MATH 321-2: MENU Real Analysis
Northwestern University, Lecture Notes

Written by Santiago Cañez

These are notes which provide a basic summary of each lecture for MATH 321-2, the second quarter of "MENU Real Analysis", taught by the author at Northwestern University. The book used as a reference is the 3rd edition of *Principle of Mathematical Analysis* by Rudin. Watch out for typos! Comments and suggestions are welcome.

Contents

Lecture 1: Darboux Sums

This course continues the study of real analysis, with the main objective being to generalize concepts you saw last quarter for $\mathbb{R}$ to other settings. After finishing one-dimensional analysis on $\mathbb{R}$ by discussing *integration*, we will switch gears to develop analysis on *function spaces*, which are spaces whose elements (or "points") are functions. Here the notion of what's called *uniform convergence* will be of key importance, and will culminate in the theory of analytic functions and (a piece of) the theory of Fourier series. Finally we will develop analysis—at least the "differential" part of analysis–on $\mathbb{R}^n$, with the main goals being to prove the *inverse* and *implicit function theorems*, which are arguably the most important (at least conceptually) theorems in all of analysis, perhaps after the Bolzano-Weierstrass theorem.

Where are we at? To put the importance of our first topic, namely integration, in the proper context, let us clarify the importance of two concepts from last quarter: continuity and differentiability. Both of these amount to giving us ways of controlling the growths of functions and understanding how well or how poorly they behave. (By "growth" here we mean information about the change $f(x) - f(y)$ in function values in terms of $x - y$.) Continuity gives us "abstract" control; that is, we can control $f(x) - f(y)$ by controlling $x - y$, but we do not have information about how to do so explicitly: given $\epsilon > |f(x) - f(y)|$ we can find $\delta > |x - y|$, but we cannot say in general what δ looks like. Continuity is really telling us about the *existence* of control.

Differentiability then gives us more explicit information about how to control the growth of f, in particular via the mean value theorem:

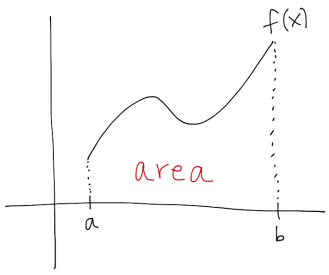
$$f(x) - f(y) = f'(c)(x - y).$$

We can express $f(x) - f(y)$ explicitly in terms of $x - y$ and a derivative term, and information about the derivative can be turned into explicit information about f. As we introduce higher orders of differentiability, we get even more explicit information about the growth of f via Taylor's theorem:

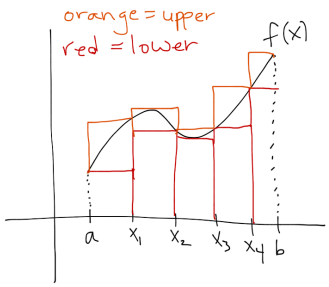
$$f(x) - f(y) = f'(y)(x - y) + \frac{f''(c)}{2}(x - y)^2, \text{ and so on.}$$

Integrability then goes in the other direction, and suggests that we give up trying to control the growth of f, at least everywhere. At points where f is continuous we already have a measure of control, and at points where f is not continuous, integrability amounts to saying that we can make such points "negligible" so that the behavior of f at such points can be ignored. This is the point of view we will work towards, and is one you will develop even further next quarter when discussing *measure theory*.

Motivation for integration. An integral should give us the area of the region under the graph of a function:



To compute this area, we approximate it using areas of rectangles as follows. First, we divide the interval $[a, b]$ into smaller pieces. Over each of these, we take a rectangle of height equal to the infimum of f over that piece, and a rectangle of height equal to the supremum of f over that piece:



The actual area we want is sandwiched between the sum of the areas of the "lower" rectangles and the sum of the areas of the "upper" rectangles. The idea is that by considering all possible such sums corresponding to all possible ways of breaking up $[a, b]$ into smaller pieces, we can get better and better approximations to the area we want. The goal is now to make this all precise.

Darboux sums. A *partition* P of $[a, b]$ is a collection of points $x_0, \ldots, x_n$ such that

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b.$$

The practical point is that this "breaks" the interval $[a, b]$ up into the smaller intervals

$$[x_0, x_1],\ [x_1, x_2], \ldots, [x_{n-2}, x_{n-1}],\ \text{and}\ [x_{n-1}, x_n].$$

For a bounded function f on $[a, b]$ and a partition P of $[a, b]$, the *lower Darboux sum* $L(f, P)$ is (letting I_k denote the k-th subinterval $[x_{k-1}, x_k]$ determined by the partition):

$$L(f, P) = \sum_{I_k} (\inf f \text{ over } I_k)(\text{length of } I_k)$$

and the *upper Darboux sum* $U(f, P)$ is

$$U(f, P) = \sum_{I_k} (\sup f \text{ over } I_k)(\text{length of } I_k).$$

Graphically, $L(f, P)$ is the sum of the areas of the "lower" rectangles in the above picture and $U(f, P)$ is the sum of the areas of the "upper" rectangles.

The sums we have defined are often referred to as lower and upper *Riemann* sums as well, but historically this approach to integration was actually developed by Darboux some number of years after Riemann gave his original approach. The two approaches are equivalent, and we will say something about Riemann's approach a bit later, but Darboux's approach is typically easier to work with, which is why modern analysis books describe it first.

Example. Suppose that $f(x) = c$ is a constant function on $[a, b]$. Then for any partition P of $[a, b]$, the supremum of f over any smaller subinterval is always c, so

$$U(f, P) = \sum_{I_k} (\sup f \text{ over } I_k)(\text{length of } I_k) = \sum_{I_k} c(\text{length of } I_k) = c \sum_{I_k} (\text{length of } I_k).$$

But the intervals I_k together make up all of $[a, b]$, so adding together their lengths gives the length of $[a, b]$. Thus

$$U(f, P) = c(b - a), \text{ and similarly } L(f, P) = c(b - a)$$

since the infimum of f over any smaller interval is also always c.

Note that this makes sense graphically: the graph of a constant function is a horizontal line, and any "lower" or "upper" rectangles we use should cover the entire region under the graph, which has area $c(b - a)$.

Another example. Consider the function f on $[0, 1]$ defined by

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q}. \end{cases}$$

No matter what partition P of $[0, 1]$ we take, the supremum of f over any subinterval is 1 since any subinterval contains a rational and the infimum of f over any subinterval is 0 since any subinterval contains an irrational. This means that

$$U(f, P) = 1 \text{ and } L(f, P) = 0$$

for any partition P of $[0, 1]$.

Nontrivial example. We compute explicitly the upper and lower sums of $f(x) = x$ on the interval $[0, b]$ determined by the partition P_n given by the points $x_k = \frac{kb}{n}$:

$$0 < \frac{b}{n} < \frac{2b}{n} < \cdots < \frac{(n-1)b}{n} < b.$$

The point of this partition is that all the partition points are evenly spaced, so that all subintervals $I_k = [x_{k-1}, x_k]$ have the same length $\frac{b}{n}$; this will be important in finding explicit values for the upper and lower sums.

Since $f(x) = x$ is strictly increasing, its supremum on $I_k = [x_{k-1}, x_k]$ is $f(x_k) = x_k$ and its infimum on I_k is $f(x_{k-1}) = x_{k-1}$. Thus

$$\begin{aligned} U(f, P_n) &= \sum_{I_k} (\sup f \text{ over } I_k)(\text{length of } I_k) \\ &= \sum_{k=1}^{n} \left(\frac{kb}{n}\right)\left(\frac{b}{n}\right) = \frac{b^2}{n^2} \sum_{k=1}^{n} k = \frac{b^2 n(n+1)}{2n^2} = \frac{b^2(n+1)}{2n} \end{aligned}$$

and

$$\begin{aligned} L(f, P_n) &= \sum_{I_k} (\inf f \text{ over } I_k)(\text{length of } I_k) \\ &= \sum_{k=1}^{n} \left(\frac{(k-1)b}{n}\right)\left(\frac{b}{n}\right) = \frac{b^2}{n^2} \sum_{k=1}^{n} (k-1) = \frac{b^2(n-1)n}{2n^2} = \frac{b^2(n-1)}{2n}, \end{aligned}$$

where we have used the fact that $1 + 2 + \cdots + \ell = \frac{\ell(\ell+1)}{2}$. Again, note that we were only able to compute this explicitly due to the fact that all subintervals in P_n had the same length.

Definition. The *upper Riemann/Darboux integral* of f over $[a, b]$ is

$$U(f) := \inf\{U(f, P) \mid P \text{ is a partition of } [a, b]\}$$

and the *lower Riemann/Darboux integral* of f over $[a, b]$ is

$$L(f) := \sup\{L(f, P) \mid P \text{ is a partition of } [a, b]\}.$$

Again, the intuition is that all upper sums overestimate the area under the graph of f, so this area should be $\leq$ the infimum of all upper sums, and all lower sums underestimate the area under the graph of f, so this area should be $\geq$ the supremum of all lower sums.

We say that f is *(Riemann/Darboux) integrable* over $[a, b]$ when the upper Darboux integral and lower Darbous integrals agree, in which case we call this common value the *integral* of f over $[a, b]$:

$$\int_{[a,b]} f := U(f) = L(f).$$

To say that the upper and lower integrals are the same means precisely the area under the graph of f is well-defined. Another common notation for the integral is, of course, $\int_a^b f(x)\,dx$.

