

Calculus

ROSS GROGAN-KAYLOR

Differential calculus

Foundations of calculus: limits and continuity

What is calculus about?

Calculus is about *measuring change* by thinking about the infinitely small. All of calculus relies on intuition about “infinitely small numbers”, or *infinitesimals*.

In *differential calculus*, the concept of infinitesimals is invoked in order to measure the speed at which change occurs. If we want to measure the speed at which the output of a function f is changing in response to a changing input, we can consider the rate of change of f at the input x to be the slope of the line that the curve f vs. x graph “becomes” as we “zoom in infinitely close” to the point $(x, f(x))$.

In *integral calculus*, we use infinitesimals to compute how much total change is caused by a given rate of change. If we want to determine how much total change has been caused by a rate of change that itself changes over time, we can think of the rate of change as being the aggregation of many rates of change that *don't* change over time, where each of these rates occurs over an “infinitely small” window of input values, and then add up all of the little changes that result.

Limits

The concept of “infinitely small numbers” is most often¹ formalized with the concept of *limits*.

If, as inputs to a function f that are *greater* than a particular input p get closer and closer to p , the outputs of f get closer and closer to a number L , then we say “the limit of f as x approaches p from the right is L ”, and write

$$\lim_{x \rightarrow p^+} f(x) = L.$$

The $+$ sign on the “ p^+ ” below the limit indicates that the limit is taken “from the right”; i.e., that the inputs that are getting closer to p are all greater than p .

Similarly, if, as inputs of f that are *less* than p get closer and closer to p , the outputs of f get closer and closer to a number M , then we say “the limit of f as x approaches p from the left is M ”, and write

$$\lim_{x \rightarrow p^-} f(x) = M.$$

The $-$ on the “ p^- ” below the limit indicates that the limit here is taken “from the left”; the inputs that are getting closer to p are all less than p .

If the limits of f at p from the left and from the right are the same, then we define the so-called *two-sided limit* to be equal to the left and right limits²:

$$\lim_{x \rightarrow p} f(x) := \lim_{x \rightarrow p^+} f(x) = \lim_{x \rightarrow p^-} f(x).$$

The lack of a $-$ or $+$ sign on the “ p ” below the limit indicates that the limit is two-sided.

With the above formalism set up, expressions involving an “infinitely small number” x can be reasoned about by considering limits of the form $\lim_{x \rightarrow 0} f(x)$; we can think of the x in such limits as being “infinitely close to zero”.

¹There is another approach to formalizing the concept of “infinitely small numbers” called *nonstandard analysis*. This approach is in some ways more intuitive because it literally does give a definition for “infinitely small number”. On the other hand, nonstandard analysis requires that one drop the assumption that “not false” is equivalent to “true”. Some may view this as a disadvantage. Nonstandard analysis also obscures the presence of “error terms” (terms that disappear in a limit taken to zero or infinity) that the limit approach makes obvious.

²The colon-equals sign $:=$ is used here instead of a regular equals sign $=$ to indicate that the combination of symbols on the left side is being *defined* to mean whatever is written on the right side. While some authors use the regular equals sign $=$ to denote definitions, I prefer to use the colon-equals $:=$ because I think it is important to disambiguate “equality by definition” from “equality resulting from mathematical reasoning”.

When limits don't exist

It is not always the case that a given limit exists.

A two-sided limit does not exist if its associated one-sided limits are not equal. When this is the case, and we have $\lim_{x \rightarrow p^-} f(x) \neq \lim_{x \rightarrow p^+} f(x)$, we say that f has a *jump discontinuity* at p .

Additionally, both two-sided and one-sided limits of a function f at a point p may fail to exist for the following reasons:

- It may be the case that as inputs approach p , outputs only grow more and more positive or more and more negative. In these cases, the limit is considered to be *infinite*, and we write something like “ $\lim_{x \rightarrow p} f(x) = \infty$ ” or “ $\lim_{x \rightarrow p} f(x) = -\infty$ ”.
- It may be the case that as inputs approach³ p , the outputs oscillate and do not settle around any particular number.

Relationship between infinite and zero limits

It's worth noting that the statement ($\lim_{x \rightarrow p} f(x) = -\infty$ or $\lim_{x \rightarrow p} f(x) = \infty$) is true exactly when $\lim_{x \rightarrow p} \frac{1}{f(x)} = 0$ is true.

Properties of limits

The following properties of limits that we present (but do not prove) are hopefully somewhat intuitive.

If $\lim_{x \rightarrow p} f(x)$ exists, then

$$\lim_{x \rightarrow p} (cf(x)) = c \lim_{x \rightarrow p} f(x)$$

If $\lim_{x \rightarrow p} f(x)$ and $\lim_{x \rightarrow p} g(x)$ both exist, then

$$\begin{aligned}\lim_{x \rightarrow p} (f(x) + g(x)) &= \lim_{x \rightarrow p} f(x) + \lim_{x \rightarrow p} g(x) \\ \lim_{x \rightarrow p} (f(x)g(x)) &= \left(\lim_{x \rightarrow p} f(x) \right) \left(\lim_{x \rightarrow p} g(x) \right) \\ \lim_{x \rightarrow p} \frac{f(x)}{g(x)} &= \frac{\lim_{x \rightarrow p} f(x)}{\lim_{x \rightarrow p} g(x)} \text{ when } \lim_{x \rightarrow p} g(x) \neq 0\end{aligned}$$

The properties $\lim_{x \rightarrow p} (cf(x)) = c \lim_{x \rightarrow p} f(x)$ and $\lim_{x \rightarrow p} (f(x) + g(x)) = \lim_{x \rightarrow p} f(x) + \lim_{x \rightarrow p} g(x)$ are collectively referred to as *the linearity of the limit operator*⁴. The last two properties are called the *product rule for limits* and the *quotient rules for limits*, respectively.

Limits at infinity

Another useful notion brought to us by the intuition of limits⁵ is that of a function's “limits at infinity”, or “end behavior”. The idea is that, for a function f , we write $\lim_{x \rightarrow \infty} f(x) = L$ if, as x gets larger and larger, $f(x)$ gets closer and closer to a number L . Similarly, we write $\lim_{x \rightarrow -\infty} f(x) = M$ if, as x gets lesser and lesser and veers off to negative infinity, $f(x)$ gets closer and closer to a number M .

Geometrically, a function's limits at infinity are the horizontal lines that its graph approaches but never touches, i.e., the function's limits at infinity correspond to *horizontal asymptotes*.

Using the definitions of limits at infinity presented above, it's possible to prove that

³Consider $f(x) = \sin(2\pi\omega(x)x)$, where $\omega(x) = (2\pi x^2)^{-1}$. Since $\lim_{x \rightarrow 0^+} \omega(x) = \infty$, the graph of f is a sine wave whose oscillation frequency approaches infinity as $x \rightarrow 0^+$. Thus, $f(x)$ oscillates wildly as $x \rightarrow 0^+$. This implies that $f(x)$ doesn't get closer to any particular value as $x \rightarrow 0^+$, i.e., that $\lim_{x \rightarrow 0^+} f(x)$ does not exist.

⁴The limit operator $\lim_{x \rightarrow p}$ can be considered to be a function that sends functions to real numbers. In general, if Φ is a function that accepts other functions as input, then Φ is said to be *linear* if, for all functions f, g and all real numbers c , we (1) have $\Phi(f + g) = \Phi(f) + \Phi(g)$ and (2) have $\Phi(cf) = c\Phi(f)$. The function $\lim_{x \rightarrow p}$ is linear in this sense.

It may seem a bit strange that such functions are called “linear functions”—after all, what do they have to do with lines? For the answer to this question, you will want to take a course in *linear algebra*. The short answer is: inputs to linear functions are themselves “linear elements”, and linear functions by definition “preserve the decomposition” of linear elements; linear functions are called what they are because they play nicely with linear elements.

⁵I'm hinting here that the technical mathematical definition of a limit at infinity ($\lim_{x \rightarrow \infty} f(x)$, $\lim_{x \rightarrow -\infty} f(x)$) is different than the technical mathematical definition of a limit at a point ($\lim_{x \rightarrow p} f(x)$). This is due to the fact that we have to make some formal interpretation of what we mean by $x \rightarrow \infty$ and $x \rightarrow -\infty$; this interpretation involves $|x|$ getting larger and larger, which is different than the “getting closer and closer” involved in the definition of limit at a point.

$$\lim_{x \rightarrow \infty} \frac{1}{x} = 0 \text{ and } \lim_{x \rightarrow -\infty} \frac{1}{x} = 0.$$

Hopefully this is intuitive enough: make x really really big, and $\frac{1}{x}$ becomes really really small. Note that the limit product rule implies

$$\lim_{x \rightarrow \infty} x^{-n} = 0 \text{ and } \lim_{x \rightarrow -\infty} x^{-n} = 0 \text{ for all integers } n > 0.$$

[Applying the squeeze theorem], we obtain

$$\lim_{x \rightarrow \infty} x^{-c} = 0 \text{ and } \lim_{x \rightarrow -\infty} x^{-c} = 0 \text{ for all } c > 0.$$

Limits at infinity of a rational function

These new facts provide us a way to take the limits at infinity of a “rational function”, which is a function defined by $x \mapsto \frac{P(x)}{Q(x)}$, where $P(x) = p_0 + p_1x + p_2x^2 + \dots + p_nx^n$ and $Q(x) = q_0 + q_1x + q_2x^2 + \dots + q_mx^m$ are degree n and m polynomials for some n and m , respectively. In your precalculus class, you may have had to memorize rules about how the degrees of P and Q determine these limits. But no memorization of silly rules is necessary: when you want to determine end behavior, take limits and see what happens!

To illustrate the point, let’s now examine the end behavior of $x \mapsto \frac{P(x)}{Q(x)}$ at positive infinity:

$$\lim_{x \rightarrow \infty} \frac{P(x)}{Q(x)} = \lim_{x \rightarrow \infty} \frac{p_0 + p_1x + p_2x^2 + \dots + p_nx^n}{q_0 + q_1x + q_2x^2 + \dots + q_mx^m} = \lim_{x \rightarrow \infty} \frac{\sum_{i=1}^n p_i x^i}{\sum_{j=1}^m q_j x^j}.$$

We now consider three cases. It is either the case that $n > m$, $n = m$, or $n < m$. We can actually ignore the case $n > m$, since we can obtain the result for this case by following an argument that is similar to the one used for the case $n < m$. So, we can assume that either $n < m$ or $n = m$, i.e. we can assume $n \leq m$.

At this point, we rewrite the inside of the limit:

$$\lim_{x \rightarrow \infty} \frac{\sum_{i=1}^n p_i x^i}{\sum_{j=1}^m q_j x^j} = \lim_{x \rightarrow \infty} \left(\frac{\sum_{i=1}^n p_i x^i}{\sum_{j=1}^m q_j x^j} \cdot \frac{x^{-n}}{x^{-n}} \right) = \lim_{x \rightarrow \infty} \left(\frac{\sum_{i=1}^n p_i x^{i-n}}{\sum_{j=1}^m q_j x^{j-n}} \right).$$

Now, we will *assume* that it is valid to apply the limit quotient rule and the linearity of the limit. (We will justify this assumption retroactively). The limit becomes

$$\frac{\lim_{x \rightarrow \infty} \sum_{i=1}^n p_i x^{i-n}}{\lim_{x \rightarrow \infty} \sum_{j=1}^m q_j x^{j-n}} = \frac{\sum_{i=1}^n p_i \lim_{x \rightarrow \infty} x^{i-n}}{\sum_{j=1}^m q_j \lim_{x \rightarrow \infty} x^{j-n}}$$

It is only valid to apply the limit quotient rule if the limit of the denominator in the above left expression is nonzero, and it is only valid to apply the linearity of the limit if, for all i and j , the limits in the sums of the above right expression exist or are finite. Now comes the retroactive justification- as we continue our computations, we will see that all of these conditions are satisfied⁶.

Consider the case in which the degree of the “top” polynomial is lesser than the degree of the “bottom” polynomial, i.e., the case in which $n < m$. First, we compute the limits in the numerator. Since $i \leq n$ and $n < m$, we have $i < n \iff i - n < 0$ for all i . Knowing that $i - n$ is negative for all i then gives us that $\lim_{x \rightarrow \infty} x^{i-n} = 0$ for all i , and so the numerator is $\sum_{i=1}^n p_i \lim_{x \rightarrow \infty} x^{i-n} = \sum_{i=1}^n (p_i \cdot 0) = 0$. As for the denominator, we know that for all j we have $j < m \iff j - m < 0$, and thus $\lim_{x \rightarrow \infty} x^{j-m} = 0$ for all $j < m$. Thus, the denominator is

$$\sum_{j=1}^m q_j \lim_{x \rightarrow \infty} x^{j-m} = \left(\sum_{j < m} q_j \lim_{x \rightarrow \infty} x^{j-m} \right) + q_m x^{m-m} = \sum_{j < m} (q_j \cdot 0) + q_m = q_m.$$

⁶In general, if we want to use a theorem that is stated in the form “(conditions) \implies (result)”, it is not valid to deduce that the conditions follow as a consequence of assuming the result- this is circular reasoning. It may seem that we are doing this here, but we are actually not. It just so happens that with limit laws, the computations involved in the result are the same computations that are examined in the testing of conditions. So, we can get ahead of ourselves and compute the result while simultaneously checking the calculations to make sure that the conditions are satisfied.

Overall, we see that when $n < m$ the limit of the rational function is $\lim_{x \rightarrow \infty} \frac{P(x)}{Q(x)} = \frac{0}{q_m} = 0$. This means that when $n < m$ the graph of $\frac{P}{Q}$ has the horizontal asymptote $y = 0$.

A similar argument as was used above shows that when $n > m$, the overall limit is $\lim_{x \rightarrow \infty} \frac{P(x)}{Q(x)} = \lim_{x \rightarrow \infty} \frac{R(x)}{S(x)}$, where $\lim_{x \rightarrow \infty} R(x) = p_n$ and $\lim_{x \rightarrow \infty} S(x) = 0$. From this it follows that $\lim_{x \rightarrow \infty} \frac{P(x)}{Q(x)} = \pm\infty$, where we determine whether \pm is $+$ or $-$ by checking whether $\frac{P}{Q}$ is eventually⁷ increasing or whether $\frac{P}{Q}$ is eventually decreasing as $x \rightarrow \infty$.

