ALGEBRAIC STRUCTURE AND GENERATIVE DIFFERENTIATION RULES

In multivariable calculus the derivative is most naturally defined by a characterizing property. Indeed, the one-variable difference-quotient is no longer meaningful. Although computing the matrix of partial derivatives is a matter of one-variable technique, proofs of the derivative properties should be intrinsic and based on the definition. Especially, because the chain rule is so important but is elaborate in coordinates, it demands a lucid intrinsic proof.

A proof in many texts relies on two preparatory results of different flavors and carries out a two-part calculation with estimates. Its heterogeneity long frustrated me as a teacher, leaving even strong students with the impression that the chain rule is complicated and that mathematics sprawls. Eventually I realized that the well known characterization of differentiability in the Landau notation makes the chain rule proof incisive. The notation phrases everything uniformly and packages the estimates, cutting straight to the real issue:

$$o(h) \text{ is a two-sided ideal of } \mathcal{O}(h).$$

Literally the previous sentence isn't quite true, failing in one small way that will be noted below, but still it is the right idea. The fact that the ideal property is the crux of the chain rule proof may also be well known, but it seems to be rare in sophomore-level texts. A proof of the chain rule using Landau notation is given at the very beginning of Nelson's book [9], for example, but there the context is already Banach space and the Fréchet derivative.

Section 1 reviews the Landau notation (originally due to Bachmann) and its basic properties. Section 2 characterizes the derivative in terms of the Landau notation. Section 3 uses the characterizing property to prove the chain rule clearly and briefly. Section 4 rehearses other benefits of phrasing differentiability in the Landau notation, for the same small startup cost.

1. Review of the Landau Notation

Fix positive integers n and m.

Definition 1.1. *Consider a function $f : U \longrightarrow \mathbb{R}^m$ where U is a neighborhood of 0 in $\mathbb{R}^n$. Then:*

- *f is an $\mathcal{O}(h)$-function if there exist $c, \delta \in \mathbb{R}^+$ such that for all $h \in U$,*
$$|h| \leq \delta \implies |f(h)| \leq c|h|.$$
More briefly, we say that such a function f is $\mathcal{O}(h)$.
- *f is an $o(h)$-function if for every $d \in \mathbb{R}^+$ there exists some $\varepsilon_d \in \mathbb{R}^+$ such that for all $h \in U$,*
$$|h| \leq \varepsilon_d \implies |f(h)| \leq d|h|.$$
More briefly, we say that such a function f is $o(h)$.

The set of $\mathcal{O}(h)$-functions is again denoted $\mathcal{O}(h)$, and similarly for $o(h)$.

Clearly $o(h) \subset \mathcal{O}(h)$.

The first of the two preparatory results mentioned in the introduction is that any linear function S is $\mathcal{O}(h)$ (i.e., a linear function has an operator norm). Indeed, the unit sphere in $\mathbb{R}^n$ is compact and S is continuous, so some $c \in \mathbb{R}^+$ exists such that $|Sh| \le c$ for $|h| = 1$; the homogeneity of S shows that $|Sh| \le c|h|$ for all nonzero h, and the inequality holds for $h = 0$ as well. Also, the only linear $o(h)$-function is zero. Indeed, for nonzero linear S, let $d = |Sh|/2$ where $|h| = 1$ and $Sh \ne 0$, and then homogeneity shows that no suitable ε_d exists. Thus the $\mathcal{O}(h)$ functions are the functions of linear decay near the origin, and the $o(h)$-functions are such functions of sub-linear decay.

The spaces $\mathcal{O}(h)$ and $o(h)$ aren't quite vector spaces because the sum of two functions $f : U \longrightarrow \mathbb{R}^m$ and $g : V \longrightarrow \mathbb{R}^m$ is defined only on $U \cap V$. The vector space structure is sensible at the level of function-elements (f, U) or, better, at the level of germs $[f]$ where f and g belong to the same germ if they agree on some neighborhood of 0. We freely suppress these issues, tacitly working with germs when we work at the level of functions, and treating $\mathcal{O}(h)$ and $o(h)$ as vector spaces after all. Strictly speaking, we should check that all properties being discussed for functions are defined at the level of germs, but this will always be clear.