Back to previous examples. In the constant function example, since all upper sums equal $c(b - a)$ the upper integral is $c(b - a)$, and since all lower sums equal to $c(b - a)$ the lower integral is also $c(b - a)$. Thus a constant function is integrable and

$$\int_a^b c\,dx = c(b - a),$$

which is the expected area under the graph of $f(x) = c$.

The function f which is 1 at rationals and 0 at irrationals has all upper sums equal to 1 and all lower sums equal to 0, so the upper integral is 1 and the lower integral is 0. Thus

$$U(f) \neq L(f),$$

so f is not integrable over $[0, 1]$. This means that the area under the graph of f is not well-defined, at least if we restrict ourselves to using finitely many rectangles to try to estimate this area. (Measure theory next quarter will give a way to "integrate" this function in a more general sense.)

For $f(x) = x$ on $[0, b]$, the computations from before give

$$\frac{b^2(n - 1)}{2n} = L(f, P_n) \leq \frac{b^2}{2} \leq U(f, P_n) = \frac{b^2(n + 1)}{2n}.$$

Since both $L(f, P_n)$ and $U(f, P_n)$ actually converge to $\frac{b^2}{2}$ as n increases, this suggests the the lower integral and upper integral of $f(x) = x$ over $[0, b]$ should both equal $\frac{b^2}{2}$, meaning that f is integrable over $[0, b]$ with integral equal to:

$$\int_0^b x\,dx = \frac{b^2}{2}.$$

Now, of course we know from calculus that this is absolutely true, but this is not something we can fully conclude just yet.

The problem is that the upper integral is supposed to be the infimum of *all* possible upper sums and the lower integral the supremum of *all* possible lower sums, and so far we only know these sums for the special partitions P_n where all partition points are evenly spaced. Knowing that the

infimum of the specific upper sums $U(f, P_n)$ is $\frac{b^2}{2}$ is not enough (yet) to say that the infimum of all possible upper sums is also $\frac{b^2}{2}$. Similarly, knowing that the supremum of the specific lower sums $L(f, P_n)$ is $\frac{b^2}{2}$ is by itself not enough to conclude that the lower integral has this same value.

This illustrates a problem with using the upper and lower integrals to check for integrability: for most random partitions P, the values of $U(f, P)$ and $L(f, P)$ are simply impossible to compute directly, and hence it is not feasible that we can directly find the supremum of all lower sums and the infimum of all upper sums. We need another way to test for integrability which avoids having to check all possible partitions. Fortunately, there is such a method, as we will discuss next time.

Riemann-Stieltjes integration. The length of an interval $[x_{i-1}, x_i]$ used in an upper or lower sum is simply

$$x_i - x_{i+1}.$$

If we instead allow for some type of "weighted" length, we get a slightly more general type of integral. To be clear, for an increasing function $\alpha : [a, b] \to \mathbb{R}$ we take the "weighted length" of $[x_{i-1}, x_i]$ to be change in α given by

$$\Delta\alpha_i := \alpha(x_i) - \alpha(x_{i-1}).$$

With this modified "length", we get modified lower and upper sums

$$L(f, P, \alpha) = \sum(\inf f)\Delta\alpha_i \quad \text{and} \quad U(f, P, \alpha) = \sum(\sup f)\Delta\alpha_i.$$

The original Darboux sums we considered were the case where $\alpha(x) = x$ is the identity function. We can then take supremums of lower sums and infimums of upper sums to get lower and upper integrals, and define a new notion of integrability—called *Riemann-Stieltjes integrability with respect to* α—from this. The Riemann-Stieltjes integral of f with respect to α, when it exists, is then denoted by

$$\int_a^b f \, d\alpha.$$

The Riemann-Stieltjes integral gives a useful way to study different types of "summations" from a single point of view, as we will briefly clarify later, but it will not play a major role for us this quarter. All properties and proofs we will give for the original Riemann/Darboux integral for the most part work exactly the same as for the Riemann-Stieltjes integral, so we are not losing much generality by mainly sticking with the $\alpha =$ identity function case, at lease for the purposes of this course. The Riemann-Stieltjes integral is more important in certain applications to probability or physics, among other things, just not so much things we will study here.

Lecture 2: Integrability

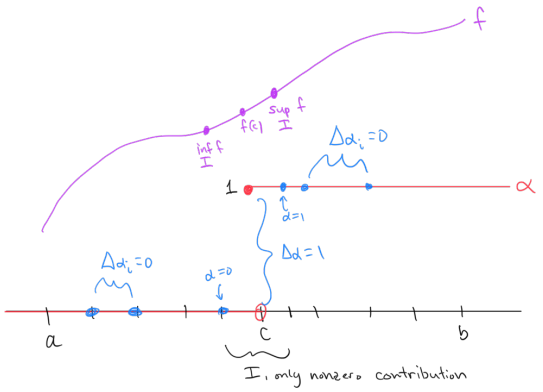
Warm-Up. Suppose $f : [a, b] \to \mathbb{R}$ is bounded and continuous at $c \in (a, b)$. We show that f is Riemann-Stieltjes integrable with respect to the step function

$$\alpha(x) = \begin{cases} 0 & x < c \\ 1 & x \geq c \end{cases}$$

and determine the value of the Riemann-Stieltjes integral $\int_a^b f \, d\alpha$. Given a partition P of $[a, b]$, the upper sum with respect to α is

$$U(f, P, \alpha) = \sum_i \left(\sup_{I_i} f \right) \Delta\alpha_i$$

where $I_i = [x_{i-1}, x_i]$ is the i-th subinterval determined by P and $\Delta\alpha_i = \alpha(x_i) - \alpha(x_{i-1})$ is the "weighted length" of this subinterval. But α is constant on any subinterval that does not contain c, so that $\Delta\alpha_i = 0$ on these subintervals:



Hence the only potentially nonzero contribution to the upper sum comes from the subinterval I that contains c, and

$$U(f, P, \alpha) = \left(\sup_I f\right) \Delta\alpha = \sup_I f$$

since $\Delta\alpha = 1$ on this interval because $\alpha(\text{right endpoint}) = 1$ and $\alpha(\text{left endpoint}) = 0$. In the same way, we get that the lower sum is $L(f, P, \alpha) = \inf_I f$.

We will see in a bit that when considering the infimum of upper sums or the supremum of lower sums, all that matters is the behavior of subintervals that get smaller and smaller. (The fact is that upper sums can only get smaller when taking *refinements* of partitions, and lower sums can only get larger.) But for small intervals we can control the values of f, and hence $\sup f$ and $\inf f$, using continuity. To be clear, for any $\epsilon > 0$ we can imagine that the subinterval I is small enough (of length smaller than an appropriate $\delta > 0$) so that $|f(x) - f(c)| < \epsilon$ on I. But then

$$f(x) \in (f(c) - \epsilon, f(c) + \epsilon) \implies U(f, P, \alpha) = \sup_I f \in [f(c) - \epsilon, f(c) + \epsilon].$$

This is true on any smaller interval as well, so we get that the infimum of all such upper sum values is also in $[f(c) - \epsilon, f(c) + \epsilon]$ (since, again, the infimum is determine solely by taking finer and finer partitions), and thus

$$U(f, \alpha) \in [f(c) - \epsilon, f(c) + \epsilon]$$

where. $U(f, \alpha) = \inf\{U(f, P, \alpha\}$ is the upper Riemann-Stieltjes integral. But this is true for all $\epsilon > 0$, so we get that $U(f, \alpha) = f(c)$. The same is true when replacing $\sup f$ by $\inf f$ and taking supremums of lower sums, so the lower integral is $L(f, \alpha) = f(c)$ as well. Thus the upper and lower integrals agree, so f is Riemann-Stieltjes integrable with respect to α and $\int_a^b f\, d\alpha = f(c)$ is the value of the Riemann-Stieltjes integral.

As we said last time, we will not do much with the Riemann-Stieltjes integral in this course, but this example gives a hint of its use: evaluation of a function at a point (at which it is continuous) can be viewed as a type of integral! If instead we wanted to obtain something like

$$\int f\, d\alpha = f(c_1) + f(c_2)$$

as the value of an integral, all we need to do is use a step function which has two jumps, one at c_1 and one at c_2. Similarly if we wanted to obtain any finite sum of values as an integral, and we

allow step functions with infinitely many jumps, it is possible to obtain an entire infinite series as the value of a certain Riemann-Stieltjes integral. The point is that (Riemann) integration is often viewed as a type of "continuous" summation, and by using Riemann-Stieltjes integration we can make this analogy clearer by viewing discrete summation (either finite or finite) as literally a type of integral as well. The Riemann-Stieljes integral gives, if nothing else, a way to study integration and summation from a common unified point of view.

Refinements. As we work towards finding an approach to integrability that does not require computing all possible upper or lower sums, we need a way to compare upper and lower sums which come from different partitions. To do so, we need the notion of a refinement. Given a partition P of $[a, b]$, a *refinement* of P is a partition P' where we take P and throw in additional partition points. This has the practical effect of taking the subintervals determined by P and breaking them up even further.