If P and Q have equal degrees and $n = m$, then the limit computation for the denominator is the same as the above. The limit computation for the numerator is analogous to the one for the denominator, so we end up with the overall limit $\lim_{x \rightarrow \infty} \frac{P(x)}{Q(x)} = \frac{p_n}{q_m}$. Thus, when $n = m$, the graph of $\frac{P}{Q}$ has a horizontal asymptote at $y = \frac{p_n}{q_m}$.

This completes our analysis of all cases. Don't worry too much about the details of this derivation, and certainly do not try to memorize this derivation! If you take away anything from the above, just remember that you do *not* have to memorize anything, and that you can determine the end behavior of a rational function by following a *process* if need be.

Continuity

In calculus, the objects of study are what are called *continuous functions*. Intuitively, continuous functions are ones whose graphs that are smooth and without gaps. We will use the mathematical tool of limits in order to define what it really means for a function to be continuous:

- We say that a function f is *continuous at a point* p if $\lim_{x \rightarrow p} f(x)$ exists and $\lim_{x \rightarrow p} f(x) = f(p)$.
- We say that a function f is *continuous everywhere*, or simply *continuous*, if it is continuous at every point.

It's instructive to think about how a function could fail to be continuous at a point p . There are two ways for a so-called *discontinuity* to occur: either it is the case that $\lim_{x \rightarrow p} f(x)$ does not exist (recall, this could be due to a jump discontinuity, an infinite limit, or an oscillating limit), or it is the case that $\lim_{x \rightarrow p} f(x) \neq f(p)$. When the latter occurs, f is said to have a *point discontinuity* at p . An example of a function that satisfies the latter condition is

$$g(x) = \begin{cases} x & x \neq 0 \\ 1 & x = 0 \end{cases}.$$

In this case, we have $\lim_{x \rightarrow 0} g(x) = 0 \neq 1 = g(0)$, so g has a point discontinuity at 0.

Limit of a composition

Suppose f and g are functions, that $\lim_{x \rightarrow p} f(x)$ exists, and that g is continuous at $\lim_{x \rightarrow p} f(x)$. In this situation, then we can compute $\lim_{x \rightarrow p} g(f(x))$ by “bringing the limit inside the application of g ”:

$$\lim_{x \rightarrow p} g(f(x)) = g(\lim_{x \rightarrow p} f(x)).$$

⁷Graphs of polynomials may alternate between increasing and decreasing for a little while, but for every polynomial there must exist a and b such that the polynomial is either strictly increasing or decreasing on $(-\infty, a)$ and either strictly increasing or decreasing on (b, ∞) .

Functions and functions evaluated on inputs

Before we dive into calculus, we need to review the difference between two concepts that are often conflated- that of a *function* and that of a *function evaluated on an input*.

- A *function* is an association between two sets in which every item from the set of “inputs” is associated with at most one item from the set of “outputs”.
- A *function evaluated on an input* is exactly what it sounds like- it is the result of taking an item from the function’s associated input set and plugging it into the function. In other words, a function evaluated on an input is an “output” of the function.

If f is a function, and x is an input, then $f(x)$ is a function evaluated on an input.

Oftentimes, people become sloppy and conflate the two notions. When someone writes, “the function a^2 ”, they are using incorrect mathematical grammar, since a^2 is not a function, but instead the the result of evaluating the function f defined by $f(x) = x^2$ on the input a . Saying “let $f(a)$ be a function” is also incorrect mathematical grammar, for the same reason.

Anonymous functions

Since it’s really tedious to write things such as “the function f defined by $f(x) = x^2$ ”, we introduce the following new notation:

“ $x \mapsto (\text{expression involving } x)$ ” means “the function sending x to (expression involving x)”.

Using the above notation, we can now say “the function $x \mapsto x^2$ ” instead of “the function f defined by $f(x) = x^2$ ”. This notation is called *anonymous function notation* because it allows us to specify functions without giving them names such as f .

Introduction to derivatives

The limit definition of the derivative

Given a function f that maps real numbers to real numbers, we define the *derivative of f* to be the function f' that is defined as follows:

$$f'(x) := \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

Note, we have used $:=$ instead of $=$ because we are presenting a definition rather than proving an equality. Also note that the Greek capital letter delta, Δ , is used because it is a symbol associated with changes in quantities; Δx is thought of as a small change in input.

There is nothing special about the variable Δx in the above definition- it is entirely possible to write the definition of the derivative without using the Δ symbol. In fact, it is common to present the definition of the derivative using the letter h in the place of Δx , as follows:

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h}.$$

Here is how to interpret what $f'(x)$ means. Since Δx goes to zero in the limit, we can imagine that Δx is an “infinitely small number”. Thus, $f(x + \Delta x) - f(x)$ is the change in the function f that results from an infinitely small change in the input x . Dividing this infinitesimally small change in f by the infinitesimally small change in input, Δx , intuitively gives the “instantaneous rate of change” of f at x that occurs due to the changing of the input from x to $x + \Delta x$. For this reason, $f'(x)$ is often called the *instantaneous rate of change of f at x* .

Secant and tangent lines

Geometrically, the argument of the limit, $\frac{f(x + \Delta x) - f(x)}{\Delta x}$, is the slope of the so-called secant⁸ line that passes through the points $(x, f(x))$ and $(x + \Delta x, f(x + \Delta x))$, since the “rise over run” ratio of this line is equal to the expression that is inside the above limit:

$$\frac{f(x + \Delta x) - f(x)}{(x + \Delta x) - x} = \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

When we take the limit $\Delta x \rightarrow 0$ of the slope of this secant line, the inputs x and $x + \Delta x$ become infinitely close, and thus the endpoints $(x, f(x))$ and $(x + \Delta x, f(x + \Delta x))$ of the secant line become infinitely close. Intuitively, it should be clear that, if it exists, the limit approached by taking $\Delta x \rightarrow 0$ is the slope of the line that touches the graph of f at the point $(x, f(x))$ and that “aligns with the slope of the function’s graph at that point”. We call this line the *tangent line*⁹ to f at $(x, f(x))$.

Differentiability

When the derivative of a function f exists at an input p , we say that f is *differentiable at p* . If f is differentiable at all inputs p , we say f is *differentiable*.

Continuity is a requirement for differentiability

[TO DO]

[diff \Rightarrow cont, but cont \neq diff]

Leibniz notation

So far, we have decided to denote the derivative of a function f with the notation f' . However, there is a much more intuitive way to notate derivatives. We define the following notation:

⁸A line is said to be *secant* to a graph if it intersects the graph at two or more points.

⁹Though we were speaking of tangent lines to gain intuition about the derivative, if we are to be formal, then we actually use the derivative to define what it means for a line to be a tangent line. The *tangent line* to the graph f at $(x, f(x))$ is the line of slope $f'(x)$ that passes through $(x, f(x))$.

$$\frac{df(g(x))}{dg(x)} := f'(g(x)) \text{ for all functions } g.$$

In particular, we have

$$\frac{df(x)}{dx} = f'(x).$$

This notation aids intuition because you can informally think of $\frac{df(x)}{dx}$ as corresponding to dividing an infinitesimal “rise” at x , $df(x)$, by the infinitesimal “run” at x , dx , to get (instantaneous slope at x) = $\frac{\text{infinitesimal rise}}{\text{infinitesimal run}} = \frac{df(x)}{dx}$. Recall that this is the idea that is formalized in the definition of the derivative as a limit.

The notations $f'(g(x))$ and $f'(x)$ are said to be *prime notation*, and the notations $\frac{df(g(x))}{dg(x)}$ and $\frac{df(x)}{dx}$ are said to be *Leibniz notation*.

Leibniz operator notation

You can write $\frac{df(g(x))}{dg(x)}$ in “operator” notation with the following new definition:

$$\frac{d}{dg(x)} f(x) := \frac{df(g(x))}{dg(x)} = f'(g(x)).$$

In particular, we have

$$\frac{d}{dx} f(x) = f'(x).$$

Leibniz vertical bar notation

When we prefer to think of f as changing in response to an input that is specifically associated with the symbol x but want to evaluate f' at a point p , we *can* do so by using the fact $\frac{df(p)}{dp} = f'(p)$. This is not intuitive, however; the presence of p in $\frac{df(p)}{dp}$ associates the letter p with the input (we would rather that the letter x be associated with the input). To remedy this situation, we define the following notation involving a vertical bar:

$$\left. \frac{df}{dx} \right|_p := f'(p).$$

One oddity of this notation is that the letter in the “denominator” of the derivative on the left side of the above is irrelevant: we have $\left. \frac{df}{da} \right|_p = \left. \frac{df}{db} \right|_p = \left. \frac{df}{dc} \right|_p = \dots = \left. \frac{df}{dz} \right|_p$.

The most common abuse of Leibniz notation.

People often write $\frac{df}{dx}$ to mean either f' or $f'(x)$.

Using $\frac{df}{dx}$ to mean f' leads to more notation problems (see “Single-variable calculus derivative notation” in the appendix), and using $\frac{df}{dx}$ to mean $f'(x)$ is also confusing because it conflicts with the practice of using $\frac{df}{dx}$ to mean f' .

In this book, we will not ever use the notation $\frac{df}{dx}$ in this book; only f' , $f'(x)$, $\frac{df(g(x))}{dg(x)}$, $\frac{df(x)}{dx}$, $\frac{d}{dx} f(x)$, and $\left. \frac{df}{dx} \right|_p$.

More on derivative notation

The advantage of prime notation; on functions versus functions evaluated at inputs.

In light of the above “most common abuse of Leibniz notation”, prime notation is best for illustrating the fact that the derivative f' of a function f is itself a function. Leibniz notation does not address this subtlety, because, “in the wild”, the symbol $\frac{df}{dx}$ can mean either f' (which is a function) or $f'(x)$ (which is a function evaluated at an input).

Advantages of Leibniz notation.

You should use Leibniz notation, because, as mentioned before, it makes derivatives *appear* to be fractions. This helps with intuition a lot.

One important example of this is that Leibniz notation can also help you remember the units of a derivative. For example, if the output of f has units of length, and the input x of f has units of time, then you can reason that the tiny increments in f and x , $df(x)$ and dx , must also have units of length and time, respectively. This lets you realize that $\frac{df(x)}{dx}$ has units of $\frac{\text{length}}{\text{time}}$.

Lastly, Leibniz notation helps you to remember the *chain rule* for derivatives. The chain rule, which we will soon learn, is: “ $\frac{dg(f(x))}{dx} = \frac{dg(f(x))}{df(x)} \frac{df(x)}{dx}$ ”. Notice that it *looks like* the $df(x)$ ’s on the right side cancel with each other to give the left side.

Differentiation formulas

Now that we have developed some sense of the various derivative notations, we are ready to derive (i.e. prove) some facts that allow us to compute derivatives of complicated expressions by means of computing derivatives of simpler sub-expressions.

In this section- and book in general- I am not a stickler about stating the conditions that must be satisfied for a fact about derivatives to be applicable. I will make note of conditions when doing derivations, but will not restate them in final “boxed” formulas. Focusing on such conditions is not really relevant for an introduction to calculus, and, anyways, the conditions are usually self-evident from the formula in question. If a derivative formula involves some derivative $g'(x)$, well, then one of the conditions is probably “ g is differentiable”!

The power rule

The derivative formula that pretty much everyone learns first is the formula for $\frac{d}{dx}x^n$, when n is an integer. We derive the formula by starting from the definition of the derivative, which tells us

$$\frac{d}{dx}x^n = \lim_{h \rightarrow 0} \frac{(x+h)^n - x^n}{h}.$$

To evaluate this limit, we need to figure out how to rewrite the product $(x+h)^n = \underbrace{(x+h)\dots(x+h)}_{n \text{ times}}$ as a sum. The binomial theorem tells us how to do this; recall that it says

$$(x+h)^n = {}_nC_0x^nh^0 + {}_nC_1x^{n-1}h^1 + {}_nC_2x^{n-2}h^2 \dots + {}_nC_nx^0h^n,$$

where ${}_nC_k = \frac{n!}{k!(n-k)!}$ is the number of unique k -element subsets of an n -element set. Noting that ${}_nC_0 = 1$ and ${}_nC_1 = n$, we have

$$\lim_{h \rightarrow 0} \frac{(x+h)^n - x^n}{h} = \lim_{h \rightarrow 0} \frac{x^n + nx^{n-1}h + ({}_nC_2x^{n-2}h^2 + \dots + {}_nC_nx^0h^n) - x^n}{h}.$$

Now, the x^n in the leftmost portion of the numerator cancels with the $-x^n$ at the rightmost portion of the numerator, so

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{(x+h)^n - x^n}{h} &= \lim_{h \rightarrow 0} \frac{nx^{n-1}h + ({}_nC_2x^{n-2}h^2 + \dots + {}_nC_nx^0h^n)}{h} \\ &= \lim_{h \rightarrow 0} \left(nx^{n-1} + ({}_nC_2x^{n-2}h + \dots + {}_nC_nx^0h^{n-1}) \right) \\ &= \lim_{h \rightarrow 0} nx^{n-1} + \left(\lim_{h \rightarrow 0} {}_nC_2x^{n-2}h + \dots + \lim_{h \rightarrow 0} {}_nC_nx^0h^{n-1} \right) \\ &= nx^{n-1} + (0 + \dots + 0) \\ &= nx^{n-1}. \end{aligned}$$

Thus

$$\frac{d}{dx}x^n = nx^{n-1} \text{ for all } x \text{ and any integer } n.$$

So, for example $\frac{d}{dx}x^2 = 2x$ and $\frac{d}{dx}x^3 = 3x^2$.

This formula can be extended to work for real numbers, not just integers¹⁰:

$$\boxed{\frac{d}{dx}x^c = cx^{c-1} \text{ for all } x \text{ and } c}$$

For example, $\frac{d}{dx}x^{\frac{1}{2}} = \frac{1}{2}x^{-\frac{1}{2}}$ and $\frac{d}{dx}x^\pi = \pi x^{\pi-1}$.

