ALGEBRAIC STRUCTURE AND GENERATIVE DIFFERENTIATION RULES

Introduction

The thesis of this writeup is that

Algebraic structure clarifies the basic properties of the derivative.

Here are two examples of arguments that warrant being clarified.

- Proving the product rule for derivatives relies on the little trick of inserting two terms that add to 0,

$$(f \cdot g)(a+h) - (f \cdot g)(a) = f(a+h)g(a+h) - f(a)g(a+h) \\ + f(a)g(a+h) - f(a)g(a),$$

  and then one needs to have in place—or to stop and establish—the fact that differentiability implies continuity. Proving the product rule this way can teach students that mathematics is esoteric and fragile, a paralyzing idea in the long run.
- Similarly but worse, proving the chain rule inevitably runs into problems. The corresponding trick of inserting two terms that multiply to 1 gives something along the lines of

$$\frac{\Delta z}{\Delta x} = \frac{\Delta z}{\Delta y} \cdot \frac{\Delta y}{\Delta x},$$

  but a complete argument must handle circumstances where $\Delta y = 0$ in some other way while being sure to cover all possibilities. The proof that one sees in a calculus course tends to be incomplete, or cluttered and unpersuasive. It can leave even strong students with the impression that mathematics quickly becomes confusingly complicated, so that the best thing is to place one's faith in authority figures and rituals.

One idea of this writeup is that to establish *generative* rules for differentiation, i.e., rules that produce new derivatives from given derivatives rather than establish any particular derivative from scratch, the traditional definition of the derivative,

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

is an awkward tool: it is a computational formula that says what the derivative *is* but does not say what the derivative *does*. By contrast, the derivative is uniquely determined by a *characterizing property*,

$$f(a+h) = f(a) + f'(a)h + \mathrm{o}(h) \qquad \text{(notation to be explained below)},$$

which precisely describes how it behaves. This writeup will argue that the characterizing property is the right tool for proving generative derivative rules. Indeed, the proofs of the product rule and the chain rule become purely mechanical, with no tricks to remember and with no subtleties. This writeup will also argue that the

characterizing property of the derivative is actually closer to our intuition than the usual limit definition.

Along with the utility of working with a characterizing property, the second idea running through this writeup brings us back to its thesis. The characterizing property of the derivative uses Landau's so-called *little-oh notation*, and once the Landau notation is in place, its properties are *assertions about algebraic structure*. Specifically the structure-type in question is called an *algebra*, a sort of hybrid of a vector space and a ring. We will see that after some normalizing reductions, generative differentiation results such as the product rule and the chain rule do little more than rephrase the structural assertions.

A person might reasonably wonder why these ideas are not commonplace if they are so handy. One answer is that the typical calculus course can not assume that the students are comfortable with algebraic structure, while a typical first semester algebra course does not demonstrate its concepts in a serious environment such as calculus. A second answer is that in fact everybody does know these ideas, provided that we interpret *everybody* and *knows* appropriately: any graduate-level course in manifold theory or algebraic geometry will introduce many of the ideas here, and then revisiting calculus in light of the ideas is a routine exercise. However, since there is no mathematical cachét in revisiting ideas for the sake of clarifying them, the improved understanding that the revised viewpoint gives may go unnoticed.

Section 1 below explains that while the functions defined about 0 form only a set, a natural equivalence relation on the functions gives the resulting set of equivalence classes the structure of a vector space: the equivalence classes can be added and scaled. These equivalence classes are called *germs* of functions because a class describes its constituent functions only locally.

Section 2 introduces the Landau notation, whose simplest cases already capture familiar ideas of boundedness and continuity. Section 3 establishes basic properties of the Landau notation as assertions about algebraic structure.

Section 4 redefines the derivative by its characterizing property and then shows that the characterizing property is equivalent to the usual definition of the derivative as a limit. Section 5 establishes basic differentiation rules using the characterizing property rather than the limit definition. Finally, sections 6 and 7 give the incisive proofs of the product rule and the chain rule. Each proof reduces its rule to a normalized case and then proves the normalized case with the characterizing property.