We are interested in what happens to upper and lower sums when taking a refinement of a given partition; in other words, what's the relation between $U(f, P)$ and $U(f, P')$, and between $L(f, P)$ and $L(f, P')$? Suppose we had a subinterval $[x_{k-1}, x_k]$ for P which was broken up into two pieces after adding one more partition point s to create P':

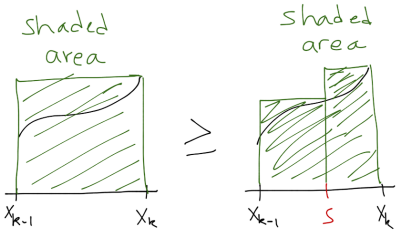
In the sum making up $U(f, P)$ we have a term which looks like

$$(\sup f \text{ over } [x_{k-1}, x_k])(\text{length of } [x_{k-1}, x_k])$$

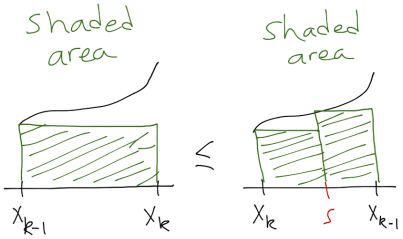
and corresponding to this in the sum making up $U(f, P')$ we have two terms which look like

$$(\sup f \text{ over } [x_{k-1}, s])(\text{length of } [x_{k-1}, s]) + (\sup f \text{ over } [s, x_k])(\text{length of } [s, x_k]).$$

But the supremum of f over all of $[x_{k-1}, x_k]$ is greater than or equal to its supremum over either smaller interval $[x_{k-1}, s]$ or $[s, x_k]$, so the first expression above is greater than or equal to the sum in the second expression; graphically we are saying that



For infimums the opposite is true: the infimum of f over all of $[x_{k-1}, x_k]$ is less than or equal to its infimum over either smaller interval, so



The same things are true for any subinterval of P which was broken up into smaller intervals in P', so we conclude that

$$U(f, P) \geq U(f, P') \text{ and } L(f, P) \leq L(f, P').$$

In other words, adding more points to your partition either decreases an upper sum or keeps it the same, and either increases a lower sum or keeps it the same. Intuitively, upper sums get "smaller" under refinements and lower sums get "bigger".

Existence of upper/lower integrals. Now we can make the comparisons we need. For any partitions P and Q of $[a, b]$, we claim that

$$L(f, P) \leq U(f, Q).$$

That is, any lower sum whatsoever is less than or equal to any upper sum whatsoever. Indeed, let $P \cup Q$ denote the partition formed by taking the points of P together with the points of Q, which is a *common* refinement of both P and Q. By the property of refinements given before, we have

$$L(f, P) \leq L(f, P \cup Q) \leq U(f, P \cup Q) \leq U(f, Q),$$

where the middle inequality comes from the fact that for a single partition, the lower sum is always less than or equal the upper sum since one uses $\inf f$ and the other $\sup f$. Thus $L(f, P) \leq U(f, Q)$ as claimed.

In particular, this means that for fixed P, $L(f, P)$ is a lower bound for the set of all upper sums, so the infimum of the set of all upper sums (i.e., the upper integral of f) exists and

$$L(f, Q) \leq U(f).$$

But then $U(f)$ is an upper bound for the set of all lower sums, so the supremum of the set of lower bounds (i.e., the lower integral) exists and $L(f) \leq U(f)$. Note that this inequality (perhaps intuitively true) was not obvious (to me at least!) at the start.

Example from last time. We can now finish off the $f(x) = x$ example from last time. Recall that for the partition P_n of $[0, b]$ consisting of equally-spaced points, we computed

$$U(f, P_n) = \frac{b^2(n+1)}{2n} \quad \text{and} \quad L(f, P_n) = \frac{b^2(n-1)}{2n}.$$

Trying to compute upper and lower sums for other not-so-nice partitions is going to be impossible, but we do not ned to do so! We have

$$\frac{b^2(n-1)}{2n} = L(f, P_n) \leq L(f) \leq U(f) \leq U(f, P_n) = \frac{b^2(n+1)}{2n},$$

where the first and third inequalities come from the definitions of $L(f)$ and $U(f)$ respectively as a supremum and an infimum, and the the middle inequality is the one that required knowledge of refinements. Taking limits throughout gives

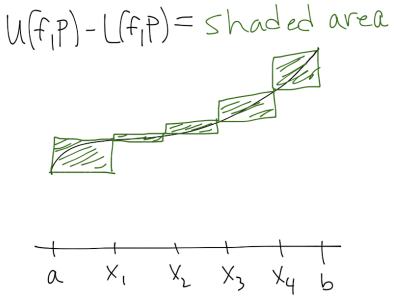
$$\frac{b^2}{2} \leq L(f) \leq U(f) \leq \frac{b^2}{2},$$

so we conclude that we must have equalities throughout. Hence $L(f) = U(f) = \frac{b^2}{2}$, so f is integrable on $[0, b]$ and $\int_a^b f = \frac{b^2}{2}$, all as expected.

Recasting integrability. We now give the way to characterize integrability without caring about the behavior of all possible upper/loser sums. The claim is that a bounded function $f : [a, b] \to \mathbb{R}$ is integrable if and only if for all $\epsilon > 0$ there exists a partition P of $[a, b]$ such that

$$U(f, P) - L(f, P) < \epsilon.$$

So, in the end we only need to consider the behavior of certain well-chosen partitions. Integrable means that the infimum of the upper sums should equal the supremum of the lower sums, and intuitively this suggests that upper and lower sums can be made arbitrarily close to one other, which is what the condition in this result says. Graphically, $U(f, P) - L(f, P)$ is the sum of the areas of the small rectangles "between" the upper and lower sums:



and the condition says that this sum of small areas can be made arbitrarily small.

Here's the proof. Suppose that f is integrable on $[a, b]$, so that the upper and lower integrals are the same:

$$U(f) = L(f).$$

Call this common value I to make notation simpler. Let $\epsilon > 0$. Then, by properties of supremums and infimums, there exists a partition P of $[a, b]$ such that

$$I - \frac{\epsilon}{2} < L(f, P)$$

and there exists a partition Q of $[a, b]$ such that

$$U(f, Q) < I + \frac{\epsilon}{2}.$$

Thus for the partition $P \cup Q$, which is a refinement of both P and Q, we have

$$U(f, P \cup Q) - L(f, P \cup Q) \leq U(f, Q) - L(f, P) < \left(I + \frac{\epsilon}{2}\right) - \left(I - \frac{\epsilon}{2}\right) = \epsilon,$$

where the second inequality follows from the fact that we replaced $U(f, Q)$ by something larger and $L(f, P)$ by something smaller. Thus $P \cup Q$ satisfies the requirement in the statement of the theorem.

Conversely, suppose that for any $\epsilon > 0$ there exists a partition P of $[a, b]$ such that $U(f, P) - L(f, P) < \epsilon$. Then for such a partition we have

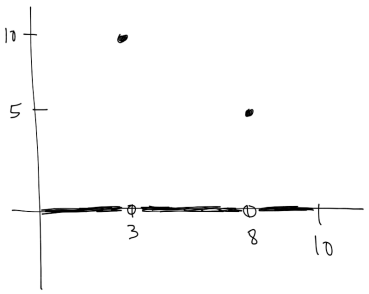
$$U(f) - L(f) \leq U(f, P) - L(f, P) < \epsilon$$

where again the first inequality follows from replacing the first term by something larger and the second by something smaller. This says that the nonnegative number

$$U(f) - L(f)$$

is smaller than any $\epsilon > 0$, and so must be zero. Hence the upper and lower integrals are the same so f is integrable on $[a, b]$.

Example. Consider the function f on $[0, 10]$ which is zero everywhere, except at 3 and 8 where $f(3) = 10$ and $f(8) = 5$:



We claim that this function is integrable and that its integral over $[0, 10]$ is 0. This makes sense intuitively: the region "under" the graph of f consists of two vertical lines (above $x = 3$ and $x = 8$) and the "area" of these two vertical lines should indeed be 0.

Taking $\epsilon > 0$, we want to find a partition P of $[0, 10]$ such that

$$U(f, P) - L(f, P) < \epsilon.$$

This difference between upper and lower sums looks like

$$U(f, P) - L(f, P) = \sum_{I_k} (\sup f - \inf f \text{ on } I_k)(\text{length of } I_k).$$

In this case, no matter what the subinterval I_k is, the infimum of f over it is zero, so the above becomes:

$$U(f, P) - L(f, P) = \sum_{I_k} (\sup f \text{ on } I_k)(\text{length of } I_k).$$

But on any I_k which does not contain 3 or 8, the supremum of f is also zero, so the above simplifies to just the sum over the intervals containing 3 and 8; say

$$U(f, P) - L(f, P) = (\sup f \text{ on } J)(\text{length of } J) + (\sup f \text{ on } K)(\text{length of } K)$$