Linearity of the derivative

Using the definition of the derivative, you can prove the following.

For all functions f and g ,

$$\begin{aligned}\frac{d}{dx}(cf(x)) &= c \frac{d}{dx}(f(x)) \text{ for all } x \text{ and } c \\ \frac{d}{dx}(f(x) + g(x)) &= \frac{d}{dx}(f(x)) + \frac{d}{dx}(g(x)) \text{ for all } x\end{aligned}$$

Recall the notion of *scaling a function by a real number* and of *adding two functions together*: if r, s are functions and d is a real number, then dr is the function defined by $(dr)(x) := dr(x)$ and $r + s$ is the function defined by $(r + s)(x) := r(x) + s(x)$. Using this notation, the above can be restated as follows.

For all functions f and g ,

$$\boxed{\begin{aligned}(cf)' &= cf' \text{ for all } c \\ (f + g)' &= f' + g'\end{aligned}}$$

The product rule

$$\frac{d}{dx}(f(x)g(x)) = \frac{df(x)}{dx}g(x) + f(x)\frac{dg(x)}{dx} \text{ for all } x$$

$$\boxed{(fg)' = f'g + fg'}$$

Compositions of functions

When given functions f and g , there is a function called the *composition of g with f* that is denoted by $g \circ f$ and defined by

$$(g \circ f)(x) := g(f(x)).$$

Put simply, $g \circ f$ is the function that represents what happens if you wrap the result of doing f and then g into a single process.

One intuitive way to evaluate $(g \circ f)(x)$ is as follows. First, forget that f is a function, pretend that f is a real number instead, and compute $g(f)$. Then substitute $f(x)$ in for f ; this yields $g(f(x))$, as desired. Symbolically, we have

$$(g \circ f)(x) = g(f)|_{f \rightarrow f(x)}.$$

In light of this, we introduce the notation $g(f)$ for function composition:

$$g(f) := g \circ f.$$

Most of the time, we will use the $g \circ f$ notation for function composition. We would use the $g(f)$ notation all of the time if we could; I just personally find that this notation adds too many parentheses to syntax.

¹⁰See the appendix for a proof.

The chain rule

The *chain rule* tells us how to relate the derivative of $g \circ f$ of the derivatives of g and f . In prime notation, it is as follows:

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

We see that the instantaneous rate of change of a composition of functions at a point is the instantaneous rate of change of the “outer function”, g , at the input that g receives from f , times the instantaneous rate of change of the “inner function”, f , at the input f receives. Thus, the rate of change of $g \circ f$ at a point is proportional to the instantaneous rates of change of g and f at the “relevant” inputs (those being $f(x)$ and x , respectively).

The above presentation of the chain rule is probably the best version of the chain rule to learn first, as it is most “to the point”, but it is not the most elegant.

The chain rule as a “cancellation of fractions”

To obtain the most elegant presentation of the chain rule, recall from the definition of Leibniz notation that we have $g'(f(x)) = \frac{dg(f(x))}{df(x)}$ and $f'(x) = \frac{df(x)}{dx}$, so the chain rule is

$$\frac{d(g \circ f)(x)}{dx} = \frac{dg(f(x))}{df(x)} \frac{df(x)}{dx}.$$

Using the definition $(g \circ f)(x) := g(f(x))$ to rewrite the left side, we have

$$\frac{dg(f(x))}{dx} = \frac{dg(f(x))}{df(x)} \frac{df(x)}{dx}.$$

Notice that it now appears as if the $df(x)$ ’s of the right side are cancelling in a multiplication of fractions to give the left side.

In summary, here are four ways to state the chain rule.

$(g \circ f)'(x) = (g' \circ f)(x)f'(x) = g'(f(x))f'(x) \text{ for all } x$ $(g \circ f)' = (g' \circ f)f'$ $\frac{d(g \circ f)(x)}{dx} = \frac{dg(f(x))}{df(x)} \frac{df(x)}{dx} \text{ for all } x$ $\frac{dg(f(x))}{dx} = \frac{dg(f(x))}{df(x)} \frac{df(x)}{dx} \text{ for all } x$
--

A bad way to state the chain rule

Most calculus textbooks state the chain rule in the following confusing way:

$$\text{If } y = f(u) \text{ and } u = g(x) \text{ are differentiable functions, then } \frac{dy}{dx} = \frac{dy}{du} \frac{du}{dx}.$$

There are several problems with this statement:

- The phrase “If $y = f(u)$ and $u = g(x)$ are differentiable functions” is meant to be interpreted by the reader as the following system of equations:

$$\begin{cases} y = f(u) \\ u = g(x) \end{cases}, \text{ where } x \text{ is a real number and } f, g \text{ are differentiable functions.}$$

It’s extremely unclear that this interpretation is intended, since no phrase that indicates x is supposed to have many possible values (like “ x is a real number” or “where x is a real number”) is included.

- The phrase “If $y = f(u)$ and $u = g(x)$ are differentiable functions” implies that y and u are functions, which is not the case; y and u are outputs of functions, and are not functions themselves; they are variables in the above-mentioned system of equations.
- Not all expressions obtained by making the substitutions specified in the system of equations are defined. Namely, if we perform the substitutions $y = f(u)$ and $u = g(x)$ in the term $\frac{dy}{du}$, we obtain $\frac{dy}{du} = \frac{df(u)}{du} = \frac{df(g(x))}{dg(x)}$, which is a notation that isn’t defined in most texts. (As we have seen, the correct definition is $\frac{df(g(x))}{dg(x)} := f'(g(x))$.)

More differentiation formulas

The quotient rule

Let f and g be functions. When x is such that $g(x) \neq 0$, we can use the product rule to compute the derivative of $x \mapsto \frac{f(x)}{g(x)}$, and so obtain a *quotient rule*:

$$\frac{d}{dx} \left(\frac{f(x)}{g(x)} \right) = \frac{d}{dx} \left(f(x) \frac{1}{g(x)} \right) = \frac{df(x)}{dx} \frac{1}{g(x)} + f(x) \frac{d}{dx} (g(x)^{-1}) \text{ when } g(x) \neq 0.$$

The chain rule gives

$$\frac{d}{dx} (g(x)^{-1}) = -g(x)^{-2} g'(x) = -\frac{g'(x)}{g(x)^2} \text{ when } g(x) \neq 0,$$

so the above is

$$\frac{d}{dx} \left(\frac{f(x)}{g(x)} \right) = f'(x) \frac{1}{g(x)} - f(x) \frac{g'(x)}{g(x)^2} = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2} \text{ when } g(x) \neq 0.$$

In all, we have that for all functions f and g :

$$\boxed{\frac{d}{dx} \left(\frac{f(x)}{g(x)} \right) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2} \text{ when } g(x) \neq 0.}$$

Derivatives of trig functions

We now compute the derivatives of the trig functions \sin , \cos , \tan . We do so by finding the derivative of \sin and then using trig identities and derivative formulas to infer the derivatives of \cos and \tan .

Before we compute the derivative of \sin , we need to know that

$$\begin{aligned} \lim_{x \rightarrow 0} \frac{\sin(x)}{x} &= 1 \\ \lim_{x \rightarrow 0} \frac{\cos(x) - 1}{x} &= 0. \end{aligned}$$

The squeeze theorem can be used to prove¹¹ the first fact, and the first fact in combination with the identity $\cos(x) = \sin(x - \frac{\pi}{2})$ can be used to prove the second fact.

Knowing these facts, we can use the formula $\sin(a+b) = \sin(a)\cos(b) + \sin(b)\cos(a)$ to compute the derivative of \sin :

$$\begin{aligned} \frac{d}{dx} \sin(x) &= \lim_{h \rightarrow 0} \frac{\sin(x+h) - \sin(x)}{h} = \lim_{h \rightarrow 0} \frac{\sin(x)\cos(h) + \sin(h)\cos(x) - \sin(x)}{h} \\ &= \lim_{h \rightarrow 0} \frac{\sin(x)(\cos(h) - 1) + \sin(h)\cos(x)}{h} = \lim_{h \rightarrow 0} \left(\frac{\cos(h) - 1}{h} \right) \sin(x) + \lim_{h \rightarrow 0} \left(\frac{\sin(h)}{h} \right) \cos(x) \\ &= 0 \cdot \sin(x) + 1 \cdot \cos(x) \\ &= \cos(x). \end{aligned}$$

¹¹[https://proofwiki.org/wiki/Limit_of_\(Cosine_\(X\)\)__-_1_over_X/Proof_3](https://proofwiki.org/wiki/Limit_of_(Cosine_(X))__-_1_over_X/Proof_3) for the proof that $\lim_{x \rightarrow 0} \frac{\cos(x) - 1}{x} = 0$.

Thus

$$\frac{d}{dx} \sin(x) = \cos(x).$$

Using the fact that $\cos(x) = \sin(\frac{\pi}{2} - x)$, we can compute the derivative of $\cos(x)$ using the chain rule and our newly acquired expression for the derivative of \sin :

$$\frac{d}{dx} \cos(x) = \frac{d}{dx} \sin\left(\frac{\pi}{2} - x\right) = \cos\left(\frac{\pi}{2} - x\right) \frac{d}{dx} \left(\frac{\pi}{2} - x\right) = -\cos\left(\frac{\pi}{2} - x\right) = -\sin(x).$$

Now, we compute the derivative of $\tan = \frac{\sin}{\cos}$. We have

$$\frac{d}{dx} \tan(x) = \frac{\frac{d}{dx} \sin(x)}{\frac{d}{dx} \cos(x)} = \frac{\cos(x) \frac{d}{dx} \sin(x) - \sin(x) \frac{d}{dx} \cos(x)}{\cos^2(x)} = \frac{\cos^2(x) + \sin^2(x)}{\cos^2(x)}$$

by the quotient rule. We know $\sin^2(x) + \cos^2(x) = 1$ for all x , so

$$\frac{d}{dx} \tan(x) = \frac{1}{\cos^2(x)} = \sec^2(x).$$

Recall that $\sec(x) := \frac{1}{\cos(x)}$.

In all, we have

$\begin{aligned} \frac{d}{dx} \sin(x) &= \cos(x) \\ \frac{d}{dx} \cos(x) &= -\sin(x) \\ \frac{d}{dx} \tan(x) &= \sec^2(x) \end{aligned}$
--

Derivatives of inverse functions

Let f be a function. We will now derive a formula that relates the derivative of f^{-1} to the derivative of f . First, notice that since $f \circ f^{-1}$ is the identity $f \circ f^{-1} = I$ defined by $I(y) = y$, we have

$$\frac{d(f \circ f^{-1})(y)}{dy} = 1.$$

(We are using the letter y here because f^{-1} maps outputs of f , which we typically denote with y , to inputs of f).

On the other hand, the chain rule gives us that

$$\frac{d(f \circ f^{-1})(y)}{dy} = \frac{df(f^{-1}(y))}{df^{-1}(y)} \frac{df^{-1}(y)}{dy}.$$

Comparing the two equations and solving for $\frac{df^{-1}(y)}{dy}$ gives

$\begin{aligned} \frac{df^{-1}(y)}{dy} &= \frac{1}{\frac{df(f^{-1}(y))}{df^{-1}(y)}} \text{ for all } y \text{ such that } \frac{df(f^{-1}(y))}{df^{-1}(y)} \neq 0 \\ (f^{-1})'(y) &= \frac{1}{f'(f^{-1}(y))} \text{ for all } y \text{ such that } f'(f^{-1}(y)) \neq 0 \end{aligned}$

If we define the notation

$$\frac{dy}{df^{-1}(y)} := \frac{df(f^{-1}(y))}{df^{-1}(y)},$$

then the first equation of the above restates as

$$\boxed{\frac{df^{-1}(y)}{dy} = \frac{1}{\frac{dy}{df^{-1}(y)}} \text{ for all } y \text{ such that } \frac{dy}{df^{-1}(y)} \neq 0}$$

This definition is especially intuitive because it can be remembered by noticing that the “numerators” of the derivatives on both sides are equal: $y = f(f^{-1}(y))$.

Derivatives of inverse trig functions

Ways to think about derivatives

[TO DO]

- conventional way we already discussed: as rate of instantaneous change/speed, and as slope of tangent line
- as densities (not necessarily mass density)
- as representing change caused by a particular action, since we often approximate $df(x) \approx \frac{df(x)}{dx}dx$ (esp. when dealing with partial derivatives)

Derivatives of exponential and logarithmic functions

In precalculus, you should have learned about the functions $x \mapsto b^x$ and $x \mapsto \log_b(x)$. You may have even learned about the number e , and about the natural logarithm $\ln := \log_e$. To properly understand these concepts, especially the concept of e , we need the tools of calculus.

Definition of exponential functions

Let's first review what an exponential function such as $x \mapsto b^x$ really is.

When x is a positive integer, we define $b^x := \underbrace{b \cdot \dots \cdot b}_{x \text{ times}}$. When x is a negative integer, we define $b^x := \frac{1}{b^{-x}}$.

When x is an arbitrary integer, we define $b^{\frac{1}{x}}$ to be such that $f(b) = b^{\frac{1}{x}}$ is the inverse function to $g(b) = b^x$ on the domain $b > 0$; that is, we define $b^{\frac{1}{x}}$ to be such that $f(g(b)) = (b^x)^{\frac{1}{x}} = b = (b^{\frac{1}{x}})^x = g(f(b))$. Sometimes, $b^{\frac{1}{x}}$ is instead written as $\sqrt[x]{b}$.

[see precalc text for more in depth discussion of why these definitions are chosen]

What about when x is a real number, and is thus potentially irrational, i.e., inexpressible as a ratio of two integers? Well, intuitively, one would expect that if x_ℓ and x_r are rational numbers with $x_\ell < x < x_r$, then $b^{x_\ell} < b^x < b^{x_r}$. If we could just keep finding rational x_ℓ and x_r that are closer and closer to x , then it would make sense to call the limit approached by b^{x_ℓ} and b^{x_r} by the name " b^x ".