1. A Vector Space Whose Elements Are Not Quite Functions

Consider any real number $a \in \mathbb{R}$. A *neighborhood of* a is an open set in $\mathbb{R}$ containing a. We want to study *local* properties of real-valued functions defined about a, meaning that any such function is defined on some neighborhood of a, but the properties don't depend on the function-behavior on any particular neighborhood, even though they do depend on more than just the function-value at a itself.

If the function $f : U_a \longrightarrow \mathbb{R}$ is defined on a neighborhood of a then its translate

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

is defined on the neighborhood U of 0, and the local properties of f_o at 0 are precisely the local properties of f at a. Thus from now on we freely take $a = 0$.

That is, the set of functions that we care about is

$$\mathcal{F} = \{f : U_f \longrightarrow \mathbb{R} \mid U_f \text{ is a neighborhood of } 0\}.$$

As algebraists we are interested not in sets but in structures. The formulas that should give $\mathcal{F}$ a vector space structure are obvious,

$$(f+g)(x) = f(x) + g(x),$$
$$(kf)(x) = k\,f(x).$$

However, the first formula is problematic because two functions $f, g \in \mathcal{F}$ need not have the same domain. Their sum $f+g$ is defined only on $U_f \cap U_g$. On the bright side, $U_f \cap U_g$ is still a neighborhood of 0.

Although we could redefine $\mathcal{F}$ to require that all functions are defined on a common neighborhood $U = U_{\mathcal{F}}$ of 0, this idea misses the goal of discussing function-properties that are local at 0. We could work with pairs (f, U) and rules such as $(f, U) + (g, V) = (f + g, U \cap V)$, but expressly dragging the domains around and manipulating them would be cumbersome. Still, shrinking function-domains whenever desirable or necessary is indeed the right idea, and so we would like machinery that does so automatically and tacitly. Thus, define an equivalence relation on $\mathcal{F}$,

$$g \sim f \quad \text{if} \quad g = f \text{ on some neighborhood } U_{f,g} \text{ of } 0.$$

An equivalence class is called a *germ*. The germ of a function f specifies more than $f(0)$ and yet does not specify f at any point other than 0, much less on any neighborhood of 0. The germ of a function f is denoted $[f]$, but we will simply let representatives denote equivalence classes.

The vector space formulas from above are sensible at the level of germs. That is, if $\tilde{f} \sim f$ and $\tilde{g} \sim g$, so that $\tilde{f} = f$ on some $U_{f,\tilde{f}}$ and $\tilde{g} = g$ on some $U_{g,\tilde{g}}$, then also $\tilde{f} + \tilde{g} = f + g$ on $U_{f,\tilde{f}} \cap U_{g,\tilde{g}}$ and $k\tilde{f} = kf$ on $U_{f,\tilde{f}}$, and so $\tilde{f} + \tilde{g} \sim f + g$ and $k\tilde{f} \sim kf$. Thus we now have an algebraic structure rather than a set, the vector space of germs of functions at 0,

$$(\mathcal{G}, +, \cdot) \quad \text{where } \mathcal{G} = \mathcal{F}/\sim.$$

2. Landau Notation

We will study substructures of $\mathcal{G}$ by first describing subsets of $\mathcal{F}$, i.e., by describing particular types of functions defined about 0. This section introduces the relevant notation. The single definition to follow captures several standard ideas and in fact improves upon them.

Definition 2.1 ($\mathcal{O}(h^\alpha)$ and $\mathrm{o}(h^\alpha)$)**.** *Let $\alpha \in \mathbb{R}_{\geq 0}$ be a nonnegative real number. Consider a function $f : U \longrightarrow \mathbb{R}$ where U is a neighborhood of 0. Then:*

- *f is an $\mathcal{O}(h^\alpha)$-function if there exist $c, \delta \in \mathbb{R}^+$ such that for all $h \in U$,*
$$|h| \leq \delta \implies |f(h)| \leq c|h|^\alpha.$$
The set of $\mathcal{O}(h^\alpha)$-functions is itself denoted $\mathcal{O}(h^\alpha)$.
- *f is an $\mathrm{o}(h^\alpha)$-function if for every $d \in \mathbb{R}^+$ there exists some $\varepsilon_d \in \mathbb{R}^+$ such that for all $h \in U$,*
$$|h| \leq \epsilon_d \implies |f(h)| \leq d|h|^\alpha.$$
The set of $\mathrm{o}(h^\alpha)$-functions is itself denoted $\mathrm{o}(h^\alpha)$.

