilde{f})_{\mathbf{0}_n}[h] + o(h)\right)\\ &= [D(\tilde{g})_{\mathbf{0}_m} \circ D(\tilde{f})_{\mathbf{0}_n}](h) + D(\tilde{g})_{\mathbf{0}_m}(o(h)) + o(o(h))\\ &= [D(\tilde{g})_{\mathbf{0}_m} \circ D(\tilde{f})_{\mathbf{0}_n}](h) + o(h).\end{aligned}$$

Thus,

$$D(g \circ f)_a = D(\tilde{g} \circ \tilde{f})_{\mathbf{0}_n} = D(\tilde{g})_{\mathbf{0}_m} \circ D(\tilde{f})_{\mathbf{0}_n} = \boxed{D(g)_{f(a)} \circ D(f)_a}.$$

❦

The product and quotient rules may be proven by differentiating xy and x/y, then using chain rule.