Fortunately, it is indeed a fact that for every real number x there is a sequence (x_n) of rational numbers converging to x . Thus, as suggested above, we *define* b^x to be the limit that is approached by b^{x_n} , where the x_n converge to x :

$$b^x := \lim_{n \rightarrow \infty} b^{x_n}, \text{ where } (x_n) \text{ is any sequence of rational numbers with } \lim_{n \rightarrow \infty} x_n = x$$

For this definition to be valid, we have to show that this limit exists for all b and that, given two distinct sequences (x_n) and (y_n) of rational numbers that converge to x , it doesn't matter which sequence we use to compute b^x , i.e., $\lim_{n \rightarrow \infty} b^{x_n} = \lim_{n \rightarrow \infty} b^{y_n}$. [See appendix]

Properties of exponential functions

We now discuss some important properties of exponential functions. For all real numbers b, x, y , we have:

- $b^{x+y} = b^x b^y$
- $(b^x)^y = b^{xy}$
- When $b > 0$, the function $x \mapsto b^x$ is continuous, one-to-one, and maps to all positive real numbers.

We already know from precalculus that the first two properties hold for rational x and y . In the appendix, we use the above definition of b^x to show that they hold for real numbers x and y . We also show the third property in the appendix.

Note that the third property breaks down when $b < 0$: when $b < 0$, the function $x \mapsto b^x$ is not continuous¹², and does not map to every positive real number¹³.

Definition of logarithms

[since $x \mapsto b^x$ is one-to-one, it has an inverse. we call the inverse \log_b . since $b^0 = 1$ and $\lim_{x \rightarrow -\infty} b^x = 0$ then $\log_b(1) = 0$ and $\lim_{x \rightarrow 0} \log_b(x) = -\infty$.]

\log_b is defined to be the inverse function to $x \mapsto b^x$. So \log_b satisfies $\log_b(b^x) = x = b^{\log_b(x)}$.

\log_b **when** $b < 0$

When $b < 0$, the function \log_b is kind of nasty.

[TO DO]

For this reason, we will only consider \log_b when $b > 0$

¹²For integer n, m , we have that $b^{\frac{n}{m}}$ is positive when n is even and that $b^{\frac{n}{m}}$ is negative when n is odd and m is odd (if n is odd and m is even, then b^n is negative and so $b^{\frac{n}{m}}$ is undefined). Thus $x \mapsto b^x$ oscillates wildly between positive and negative numbers when $b < 0$ in an arbitrary small window of inputs- it cannot be continuous.

¹³ b^x cannot be a positive odd power of b when $b < 0$

Properties of logarithms

Once it is shown that the above “Properties of exponential functions” hold for real numbers, it is easy to show that the following corresponding properties of logarithms immediately hold. For all real numbers $b, x, y > 0$, we have:

- $\log_b(xy) = \log_b(x) + \log_b(y)$
- $\log_b(x^y) = y \log_b(x)$
- $x \mapsto \log_b(x)$ is continuous, one-to-one, and maps to *all* real numbers.

[For intuition that helps with interpreting what these properties are saying, see precalc text.]

Note, we require $b > 0$ so that the third bullet point of “Properties of exponential functions” is true, as, if *that* bullet point is true, then the third bullet point of the immediate above must be true too. We require $x, y > 0$ because x, y are elements of the range of $x \mapsto b^x$ (the range of $x \mapsto b^x$ is the positive real numbers).

Changes of base

Change of base for exponential functions

The *change of base for exponential functions* is the fact that, for any real numbers a, b with $b > 0$,

$$a^x = b^{\log_b(a^x)} = b^{\log_b(a)x}.$$

That is, for any a, b with $b > 0$, we have

$$a^x = b^{dx}, \text{ where } d = \log_b(a).$$

So, it doesn’t really matter what base one uses in an exponential function, since any base can be “converted” to another base with the above formula.

Change of base for logarithms

We can derive a way to change the base of a logarithm from our knowledge of the change of base for exponential functions. Start with the change of base of exponential functions, and take \log_b of both sides to obtain $\log_b(a^x) = dx$. Let $y = a^x$ to obtain that $\log_b(y) = d \log_a(y)$, where $d = \log_b(a)$. Swapping a and b and replacing y with x , we have the *change of base for logarithms*:

$$\log_a(x) = d \log_b(x), \text{ where } d = \log_a(b).$$

The change of base for logarithms is often stated as

$$\log_b(x) = \frac{\log_a(x)}{\log_a(b)}.$$

Note that the change of base for logarithms implies that when $x > 0$,

$$\log_x(b) = \frac{\log_a(b)}{\log_a(x)} = \frac{1}{\log_b(x)}.$$

Derivatives of exponential functions

Given a real number $b > 0$, let’s take the derivative of b^x with respect to x .

$$\frac{d}{dx} b^x = \lim_{h \rightarrow 0} \frac{b^{x+h} - b^x}{h} = \lim_{h \rightarrow 0} \frac{b^x(b^h - 1)}{h} = \left(\lim_{h \rightarrow 0} \frac{b^h - 1}{h} \right) b^x$$

Notice that the limit in this last expression can be rewritten:

$$\lim_{h \rightarrow 0} \frac{b^h - 1}{h} = \lim_{h \rightarrow 0} \frac{b^{0+h} - b^0}{h} = \left(\frac{d}{dx} b^x \right) \Big|_0.$$

Thus, the derivative of b^x with respect to x is

$$\frac{d}{dx} b^x = \left(\frac{d}{dx} b^x \right) \Big|_0 b^x.$$

That is, the derivative of b^x with respect to x is some constant times b^x . In order to obtain a more useful expression for $\frac{d}{dx} b^x$, we must evaluate the expression $\left(\frac{d}{dx} b^x \right) \Big|_0$. A natural way to do so is to first consider the b for which¹⁴ we have $\left(\frac{d}{dx} b^x \right) \Big|_0 = 1$.

For the sake of having concrete notation, we will *define* e to be the real number such that $\left(\frac{d}{dx} e^x \right) \Big|_0 = 1$. Then

$$\frac{d}{dx} e^x = \left(\frac{d}{dx} e^x \right) \Big|_0 e^x = 1 \cdot e^x = e^x.$$

Now, we use change of base for exponential functions in order to determine $\frac{d}{dx} b^x$ by making use of the fact that we know the derivative of e^x . We have

$$\frac{d}{dx} b^x = \frac{d}{dx} e^{\log_e(b^x)} = \frac{d}{dx} e^{\log_e(b)x} = \log_e(b) e^{\log_e(b)x} = \log_e(b) b^x.$$

In all,

$$\frac{d}{dx} b^x = \log_e(b) b^x.$$

In some sense, e is the “most natural exponent” because it is the one for which $\frac{d}{dx} e^x = 1 \cdot e^x$. For this reason, the logarithm of base e is called the *natural logarithm*, and is denoted by \ln (for “logarithmus naturalis”). Formally, we define $\ln := \log_e$. So the above is restated as

$$\frac{d}{dx} b^x = \ln(b) b^x$$

Alternate expression for e^x

We now make a series of observations that lead us to an alternate expression for e^x . To start, recall that we *defined* e to be such that $\left(\frac{d}{dx} e^x \right) \Big|_0 = 1$. Also recall that, for any function f , the derivative of the inverse function f^{-1} is given by:

$$\frac{df^{-1}(y)}{dy} = \frac{1}{\frac{df(f^{-1}(y))}{df^{-1}(y)}} \text{ for all } y \text{ such that } \frac{df(f^{-1}(y))}{df^{-1}(y)} \neq 0.$$

We can substitute $f(x) = e^x$ and $y = 1$ into the above to obtain

$$\frac{d \ln(1)}{d1} = \frac{1}{\frac{e^0}{d0}},$$

which is more nicely written as

$$\left(\frac{d}{dy} \ln(y) \right) \Big|_1 = \frac{1}{\frac{d}{dx} (e^x) \Big|_0}.$$

¹⁴Does such a b even exist? Does $\left(\frac{d}{dx} b^x \right) \Big|_0 = \lim_{h \rightarrow 0} \frac{b^h - 1}{h}$ even converge for all b ? The answer to both questions is yes, but showing so is quite technical.

Since $\frac{d}{dx}(e^x)|_0 = e^x|_0 = e^0 = 1$, we have

$$\left(\frac{d}{dy} \ln(y)\right)\Big|_1 = 1.$$

Computing the derivative of $\ln(x)$ at $x = 1$ explicitly, we have

$$\lim_{h \rightarrow 0} \frac{\ln(1+h) - \ln(1)}{h} = \lim_{h \rightarrow 0} \frac{\ln(1+h)}{h} = \lim_{h \rightarrow 0} \ln((h+1)^{\frac{1}{h}}) = 1.$$

Interestingly enough, a not yet encountered expression for e jumps out at us from this above equation. Apply the function $x \mapsto e^x$ to both sides of the above (using continuity to bring the application of $x \mapsto e^x$ inside the limit) to obtain

$$\lim_{h \rightarrow 0} (h+1)^{\frac{1}{h}} = e,$$

and then substitute $k := \frac{1}{h}$ to obtain

$$\lim_{k \rightarrow \infty} \left(\frac{1}{k} + 1\right)^k = e.$$

Substituting n in for k , we see that we have shown

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n.$$

We now use this new expression for e to determine a new expression for e^x . Applying the function $x \mapsto e^x$ to both sides of the above, we see that for all x ,

$$e^x = \left(\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n\right)^x = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^{nx}.$$

Substitute $p := nx$ to obtain

$$e^x = \lim_{p \rightarrow \infty} \left(1 + \frac{x}{p}\right)^p.$$

Overall, we have shown

$$\boxed{\begin{aligned} e &= \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n \\ e^x &= \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n \end{aligned}}$$

e and compound interest

Calculus textbooks are overly fond of discussing how to interpret the fact that $e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n$ in the setting of interest rates, so I include the below for completeness.

The argument of the limit, $\left(1 + \frac{1}{n}\right)^n$, can be interpreted as the factor by which an initial quantity grows when the growth of the quantity occurs in n stages and where the quantity grows by a factor of $1 + \frac{1}{n}$ in each stage. For example, if you put \$100 in the bank with a yearly interest rate of $\frac{1}{5} = 20\%$, and let this money sit in the bank for 5 years, then your \$100 would have grown by a factor of $\left(1 + \frac{1}{5}\right)^5 = (120\%)^5 = 2.48832 = 248.832\%$ at the end of the 5 years. So, you would have $\$100 \cdot 2.48832 = \248.832 at the end of the 5 years.

We say that the money *compounds* whenever it grows as a result of applying the interest rate. In the above scenario, the money is compounded $n = 5$ times, once at the end of each year. Taking the limit as $n \rightarrow \infty$ represents using a very small interest rate (an interest rate of $\frac{1}{n}$) infinitely often. Thus, e can be interpreted as the factor by which quantities grow when subject to so-called *continuously compounded* interest.

Common pedagogical problems with e , e^x and $\ln(x)$

You should be warned that the usual ways to present e , e^x and $\ln(x)$ are all incredibly problematic. In fact, the unproblematic derivation and explanation of e , e^x and $\ln(x)$ presented here is so uncommon that I have never read of it in another math textbook, and have only seen it discussed in an online blog by Paramanand Singh.

The problem with these usual ways is that every one of them starts by assuming the truth of one of our “end results” (such as $e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n$ or $e^x = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n$, or even facts we haven’t learned of yet, like “ $e = \sum_{n=0}^{\infty} \frac{1}{n!}$ ”, “ $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ ”, and “ $\ln(x) = \int_1^x \frac{1}{x} dx$ ”) and then uses said “end result” as a starting point from which to establish the properties of e , e^x and $\ln(x)$. This is unintuitive and unnatural- who could reasonably come up with the idea to use one of these “end results” as a starting point without doing a ridiculous amount of prior investigation? Well-explained math is math that feels at least somewhat plausible to discover, and starting with one of the mentioned “end results” is not of this spirit.

Derivatives of logarithms

Analytical geometry

- Increasingness, decreasingness, local/global extrema
 - Extreme value theorem: if an extremum of f occurs at x , then $\left.\frac{df}{dx}\right|_x = 0$.
 - * The converse of extreme value theorem is not true: it is not necessarily true that if $\left.\frac{df}{dx}\right|_x = 0$, then x is an extreme value of f .
- Concavity, inflection points
- Horizontal and vertical asymptotes

Misc.

- Mean value theorem
 - Isn’t “misc” if you’re proving calculus, but is “misc” if you’re using calculus.
- L’Hopital’s rule

Integral calculus

Riemann sums and definite integrals

Riemann sums

Given the derivative f' of some function f , we can approximate the change in f from $x = a$ to $x = b$, which is $f(b) - f(a)$, as follows. First, we break up the interval $[a, b]$ into many intervals:

$$[x_1, x_2] = [a, x_2], [x_2, x_3], \dots, [x_{n-1}, x_n] = [x_{n-1}, b], \text{ where } x_{i+1} > x_i.$$

Next, we approximate f' as a piecewise function that is constant on each piece. For a particular “piece” $[x_i, x_{i+1}]$, we pick one of the values that f' takes on within this interval, and then define our piecewise approximation to be equal to this value when evaluated on any input inside the same interval.

Lastly, we apply the formula (change in output) = (rate of change) · (change in input), which is true for constant rates of change, on each piece, and sum up the changes in output of all the pieces to obtain an approximation of the total change in f from $x = a$ to $x = b$. The approximation gets better as we use more pieces.

Symbolically, all the previous steps combine to tell us that the change in f from $x = a$ to $x = b$ is approximately:

$$f(b) - f(a) \approx \sum_{i=1}^n \left. \frac{df}{dx} \right|_{x=p_i} (\Delta x)_i, \text{ where } (\Delta x)_i = x_{i+1} - x_i,$$

and where we choose p_i in one of the following ways:

$$p_i = x_i \text{ for a left Riemann sum}$$

$$p_i = x_i + \frac{(\Delta x)_i}{2} \text{ for a midpoint Riemann sum}$$

$$p_i = x_{i+1} \text{ for a right Riemann sum.}$$

The choice of p_i determines which value our piecewise approximation takes on in each piece; for example, if a left Riemann sum is chosen, then the piecewise approximation takes on the “leftmost” value of f within that piece.