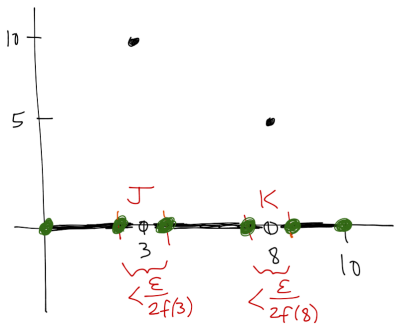
where J is the subinterval containing 3 and K the subinterval containing 8. The supremum of f on J is $f(3) = 10$ and its supremum on K is $f(8) = 5$, so

$$U(f, P) - L(f, P) = 10(\text{length of } J) + 5(\text{length of } K).$$

This is the expression we want to make smaller than ϵ, and we can do so by constructing our partition P in such a way that the lengths of J and K are small enough that they balance out the values of $f(3)$ and $f(10)$! In particular, if the length of J was smaller than $\frac{\epsilon}{2 \cdot 10}$ and the length of K smaller than $\frac{\epsilon}{2 \cdot 5}$ (note the $\frac{\epsilon}{2}$-trick which is used here) then the above becomes

$$U(f, P) - L(f, P) = 10(\text{length of } J) + 5(\text{length of } K) < 10\left(\frac{\epsilon}{2 \cdot 10}\right) + 5\left(\frac{\epsilon}{2 \cdot 5}\right) = \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

as required. Thus, given $\epsilon > 0$, picking an interval J around 3 of length smaller than $\frac{\epsilon}{20}$ and an interval K around 8 of length smaller than $\frac{\epsilon}{10}$:



gives a partition P (the points of which are the green points in the picture above, consisting of the endpoints of all subintervals used) such that $U(f,P) - L(f,P) < \epsilon$, so f is integrable on $[0,10]$. (In this argument we are implicitly assuming that J and K do not overlap, but this can always be made to hold by shrinking J and K if need be.) Since all lower sums are 0, the lower integral is 0, so

$$\int_0^{10} f(x)\,dx = 0.$$

Integrability as control. This idea of constructing a partition by making subintervals small enough to balance out some "bad" behavior of a function is the crucial technique used in many integration problems and results, and is what I meant back on the first day in saying that integrability amounts to saying that the behavior of a function near points where it cannot be controlled can be made negligible. The same type argument as above works for any function that is piecewise constant, no matter how many (finitely many) "jumps" it has, nor the specific values of the constant on each piece.

Lecture 3: More on Integration

Warm-Up 1. We show that if $f : [a,b] \to \mathbb{R}$ is integrable, then $|f|$ is as well. This uses what we'll call the *reverse triangle inequality*:

$$|f(x)| - |f(y)| \leq |f(x) - f(y)|.$$

(If you have not seen this before, it follows from the usual triangle inequality via

$$|f(x)| = |(f(x) - f(y)) + f(y)| \leq |f(x) - f(y)| + |f(y)|$$

and rearranging terms.) This holds on any subinterval of $[a,b]$, so taking supremums gives

$$\sup|f| - \inf|f| \leq \sup f - \inf f$$

on any interval. Thus for given $\epsilon > 0$, we use integrability of f to pick a partition P of $[a,b]$ such that

$$U(f,P) - L(f,P) < \epsilon,$$

and then we get

$$\begin{aligned} U(|f|,P) - L(|f|,P) &= \sum (\sup|f| - \inf|f|)\,\text{length} \\ &\leq \sum (\sup f - \inf f)\,\text{length} \end{aligned}$$

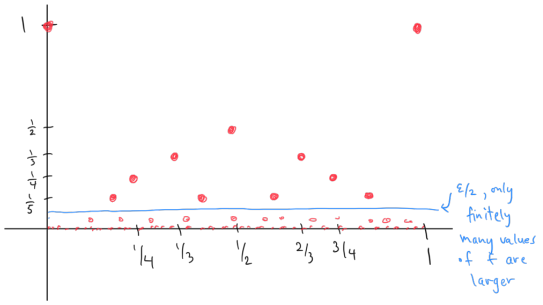
$$= U(f,P) - L(f,P)$$
$$< \epsilon,$$

so $|f|$ is integrable on $[a,b]$ as claimed.

Warm-Up 2. My favorite function of all time is the function $f : [0,1] \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 0 & x \notin \mathbb{Q} \\ \frac{1}{q} & x = \frac{p}{q} \in \mathbb{Q} \text{ in reduced form.} \end{cases}$$

(Since 0 can be written as $\frac{0}{q}$ for any q, by convention we take $0 = \frac{0}{1}$ so that $f(0) = 1$.) The graph of f looks something like



If you have not encountered this function before, for a nice exercise you should try to show that f is discontinuous at each rational but actually continuous (!!!) at each irrational. (This is why this is my favorite function!) Here we show that f is integrable and determine the value of $\int_0^1 f$. Given any partition P of $[0,1]$, we have $L(f,P) = 0$ since any subinterval will always contain irrationals. Thus the lower integral $L(f) = \sup\{L(f,P)\}$ is 0, and hence $\int_0^1 f$—once we know it exists—has the value zero.

The difference $U(f,P) - L(f,P)$ looks like

$$U(f,P) - L(f,P) = U(f,P) = \sum_k (\sup f \text{ on } I_k)(\text{length } I_k) \leq \sum_k (\text{length } I_k) = 1$$

since $\sup f \leq 1$ always; however, this is not going to help us if we want to make this expression smaller than ϵ. The idea we will use is the same one you would use to show that f is continuous at each irrational: given some $\epsilon > 0$, there are only finitely many rationals $r \in [0,1]$ satisfying $f(r) \geq \epsilon$. For all other rationals, we have $f(r) < \epsilon$, and this will give us a way to bound $\sup f$, at least over subintervals which contain none of the rationals where $f(r) \geq \epsilon$. Thus, we break up our entire sum into two pieces—the piece over the intervals J containing a rational where $f(r) \geq \epsilon$, and a piece over the intervals K containing no such rationals:

$$U(f,P) = \sum_J (\sup f)(\text{length}) + \sum_K (\sup f)(\text{length}).$$

Actually, based on this breaking up into two pieces, we actually go back and replace the previous ϵ by $\frac{\epsilon}{2}$. That is, we consider the rationals where $f(r) \geq \frac{\epsilon}{2}$ and denote the intervals containing such a rational by J and the others by K. Over each K, $\sup f \leq \frac{\epsilon}{2}$ so the entire second piece above is bounded by

$$\sum_K (\sup f)(\text{length}) \leq \sum_K \frac{\epsilon}{2}(\text{length}) = \frac{\epsilon}{2} \sum_K (\text{length}) \leq \frac{\epsilon}{2}$$

since adding up all the lengths of the K intervals can't give more than the total length of $[0,1]$. The goal is now to bound the first piece of $U(f,P)$ above also by $\frac{\epsilon}{2}$, giving us $U(f,P) < \epsilon$ in the end. But there are only finitely many rationals satisfying $f(r) \geq \frac{\epsilon}{2}$, so if we construct our partition to surround each of these rationals by a small enough interval, we can make the sum over the intervals K containing smaller than whatever we'd like. Here's our proof.

Let $\epsilon > 0$ and denote the finitely many rationals r such that $f(r) \geq \frac{\epsilon}{2}$ by $r_1, r_2, \ldots, r_n$. For each r_k, take an interval J_k around it such that

$$\text{length of } J_k < \frac{\epsilon}{2n},$$

and if need be make this intervas even smaller to guarantee that each J_k only contains one of the r_i's. (We implicitly assumed this in our scratch work above.) Take P to be the partition of $[0,1]$ consisting of $0, 1$, and the endpoints of all the intervals J_k.

Then $L(f,P) = 0$ so

$$U(f,P) - L(f,P) = U(f,P) = \sum_{k=1}^{n} (\sup f \text{ on } J_k)(\text{length } J_k) + \sum_{K} (\sup f \text{ on } K)(\text{length } K)$$

where the second sum is over the subintervals K which contain none of $r_1, \ldots, r_n$. On each K, $\sup f \leq \frac{\epsilon}{2}$ while on each J_k, $\sup f \leq 1$. Thus

$$\begin{aligned} U(f,P) &\leq \sum_{k=1}^{n} (\text{length } J_k) + \sum_{K} \frac{\epsilon}{2} (\text{length } K) \\ &< \sum_{k=1}^{n} \frac{\epsilon}{2n} + \frac{\epsilon}{2} \sum_{K} (\text{length } K) \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \\ &= \epsilon. \end{aligned}$$

Hence for this partition we have $U(f,P) - L(f,P) < \epsilon$, so we conclude that f is integrable on $[0,1]$ as claimed. Since all lower sums are equal to 0, the value of its integral over $[0,1]$ is zero.

Continuous implies integrable. We now prove that if $f : [a,b] \to \mathbb{R}$ is continuous, then f is integrable. Actually, what we need to use here is the fact that f is uniformly continuous since it is continuous on a compact domain. Given $\epsilon > 0$, we want a partition P of $[a,b]$ such that

$$U(f,P) - L(f,P) = \sum_{k} (\sup f - \inf f)(\text{length } I_k) < \epsilon.$$

Now, in order to bound $\sup f - \inf f$ over a subinterval I_k we want to able to bound expressions of the form

$$|f(x) - f(y)| \text{ for } x, y \in I_k.$$

But this we can do using uniform continuity of f, which says that $|f(x) - f(y)|$ can be bounded by whatever positive number we want as soon as x and y are close enough to each other. (We need uniform continuity and not just continuity since both x and y can vary.)