The notations $\mathcal{O}(h^\alpha)$ and $\mathrm{o}(h^\alpha)$ are a bit sloppy in that strictly speaking the quantity h^α is defined for negative h only if the nonnegative real number α is an integer. Writing $\mathcal{O}(|h|^\alpha)$ and $\mathrm{o}(|h|^\alpha)$ would be better, but the absolute value signs can be tacitly understood, and in any case our primary concern here is with $\alpha = 0$ and $\alpha = 1$.

For $\alpha = 0$, the definition of an $\mathcal{O}(1)$-function says that *for some* envelope determined by the lines of height $\pm c$ where $c \in \mathbb{R}^+$, the graph of the function lies inside the envelope if we look in a narrow enough vertical strip. That is, an $\mathcal{O}(1)$-function is *bounded near the origin*. For example, any constant function is an $\mathcal{O}(1)$-function. An $\mathcal{O}(1)$-function need not take 0 to 0, but this situation is exceptional: every $\mathcal{O}(h^\alpha)$-function where $\alpha > 0$ and every $\mathrm{o}(h^\alpha)$-function where $\alpha \geq 0$ does take 0 to 0.

Again for $\alpha = 0$, the definition of an $\mathrm{o}(1)$-function says that *given any* envelope determined by the lines of height $\pm d$ where $d \in \mathbb{R}^+$, the graph of the function lies inside the envelope if we look in a narrow enough vertical strip. The required narrowness can change if the envelope changes. For example, the function $f(h) = h$ is an $\mathrm{o}(1)$-function (with $\varepsilon_d = d$ for any $d \in \mathbb{R}^+$), but the only constant $\mathrm{o}(1)$-function is the zero function. We will see below that the $\mathrm{o}(1)$-functions are precisely the functions that take 0 to 0 and are continuous at 0.

For $\alpha = 1$, the definition of an $\mathcal{O}(h)$-function says that for some envelope determined by the lines of *slope* $\pm c$ where $c \in \mathbb{R}^+$, the graph of the function lies inside the envelope if we look in a narrow enough vertical strip. For example, any function $f(h) = kh$ where $k \in \mathbb{R}$ is an $\mathcal{O}(h)$-function. The function $f(h) = |h|^{1/2}$ is an $\mathrm{o}(1)$-function but not an $\mathcal{O}(h)$-function.

Again for $\alpha = 1$, the definition of an $\mathrm{o}(h)$-function says that given any envelope determined by the lines of slope $\pm d$ where $d \in \mathbb{R}^+$, the graph of the function lies inside the envelope if we look in a narrow enough vertical strip. The required narrowness can change if the envelope changes. For example, the function $f(h) = h^2$ is an $\mathrm{o}(h)$-function (with $\varepsilon_d = d$ for any $d \in \mathbb{R}^+$), but the function $f(h) = kh$ where $k \in \mathbb{R}$ is an $\mathrm{o}(h)$-function only for $k = 0$.

The containment relations among the functions of sub-linear decay, of linear decay, and of sub-constant decay are

$$\mathrm{o}(h) \subset \mathcal{O}(h) \subset \mathrm{o}(1),$$

and more generally,

$$\mathrm{o}(h^\alpha) \subset \mathcal{O}(h^\alpha) \subset \mathrm{o}(h^\beta) \qquad (\alpha > \beta \geq 0).$$