We have used the words “Riemann sum” in the above definition of p_i , but not explained what “Riemann sum” means. What is a Riemann sum? A *Riemann sum of a function g from a to b* is a sum of the form

$$\sum_{i=1}^n g(p_i)(\Delta x)_i, \text{ where } (\Delta x)_i = x_{i+1} - x_i,$$

where the intervals $[x_i, x_{i+1}]$ are chosen in the same way as above (in particular, we must have $x_1 = a$ and $x_n = b$), and where the p_i are chosen in one of the ways described above.

So, the above sum involving $\left. \frac{df}{dx} \right|_{x=p_i}$ is a Riemann sum of $x \mapsto \left. \frac{df}{dx} \right|_x = \frac{df(x)}{dx}$ from a to b ; it is a Riemann sum of f' from a to b .

Riemann sums and area

If one has a function g and draws a graph of g vs. x , they can notice that $g(p_i)(\Delta x)_i$ is the width of the rectangle with height $g(p_i)$ and width $(\Delta x)_i$. So, a Riemann sum of g from a to b approximates the area under the graph of g that is above the y -axis and between the vertical lines $x = a$ and $x = b$.

The definite integral

We define the *definite integral from a to b* of a function g to be the following limit of a Riemann sum:

$$\int_a^b g(x)dx := \lim_{\max(\Delta x)_i \rightarrow 0} \sum_{i=1}^n g(p_i)(\Delta x)_i, \text{ where } p_i \text{ is any of the above.}$$

In the limit, the maximum input interval $\max((\Delta x)_i)$ approaches 0, so all other input intervals are forced to approach 0 as well. Thus, the above limit corresponds to approximating the change in g when g is approximated as a piecewise function of “infinitely many” pieces.

Since a Riemann sum of $x \mapsto \frac{df(x)}{dx}$ from a to b approximates the change in f over this input interval,

$$f(b) - f(a) \approx \sum_{i=1}^n \left. \frac{df}{dx} \right|_{x=p_i} (\Delta x)_i,$$

it stands to reason that the definite integral of $x \mapsto \frac{df(x)}{dx}$ from a to b , being an “infinitely accurate Riemann sum”, will be equal to this change:

$$f(b) - f(a) = \int_a^b \left. \frac{df}{dx} \right|_x dx = \int_a^b \frac{df(x)}{dx} dx.$$

This is indeed true.

The definite integral and area

In the previous section, we noted that “a Riemann sum of g from a to b approximates the area under the graph of g that is above the y -axis and between the vertical lines $x = a$ and $x = b$ ”. Since an integral is intuitively an “infinitely accurate Riemann sum”, the exact area under the graph of g from a to b is $\int_a^b g(x) dx$.

Antiderivatives and the fundamental theorem of calculus

Our investigation of Riemann sums and definite integrals showed that we can compute the change in a function f over an input interval $x = a$ to $x = b$ as

$$f(b) - f(a) = \int_a^b \frac{df(x)}{dx} dx.$$

Since any continuous function is the derivative of some other continuous function, we can rewrite the above by interpreting $x \mapsto \frac{df(x)}{dx}$ to be an arbitrary continuous function g :

$$? = \int_a^b g(x) dx$$

We used a question mark in the above because it is somewhat unclear how to convert the previous left side of the equation, $f(b) - f(a)$, into an expression that involves g rather than f .

In order to figure out what $?$ is, notice that in the equation $f(b) - f(a) = \int_a^b \frac{df(x)}{dx} dx$, the function on the right side, $x \mapsto \frac{df(x)}{dx}$, is the derivative of the function on the left side, f . So it stands to reason that in the above equation, the function g on the right side is the derivative of whatever function is involved in the left side. Therefore, we will guess that

$$G(b) - G(a) = \int_a^b g(x) dx, \text{ where for all } x \text{ we have } \frac{dG(x)}{dx} = g(x).$$

In order to prove that this guess is correct, we must think more about the function G .

In general, any function H that satisfies $H' = g$ is said to be an *antiderivative of g* . (Thus, the G above is an antiderivative of g). Observe that there are many antiderivatives of any function, since, if H is an antiderivative of g and c is a constant, then $H + c$ must be an antiderivative of g as well¹⁵.

In fact, one can show that *any* two antiderivatives of a function must differ by a constant¹⁶. This is very important- it means that if you’ve found one antiderivative of a function, you’ve really found all its

¹⁵For all x , we have $\frac{d}{dx}(H(x) + c) = \frac{dH(x)}{dx} + 0 = \frac{dH(x)}{dx} = g(x)$.

¹⁶Suppose F and G are both antiderivatives of g . Then $F' = G'$, so $F' - G' = 0$. Using the linearity of the derivative, we have $(F - G)' = 0$. It follows from the mean value theorem that if a function’s derivative is zero for all inputs, then the function must be a constant function. Applying this fact to the function $F - G$, we see that $F - G$ must be a constant function. That is, F and G differ by a constant.

antiderivatives: the collection of all antiderivatives is equal to the collection of functions obtained by adding constants to the one antiderivative you found.

Now, we can justify our above guess; we can justify why the old left side $f(b) - f(a)$ is equal to the new left side $G(b) - G(a)$. The justification is as follows. Since f and G are antiderivatives of $f' = g$, they differ by a constant: for each x , we have $f(x) = g(x) + d$ for some constant d . This implies that $f(b) - f(a) = (G(b) + d) - (G(a) + d) = G(b) - G(a)$, as we claimed.

We now restate the equation above, with the only change being that we use the letters f and F instead of g and G . This equation is called the *the second part of the fundamental theorem of calculus*. We will abbreviate it as *FTC 2*.

$$\text{If } F \text{ is an antiderivative of } f \text{ then } \int_a^b f(x)dx = F(b) - F(a).$$

If x is substituted for b , FTC 2 can be written as

$$\text{If } F \text{ is an antiderivative of } f \text{ then } F(x) = \int_a^x f(x)dx + F(a) \text{ for all } x.$$

This shows that any antiderivative can be written as a definite integral plus a constant. For this reason, the taking of an antiderivative is notated with an integral symbol without limits of integration:

$$“F(x) = \int f(x)dx \text{ for all } x” \text{ means “} F \text{ is an antiderivative of } f”.$$

Such integrals without limits are called *indefinite*. With this new notation, FTC 2 can be written as

$$\text{If } F(x) = \int f(x)dx \text{ for all } x \text{ then } \int_a^b f(x)dx = F(b) - F(a).$$

The *first part of the fundamental theorem of calculus* follows as a logical consequence of the second part. From above, we have

$$\text{If } F \text{ is an antiderivative of } f \text{ then } F(x) = \int_a^x f(x)dx + F(a) \text{ for all } x.$$

Take the derivative of both sides of the equation with respect to x to obtain the *first part of the fundamental theorem of calculus*, which we abbreviate as *FTC 1*:

$$f(x) = \frac{d}{dx} \int_a^x f(x).$$

We restate FTC 2 and FTC 1 for convenience:

$$\begin{aligned} \text{If } F(x) = \int f(x)dx \text{ for all } x \text{ then } \int_a^b f(x)dx &= F(b) - F(a), \\ \frac{d}{dx} \int_a^x f(x) &= f(x) \end{aligned}$$

FTC 2 and FTC 1 formalize the sense in which antiderivatives and definite integrals are related. FTC 2 says that definite integrals can be computed by evaluating antiderivatives. FTC 1 says that the function $x \mapsto \int_a^x f(x)dx$, which involves a definite integral, is an antiderivative of f .

Note that the previous sentence implies:

Every continuous function has an antiderivative.

Contrastingly, recall that it is *not* the case that every continuous function has a derivative.

Why FTC 2 before FTC 1?

We've proved FTC 2 implies FTC 1. It's also possible to prove FTC 1 implies FTC 2.

In my opinion, FTC 2 is more intuitive than FTC 1. FTC 1 says “the rate of change of the ‘accumulation’ of f is equal to f itself”, which seems a bit complicated for someone to intuit. FTC 2, which says that adding up many “infinitesimal” changes gives an overall change, seems to be a much easier discovery. Since FTC 1 is easily seen to follow from FTC 2, I think that treating FTC 2 as the fundamental principle and FTC 1 as a corollary is the way to go.

Integral notation

Before we proceed further, we should define some conventions for notating integrals. So far, the only conventions we've established are that the definite integral of a function f from a to b is written as $\int_a^b f(x)dx$ and that the evaluation at x of an antiderivative of a function f is written as $\int f(x)dx$.

We now lay out new conventions analogous to prime notation¹⁷ for derivatives:

$$\int f := \text{an antiderivative of } f, \quad \int_a^b f := \text{the definite integral of } f \text{ from } a \text{ to } b.$$

With these new conventions, we can restate the already established integral notation more clearly:

$$\int f(x)dx := \left(\int f \right) \Big|_x, \quad \int_a^b f(x)dx := \int_a^b f.$$

We also define integral notation analogous to the Leibniz notation $\frac{df(g(x))}{dg(x)} := f'(g(x))$:

$$\int f(g(x))dg(x) := \left(\int f \right) \Big|_{g(x)}, \quad \int_a^b f(g(x))dg(x) := \left(\int_a^b f \right) \Big|_{g(x)}.$$

Lastly, we define the vocabulary word *integrand*. The integrand of an integral is just the argument of that integral. So, the integrand of $\int f$ is f and the integrand of $\int f(x)dx$ is $f(x)$.

Integration formulas

[Pull from introduction of next section to write intro to this section]

The power rule

Linearity of the integral

Integrals of trig functions

Integrals of inverse functions

Integrals via inverse trigonometry

Integrals of exponential and logarithmic functions

When $x > 0$, we have $\int \frac{1}{x}dx = \ln(x)$. We don't have $\int \frac{1}{x}dx = \ln(x)$ when $x < 0$ because $\ln(x)$ isn't even defined when $x < 0$.

To find an antiderivative for $x \mapsto \frac{1}{x}$ for all x , notice that $\ln(|x|)$ is defined for all x . Is $x \mapsto \ln(|x|)$ an antiderivative of $x \mapsto \frac{1}{x}$, though? We do a quick check, and find out it is:

$$\frac{d}{dx} \ln(|x|) = \frac{1}{|x|} \frac{d|x|}{dx} = \frac{1}{|x|} \frac{d}{dx} \begin{pmatrix} x & x \geq 0 \\ -x & x < 0 \end{pmatrix} = \frac{1}{|x|} \begin{pmatrix} 1 & x \geq 0 \\ -1 & x < 0 \end{pmatrix} = \frac{1}{x}.$$

So in all, we have

¹⁷The notations $\int f$, $\int_a^b f$, and f' are analogous because they are concern functions, not functions evaluated on inputs.

$$\boxed{\int \frac{1}{x} dx = \ln(|x|)}$$

People often write $\ln(|x|)$ as simply $\ln |x|$.

etc.

$$\int_a^b = - \int_b^a$$

The inverse chain rule

[change below text to introduce the idea of the “inverse” chain rule]

In this section, we derive several integration formulas. We often do so by starting with a differentiation formula, taking the antiderivative of both sides, and interpreting some derivative that appears in the resulting expression to be “just a regular function”. Such interpretations are valid because FTC 1 tells us that every continuous function has an antiderivative. Because integral formulas make use of this interpretation in their derivations, they all require that certain derivatives are continuous; i.e., they all require that certain functions are “continuously differentiable”.

In spite of the above note about “continuous differentiability”, I will not be a stickler about stating the conditions that must be satisfied for a fact about integrals to be applicable, just as was the case when deriving derivative formulas. I will make note of conditions when doing derivations, but will not restate them in final “boxed” formulas. Conditions for integral formulas are in general less self-evident than those for derivative formulas, but you can still determine good guesses as to what the conditions are by starting with the corresponding derivative formula and integrating both sides.

Change of variables theorem

First, we will derive the fact for integrals that corresponds to the chain rule for derivatives. Recall that the chain rule can be stated as

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

Integrate both sides to obtain

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

Now, we will think of g' as being “just a regular function”. To formalize this, substitute $h := g'$. This implies¹⁸ $g = \int h$, and we have

$$\left(\int h \right) \circ f = \int (h \circ f) f'.$$

Now, substitute the letter g in for the letter h , and “flip” the equation around to obtain our primary result.

$$\int (g \circ f) f' = \left(\int g \right) \circ f.$$

Using the integral notation analogous to Leibniz notation, the primary result can be restated as

$$\boxed{\int g(f(x)) \frac{df(x)}{dx} dx = \int g(f(x)) df(x)}$$

¹⁸We need to justify that h has an antiderivative here. In order for this to be guaranteed, we actually needed to assume at the beginning that g is *continuously differentiable*, i.e., that g' is continuous. This is a pretty technical detail, so we don’t mention it outside this footnote.

This theorem makes the idea of “canceling differentials” within integrals rigorous, since it appears as if the dx in the “denominator” of the integrand on the left side is canceled by the dx that is the placeholder of the left side’s integrand¹⁹.

The theorem is often stated with the substitutions $g = f$ and $f = u$:

$$\int f(u(x)) \frac{du(x)}{dx} dx = \int f(u(x)) du(x)$$

Separation of variables

One important application of the change of variables theorem is the solving of so-called *separable differential equations*, which are equations of the form

$$\frac{df(x)}{dx} = f(x)g(f(x)) \text{ for all } x.$$

(In general, a *differential equation* is an equation that involves a function and one or more of its derivatives). To solve the above equation, we will apply the change of variables theorem and “cancel differentials”. Rephrasing the original equation, we have

$$\frac{1}{g(f(x))} \frac{df(x)}{dx} = f(x) \text{ for all } x \text{ such that } g(f(x)) \neq 0.$$

Integrate both sides to obtain

$$\int \frac{1}{g(f(x))} \frac{df(x)}{dx} dx = \int f(x) dx \text{ for all } x \text{ such that } g(f(x)) \neq 0.$$

Now use the change of variables theorem to “cancel the dx ’s on the left side”:

$$\begin{array}{l} \text{The solution to } \frac{df(x)}{dx} = f(x)g(f(x)) \text{ for all } x \text{ is} \\ \int \frac{1}{g(f(x))} df(x) = \int f(x) dx \text{ for all } x \text{ such that } g(f(x)) \neq 0. \end{array}$$

In practice, we (attempt to) solve the separable differential equation by explicitly computing the above integrals and then using algebra to solve the above equation for f .