So for the positive number $\frac{\epsilon}{2(b-a)}$ (why this? we'll see) there exists $\delta > 0$ such that

$$|x - y| < \delta \text{ implies } |f(x) - f(y)| < \frac{\epsilon}{2(b-a)}.$$

Thus if we construct our partition so that each subinterval as length smaller than δ, then any points in this subinterval are close enough to guarantee that $|f(x) - f(y)| < \frac{\epsilon}{2(b-a)}$ on that subinterval, giving

$$\sup f - \inf f \leq \frac{\epsilon}{2(b-a)} \text{ on that subinterval.}$$

Here's the proof. Let $\epsilon > 0$. Since f is continuous on $[a, b]$, it is uniformly continuous on $[a, b]$, so there exists $\delta > 0$ such that

$$\text{if } |x - y| < \delta, \text{ then } |f(x) - f(y)| < \frac{\epsilon}{2(b-a)} \text{ for all } x, y \in [a, b].$$

Let P be a partition of $[a, b]$ such that the lengths of all subintervals are smaller than δ. Then we get $\sup f - \inf f \leq \frac{\epsilon}{2(b-a)}$ on each subinterval, so

$$\begin{aligned} U(f, P) - L(f, P) &= \sum (\sup f - \inf f) \text{ length} \\ &\leq \sum \frac{\epsilon}{2(b-a)} \text{ length} \\ &= \frac{\epsilon}{2(b-a)} \sum \text{length} \\ &= \frac{\epsilon}{2(b-a)} (b - a) \\ &= \frac{\epsilon}{2} \\ &< \epsilon. \end{aligned}$$

(The desire to get a strict inequality $< \epsilon$ rather than $\leq \epsilon$ is why we threw in the extra factor of 2 in the denominator at the start; it's not really needed.)

Piecewise continuous example. We show that the function $f : [0, 3] \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} e^x & \text{if } 0 \leq x < 1 \\ 30 & x = 1 \\ \cos \frac{1}{x} & \text{if } 1 < x < 2 \\ -100 & x = 2 \\ -x^4 & \text{if } 2 < x \leq 3 \end{cases}$$

is integrable. (This is not the function we looked at in class, but the idea is exactly the same. I just happened to have this particular example already nicely typed up from elsewhere.)

Let $\epsilon > 0$ and let M be a bound on f over $[0, 3]$. Pick intervals J and K around 2 and 3 respectively, each of length less than $\frac{\epsilon}{2M \cdot 5}$ and small enough so that they do not overlap. On the interval $[0, \text{left endpoint of } J]$, $f = e^x$ is continuous so f is integrable here and there exists a partition P_1 of this interval such that

$$U(e^x, P_1) - U(e_x, P_2) < \frac{\epsilon}{5}.$$

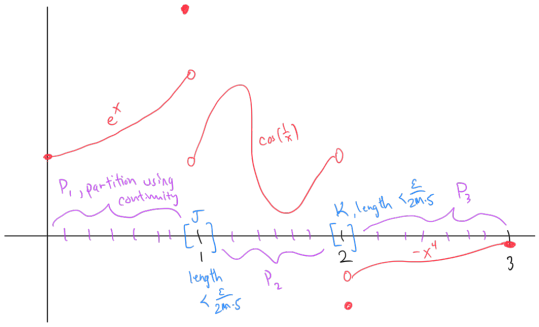
Similarly, $f = \cos \frac{1}{x}$ is continuous on the interval from the right endpoint of J to the left endpoint of K, so there exists a partition P_2 of this interval such that

$$U(\cos \tfrac{1}{x}, P_2) - L(\cos \tfrac{1}{x}, P_2) < \frac{\epsilon}{5},$$

and $f = -x^4$ is continuous on [right endpoint of $K, 3$], so there is a partition P_3 of this interval such that

$$U(-x^4, P_3) - L(-x^4, P_3) < \frac{\epsilon}{5}.$$

Altogether, this looks like



Let P be the partition of $[0, 3]$ consisting of $2, 3$, all points making up P_1, all points making up P_2, and all points making up P_3. The the difference $U(f, P) - L(f, P)$ can then be broken up into five pieces, consisting of contributions from

$$P_1,\ J,\ P_2,\ K,\ \text{and } P_3.$$

The contributions on the pieces coming from P_1, P_2, and P_3 are each smaller than $\frac{\epsilon}{5}$ by choice of these partitions. The contributions from J and K are smaller than

$$(\sup f - \inf f)\,\text{length} < 2M \frac{\epsilon}{2M \cdot 5} = \frac{\epsilon}{5}$$

since $|f(x) - f(y)| \leq 2M$ for all x, y because M is a bound on f, and by the choice of the lengths of J and K. Putting it all together gives

$$U(f, P) - L(f, P) = \sum \text{contributions} < \frac{\epsilon}{5} + \frac{\epsilon}{5} + \frac{\epsilon}{5} + \frac{\epsilon}{5} + \frac{\epsilon}{5} = \epsilon.$$

Hence f is integrable on $[0, 3]$ as claimed. (The same idea works for any function which is piecewise continuous, no matter how many "pieces" it consists of and what those pieces look like.)

Lecture 4: Fundamental Theorem of Calculus

Warm-Up. Suppose $f : [a, b] \to \mathbb{R}$ is integrable and for fixed $k \in \mathbb{N}$ set

$$D_k := \left\{ c \in [a, b] \;\middle|\; \inf_{\delta > 0} \left(\sup_{x,y \in [c-\delta, c+\delta]} |f(x) - f(y)| \right) \geq \frac{1}{k} \right\}.$$

We claim that for any $\epsilon > 0$ there exist finitely many intervals covering D_k whose total sum of lengths is smaller than ϵ:

$$\sum \text{length} < \epsilon.$$

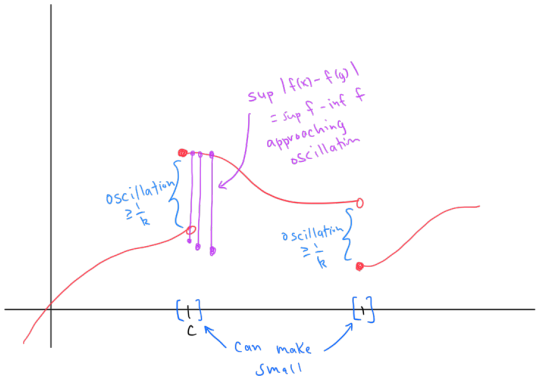
Let us first give some context. The number

$$\inf_{\delta > 0} \left(\sup_{x,y \in [c-\delta, c+\delta]} |f(x) - f(y)| \right)$$

is called the *oscillation* of f at c and measures the extent (or not) to which f fails to be continuous at c. Indeed, the supremum

$$\sup_{x,y\in[c-\delta,c+\delta]} |f(x)-f(y)| = \sup_{x\in[c-\delta,c+\delta]} f(x) - \inf_{x\in[c-\delta,c+\delta]} f(x)$$

measures by how much the values of f can differ on the interval $[c-\delta, c+\delta]$ around c, and by taking the infimum of these as $\delta > 0$ decreases (note that the supremums can only get smaller as δ gets smaller since we are taking the supremum of a smaller set) we are measuring by how much f can change arbitrarily close to c:



The oscillation of f at c measures the "jump" (if there is one) of f at c, and indeed to say that f is continuous at c means precisely that the oscillation is zero:

$$\inf_{\delta>0}\left(\sup_{x,y\in[c-\delta,c+\delta]} |f(x)-f(y)|\right) = 0 \iff f \text{ is continuous at } c.$$

To be clear, the definition of continuity via

$$|x-c| < \delta \implies |f(x)-f(c)| < \epsilon$$

says precisely that

$$\sup_{x,y\in[c-\delta,c+\delta]} |f(x)-f(y)| \leq 2\epsilon$$

using $|f(x)-f(y)| \leq |f(x)-f(c)| + |f(c)-f(y)|$, so the infimum of such supremums as $\epsilon > 0$ gets smaller will be zero. Thus, the set D_k consists of points at which f is not continuous because the oscillation at those points is at least $\frac{1}{k} > 0$. The point of this problem is to show that the set of such points is "small" in the sense that it can be covered by appropriately small intervals. This fits into the idea that integrability amounts to saying that the behavior near poorly-behaved points can be made "negligible", which we will formally make precise next time.

Back to the Warm-Up. Let $\epsilon > 0$ and pick, using integrability, a partition P of $[a,b]$ such that

$$U(f,P) - L(f,P) < \frac{\epsilon}{k}.$$

Among the subintervals determined by P are the ones that contain an element of D_k, and on these specific subintervals we have

$$\sup f - \inf f = \sup |f(x)-f(y)| \geq \frac{1}{k}$$

since the oscillation at elements of D_k is at least $\frac{1}{k}$. Thus if we extract these subintervals along from the entire sum $U(f,P) - L(f,P)$, we get

$$\sum_{\text{subintervals } \cap D_k \neq \emptyset} \frac{1}{k} \cdot \text{length} \leq \sum_{\text{subintervals } \cap D_k \neq \emptyset} (\sup f - \inf f) \cdot \text{length} \leq U(f,P) - L(f,P) < \frac{\epsilon}{k}$$

where the second inequality follows since $U(f,P) - L(f,P)$ includes more nonnegative term than the sum right before. After clearing the constant k in the first and final terms, we get

$$\sum_{\text{subintervals } \cap D_k \neq \emptyset} \text{length} < \epsilon$$

as desired. (Next time we will define the notion of a set having "measure zero", and see that this result means that D_k has measure zero.)