The reader may be wondering about the weak inequalities such as $|f(h)| \leq c|h|^\alpha$ in Definition 2.1, rather than the corresponding strict inequalities that would look more similar to epsilon-delta type inequalities from calculus. For $\mathcal{O}(1)$-functions there is no difference: certainly the strict inequality implies the weak one, and if the weak inequality holds for c then the strict one holds for $2c$. There is no difference for $\mathrm{o}(1)$-functions either. But for $\alpha > 0$ the weak inequality is essential to the definition of $\mathcal{O}(h^\alpha)$-functions and of $\mathrm{o}(h^\alpha)$-functions, since for $h = 0$ the strict inequality is impossible while the weak inequality says that $f(0) = 0$. On the other hand, the inequalities in the usual epsilon-delta definitions from calculus can be weak rather than strict with no effect on the outcome. We will freely use this fact without comment from now on.

We end this section with the promised characterization of o(1)-functions.

Proposition 2.2 (Sub-Constant Decay and Continuity)**.** *Consider a function* $f : U \longrightarrow \mathbb{R}$ *where* U *is a neighborhood of* 0*. Then* f *is an* o(1)*-function if and only if* $f(0) = 0$ *and* f *is continuous at* 0*.*

The proposition is clear geometrically in light of the interpretation of an o(1)-function as having sub-constant decay. The issues in play are symbolic. In many calculus courses the definition of continuity at a point involves verifying three conditions: the function takes a value at the point, the function has a limit as its inputs approach the point, and the limit equals the value. The proposition says that the o(1)-definition encodes all three conditions.

Proof. Suppose that $f : U \longrightarrow \mathbb{R}$ satisifes $\lim_{h\to 0} f(h) = 0$. Then given any $d \in \mathbb{R}^+$ we have a corresponding $\varepsilon_d \in \mathbb{R}^+$ such that for all $h \in U$,

$$0 < |h| \leq \varepsilon_d \implies |f(h)| \leq d.$$

If also $f(0) = 0$ then in fact given d we have ε_d such that for all h,

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

i.e., f is an o(1)-function. For the converse, the previous inequality implies the one before it, and $f(0) = 0$ for an o(1)-function f because $|f(0)| \leq d$ for any $d \in \mathbb{R}^+$. □

3. Landau Notation and Algebraic Structure

For any $\alpha \geq 0$, the property of being an $\mathcal{O}(h^\alpha)$-function is local, and so it holds across $\sim$-equivalence classes in the set $\mathcal{F}$ from section 1. That is, either all functions in an equivalence class are $\mathcal{O}(h^\alpha)$-functions or none of them is. The same result holds with $\mathrm{o}(h^\alpha)$ in place of $\mathcal{O}(h^\alpha)$. To rephrase, the $\mathcal{O}(h^\alpha)$ property and the $\mathrm{o}(h^\alpha)$ property both descend to germs.

For any $\alpha \geq 0$, if $f : U \longrightarrow \mathbb{R}$ and $g : V \longrightarrow \mathbb{R}$ are $\mathcal{O}(h^\alpha)$-functions then so again is $f + g : U \cap V \longrightarrow \mathbb{R}$ and so is $kf : U \longrightarrow \mathbb{R}$ for any $k \in \mathbb{R}$; if f and g are $\mathrm{o}(h^\alpha)$-functions then so again are $f + g$ and kf for any k. That is, working at the level of germs rather than functions to take into account shrinking function-domains as necessary,

$$\mathcal{O}(h^\alpha)/\sim \quad \text{and} \quad \mathrm{o}(h^\alpha)/\sim \quad \text{form vector spaces for all } \alpha \geq 0.$$

Let $f : U \longrightarrow \mathbb{R}$ and $g : V \longrightarrow \mathbb{R}$ be o(1)-functions. For some $d \in \mathbb{R}^+$ we have $[-d, d] \in V$ and thus the composition $g \circ f$ is defined on $(-\varepsilon_{f,d}, \varepsilon_{f,d}) \cap U$. Composition of o(1)-functions clearly descends to germs. Since

$$\mathrm{o}(h^\alpha) \subset \mathcal{O}(h^\alpha) \subset \mathrm{o}(1) \qquad (\alpha > 0),$$

the results of this paragraph apply to $\mathrm{o}(h^\alpha)$-functions for all $\alpha \geq 0$ and to $\mathcal{O}(h^\alpha)$-functions for all $\alpha > 0$. The one exception is that composition is not generally sensible in $\mathcal{O}(1)$.