Mnemonic for separation of variables

The solution to the differential equation

$$\frac{df(x)}{dx} = f(x)g(f(x)) \text{ for all } x$$

can be remembered by the mnemonic of doing algebra to “put the f ’s and $df(x)$ ’s on one side and put the x ’s and dx ’s on the other side”. We divide both sides by $g(f(x))$ to get the f ’s and $df(x)$ ’s on the left side, and then “multiply both sides by dx ” to get the x ’s on the right side:

$$\frac{df(x)}{g(f(x))} = f(x)dx \text{ for all } x \text{ such that } g(f(x)) \neq 0.$$

Now we “integrate both sides” of the above mnemonic expression to obtain

$$\int \frac{1}{g(f(x))} df(x) = \int f(x) dx \text{ for all } x \text{ such that } g(f(x)) \neq 0,$$

exactly as before.

¹⁹A quicker proof, which you might discover if you suspected the primary result to be true is as follows. First, notice that the chain rule implies $\left(\left(\int g\right) \circ f\right)' = (g \circ f)f'$. Then integrate both sides.

More integration formulas

Integration by parts

- Is the “reverse product rule”.
- $(fg)' = f'g + g'f$. Integrate both sides to obtain $fg = \int (f'g + g'f) = \int f'g + \int g'f$. Now subtract one of the integrals on the right side to obtain either

$$\int f'g = fg - \int fg'$$

or

$$\int fg' = fg - \int f'g.$$

It doesn't matter which of the above equations you use, as swapping f and g in one of the equations yields the other. We will consider the second equation,

$$\int fg' = fg - \int f'g.$$

Substitute $h = g'$, so $g = \int h$. Then

$$\int fh = fH - \int f'H, \text{ where } H = \int h.$$

Rewriting the above with the letters f, g , and G , we have

$$\boxed{\int fg = fG - \int f'G, \text{ where } G = \int g}$$

When using the above formula, pick f to be a function whose derivative is computable, and pick G to be a function whose integral is computable.

[insert spaces in integrals]

- We now present the most popular way to state the integration by parts formula. Above, we showed that

$$\int f(x)g'(x) dx = f(x)g(x) - \int g(x)f'(x) dx \text{ for all } x$$

If we use the mnemonics “ $df(x) = f'(x)dx$ ” and “ $dg(x) = g'(x)dx$ ”, we can write this formula as:

$$\int f(x)dg(x) = f(x)g(x) - \int g(x)df(x),$$

where “ $dg(x) = g'(x)dx$ ” and “ $df(x) = f'(x)dx$ ”

Note that the meanings of $df(x)$ and $dg(x)$ in the above are *not* the same as the meanings that make most sense²⁰ Additionally, it is more work to memorize the above formula and the associated meanings of $df(x)$ and $dg(x)$ than it is to derive the above boxed formula from the product rule. Therefore, *please ignore this formula*. It is included only for completeness.

Unfortunately, the above formula is most commonly written with $u := f$ and $v := g$, in the even more sloppy notation “ $\int u dv = uv - \int v du$ ”.

²⁰The definition $\int f(g(x))dg(x) := (\int f)|_{g(x)}$, which was given earlier, is what makes most sense.

Multivariable calculus

References

- *The Calculus of Several Variables* by Robert C Rogers of Nagoya University in Japan: <https://www.math.nagoya-u.ac.jp/~richard/teaching/s2016/Ref2.pdf>
- *Vector Calculus, Linear Algebra, and Differential Forms* by Hubbard and Hubbard
- *Calculus on Manifolds* by Michael Spivak

Differentiation

Introduction

Let's compare two approaches to multivariable calculus:

1. (General differentiability first). Some linear algebra is presented. Partial and directional derivatives of a function $\mathbb{R}^n \rightarrow \mathbb{R}^m$ is defined and general differentiability of functions $\mathbb{R}^n \rightarrow \mathbb{R}^m$ is discussed. Functions $\mathbb{R} \rightarrow \mathbb{R}^n$ and $\mathbb{R}^n \rightarrow \mathbb{R}$ are treated as special cases, and things such as the gradient ∇f are discussed.
2. (Elementary concepts first). Some linear algebra is presented. Functions $\mathbb{R} \rightarrow \mathbb{R}^n$ and $\mathbb{R}^n \rightarrow \mathbb{R}$ are discussed. The directional derivative of a function $\mathbb{R}^n \rightarrow \mathbb{R}^m$ is defined, and general differentiability of functions $\mathbb{R}^n \rightarrow \mathbb{R}^m$ is discussed.

I used to think that general differentiability first was the better way, because the difference between general differentiability and the existence of directional derivatives is subtle enough that it should be established early on. I now better understand this difference, and think it's possible to correctly motivate general differentiability even after one has only been dealing with existence of partial derivatives for a long time.

My increased understanding is due to learning that general differentiability is equivalent to having directional derivatives defined in all directions *and* having these directional derivatives vary linearly with the direction vector. When one knows this, the difference between the existence directional derivatives and general differentiability is clear. One could argue, that to understand the generalized directional derivatives invoked in this fact, one should have more prior experience with “subconcepts” like directional and partial derivatives of scalar-valued functions and scalar derivatives of vector-valued functions!

After further thought, though, I don't really buy this. The definitions of general directional and partial derivatives for functions $\mathbb{R}^n \rightarrow \mathbb{R}^m$ seem pretty immediately understandable:

$$\begin{aligned} \mathbf{D}_{\mathbf{v}}\mathbf{f}|_{\mathbf{p}} &:= \lim_{h \rightarrow 0} \frac{\mathbf{f}(\mathbf{p} + h\mathbf{v}) - \mathbf{f}(\mathbf{p})}{h} \\ \partial_i \mathbf{f} &:= \mathbf{D}_{\mathbf{e}_i} \mathbf{f}. \end{aligned}$$

And when one considers that, since knowledge of some linear algebra (linear transformations, dot product, and cross product) is necessary to study planes, then approaches that centralize linear algebra (instead of using it at the beginning and end, but not middle) should probably be preferred, general differentiability again seems much better.

But I think one argument in favor of elementary concepts first is just so powerful it can't be beaten. It's not that elementary concepts help understanding general differentiability so much, it's that they're *so elementary* it feels unnatural to postpone them for so long! I can't imagine telling someone studying physics that they absolutely must understand general differentiability before studying trajectories or partial derivatives, for example. Those concepts are just too basic to not approach right away.

On linear algebra

Actually, when I remember that integral multivariable calculus requires knowing about the determinant, it seems to me that an introduction to linear algebra should simply be a prerequisite for multivariable calculus. As alluded to before, the syllabus would be:

- \mathbb{R}^n
- Subspaces of \mathbb{R}^n ; span, bases, linear independence

- Linear functions $\mathbb{R}^n \rightarrow \mathbb{R}^m$ and the matrix-vector product (first as a linear combination of columns, then in terms of entries)
 - Some recognition that $\frac{d}{dx}$ and \int are linear functions (but avoid defining vector space!)
- Composition of linear functions and the matrix-matrix product (first as a matrix with columns given by matrix vector products, then in terms of entries)
- The dot product
 - Intuition rather than formal proof should be used to justify facts about rotations and projections used to prove that the algebraic formula implies the geometric formula
- The determinant
- Intuitive presentation of orientation
- The cross product

Outline

First, study functions $\mathbb{R} \rightarrow \mathbb{R}^n$.

- Define derivatives and integrals of functions $\mathbb{R}^n \rightarrow \mathbb{R}$.
- Study space curves and derive the Frenet frame.

Next, study functions $\mathbb{R}^n \rightarrow \mathbb{R}$.

- Define the directional derivative of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ as $\left. \frac{\partial f}{\partial \mathbf{v}} \right|_{\mathbf{p}} := \left. \frac{d(f \circ \mathbf{x})(t)}{dt} \right|_{t_0}$, where \mathbf{x} is a curve satisfying $\mathbf{x}(t_0) = \mathbf{p}$ and $\left. \frac{d\mathbf{x}(t)}{dt} \right|_{t_0} = \mathbf{v}$.
 - Prove that the directional derivative doesn't depend on the curve chosen.
- Prove that the limit expression, $\left. \frac{\partial f}{\partial \mathbf{v}} \right|_{\mathbf{p}} = \lim_{h \rightarrow 0} \frac{f(\mathbf{p} + h\mathbf{v}) - f(\mathbf{p})}{h}$, of the directional derivative, follows.
- For $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $\mathbf{x} = (x_1, \dots, x_n)^\top$, define $\frac{\partial f(\mathbf{x})}{\partial x_i} := \left. \frac{\partial f}{\partial \mathbf{e}_i} \right|_{\mathbf{x}}$ and $\partial_i f := \left(\mathbf{x} \mapsto \frac{\partial f(\mathbf{x})}{\partial x_i} \right)$ where $\mathbf{x} = (x_1, \dots, x_n)$.
 - Aside: for $\mathbf{f} = (f_1, \dots, f_n)^\top : \mathbb{R} \rightarrow \mathbb{R}^n$ and $\mathbf{x} = (x_1, \dots, x_n)^\top$, define $\frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_i} := \left(\frac{\partial f_1(\mathbf{x})}{\partial x_i}, \dots, \frac{\partial f_n(\mathbf{x})}{\partial x_i} \right)^\top$ and $\partial_i \mathbf{f} := \left(\mathbf{x} \mapsto \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_i} \right)$.
- When we have $\mathbf{x} : \mathbb{R} \rightarrow \mathbb{R}^n$ and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ then we can obtain $f \circ \mathbf{x} : \mathbb{R} \rightarrow \mathbb{R}$. Prove the multivariable chain rule, $\frac{d(f \circ \mathbf{x})(t)}{dt} = \sum_{i=1}^n \frac{\partial f(\mathbf{x})}{\partial x_i} \frac{dx_i(t)}{dt}$.
- Define $\nabla_{\mathbf{x}} f(\mathbf{x}) := \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \dots, \frac{\partial f(\mathbf{x})}{\partial x_n} \right)^\top$ and $\nabla f := (\mathbf{x} \mapsto \nabla_{\mathbf{x}} f(\mathbf{x}))$, so that the multivariable chain rule can be restated as $\left. \frac{d(f \circ \mathbf{x})(t)}{dt} \right|_{\mathbf{p}} = (\nabla f)|_{\mathbf{p}} \cdot \left. \frac{d\mathbf{x}(t)}{dt} \right|_{t_0}$.
- Prove that the directional derivative is $\left. \frac{\partial f}{\partial \mathbf{v}} \right|_{\mathbf{p}} = (\nabla f)|_{\mathbf{p}} \cdot \left. \frac{d\mathbf{x}(t)}{dt} \right|_{t_0} = \nabla f|_{\mathbf{p}} \cdot \mathbf{v}$.
- Remark that $(\mathbf{v}, f) \mapsto \left. \frac{\partial f}{\partial \mathbf{v}} \right|_{\mathbf{p}}$ is linear in both arguments.
- Use the fact that $\left. \frac{\partial f}{\partial \mathbf{v}} \right|_{\mathbf{p}} = \nabla f|_{\mathbf{p}} \cdot \mathbf{v}$ to prove that ∇f is the direction of greatest increase in f , and (somewhat surprisingly) that $-\nabla f$ is the direction of greatest decrease in f .

Use the previous results in differential calculus to generalize to functions $\mathbb{R}^n \rightarrow \mathbb{R}^m$.

- Use the chain rule for partial derivatives to prove a new chain rule: $\frac{d(\mathbf{f} \circ \mathbf{x})(t)}{dt} = \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \frac{d\mathbf{x}(t)}{dt}$, where $\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}$ is the matrix whose i th row is $\nabla_{\mathbf{x}} f_i$. Define $\mathbf{Jf}|_{\mathbf{x}} := \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}$.
- Aside on how since the i th column of $\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}$ is $\frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_i}$, we have $\frac{d(\mathbf{f} \circ \mathbf{x})(t)}{dt} = \sum_{i=1}^n \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_i} \frac{dx_i(t)}{dt}$.

- Define the directional derivative of $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ similarly to before, so $\left. \frac{\partial \mathbf{f}}{\partial \mathbf{v}} \right|_{\mathbf{p}} := \left. \frac{d(\mathbf{f} \circ \mathbf{x})}{dt} \right|_{t_0}$, where \mathbf{x} is a curve satisfying $\mathbf{x}(t_0) = \mathbf{p}$ and $\left. \frac{d\mathbf{x}(t)}{dt} \right|_{t_0} = \mathbf{v}$.
- Prove that $\left. \frac{\partial \mathbf{f}}{\partial \mathbf{v}} \right|_{\mathbf{p}}$ exists for all $\mathbf{v} \in \mathbb{R}^n$ iff $\mathbf{Jf}|_{\mathbf{p}}$ exists.
- Prove that if $\mathbf{Jf}|_{\mathbf{p}}$ exists, then $\left. \frac{\partial \mathbf{f}}{\partial \mathbf{v}} \right|_{\mathbf{p}} = \mathbf{Jf}|_{\mathbf{p}} \mathbf{v}$.

Then, talk about local linear approximations.

- (Theorem from Hubbards). (p. 106). $f : \mathbb{R} \rightarrow \mathbb{R}$ is differentiable at $p \in \mathbb{R}$ iff $(\Delta f)_p(h) := f(p+h) - f(p)$ is approximately a linear function, in the sense that f is differentiable at p iff there is a linear function $h \mapsto hf'(p) : \mathbb{R} \rightarrow \mathbb{R}$ such that $\lim_{h \rightarrow 0} \frac{(\Delta f)_p(h) - hf'(p)}{h} = 0$.
- (Defn from Spivak). (p. 16). \mathbf{f} is *differentiable at* $\mathbf{p} \in \mathbb{R}^n$ iff there exists a linear function $\mathbf{Df}|_{\mathbf{p}} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that $\lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{f}(\mathbf{p}+\mathbf{h}) - \mathbf{f}(\mathbf{p}) - \mathbf{Df}|_{\mathbf{p}}(\mathbf{h})\|}{\|\mathbf{h}\|} = 0$. The function $\mathbf{Df}|_{\mathbf{p}}$ is called the *total derivative*.
 - (Original remark). Note that we haven't yet shown $\mathbf{Df}|_{\mathbf{p}}$ is unique.
- Prove that a function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is differentiable at $\mathbf{p} \in \mathbb{R}^n$ iff all directional derivatives at \mathbf{p} exist and iff $\mathbf{v} \mapsto \left. \frac{\partial \mathbf{f}}{\partial \mathbf{v}} \right|_{\mathbf{p}}$ is linear for all $\mathbf{v} \in \mathbb{R}^n$.
- (Theorem from Spivak). (p. 30). If $\mathbf{f} = (f_1, \dots, f_n)^\top$ is differentiable at \mathbf{p} , then $\mathbf{Jf}|_{\mathbf{p}}$ exists.
 - (Remark from Hubbard). If $\mathbf{Df}|_{\mathbf{p}}$ exists then it is unique, since it's represented by a matrix.