Properties of integrals. Integrals, as we've defined them, have all the properties one might expect from a calculus course:

- (linearity) If $f, g : [a,b] \to \mathbb{R}$ are integrable, then $f+g$ is integrable on $[a,b]$ and

$$\int_a^b (f+g) = \int_a^b f + \int_a^b g,$$

  and if $c \in \mathbb{R}$, then cf is integrable on $[a,b]$ and

$$\int_a^b cf = c\int_a^b f.$$

- (domain splitting) If $c \in (a,b)$, then f is integrable on $[a,b]$ if and only if f is integrable on $[a,c]$ and $[c,b]$, and

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

- (monotonicity) If f and g are integrable on $[a,b]$ and $f(x) \leq g(x)$ for all $x \in [a,b]$, then

$$\int_a^b f \leq \int_a^b g.$$

- (absolute value) If f is integrable on $[a,b]$, then $|f|$ is integrable on $[a,b]$ and

$$\left|\int_a^b f\right| \leq \int_a^b |f|.$$

The first property in "lineraity" is proved in Rudin, and comes down to making appropriate estimates between upper/lower sums for $f+g$ versus those for f and g alone. This is not trivial since in general we can only guarantee

$$U(f+g,P) \leq U(f,P) + U(g,P) \text{ since } \sup(f(x)+g(x)) \leq \sup f(x) + \sup g(x)$$

instead of equalities, and the opposite inequalities for lower sums, but we leave the details to Rudin. (Even better, do it on your own!) The scalar multiplication property is more straightforward and follows from

$$\sup cf(x) = c\sup f(x) \quad \text{and} \quad \inf cf(x) = c\inf f(x) \quad \text{for } c \geq 0,$$

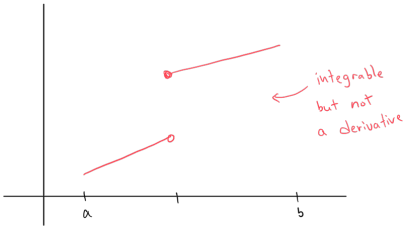
while for $c < 0$ the same is true except sup and inf get interchanged since $\sup(-f(x)) = -\inf f(x)$.

The domain splitting property is on the homework and comes down to carefully comparing partitions of $[a, c]$ and $[c, b]$ to those of $[a, b]$. Monotonicity follows from the fact that the $\sup f$ on an interval is less than or equal to $\sup g$ on an interval, so $U(f, P) \leq U(g, P)$ for any partition. As for absolute value, we showed that $|f|$ is integrable as a Warm-Up last time, and the integral bound comes from applying monotonicity to

$$-|f| \leq f \leq |f| \implies -\int |f| \leq \int f \leq \int |f|.$$

Integrability vs anti-differentiability. And yet, not everything about integration behaves exactly as you would naively expect from a calculus course, where in particular computing an integral comes down to finding an antiderivative, so that "integration" and "anti-differentiation" are synonymous. This is not true in general, which is why the result that specifies when it is true—the fundamental theorem of calculus—is indeed "fundamental".

To see the distinction between integration and anti-differentiation, first consider a function like



This function is integrable via standard methods (i.e., surround the discontinuity by a small enough interval and use continuity to partition everything else), and yet does not have an anti-derivative, specifically because it does not have the intermediate value property as all derivatives (from last quarter) do. So, integrability in general does not imply existence of an anti-derivative. Second, the function

$$f(x) = \begin{cases} x^2 \sin(\frac{1}{x^2}) & x \neq 0 \\ 0 & x = 0 \end{cases}$$

is differentiable (use a limit to check differentiability at 0) and

$$f'(x) = \begin{cases} 2x \sin(\frac{1}{x^2}) - \frac{2}{x} \cos(\frac{1}{x^2}) & x \neq 0 \\ 0 & x = 0. \end{cases}$$

Thus f' has an anti-derivative, namely f, but f' is not integrable on, say $[-1, 1]$ because it is unbounded here. Thus, existence of an anti-derivative in general does not imply integrability

Fundamental Theorem of Calculus, I. The fundamental theorem of calculus clarifies the relation between integrals and anti-derivatives. There are typically two versions, one which says what happens when you "integrate a derivative" and the second when you "differentiate an integral". Here we state the first: If f is differentiable and f' is integrable on $[a, b]$, then

$$\int_a^b f'(x)\, dx = f(b) - f(a).$$

A key assumption here is that f' is integrable since, as we saw in the previous example, this is not guaranteed. So, indeed the method of finding an anti-derivative and evaluating at endpoints does give the correct value of the integral, under the correct assumption.

For the proof, take any partition $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$ of $[a, b]$. Write the difference $f(b) - f(a)$ as a telescoping sum by adding and subtracting all intermediate $f(x_i)$:

$$\begin{aligned} f(b) - f(a) &= f(x_n) - f(x_{n-1}) + f(x_{n-1}) - f(x_{n-2}) + \cdots - f(x_1) + f(x_1) - f(x_0) \\ &= \sum_{i=1}^{n} [f(x_i) - f(x_{i-1})]. \end{aligned}$$

Now, here's the magic: by the mean value theorem, for each i we have

$$f(x_i) - f(x_{i-1}) = f(c_i)(x_i - x_{i-1})$$

for some $c_i \in (x_{i-1}, x_i)$, so

$$f(b) - f(a) = \sum_{i=1}^{n} [f(x_i) - f(x_{i-1})] = \sum_{i=1}^{n} f'(c_i)(x_i - x_{i-1}).$$

The sum on the right sits between the upper and lower sums for f' for the partition P since $\inf f \leq f'(c_i) \leq \sup f$, so we get

$$L(f', P) \leq f(b) - f(a) \leq U(f', P)$$

and thus $L(f') \leq f(b) - f(a) \leq U(f')$. Since f' is integrable, $L(f) = U(f)$ so we have equality throughout and hence $\int_a^b f'(x)\,dx = U(f) = L(f) = f(b) - f(a)$ as claimed.

Fundamental Theorem of Calculus, II. For the second statement, suppose $f : [a, b] \to \mathbb{R}$ is integrable and define $F : [a, b] \to \mathbb{R}$ by

$$F(x) = \int_a^x f(t)\,dt.$$

The claim is that any point $x_0 \in (a, b)$ at which f is continuous, F is differentiable at $F'(x_0) = f(x_0)$. (So, "differentiating the integral gives the integrand.") To prove this we simply verify that the limit defining $F'(x_0)$ exists and has value $f(x_0)$:

$$\lim_{h \to 0} \frac{F(x_0 + h) - F(x_0)}{h} = f(x_0).$$

For fixed $\epsilon > 0$, this means we need $\delta > 0$ such that

$$0 < |h| < \delta \implies \left| \frac{F(x_0 + h) - F(x_0)}{h} - f(x_0) \right| < \epsilon.$$

Using domain splitting, we have

$$F(x_0 + h) - F(x_0) = \int_a^{x_0+h} f(t)\,dt - \int_a^{x_0} f(t)\,dt = \int_{x_0}^{x_0+h} f(t)\,dt.$$

(One small detail is that h is allowed to be negative, in which case $x_0 + h < x_0$ so that the bounds on the integral are in the wrong order. We use the common convention that by such an integral we mean the negative of the integral with the bounds in the correct order: if $c < d$, then $\int_d^c f := -\int_c^d f$.) By writing the constant $f(x_0)$ as $f(x_0) = \frac{1}{h}\int_{x_0}^{x_0+h} f(x_0)\,dt$, we thus have

$$\left| \frac{F(x_0 + h) - F(x_0)}{h} - f(x_0) \right| = \left| \frac{1}{h} \int_{x_0}^{x_0+h} f(t)\,dt - \frac{1}{h} \int_{x_0}^{x_0+h} f(x_0)\,dt \right|$$

$$= \left| \frac{1}{h} \int_{x_0}^{x_0+h} (f(t) - f(x_0))\, dt \right|.$$

Now we start bounding:

$$\left| \frac{1}{h} \int_{x_0}^{x_0+h} (f(t) - f(x_0))\, dt \right| \leq \frac{1}{|h|} \int_{\min\{x_0, x_0+h\}}^{\max\{x_0, x_0+h\}} |f(t) - f(x_0)|\, dt,$$

where we use the min and max bounds we do since h might be negative. All we need now is to appropriate bound $|f(t) - f(x_0)|$ using continuity, and we are good to go. Since f is continuous at x_0, we can pick $\delta > 0$ such that

$$|t - x_0| < \delta \implies |f(t) - f(x_0)| < \epsilon.$$

For $|h|$ smaller than this δ, all t between $\min\{x_0, x_0 + h\}$ and $\max\{x_0, x_0 + h\}$ satisfy $|t - x| < \delta$, so we get

$$\left| \frac{F(x_0 + h) - F(x_0)}{h} - f(x_0) \right| \leq \frac{1}{|h|} \int_{\min\{x_0, x_0+h\}}^{\max\{x_0, x_0+h\}} |f(t) - f(x_0)|\, dt < \frac{1}{|h|} \int_{\min\{x_0, x_0+h\}}^{\max\{x_0, x_0+h\}} \epsilon\, dt = \epsilon$$

where we use the fact that the length of in the interval from $\min\{x_0, x_0 + h\}$ to $\max\{x_0, x_0 + h\}$ is $|h|$. Thus

$$\lim_{h \to 0} \frac{F(x_0 + h) - F(x_0)}{h} = f(x_0)$$

as claimed. (Note that along the way, from

$$F(x_0 + h) - F(x_0) = \int_{x_0}^{x_0+h} f(t)\, dt$$

we see that, regardless of whether or not it is differentiable, F is always (uniformly) continuous: if $M > 0$ is a bound on f, then

$$|F(x_0 + h) - F(x_0)| \leq \int_{\min\{x_0, x_0+h\}}^{\max\{x_0, x_0+h\}} |f(t)|\, dt \leq \int_{\min\{x_0, x_0+h\}}^{\max\{x_0, x_0+h\}} M\, dt = M|h|,$$

which implies uniform continuity on $[a, b]$.)