However, the fact that a composition of two functions of some given type is defined does not imply that it again has the same type. Indeed, one finds that if $f, g \in \mathcal{O}(h^\alpha)$ where $\alpha > 0$ then their composition, after suitably shrinking the domain of f, need lie only in $\mathcal{O}(h^{\alpha^2})$. Loosely the calculation is that for small enough h,

$$|(g \circ f)(h)| = |g(f(h))| \leq c_g |f(h)|^\alpha \leq c_g (c_f |h|^\alpha)^\alpha = c_g c_f^\alpha |h|^{\alpha^2}.$$

Thus $\mathcal{O}(h^\alpha)$ is closed under composition only when $\mathcal{O}(h^{\alpha^2}) \subset \mathcal{O}(h^\alpha)$, i.e., only when $\alpha^2 \geq \alpha$, i.e., only when $\alpha \geq 1$. A similar argument holds for $\mathrm{o}(h^\alpha)$, except that now also $\alpha = 0$ is allowed, so that $\mathrm{o}(1)$ is closed under composition along with $\mathrm{o}(h^\alpha)$ for $\alpha \geq 1$.

Thus the vector spaces $\mathcal{O}(h^\alpha)/\sim$ and $\mathrm{o}(h^\alpha)/\sim$ for $\alpha \geq 1$, along with the vector space $\mathrm{o}(1)/\sim$, carry the additional operation of composition. That is, these vector spaces also form rings, the rings being noncommutative and only $\mathcal{O}(h)/\sim$ and $\mathrm{o}(1)/\sim$ containing a multiplicative identity. As structures that carry addition, scalar multiplication, and multiplication, $\mathcal{O}(h^\alpha)/\sim$ and $\mathrm{o}(h^\alpha)/\sim$ for $\alpha \geq 1$, and $\mathrm{o}(1)/\sim$, form algebras.

The following algebraic result will turn out to be the crux of the chain rule for derivatives.

Proposition 3.1. $\mathrm{o}(h)/\sim$ *is a two-sided ideal of* $\mathcal{O}(h)/\sim$.

Proof. Let $f : U \longrightarrow \mathbb{R}$ be an $\mathcal{O}(h)$-function, and let $g : V \longrightarrow \mathbb{R}$ be an $\mathrm{o}(h)$-function. Thus we have c and δ, and for any $d \in \mathbb{R}^+$ we have ε_d. In what follows, shrink the domain of f or the domain of g as necessary for the relevant composition to be defined.

First we show that $g \circ f$ is an $\mathrm{o}(h)$-function. The idea is that for small enough h, also $|f(h)| \leq c|h|$ is small, and so $|g(f(h))| \leq \tilde{d}|f(h)| \leq \tilde{d}c|h|$ for any $\tilde{d} \in \mathbb{R}^+$, and we can make $\tilde{d}c$ as small as we want by suitable choice of $\tilde{d}$. For the quantitative argument, let $d \in \mathbb{R}^+$ be given. Define

$$\tilde{d} = d/c \quad \text{and} \quad \rho_d = \min\{\varepsilon_{\tilde{d}}/c, \delta\}.$$

Then for all $h \in U$,

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

Second we show that $f \circ g$ is an $\mathrm{o}(h)$-function. The idea is that for any $\tilde{d} \in \mathbb{R}^+$, for small enough h, also $|g(h)| \leq \tilde{d}|h|$ is small, and so $|f(g(h))| \leq c|g(h)| \leq c\tilde{d}|h|$, and we can make $c\tilde{d}$ as small as we want by suitable choice of $\tilde{d}$. For the quantitative argument, let $d \in \mathbb{R}^+$ be given. Define

$$\tilde{d} = d/c \quad \text{and} \quad \rho_d = \min\{\varepsilon_{\tilde{d}}, \delta/\tilde{d}\}.$$

Then for all $h \in U$,

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

□

The proof just given is the most quantifier-intensive work in this body of ideas. The reader who wants to practice this sort of thing could similarly quantify the earlier informal arguments that the composition of $\mathcal{O}(h^\alpha)$-germs is $\mathcal{O}(h^{\alpha^2})$ for $\alpha > 0$ and similarly for $\mathrm{o}(h^\alpha)$-germs and $\alpha \geq 0$. However, in a calculus course, everything

could be done much more digestibly by not quantifying *for small enough* h and by downplaying the need to work at the level of germs rather than functions.