For the discussion on differentiability versus continuous differentiability, use Spivak's clear theorems and the Hubbards' examples.

- (Remark from Spivak). (p. 30). If \mathbf{f} 's first partial derivatives all exist at \mathbf{p} then it is not necessarily the case that \mathbf{f} is differentiable at \mathbf{p} .
 - (Example from Hubbard). (p. 122). $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ defined by $f(x, y) = \frac{x^2 y}{x^2 + y^2}$ when $(x, y) \neq (0, 0)$ and $f(x, y) = 0$ when $(x, y) = (0, 0)$ is continuous and has partial derivatives but isn't differentiable at $(0, 0)$.
- (Original). Even if all directional derivatives of \mathbf{f} exist at \mathbf{p} , it is not necessarily the case that \mathbf{f} is differentiable at \mathbf{p} .
 - (Original example). Let $A = \{(x, y) \mid \exists x \ y = x^2\} - \{(0, 0)\}$, and consider the characteristic function $\chi_A : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ of A . Then the directional derivative $\left. \frac{\partial \chi_A}{\partial \mathbf{v}} \right|_{\mathbf{0}}$ of χ_A at $\mathbf{0}$ exists for all $\mathbf{v} \in \mathbb{R}^2$ but χ_A is not differentiable at $\mathbf{0}$.
- (Remark from Hubbards). (p. 123). If \mathbf{f} is differentiable at \mathbf{p} , then $\mathbf{Df}|_{\mathbf{p}}$ is not guaranteed to be a good approximation.
 - (Example from Hubbard). (p. 123). $f : \mathbb{R} \rightarrow \mathbb{R}$ defined by $f(x) = \frac{1}{2}x + x^2 \sin\left(\frac{1}{x}\right)$ when $x \neq 0$ and $f(x) = 0$ when $x = 0$ is continuous and differentiable, but its derivative is not a good approximation of it at the origin, since it does not correctly indicate whether f is increasing or decreasing: "although the derivative at 0 is positive, the function is not increasing in any neighborhood of 0 since on any such interval the derivative is negative."
- (Defn from Spivak). (p. 31). $\mathbf{f} = (f_1, \dots, f_n)^\top$ is *continuously differentiable at* $\mathbf{p} \in \mathbb{R}^n$ iff all partial derivatives $\partial_i f_j$ are continuous at \mathbf{p} and every partial derivative $\partial_i f_j$ is defined in some open set containing \mathbf{p} .

- (Theorem from Spivak). (p. 31). If \mathbf{f} is continuously differentiable at \mathbf{p} then \mathbf{f} is differentiable at \mathbf{p} .

Now we state rules about $\mathbf{Df}|_{\mathbf{p}}$.

- (Theorem from Spivak). (p. 20). Rules about $\mathbf{Df}|_{\mathbf{p}}$.
- (Theorem from Spivak). (p. 32). If $\mathbf{f} = (f_1, \dots, f_m)^\top : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is such that each f_i is continuously differentiable at \mathbf{p} and $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is differentiable at $\mathbf{f}(\mathbf{p})$, then $\partial_i(f \circ \mathbf{f})|_{\mathbf{p}} = \sum_{j=1}^m (\partial_j f)|_{\mathbf{f}(\mathbf{p})} \partial_i(f_j)|_{\mathbf{p}}$.
 - (Original). In the situation of the above theorem, define the notation $\frac{\partial f(\mathbf{f}(\mathbf{x}))}{\partial f_i(\mathbf{x})} := \partial_i f|_{\mathbf{f}(\mathbf{x})}$. With this notation, the above theorem becomes $\left. \frac{\partial(f \circ \mathbf{f})(\mathbf{x})}{\partial x_i} \right|_{\mathbf{p}} = \sum_{j=1}^m \left. \frac{\partial f(\mathbf{f}(\mathbf{x}))}{\partial f_j(\mathbf{x})} \right|_{\mathbf{p}} \left. \frac{\partial f_j(\mathbf{x})}{\partial x_i} \right|_{\mathbf{p}}$.

Integration

Theorem 1. (Fubini's theorem).

On p. 61 of *Calculus on Manifolds*, Spivak notes that if $C \subseteq A \times B$, then

$$\int_C f = \int_{A \times B} \chi_C f = \int_A \left(\int_B \chi_C(x, y) f(x, y) dy \right) dx,$$

and that the “main difficulty in deriving expressions for $\int_C f$ will be determining $C \cap (\{x\} \times B)$ for $x \in A$ ”. [I think that by “determining $C \cap (\{x\} \times B)$ for $x \in A$ ”, Spivak means “determining inequalities in x, y that describe the $C \cap (\{x\} \times B)$.”]

Derivation 2. (Change of variables).

Let $U, V \subseteq \mathbb{R}^n$ be open balls in \mathbb{R}^n and let $\mathbf{F} : U \subseteq \mathbb{R}^n \rightarrow V \subseteq \mathbb{R}^n$ be a sufficiently differentiable invertible function. Consider the integral

$$\int_V f \, d\text{vol},$$

where $f : V \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ is also sufficiently differentiable.

Now partition V into subsets $\{V_i\}_{i=1}^n$ and select a point \mathbf{y}_i from each V_i . We have

$$\int_V f \, d\text{vol} \approx \sum_{i=1}^n f(\mathbf{y}_i) \text{vol}(V_i).$$

Since \mathbf{F} is invertible, there are \mathbf{x}_i and subsets $\{U_i\}_{i=1}^n$, with $\mathbf{x}_i \in U_i$, partitioning U for which $\mathbf{y}_i = \mathbf{F}(\mathbf{x}_i)$ and $V_i = \mathbf{F}(U_i)$. The thus above becomes

$$\sum_i f(\mathbf{F}(\mathbf{x}_i)) \text{vol}(\mathbf{F}(U_i)).$$

Since the Jacobian is the best linear approximation to \mathbf{F} , when \mathbf{x} is close to \mathbf{x}_i , we have $\mathbf{F}(\mathbf{x}) \approx \mathbf{F}(\mathbf{x}_i) + \frac{\partial \mathbf{F}}{\partial \mathbf{x}_i}(\mathbf{x} - \mathbf{x}_i)$. Since translating a region preserves its volume, the image of each U_i under this approximation has²¹ the volume $\text{vol}\left(\frac{\partial \mathbf{F}}{\partial \mathbf{x}_i}\Big|_{U_i}\right)$, and so the above is approximately equal to

$$\sum_i f(\mathbf{F}(\mathbf{x}_i)) \text{vol}\left(\frac{\partial \mathbf{F}}{\partial \mathbf{x}_i}\Big|_{U_i}\right) = \sum_i f(\mathbf{F}(\mathbf{x}_i)) \det\left(\frac{\partial \mathbf{F}}{\partial \mathbf{x}_i}\right) \text{vol}(U_i) \approx \int_U f(\mathbf{F}(\mathbf{x})) \det\left(\frac{\partial \mathbf{F}}{\partial \mathbf{x}}\right) d\text{vol}.$$

Thus

$$\int_V f \, d\text{vol} = \int_U f(\mathbf{F}(\mathbf{x})) \det\left(\frac{\partial \mathbf{F}}{\partial \mathbf{x}}\right) d\text{vol}$$

Vector calculus

Definition 3. (Divergence).

Let S be a closed 3-dimensional surface containing $\mathbf{p} \in \mathbb{R}^3$.

$$\text{div}(\mathbf{F})|_{\mathbf{p}} := \lim_{\text{vol}(S) \rightarrow 0} \frac{\int_{\partial S} \mathbf{F} \cdot \mathbf{n} dA}{\text{vol}(S)}.$$

Definition 4. (Curl).

Let S be a closed 2-dimensional surface containing $\mathbf{p} \in \mathbb{R}^3$.

$$\text{curl}(\mathbf{F})|_{\mathbf{p}} := \lim_{\text{vol}(S) \rightarrow 0} \frac{\int_{\partial S} \mathbf{F} \cdot d\mathbf{r}}{\text{vol}(S)}.$$

²¹ $\text{vol}(\mathbf{F}(\mathbf{x}_i) + \frac{\partial \mathbf{F}}{\partial \mathbf{x}_i}(\mathbf{x} - \mathbf{x}_i)\Big|_{U_i - \mathbf{x}_i}) = \text{vol}(\mathbf{F}(\mathbf{x}_i) + \frac{\partial \mathbf{F}}{\partial \mathbf{x}_i}\Big|_{U_i}) = \text{vol}\left(\frac{\partial \mathbf{F}}{\partial \mathbf{x}_i}\Big|_{U_i}\right).$

Theorem 5. (Divergence formula).

$$\operatorname{div}(\mathbf{F}) = \nabla \cdot \mathbf{F}.$$

Proof. <http://www.supermath.info/CalculusIIIvectorcalculus2011.pdf>
Proof of 2-dimensional case. □

Theorem 6. (Curl formula).

$$\operatorname{curl}(\mathbf{F}) = \nabla \times \mathbf{F}.$$

Proof. <http://www.supermath.info/CalculusIIIvectorcalculus2011.pdf> □

Definition 7. (Irrotational vector field).

A vector field \mathbf{F} is called *irrotational* iff $\nabla \times \mathbf{F} = \mathbf{0}$.

Theorem 8. (Irrotational vector field).

A vector field \mathbf{F} is irrotational iff the following equivalent conditions hold:

1. $\int_C \mathbf{F} \cdot d\mathbf{r}$ is independent of C
2. If C is a closed curve then $\int_C \mathbf{F} \cdot d\mathbf{r} = 0$.
3. There exists V such that $\mathbf{F} = -\nabla V$

Remark 9. (Irrotational equals conservative).

A vector field is called *conservative* iff for all closed curves C we have $\int_C \mathbf{F} \cdot d\mathbf{r} = 0$. The above theorem shows that the set of irrotational vector fields is equal to the set of conservative vector fields.

Definition 10. (Solenoidal vector field).

A vector field \mathbf{F} is called *solenoidal* iff $\nabla \cdot \mathbf{F} = 0$.

Theorem 11. (Solenoidal vector field).

A vector field \mathbf{F} is solenoidal iff the following equivalent conditions hold:

1. $\int_S \mathbf{F} \cdot \hat{\mathbf{n}} dA$ is independent of S if ∂S is fixed
2. If S is a closed surface then $\int_S \mathbf{F} \cdot \hat{\mathbf{n}} dA = 0$.
3. There exists \mathbf{A} such that $\mathbf{F} = \nabla \times \mathbf{A}$

Appendix

The power rule for real number exponents

Unfortunately, the following proof requires knowledge of the calculus of exponential and logarithmic functions, which one does not know when they first encounter the power rule.

By the change of base for exponential functions, we have $x^c = e^{c \ln(x)}$. Thus

$$\frac{d}{dx} x^c = \frac{d}{dx} e^{c \ln(x)} = e^{c \ln(x)} \cdot \frac{c}{x} = x^c \cdot \frac{c}{x} = cx^{c-1}.$$

Single-variable calculus derivative notation

There are many types of derivative notation used in single-variable calculus. In this book, I wanted to establish notation for single-variable derivatives that is as intuitive and unambiguous as possible. The first criterion I prioritized for derivative notation was:

1. The notation must be some sort of Leibniz notation, since the chain rule is most elegantly stated when it appears to be a cancellation of fractions.

Without further ado, the first candidate notation is:

$$\begin{aligned}\frac{df}{dx} &:= f' \\ \frac{dg}{df} &:= g' \circ f \text{ when } f \text{ is a function.}\end{aligned}$$

The chain rule is stated pleasingly in this notation as

$$\frac{dg(f)}{dx} = \frac{dg}{df} \frac{df}{dx}.$$

(Recall that $g(f)$ is alternative notation for $g \circ f$).

Unfortunately, this notation does have a problem. Recall the inverse derivative rule in prime notation:

$$(f^{-1})'(x) = \frac{1}{(f' \circ f^{-1})(x)}.$$

If we rewrite this rule in our first candidate notation, we obtain

$$\frac{df^{-1}}{dx} = \frac{1}{\frac{df}{df^{-1}}}.$$

Making the intuitive definition $\frac{dx}{df^{-1}} := \frac{1}{\frac{df}{df^{-1}}}$ and taking the reciprocal of both sides of the above then yields the result

$$\frac{dx}{df^{-1}} = \frac{df}{df^{-1}},$$

which is not at all intuitive.

The second candidate notation is:

$$\frac{dg(f(x))}{df(x)} := g'(f(x)).$$

With this notation, the chain rule is

$$\frac{dg(f(x))}{dx} = \frac{dg(f(x))}{df(x)} \frac{df(x)}{dx},$$

which is again pleasing, although more cluttered than the statement of the chain rule in the first notation. The inverse derivative rule in the second notation is

$$\frac{df^{-1}(x)}{dx} = \frac{1}{\frac{df(f^{-1}(x))}{df^{-1}(x)}}.$$

If, similarly to before, we make the intuitive definition $\frac{dx}{df^{-1}(x)} := \frac{1}{\frac{df^{-1}(x)}{dx}}$ and take the reciprocal of both sides of the above, we now get an intuitive result:

$$\frac{dx}{df^{-1}(x)} = \frac{df(f^{-1}(x))}{df^{-1}(x)}.$$

This result is intuitive and easily remembered because the “numerators” of each side (x and $f(f^{-1}(x))$) are equal.