Lecture 5: Riemann-Lebesgue Theorem

Warm-Up. Let $f : \mathbb{R} \to \mathbb{R}$ be the function defined by

$$f(x) = \begin{cases} \sin \frac{1}{x} & x \neq 0 \\ 0 & x = 0 \end{cases}$$

and define $F : \mathbb{R} \to \mathbb{R}$ by $F(x) = \int_0^{x^2} f(t)\, dt$. We claim that F is differentiable everywhere. Note that f is integrable on any closed interval since it is only discontinuous at one point, so the integral defining F exists.

Since f is continuous at any $x \neq 0$, the fundamental theorem of calculus together with the chain rule imply that F is differentiable at all $x \neq 0$. To be clear, the function

$$G(x) = \int_0^x f(t)\, dt$$

is differentiable at $x \neq 0$ by the fundamental theorem, and hence $F(x) = G(x^2)$ is as well by the chain rule. (Note that we need to know $x^2 \neq 0$ for $x \neq 0$ in order for the chain rule to apply here.)

Thus all that's left is to check differentiability at 0. For $x \neq 0$ we have:

$$\frac{F(x) - F(0)}{x - 0} = \frac{F(x)}{x} = \frac{1}{x} \int_0^{x^2} f(t)\, dt.$$

Since

$$\left| \frac{1}{x} \int_0^{x^2} f(t)\, dt \right| \leq \frac{1}{|x|} \int_0^{x^2} |f(t)|\, dt \leq \frac{1}{|x|} \int_0^{x^2} 1\, dt = |x|$$

and the right side goes to 0 as $x \to 0$, the squeeze theorem implies that

$$\lim_{x \to 0} \frac{F(x) - F(0)}{x - 0} = \lim_{x \to 0} \frac{1}{x} \int_0^{x^2} f(t)\, dt$$

exists and equals zero, so F is differentiable at zero as well. Thus F is differentiable everywhere.

Other topics in integration. All the other main integration results you might recall from a calculus course can now be rigorously proved. Integration by parts, for example, is a quick consequence of the fundamental theorem of calculus.

Integration by substitution, or "change of variables", also holds, although in full generality requires some real effort to prove. The general setting is that of a continuously differentiable function $\phi : [a, b] \to \mathbb{R}$ with positive derivative and an integrable function f on the image interval $[\phi(a), \phi(b)]$. Change of variables should then say that

$$\int_{\phi(a)}^{\phi(b)} f(u)\, du = \int_a^b f(\phi(x))\phi'(x)\, dx.$$

(A small modification is needed in instead $\phi' < 0$.) In a calculus course, you would naively set $u = \phi(x)$ and then take $du = \phi'(x)\, dx$ to turn the integral on the right into the integral on the left. But to actually prove that this works, in particular when f is only integrable and not necessarily continuous, takes more care. We will leave the details to a discussion section, but here is the basic strategy. An upper sum for the integral on the left looks like

$$\sum_i (\sup f \text{ on } [u_{i-1}, u_i])(u_i - u_{i-1}).$$

Under the assumptions of change of variables, we can write each u_i as $u_i = \phi(x_i)$ for some x_i, and we get that the sum above looks like

$$\sum_i (\sup f \text{ on } [\phi(x_{i-1}), \phi(x_i)])(\phi(x_i) - \phi(x_{i-1})).$$

Using the mean value theorem we can write each $\phi(x_i) - \phi(x_{i-1})$ as

$$\phi(x_i) - \phi(x_{i-1}) = \phi'(c_i)(x_i - x_{i-1}),$$

and the upper sum from before becomes

$$\sum_i (\sup f \circ \phi \text{ on } [x_{i-1}, x_i])\, \phi'(c_i)(x_i - x_{i-1}).$$

This final sum is now looks *almost* like an upper sum for the integral on the right in the change of variables formula, except with $\sup(f \circ \phi)\,\phi'(c_i)$ instead of $\sup[(f \circ \phi) \cdot \phi')$. The goal is now is to make a appropriate estimates between this and an actual upper sum for the integral on the right, and between lower sums as well, to get the equality of integrals. Again, you'll look at the details in discussion.

But note that, from another perspective, the rewritten upper sum

$$\sum_i (\sup f \text{ on } [\phi(x_{i-1}), \phi(x_i)])(\phi(x_i) - \phi(x_{i-1}))$$

for the integral on the left is precisely a Riemann-Stieltjes sum with respect to the "weight" function ϕ. Indeed, the integral

$$\int_{\phi(a)}^{\phi(b)} f(u)\,du$$

can be viewed as the Riemann-Stieltjes of f with respect to ϕ, and the point of change of variables is to then say that this Riemann-Stieltjes integral is equivalent to an ordinary Riemann integral for the function $(f \circ \phi)\phi'$. We saw before that ordinary summations can be viewed as Riemann-Stieltjes integrals, and this is now the other extreme where a Riemann-Stieltjes integral is an ordinary integral. The general Riemann-Stieltjes integral lives somewhere inbetween, providing a unified approach to all of these concepts.

One final general comment to make pertaining to integration is that, as we've mentioned before, the definition we have given in terms of upper and lower sums is due to Darboux and not Riemann. Riemann's original definition used *Riemann sums* (no upper/lower adjective), where rather than using supremums and infimums to give the "heights" of rectangles, we use the value of the function at some "sample" points. The definition of "integrable" becomes a bit more involved to state in Riemann's original approach, but in the end the two approaches are equivalent, as you will show on a homework problem.

Integrability vs continuity. We've mentioned (and seen in examples) the idea that the set of points at which a function is discontinuous is in a sense "negligible" as far as integration is concerned since we can make their contributions "small" by picking appropriately small intervals around them. We finish our discussion of integration by making this precise, and clarifying just how different (or similar?) the notions of "integrable" and "continuous" are.

If $f : [a, b] \to \mathbb{R}$ is a function, set $D(f)$ to be the set of points at which f is not continuous:

$$D(f) := \{x \in [a, b] \mid f \text{ is not continuous at } x\}.$$

If f were integrable, we could try to "split" its integral up into one which takes place over the subset of points of $[a, b]$ where f is continuous and another over the subset $D(f)$ where it is not:

$$\int_{[a,b]} f \text{ "}=\text{"} \int_{[a,b]\setminus D(f)} f + \int_{D(f)} f.$$

Of course, we have only defined what it means for a function to be integrable on an interval, so integrating over arbitrary subsets of $\mathbb{R}$ (as in the two expressions on the right) does not make sense, in this course at least. So take what we are saying with a grain of salt—it is only meant to provide some context! (Next quarter you will see how to make integration over more general subsets precise, in which case the above "equalty" will become a literal equality.)

Continuous functions are always integrable, so the $\int_{[a,b]\setminus D(f)} f$ term above should exist since f is continuous at points of $[a, b]\setminus D(f)$. Thus, morally, the only thing which determines whether or

not $\int_{[a,b]} f$ exists is whether or not $\int_{D(f)} f$ exists. But on this domain f is very poorly behaved since it is nowhere continuous, so the only hope of having this integral exist is for $D(f)$ to be "small", where the precise notion of "small" is provided by a set of *measure zero*.

Measure zero. Intuitively, sets of measure zero are the subsets of $\mathbb{R}$ which have zero "length". Here is the definition: a subset $Z \subseteq \mathbb{R}$ has *measure zero* if for any $\epsilon > 0$ there exists a countable collection $I_1, I_2, I_3, \ldots$ of intervals which cover Z such that

$$\sum_{i=1}^{\infty} \text{length}(I_i) \leq \epsilon.$$

This sum is called the *total length* of the collection $\{I_i\}$.

Let us wrap our heads around this definition. Given a countable collection of intervals, its total length is exactly what is sounds like: we are just adding up the lengths of all intervals in the collection. (Of course we should only consider collections where this sum actually exists, i.e. such that $\sum_i I_i$ converges.) If a set Z is covered by such a collection, clearly the "length" of Z should be smaller than or equal to the total length of the collection. The above definition says that a set has measure zero when its "length" is smaller than or equal to any $\epsilon > 0$, so that the "length" of a set of measure zero, if such a thing makes sense, should actually be zero.

Examples. Any finite subset of $\mathbb{R}$ has measure zero. Indeed, suppose that $Z = \{x_1, \ldots, x_n\}$ is a finite subset of $\mathbb{R}$ and fix $\epsilon > 0$. For each i, let $I_i = (x_i - \frac{\epsilon}{2n}, x_i + \frac{\epsilon}{2n})$ be the interval of radius $\frac{\epsilon}{2n}$ around x_i. Then the the collection $\{I_i\}$ covers Z and its total length is

$$\sum_i \text{length}(I_i) = \sum_{i=1}^{n} \frac{\epsilon}{n} = \epsilon.$$

Hence Z has measure zero. Note that this makes sense intuitively: the "length" of a single point is zero, and the "length" of Z is obtained by adding together the (finitely many!) lengths of $\{x_i\}$.