From now on, a symbol-string such as $\mathcal{O}(h)$ can denote either a set of functions or the corresponding algebra of germs, no longer expressly showing the quotient-by-$\sim$ in the notation. Furthermore, following standard usage, a symbol such as $\mathcal{O}(h)$ also can denote one particular-but-unspecified $\mathcal{O}(h)$-function. When the notation connotes a particular function, the equality sign is not symmetric. For example, the assertion $h^2 = \mathrm{o}(h)$, meaning that h^2 is an $\mathrm{o}(h)$-function, is true, but the assertion $\mathrm{o}(h) = h^2$, meaning that any $\mathrm{o}(h)$-function must be h^2, is false. For another example, the notation $\mathcal{O}(h) = \mathrm{o}(1)$, meaning that any $\mathcal{O}(h)$-function is also an $\mathrm{o}(1)$-function, is true, but the assertion $\mathrm{o}(1) = \mathcal{O}(h)$, meaning that any $\mathrm{o}(1)$-function is also an $\mathcal{O}(h)$-function, is false.

4. Differentiability in Terms of a Characterizing Property

Let a be a real number. Let $f : U_a \longrightarrow \mathbb{R}$ be a function defined on a neighborhood of a. Usually the condition that f is differentiable at a with derivative $f'(a) = r$ is taken to be

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

We claim that an equivalent condition is the characterizing property

$$\boxed{f(a+h) = f(a) + rh + \mathrm{o}(h)}.$$

Both the usual definition and the characterizing property are obviously sensible at the level of germs.

Although a is arbitrary, letting $U = U_a - a$ (a neighborhood of 0) and replacing f by

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

quickly normalizes the situation to $a = 0$ and $f(0) = 0$, i.e., to germs that take 0 to 0. That is (exercise),

> *f is differentiable at a with derivative $f'(a) = r$ if and only if f_o is differentiable at 0 with the same derivative $f_o'(0) = r$.*

For the remainder of this section and for the next section, we freely make the normalizing assumptions. The general arguments are no different other than a few more terms cluttering up the formulas.

Normalized, the two properties that we claim to be equivalent are the *limit of difference-quotients* definition,

$$\lim_{h\to 0} \frac{f(h)}{h} = r,$$

and the characterizing property,

$$f(h) = rh + \mathrm{o}(h).$$

The characterizing property says that the line rh approximates $f(h)$ sub-linearly for small h, and we will see easily in the next section that such a close approximation can hold for only one value of r. That is, the property characterizes r as *the slope of an existing, unique sub-linearly-approximating line to the graph of f at the origin.* In other words, r satisfies what is obviously the natural definition of the tangent slope. By contrast, the usual calculus-class notion of the tangent slope as the limit

of secant slopes is artificial as a *definition*—making it the definition only short-circuits any discussion of why it is giving the right thing—although it is useful as a computing *mechanism*.

To see that the two conditions defining the derivative are equivalent, rewrite the first condition as

$$\lim_{h\to 0} \frac{f(h) - rh}{h} = 0.$$

This condition is that given any $d \in \mathbb{R}^+$ there exists some $\varepsilon_d \in \mathbb{R}^+$ such that for all $h \in U$,

$$0 < |h| \leq \varepsilon_d \implies \left| \frac{f(h) - rh}{h} \right| \leq d.$$

An equivalent condition is that for all $h \in U$,

$$0 < |h| \leq \varepsilon_d \implies |f(h) - rh| \leq d|h|.$$

However, the inequality in the right side of the previous display holds for $h = 0$ as well. Thus the condition need not exclude $h = 0$: equivalently, for all $h \in U$,

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

This last condition is precisely that $f(h) - rh = \mathrm{o}(h)$.