Let $f : U \longrightarrow \mathbb{R}^m$ and $g : V \longrightarrow \mathbb{R}^\ell$ be $\mathcal{O}(h)$-functions, with V a neighborhood of 0 in $\mathbb{R}^m$. Definition 1.1 quickly shows that after shrinking U if necessary, the composition $g \circ f$ is defined. Composition of $\mathcal{O}(h)$-functions descends to germs.

The proof of the next result is the only quantifier-intensive moment in this note. After the proof we will briefly discuss how it could be lightened for a calculus course.

Proposition 1.2. *Let f and g be composable $\mathcal{O}(h)$-functions as in the previous paragraph. Then their composition $g \circ f$ is again $\mathcal{O}(h)$. If either f or g is $o(h)$ then so is $g \circ f$. In symbols, $\mathcal{O}(\mathcal{O}(h)) = \mathcal{O}(h)$, $\mathcal{O}(o(h)) = o(h)$, and $o(\mathcal{O}(h)) = o(h)$.*

Proof. For example, suppose that f is $\mathcal{O}(h)$ and g is $o(h)$. Thus we have c and δ, and for any $d \in \mathbb{R}^+$ we have ε_d. To show that $g \circ f$ is $o(h)$, let $d \in \mathbb{R}^+$ be given. Define $\tilde{d} = d/c$ and $\rho_d = \min\{\varepsilon_{\tilde{d}}/c, \delta\}$. Then for all $h \in U$,

$$\begin{aligned} |h| \le \rho_d \implies & |f(h)| \le c|h| \le \varepsilon_{\tilde{d}} && \text{since } |h| \le \delta \text{ and } |h| \le \varepsilon_{\tilde{d}}/c \\ \implies & |g(f(h))| \le \tilde{d}|f(h)| \le \tilde{d}c|h| && \text{since } |f(h)| \le \varepsilon_{\tilde{d}} \text{ and } |f(h)| \le c|h| \\ \implies & |g(f(h))| \le d|h| && \text{since } \tilde{d}c = d. \end{aligned}$$

The other arguments are similar. □

The proof just given could be made more digestible by leaving h small but unquantified in Definition 1.1 and in the argument: the estimate $|g(f(h))| \le \tilde{d}c|h|$ with c fixed and $\tilde{d}$ small is quick and persuasive. Also, the fact that linear functions are $\mathcal{O}(h)$ can be proved with no reference to compactness, and germs can be elided.

2. The Derivative Via a Characterizing Property

Let $a \in \mathbb{R}^n$ be a point, and let $f : U_a \longrightarrow \mathbb{R}^m$ be defined on a neighborhood of a. Let $S : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be linear. The condition that f is differentiable at a with derivative $f'(a) = S$ is

$$f(a + h) = f(a) + Sh + o(h).$$

This characterization is sensible at the level of germs. The second of the two preparatory results mentioned in the introduction is that $f(a + h) - f(a) = \mathcal{O}(h)$.

This follows immediately from the characterizing property and from the facts that any linear function is $\mathcal{O}(h)$, that $o(h) \subset \mathcal{O}(h)$, and that $\mathcal{O}(h)$ is closed under addition. Our pending proof of the chain rule tacitly uses the just-quoted facts, but it does not use the second preparatory result.

3. The Chain Rule

We are given $U_a \subset \mathbb{R}^n$ and $f : U_a \longrightarrow \mathbb{R}^m$ differentiable at a, and $g : V_{f(a)} \longrightarrow \mathbb{R}^\ell$ differentiable at $f(a)$. Let $S = f'(a)$ and $T = g'(f(a))$. Thus we have the conditions

$$\begin{aligned} f(a+h) &= f(a) + Sh + o(h), \\ g(f(a)+k) &= g(f(a)) + Tk + o(k). \end{aligned}$$

We may assume that $f(U_a) \subset V_{f(a)}$, so that the composition $g \circ f$ is defined. To show that it is differentiable at a with derivative TS, compute that

$$\begin{aligned} g(f(a+h)) &= g(f(a) + Sh + o(h)) && \text{by the first condition} \\ &= g(f(a)) + TSh + T(o(h)) + o(Sh + o(h)) && \text{by the second.} \end{aligned}$$