In addition to the above, I considered two more criteria when evaluating notations.

2. The notation must be compatible with the preferred derivative notation for the i th partial derivative of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, which is $\frac{\partial f(x_1, \dots, x_i, \dots, x_n)}{\partial x_i}$ and *not* $\frac{\partial f}{\partial x_i}$. (The former is preferred because it requires no convention to establish that x_i is to be treated as the i th argument of f , and thus works better when dealing with concrete examples: the meaning of $\frac{\partial f(a, b, c)}{\partial a}$ is clear, while the meaning of $\frac{df}{da}$ requires context.)
3. The notation should allow for statements involving functions rather than functions evaluated on inputs.

The first notation satisfies (3) but not (2), and the second notation satisfies (2) but not (3). Since the first notation also has the extremely glaring issue regarding the inverse derivative rule discussed above, the second notation is to me the clear choice.

Conflation of functions

Suppose you are doing some calculations in which you know some quantity f as a function of position x and position x as a function of time t . Formally, you have a function f_x that maps position to quantity and a function x_t that maps time to position. From these functions, we can obtain a function f_t that maps time to quantity: we have $f_t(t) := f_x(x_t(t))$, i.e. $f_t = f_x \circ x_t$.

In this sort of situation, where it is clear from context that the particular letter of the alphabet x is associated with the input of f , the following definitions are implicitly understood:

$$\begin{aligned} \frac{df(x)}{dx} &:= \frac{df_x(x)}{dx} \\ \frac{df(t)}{dt} &:= \frac{df_t(t)}{dt} = \frac{df_x(x(t))}{dt} = \frac{df_x(x(t))}{dx(t)} \frac{dx(t)}{dt}. \end{aligned}$$

And, outside this book “in the wild”, where it is common to see $\frac{df}{dx}$ meaning both f' and $f'(x)$, there are these additional implicitly understood definitions:

$$\begin{aligned} \frac{df}{dx} &:= f'_x \text{ or } f'_x(x) \\ \frac{df}{dt} &:= f'_t \text{ or } f'_t(t) \end{aligned}$$

That is, the meaning of $\frac{df(\alpha)}{d\alpha}$ and $\frac{df}{d\alpha}$ depend on what letter of the alphabet is plugged in for α .

The usefulness of this convention is that, since it encourages thinking of functions first and leaving their inputs as an afterthought, it facilitates creativity in what exactly a function’s input is imagined to be. For instance, in physics, the creative choice to think of velocity as being a function of position (rather than of time, which is more typical) lets us reinterpret force $F = ma = m \frac{dv}{dt}$ as $F = m \frac{dv}{dx} \frac{dx}{dt}$, which is very convenient for computing the integral of force with respect to displacement:

$$\int F \, dx = \int m \frac{dv}{dx} \frac{dx}{dt} dx = \int m \frac{dv}{dx} v \, dx = \int m v \frac{dv}{dx} dx = \int m v \, dv = \frac{1}{2} m v^2.$$

We will not use this convention in this book because it is confusing. But, being an extremely common convention in physics, it deserves mention.

Properties of exponential functions

[the sections regarding exponential functions and e draw heavily from <https://paramanands.blogspot.com/2014/05/theories-of-exponential-and-logarithmic-functions-part-3.html#.YSz6iN901PY>]

Show $b^{x+y} = b^x b^y$ and that $x \mapsto b^x$ is continuous. For continuity, it suffices to show continuity at $x = 0$ and then use $b^{x+y} = b^x b^y$.

Justification that e exists and other related technical facts

Here, we prove that there is a real number $b > 0$ for which $\lim_{h \rightarrow 0} \frac{b^h - 1}{h} = 1$. Knowing this fact justifies defining e to be this number.

In order to prove e exists, we will prove the following three facts:

1. $\lim_{h \rightarrow 0} \frac{b^h - 1}{h}$ converges for all nonzero real numbers b .
2. The function defined by $f(b) := \lim_{h \rightarrow 0} \frac{b^h - 1}{h}$ is invertible. In other words,
 - 2.1. f is one-to-one.
 - 2.2. For every real number $y > 0$ there exists a real number $b > 0$ such that $y = f(b)$.
3. $e = f^{-1}(1) > 0$.

Before we prove (1), (2), or (3), first note that that $f(1) = 0$ and that f is continuous wherever it is defined²².

Now we prove that these facts are true.

1. $(\lim_{h \rightarrow 0} \frac{b^h - 1}{h} \text{ converges for all nonzero real numbers } b)$.

Clearly, the limit does not exist when $b = 0$.

First, we will show that the left-sided limit $\lim_{h \rightarrow 0^-} \frac{b^h - 1}{h}$ exists. Due to the monotone convergence theorem, we know that the left-sided limit must exist if it is the case that for all $b \neq 0$ the function $h \mapsto \frac{b^h - 1}{h}$ is bounded and monotone on $(0, 1) - \{0\}$. In the below we prove that this is indeed the case.

- (Monotonicity on $(0, 1) - \{0\}$). Using $a = h, r = \frac{1}{n}, s = \frac{1}{n+1}$ in inequality (2), we see that for $b > 0$ the function $n \mapsto n(b^{\frac{1}{n}} - 1)$ is decreasing. Substituting $h = \frac{1}{n}$ shows that $h \mapsto \frac{b^h - 1}{h}$ is an increasing function for all $b > 0$. A similar argument shows that $h \mapsto \frac{b^h - 1}{h}$ is a decreasing function of h for all $b < 0$. Thus $h \mapsto \frac{b^h - 1}{h}$ is monotone on $(0, 1) - \{0\}$.
- (Boundedness on $(0, 1)$). From inequality (11), we know that $b^{h-1}(b-1) < \frac{b^h - 1}{h} < b - 1$ for all $b > 1, h \in (0, 1)$. Thus g is bounded on $(0, 1)$.

Now we show that the the left-sided limit is the same as the right-sided limit. Consider the left sided limit $\lim_{h \rightarrow 0^-} \frac{b^h - 1}{h}$, and substitute $k = -h$ so that $k \rightarrow 0^+$. We have

$$\begin{aligned} \lim_{h \rightarrow 0^-} \left(\frac{b^h - 1}{h} \right) &= \lim_{k \rightarrow 0^+} \left(\frac{b^{-k} - 1}{-k} \cdot \frac{b^k}{b^k} \right) = \lim_{k \rightarrow 0^+} \left(\frac{1 - b^k}{-k b^k} \right) = \lim_{k \rightarrow 0^+} \left(\frac{b^k - 1}{k b^k} \right) \\ &= \frac{\lim_{k \rightarrow 0^+} \left(\frac{b^k - 1}{k} \right)}{\lim_{k \rightarrow 0^+} (b^k)} = \frac{\lim_{k \rightarrow 0^+} \left(\frac{b^k - 1}{k} \right)}{1} = \lim_{h \rightarrow 0^+} \left(\frac{b^h - 1}{h} \right). \end{aligned}$$

2. (f is invertible).

Recall from the above that f is invertible if

- 2.1. f is one-to-one.
- 2.2. For every real number $y > 0$ there exists a real number $b > 0$ such that $y = f(b)$.

²² f is continuous wherever it is defined because $b \mapsto \frac{b^h - 1}{h}$ is a composition of continuous functions of b and because $\lim_{b \rightarrow b_0}$ and $\lim_{h \rightarrow 0}$ commute.

We already know that f is one-to-one because we showed that f is monotone on our way to showing (1). So, we just need to show (2.2).

In order to show (2.2), we'll need to first glean some inspiration from wishful thinking: if it *were* true that f was invertible, then we could follow the discussion of the earlier section "Derivatives of exponential functions" to obtain that $\frac{d}{dx} b^x = f(b)b^x = \log_e(b)b^x$, where $e := f^{-1}(1)$. So, we would have $f(b) = \log_e(b)$ for all nonzero b , i.e. $f = \log_e$.

Now, we don't *actually* know that f is invertible, so we can't just assume that $f = \log_e$. However, what we *can* do is try to show that f satisfies properties of logarithms, and then try to use those properties to show that (2.2) is true.

It's pretty straightforward to show that $f(ab) = f(a) + f(b)$ and $f(\frac{1}{b}) = -f(b)$ by using properties of limits. From this it follows that $f(b^x) = xf(b)$ for integer x .

Now, we will use the fact that $f(b^x) = xf(b)$ for integer x to show that $\lim_{b \rightarrow \infty} f(b) = \infty$ and $\lim_{b \rightarrow 0^+} f(b) = -\infty$. Once we know this, we are done, since if $\lim_{b \rightarrow 0^+} f(b) = -\infty$, $\lim_{b \rightarrow \infty} f(b) = \infty$, and as f is continuous, then the intermediate value theorem guarantees that for every real number $y > 0$ there is a real number $b > 0$ such that $y = f(b)$.

- If we show that $\lim_{b \rightarrow \infty} f(b) = \infty$, then $\lim_{b \rightarrow 0^+} f(b) = -\infty$ follows because $f(\frac{1}{b}) = -f(b)$. Before we show $\lim_{b \rightarrow \infty} f(b) = \infty$, we need to show that $b \mapsto f(b)$ is increasing for $b > 1$.
 – ($b \mapsto f(b)$ is increasing for $b > 1$). If $1 < a < b$ then $\frac{a}{b} < 1$, so $f(\frac{a}{b}) = f(a) - f(b) < 0$, which gives $f(a) < f(b)$.
- ($\lim_{b \rightarrow \infty} f(b) = \infty$). Let $L > 0$. We need to show that there is a b such that $f(b) > L$.
 Choose any $a > 1$ so that $f(a) > 0$. Then $L = nf(a)$ for some integer $n > 0$. I.e., $L = nf(a) = f(a^n)$. Since f is increasing for $b > 1$, and as $a^n > a > 1$, any $b > a^n$ will work: $f(b) > f(a^n) = L$.

3. To show that $f^{-1}(1) > 0$, it suffices to show that $f(b) > 0$ when $b > 1$.

Since we know $b^{h-1}(b-1) < \frac{b^h-1}{h} < (b-1)$ from inequality (11), we can take limits of both sides and then use the continuity of $x \mapsto b^x$ to obtain $\frac{b-1}{b} < f(b) < b-1$. This shows that $f(b) > 0$ when $b > 1$.

Inequalities

Let $a > 1$ be a real number and let $r > 0$ be an integer. Since $a^r > a^i$ for each i , we have $ra^r > 1 + a + a^2 + \dots + a^{r-1}$. Multiply both sides by $a-1 > 0$ and simplify to see $ra^r(a-1) > a^r - 1$. Add $r(a^r - 1)$ to both sides and factor to get $r(a^{r+1} - 1) > (r+1)(a^r - 1)$. Thus $\frac{a^{r+1}-1}{r+1} > \frac{a^r-1}{r}$. A similar argument shows $\frac{1-b^{r+1}}{r+1} < \frac{1-b^r}{r}$ for all real numbers $b \in (0, 1)$. In all,

$$\frac{a^{r+1}-1}{r+1} > \frac{a^r-1}{r} \text{ and } \frac{1-b^{r+1}}{r+1} < \frac{1-b^r}{r}, \quad (1)$$

for all real numbers a, b with $a > 1, b \in (0, 1)$ and for all integers r, s with $0 < s < r$.

It follows that if r, s are positive integers with $r > s$ then

$$\frac{a^r-1}{r} > \frac{a^s-1}{s} \text{ and } \frac{1-b^r}{r} < \frac{1-b^s}{s}, \quad (2)$$

for all real numbers a, b with $a > 1$ and $b \in (0, 1)$.

These inequalities can be extended to hold for rational r, s such that $0 < s < r$.

Use $s = 1$ in each equation from (2) to obtain the first line of the below, and $r = 1$ in each equation from (2) to obtain the second line of the below:

$$a^r - 1 > r(a - 1) \text{ and } 1 - b^r < r(1 - b) \text{ for } r > 1 \quad (3)$$

$$a^s - 1 < s(a - 1) \text{ and } 1 - b^s > s(1 - b) \text{ for } s \in (0, 1). \quad (4)$$

Again consider $a > 1$ and $b \in (0, 1)$. Then $\frac{1}{a} \in (0, 1)$ and $b > 1$, so we can substitute $b = \frac{1}{a}$ and $a = \frac{1}{b}$ into the previous two inequalities. After simplifying, we obtain

$$a^r - 1 < ra^{r-1}(a - 1) \text{ and } 1 - b^r > rb^{r-1}(1 - b) \quad (5)$$

$$a^s - 1 > sa^{s-1}(a - 1) \text{ and } 1 - b^s < sb^{s-1}(1 - b) \quad (6)$$

for all real numbers a, b with $a > 1, b \in (0, 1)$ and all rational r, s with $0 < s < r$.

Combining together the inequalities from (3), (4), (5), (6), (e.g. combine the first inequality of (3), which involves $a^r - 1$, with the first equation of (5), which also involves $a^r - 1$), we obtain

$$ra^{r-1}(a-1) > a^r - 1 > r(a-1) \text{ and } sa^{s-1}(a-1) < a^s - 1 < s(a-1) \quad (7)$$

$$rb^{r-1}(1-b) < 1 - b^r < r(1-b) \text{ and } sb^{s-1}(1-b) > 1 - b^s > s(1-b). \quad (8)$$

for all $a, b \in \mathbb{R}$ with $a \in (0, 1), b > 1$ and for all rational r, s with $r > 1, s \in (0, 1)$.

Multiplying each side of these inequalities by -1 , we have

$$ra^{r-1}(1-a) < 1 - a^r < r(1-a) \text{ and } sa^{s-1}(1-a) > 1 - a^s > s(1-a) \quad (9)$$

$$rb^{r-1}(b-1) > b^r - 1 > r(b-1) \text{ and } sb^{s-1}(b-1) < b^s - 1 < s(b-1). \quad (10)$$

for all $a, b \in \mathbb{R}$ with $a \in (0, 1), b > 1$ and for all rational r, s with $r > 1, s \in (0, 1)$.

Dividing the second equation of (10) by s , we have in particular that

$$b^{s-1}(b-1) < \frac{b^s - 1}{s} < b - 1 \text{ for all real } b > 1 \text{ and rational } s \in (0, 1). \quad (11)$$