In avoiding division, and thus avoiding the need to exclude $h = 0$, the $\mathrm{o}(h)$-characterization of the derivative is tidier than the definition of the derivative as a limit. The tidy $\mathrm{o}(h)$-characterization of the derivative is what allows the pending proofs of the product rule and the chain rule to be neater than is possible with the usual set-up. But first, as a warm-up, we use the $\mathrm{o}(h)$-characterization of the derivative to establish some basic rules.

5. Basic Differentiation Rules Via the Characterizing Property

Let f be a germ that takes 0 to 0, and suppose that f is differentiable at 0 with derivative $f'(0) = r$, i.e.,

$$f(h) = rh + \mathrm{o}(h).$$

- We show that f is continuous at 0. By Proposition 2.2, since $f(0) = 0$ the assertion rephrases as $f \in \mathrm{o}(1)$. The result is immediate thanks to the containments $\mathrm{o}(h) \subset \mathcal{O}(h) \subset \mathrm{o}(1)$ and the closure of $\mathrm{o}(1)$ under addition at the level of germs,

$$f(h) = rh + \mathrm{o}(h) = \mathcal{O}(h) + \mathrm{o}(h) = \mathrm{o}(1).$$

- We show that the derivative of f at 0 is unique. The result is immediate because the only function kh in $\mathrm{o}(h)$ is the zero function. Specifically, if

$$f(h) = rh + \mathrm{o}(h) \quad \text{and} \quad f(h) = r'h + \mathrm{o}(h),$$

  then subtracting the two conditions gives $(r - r')h = \mathrm{o}(h)$, and thus $r' = r$.
- Keeping f as above, suppose also that similarly g is a germ that takes 0 to 0 and is differentiable at 0 with derivative $g'(0) = s$. We show that

$$f + g \quad \text{is differentiable at 0 with derivative } (f+g)'(0) = r + s.$$

  The issue is simply that $\mathrm{o}(h)$ is closed under addition at the level of germs,

$$(f+g)(h) = f(h) + g(h) = rh + \mathrm{o}(h) + sh + \mathrm{o}(h) = (r+s)h + \mathrm{o}(h).$$

- Also, take any constant $k \in \mathbb{R}$. We show that that

  $kf : U \longrightarrow \mathbb{R}$ is differentiable at 0 with derivative $(kf)'(0) = kr$.

  The issue is simply that $\mathrm{o}(h)$ is closed under scalar multiplication at the level of germs,

$$(kf)(h) = k(f(h)) = k(rh + \mathrm{o}(h)) = krh + \mathrm{o}(h).$$

6. The Product Rule Via the Characterizing Property

The product rule,

$$(fg)'(a) = f(a)g'(a) + f'(a)g(a),$$

refers to the values $f(a)$ and $g(a)$ along with the derivatives $f'(a)$ and $g'(a)$. We are assuming that $f, g : U_a \longrightarrow \mathbb{R}$ are differentiable at a. We begin by normalizing the situation to $a = 0$, $f(0) = g(0) = 0$.

As usual, define $U = U_a - a$ and define

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

so that $f_o(0) = 0$. We have seen that $f_o'(0)$ exists and equals $f'(a)$. An identical discussion holds with g in place of f. For any point $h \in U$, compute (using the fact that $g'(a) = g_o'(0)$ and $f'(a) = f_o'(0)$ for the last equality)

$$\begin{aligned}
&(f \cdot g)(a + h) - (f \cdot g)(a) - \big(f(a)g'(a) + f'(a)g(a)\big)h \\
&\qquad = \big(f_o(h) + f(a)\big)\big(g_o(h) + g(a)\big) - f(a)g(a) - f(a)g'(a)h - f'(a)g(a)h \\
&\qquad = (f_o \cdot g_o)(h) + f(a)\big(g_o(h) - g'(a)h\big) + g(a)\big(f_o(h) - f'(a)h\big) \\
&\qquad = (f_o \cdot g_o)(h) + \mathrm{o}(h).
\end{aligned}$$