But $T(o(h)) = \mathcal{O}(o(h))$ and $o(Sh + o(h)) = o(\mathcal{O}(h))$, so the previous display is

$$(g \circ f)(a+h) = (g \circ f)(a) + TSh + \mathcal{O}(o(h)) + o(\mathcal{O}(h)).$$

The rules $\mathcal{O}(o(h)) = o(h)$ and $o(\mathcal{O}(h)) = o(h)$ and $o(h) + o(h) = o(h)$ complete the proof of the chain rule,

$$(g \circ f)(a+h) = (g \circ f)(a) + TSh + o(h).$$

In the chain rule we may assume that all the domain and codomain spaces have the same dimension: simply add unused variables or trivial output component-functions as necessary. The identity function now lies in $\mathcal{O}(h)$, and Proposition 1.2 *almost* says that $\mathcal{O}(h)$ forms an algebra and $o(h)$ is a two-sided ideal of $\mathcal{O}(h)$. The ideal properties prove the chain rule. The only reason that this doesn't quite work is that only one of the two distributive laws holds in $\mathcal{O}(h)$.

4. Related Comments

Setting up the Landau notation and its properties is admittedly a cost to weigh against the benefit of a clear chain rule proof. This section will argue that the notation offers enough other benefits to tilt the balance decisively in its favor.

First, with computing power now ubiquitous, algorithmic thinking in mathematics is crucial. The Landau notation is fundamental to the analysis of algorithms, and so the sooner students see it the better.

Second, basic properties of the Landau notation entail basic properties of the derivative. For tidiness, work in local coordinates: given a function $f : U_a \longrightarrow \mathbb{R}^m$, let $U = U_a - a$ and define

$$f_o : U \longrightarrow \mathbb{R}^m, \quad f_o(h) = f(a+h) - f(a);$$

then f is differentiable at a with $f'(a) = S$ if and only if f_o is differentiable at 0 with $f_o'(0) = S$ as well. The characterizing property for a normalized function (itself now denoted f) is

$$f(h) = Sh + o(h).$$

Especially, f is $\mathcal{O}(h)$. Define $o(1)$ similarly to $o(h)$ except that for every $d \in \mathbb{R}^+$ there exists some $\varepsilon_d \in \mathbb{R}^+$ such that for all suitable h,

$$|h| \leq \varepsilon_d \implies |f(h)| \leq d.$$

Thus $\mathcal{O}(h) \subset o(1)$. The $o(1)$ condition captures continuity in local coordinates: a function $f : U \longrightarrow \mathbb{R}^m$ that takes 0 to 0 is $o(1)$ if and only if it is continuous at 0. The vector space properties of $o(1)$ give the linearity of continuity. Now, let f and g take 0 to 0 and be differentiable at 0, and let k be any real number. The normalized characterizing property instantly shows that:

- f is continuous at 0 because $\mathcal{O}(h) \subset o(1)$.
- $f'(0)$ is unique because the only linear $o(h)$-function is zero.
- $f+g$ and kf are differentiable at 0 with derivatives $f'(0)+g'(0)$ and $kf'(0)$ because $o(h)$ is a vector space.

These results follow quickly from any reasonable definition of the derivative, but they come especially gracefully from the characterization. And now the main point of this note, that

- $g \circ f$ is differentiable at 0 with derivative $g'(0)f'(0)$ because $o(h)$ is a two-sided ideal of $\mathcal{O}(h)$,

gives a concrete sense of the relative difficulty of the chain rule in comparison to the other results. Incidentally, once one has verified that the chain rule reduces to the local coordinates case, its proof is yet tidier:

$$g(f(h)) = g(Sh + o(h)) = TSh + T(o(h)) + o(Sh + o(h)),$$

and the right side is $TSh + o(h)$ as before.

To discuss the product rule of one-variable calculus, let our variables and functions be scalar-valued. The usual proof relies on a little trick of inserting two terms that add to 0, and then one needs to have in place—or to stop and establish—that differentiability implies continuity. This proof can give students the impression that mathematical argument is esoteric and fragile. To prove the product rule with the characterizing property instead, let f and g be differentiable at a, and let f_o and g_o be the corresponding normalized functions. A mechanical calculation, using the characterizing property and the equalities $f'(a) = f_o'(0)$ and $g'(a) = g_o'(0)$, shows that

$$(f \cdot g)(a+h) - (f \cdot g)(a) - \big(f(a)g'(a) + f'(a)g(a)\big)h = (f_o \cdot g_o)(h) + o(h).$$

Thus the product rule reduces to showing that $(f_o \cdot g_o)(h)$ is $o(h)$. Since differentiable germs are $\mathcal{O}(h)$ in local coordinates, it suffices to show that the product of two $\mathcal{O}(h)$-germs is $o(h)$, and this follows from Definition 1.1. Alternatively, we can write

$$f_o(h)g_o(h) = \big(sh + o(h)\big)\big(th + o(h)\big) = sth^2 + sh\,o(h) + th\,o(h) + o(h)\,o(h),$$

and each term is $o(h)$, so that their sum is $o(h)$ as well. The characterizing property also works in the original coordinates, but then the clutter until many terms inevitably cancel obscures the main point that the product of differentiable germs is small. Realizing that the local result $(f_o \cdot g_o)'(0) = 0$ (with no assumption that $f_o'(0) = 0$ or $g_o'(0) = 0$) gives the full product rule clarifies that the usual proof intermixes normalization and analysis. A standard picture in this context shows that the incremental area-growth of an $(f \cdot g)$-rectangle comes mostly from the incremental translation of two sides, but also there is a new corner-area that is an order of magnitude smaller and hence insignificant in the limit.

As another example, the proof of the so-called first fundamental theorem of calculus is far less cluttered: if f is continuous on an interval then

$$\int_a^{x+h} f - \int_a^x f - f(x)h = \int_x^{x+h} (f - f(x)) = \int_x^{x+h} o(1) = o(h),$$

and so the derivative of $\int_a^x f$ is $f(x)$.

Finally, I believe that already in a one-variable calculus course, the characterizing property is how to define the derivative. To discuss what it means for the graph of a function to have tangent slope s at a point, we should work in local coordinates, and then we should subtract sh as is done to reduce the Mean Value Theorem to Rolle's Theorem. So the question is what it means for the graph of a function that takes 0 to 0 to be horizontal at the origin. A natural answer is that for any positive real number c, however small, the region between the lines of slope $\pm c$ contains the graph of f close enough to the origin. That is, f is $o(h)$. Surely the geometric intuition here is at least as clear as defining the tangent slope as the limit of secant slopes. With the derivative characterized, introducing the limit as a computing mechanism costs little since their equivalence is immediate. Then one can establish derivative properties with the characterization and compute derivatives with the limit. In hindsight, my attempts as a calculus student long ago to think clearly about the derivative amounted to trying to discern these issues.

References

[1] Tom M. Apostol, *Calculus, Volume II*, second edition, John Wiley and Sons, 1969.

[2] N. Bourbaki, *Variétés differentielles et analytique*, Hermann, 1967

[3] R. Creighton Buck, *Advanced Calculus*, third edition, McGraw-Hill, 1978

[4] Henri Cartan, *Differential Calculus*, Hermann/Houghton Mifflin, 1971

[5] John H. Hubbard and Barbara Burke Hubbard, *Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach*, third edition, Matrix Editions, 2007

[6] Edmund Landau, *Differential and Integral Calculus* (translated by Melvin Hausner and Martin Davis), Chelsea, 1951

[7] Lynn H. Loomis and Shlomo Sternberg, *Advanced Calculus*, Addison-Wesley, 1968

[8] Jerrold E. Marsden and Anthony J. Tromba, *Vector Calculus*, second edition, W. H. Freeman and Company, 1981

[9] Edward Nelson, *Topics in Dynamics I: Flows* (Mathematical Notes series), Princeton University Press and Tokyo University Press, 1969

[10] Walter Rudin, *Principles of Mathematical Analysis*, third edition, McGraw-Hill, 1976

[11] Michael Spivak, *Calculus on Manifolds*, W. A. Benjamin, Inc., 1965