Thus the product rule reduces to showing that $(f_o \cdot g_o)(h) = \mathrm{o}(h)$. To rephrase, we need to show that the product of differentiable germs that take 0 to 0 is $\mathrm{o}(h)$. The argument is the first payoff of this writeup:

Given germs f and g that take 0 to 0 such that for some r and s,

$$f(h) = rh + \mathrm{o}(h) \quad \text{and} \quad g(h) = sh + \mathrm{o}(h),$$

we want to show that

$$(f \cdot g)(h) = \mathrm{o}(h).$$

Compute that

$$f(h)g(h) = (rh + \mathrm{o}(h))(sh + \mathrm{o}(h)) = rsh^2 + rh\,\mathrm{o}(h) + sh\,\mathrm{o}(h) + \mathrm{o}(h)\,\mathrm{o}(h).$$

Each term is readily seen to be $\mathrm{o}(h)$, and hence so is their sum.

Of course 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 scalar-valued 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 normalizing and analysis.

7. The Chain Rule Via the Characterizing Property

We have $f : U_a \longrightarrow \mathbb{R}$ differentiable at a and $g : V_{f(a)} \longrightarrow \mathbb{R}$ differentiable at $f(a)$. Let $r = f'(a)$ and let $s = g'(f(a))$. Thus we have the conditions

$$f(a+h) = f(a) + rh + \mathrm{o}(h),$$
$$g(f(a)+k) = g(f(a)) + sk + \mathrm{o}(k).$$

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 sr, compute that

$$\begin{aligned} g(f(a+h)) &= g(f(a) + rh + \mathrm{o}(h)) && \text{by the first condition} \\ &= g(f(a)) + srh + s\,\mathrm{o}(h) + \mathrm{o}(rh + \mathrm{o}(h)) && \text{by the second.} \end{aligned}$$

But $s\,\mathrm{o}(h) = \mathcal{O}(\mathrm{o}(h))$ and $\mathrm{o}(rh + \mathrm{o}(h)) = \mathrm{o}(\mathcal{O}(h))$, so we have

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

The ideal properties $\mathcal{O}(\mathrm{o}(h)) = \mathrm{o}(h)$ and $\mathrm{o}(\mathcal{O}(h)) = \mathrm{o}(h)$ complete the proof,

$$(g \circ f)(a+h) = (g \circ f)(a) + srh + \mathrm{o}(h).$$

Closing Comments

Admittedly, a graceful proof of the one-variable chain rule is underwhelming if it requires so much environmental set-up. Still, the environment is worthwhile in its own right because the Landau notation is ubiquitous in mathematics, and the environment adds substance to an algebra class because the Landau notation specifies non-toy algebraic structures. But also, the ideas in this writeup scale essentially verbatim to multivariable calculus (exercise), where they are essential. The multivariable derivative must be defined by a characterizing property since no difference-quotient definition is possible. And the multivariable chain rule for derivatives is complicated in coordinates, so that its neat intrinsic form—identical symbolically to the one-variable case—can not be intuitive at the level of coordinate-dependent formulas.

Multivariable versions of the chain rule proof just given are found in many texts, but seldom phrased entirely in the Landau notation. Instead, the exposition may rely on two preparatory results of different flavors and then carry out a two-part calculation with estimating bounds. The process, seemingly piecemeal and heterogeneous, long frustrated me as a teacher. It leaves even strong students with the impression that the chain rule is complicated and that mathematics sprawls. With the Landau notation, the two preparatory results phrase uniformly as the facts that $rh = \mathcal{O}(h)$ and that a differentiable function is $\mathcal{O}(h)$ in local coordinates, and the two-part calculation is simply the algebraic fact that $\mathrm{o}(h)$ is a two-sided ideal of $\mathcal{O}(h)$. The gain in clarity and brevity for little startup cost feels real to me. With these ideas in mind, it is striking to glance over versions of the proof, sometimes several pages long, and see the underlying simplicity.