More interestingly, any countable set (such as $\mathbb{Q}$) has measure zero. We have already shown this for countable sets which are finite, so suppose that Z is countably infinite. Since Z is countable, we can list its elements as

$$x_1, x_2, x_3, \ldots.$$

Let $\epsilon > 0$ and for each i let I_i be an interval of length $\frac{\epsilon}{2^i}$ around x_i; so, I_1 is an interval of length $\frac{\epsilon}{2}$ around x_1, I_2 is an interval of length $\frac{\epsilon}{4}$ around x_2, I_3 has length $\frac{\epsilon}{8}$ around x_3, and so on. Then the collection $\{I_i\}$ covers Z and its total length is

$$\sum_{n=1}^{\infty} \text{length}(I_i) = \sum_{n=1}^{\infty} \frac{\epsilon}{2^i} = \epsilon \sum_{n=1}^{\infty} \left(\frac{1}{2}\right)^i = \epsilon,$$

where we use properties of geometric series to compute the final sum. Thus Z has measure zero.

The same argument shows more generally that if $Z_1, Z_2, Z_3, \ldots$ is a collection of countably many sets of measure zero, then their union $\bigcup_i Z_i$ has measure zero as well: we simply pick, for each i, a covering of Z_i of total length smaller than $\frac{\epsilon}{2^i}$, and take the union of all of these to get a big covering of $\bigcup_i Z_i$ of total length smaller than ϵ. Another basic result is that any subset of a set of measure zero has measure zero itself, since a covering for the larger set is also a covering for the subset.

Riemann-Lebesgue. The *Riemann-Lebesgue theorem* sates that a bounded function $f : [a, b] \to \mathbb{R}$ is (Riemann) integrable if and only if its set of discontinuity points $D(f)$ has measure zero, giving

us our sought-after characterization of integrability. In measure theory, to say that a property holds everywhere except for some set of measure zero is to say that it holds *almost everywhere*, so being Riemann integrable is equivalent to being "continuous almost everywhere".

You will prove the backwards direction of Riemann-Lebesgue on the homework, where the idea is the same as the one we've already seen elsewhere: use the measure zero condition to make contributions from subintervals where f might be discontinuous small, and use continuity to control the behavior over the other subintervals. The forward direction is essentially what we proved in the Warm-Up last time. Indeed, for each $k \in \mathbb{N}$ set

$$D_k := \left\{ c \in [a,b] \;\middle|\; \inf_{\delta > 0} \left(\sup_{x,y \in [c-\delta, c+\delta]} |f(x) - f(y)| \right) \geq \frac{1}{k} \right\}$$

to be the set of points at which the oscillation of f is at least $\frac{1}{k}$. Since continuity is equivalent to having zero oscillation, and any positive oscillation is at least as large as $\frac{1}{k}$ for some $k \in \mathbb{N}$, we have

$$D(f) = \bigcup_k D_k.$$

If f is integrable, the Warm-Up from last time shows that each D_k has measure zero (we didn't phrase the Warm-Up in terms of "measure zero", but this is indeed what we proved if you go back and check the statement), so $D(f)$ has measure zero as well since it is the countable union of sets of measure zero.

Examples. Since the discontinuity set of a continuous function is empty and the empty set has measure zero, the Riemann-Lebesgue theorem immediately implies that continuous functions on closed intervals are always integrable. A piecewise continuous function has a finite set of discontinuity points, so since finite sets always have measure zero, Riemann-Lebesgue again implies that a piecewise continuous function on $[a,b]$ is integrable.

The function $f : [0,1] \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \in \mathbb{R} \backslash \mathbb{Q}. \end{cases}$$

is discontinuous everywhere, so $D(f) = [0,1]$, which does not have measure zero and hence f is not integrable. (Actually, it is not so obvious that $[0,1]$ does not have measure zero according to our definition of "measure zero", but this is true, although it takes some care to prove correctly. This is better saved for next quarter when you will learn about measure theory proper.) My favorite function on $[0,1]$ is discontinuous at each rational, so its set of discontinuity points is countable and hence has measure zero. Thus my favorite function on $[0,1]$ is integrable, as we already knew.

Quick integrability results. Various integrability results we've seen are now easy consequences of Riemann-Lebesgue, although Riemann-Lebesgue is tougher to prove than any of these specific results. If $f, g : [a,b] \to \mathbb{R}$ are both continuous at x, then so is fg, so fg fails to be continuous possibly only at points where f is discontinuous or g is discontinuous, and hence

$$D(fg) \subseteq D(f) \cup D(g).$$

If f and g are integrable, $D(f)$ and $D(g)$ has measure zero, so $D(f) \cup D(g)$ has measure zero, and hence so does $D(fg)$, so fg is integrable. Similarly, $D(f+g) \subseteq D(f) \cup D(g)$, so $f+g$ is integrable.

If g is continuous, then $g \circ f$ is continuous at any point where f is continuous, so $D(g \circ f) \subseteq D(f)$ and we get integrability of $g \circ f$ assuming integrability of f. The Riemann-Lebesgue theorem is indeed quite powerful, at least after we've proven it, which again you'll do (the remaining direction) on the homework.

Lecture 6: Uniform Convergence

Warm-Up. Suppose $f, g : [a, b] \to \mathbb{R}$ are integrable and satisfy $f(x) < g(x)$ for all $x \in [a, b]$. We claim that then

$$\int_a^b f < \int_a^b g.$$

This seems like something which is completely intuitive, but giving a proper proof actually requires some real work, essentially amounting to proving the Riemann-Lesbesgue theorem or something close to it. We absolutely have $\int_a^b f \leq \int_a^b g$ by monotonicity of the integral, but to see that integration also preserves *strict* inequalities requires that we move beyond looking at upper/lower sums alone. The issue is that a strict inequality $f(x) < g(x)$ only guarantees $\sup f \leq \sup g$ for example, so that $U(f, P)$ might be equal to $U(g, P)$. Even if we could guarantee that $U(f, P) < U(g, P)$ for all partitions, we could still have

$$\inf\{U(f, P)\} = \inf\{U(g, P)\}.$$

Note that $g - f > 0$ in our setup, and so by inearity of the integral our claim is equivalent to

$$\int_a^b (g - f) > 0.$$

Here is the key point: such a strict inequality holds in the setting of *continuity*, and integrable implies continuous almost everywhere! That is, $g - f > 0$ is integrable and hence continuous at at least one point $x_0 \in [a, b]$ by Riemann-Lebesgue since the set of points where it is discontinuous as measure zero and hence cannot be all of $[a, b]$. But for a nonnegative function, being positive at a point at which it is continuous is in fact enough to guarantee that its integral is positive, as you are asked to show on the homework. Hence since $g - f$ is nonnegative, continuous at x_0, and $g(x_0) - f(x_0) > 0$, we do get that

$$\int_a^b (g - f) > 0$$

as claimed. (Being continuous *somewhere* is key, so you can get around using the full-blown Riemann-Lebesgue theorem by showing that there is at least point where $g - f$ is continuous, and for this the Warm-Up of Lecture 4 is enough.)

Pointwise convergence. Our focus now shifts to the study of *function spaces*, which are "spaces" whose elements are functions. We aim to develop many of the same notions we've already seen before—such as continuity and compactness—in this setting.

As a first step, we want a notion of what it means for one function to be "close" to another, or for a sequence of functions to get "closer and closer" (i.e., "converge") to a fixed function. (A sequence of functions is just a sequence $f_1, f_2, f_3, \ldots$ where each $f_n : X \to \mathbb{R}$ is a real-valued function on some common domain X.) Here is our first version of convergence:

> A sequence (f_n) of functions $f_n : X \to \mathbb{R}$ *converges pointwise* to a function $f : X \to \mathbb{R}$ if for every $x \in X$, the sequence of values $(f_n(x))$ converges to the number $f(x)$ in $\mathbb{R}$

So, $f_n \to f$ pointwise if, point-by-point, the values of f_n get arbitrarily close to the values of f. If so, we call f the *pointwise limit* of the sequence (f_n). Note that here we are not varying the inputs into the function: for each fixed $x \in X$, we evaluate all f_n at this one point, and ask for convergence of the resulting sequence in $\mathbb{R}$. Note also that pointwise limits, if they exist, are unique because limits in $\mathbb{R}$ are unique, so that for each x there is only to which thing $(f_n(x))$ could converge.

Examples. The sequence (f_n) defined by $f_n(x) = \frac{1}{n}\sin x$ converges pointwise on $\mathbb{R}$. Indeed, for fixed $x \in \mathbb{R}$ we have

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \tfrac{1}{n}\sin x = 0,$$

so $f_n(x) \to 0$ for all $x \in \mathbb{R}$. The pointwise limit of this sequence is thus the constant zero function.

The sequence (g_n) defined by

$$g_n(x) = \cos(\tfrac{x}{n}) + \sqrt{x^2 + \tfrac{1}{n}}$$

converges pointwise on $[-1, 1]$. For fixed $x \in \mathbb{R}$, we have $\frac{x}{n} \to 0$ and $x^2 + \frac{1}{n} \to x^2$, so continuity of cosine and the square root function give

$$\cos(\tfrac{x}{n}) + \sqrt{x^2 + \tfrac{1}{n}} \to \cos(0) + \sqrt{x^2} = 1 + |x|.$$

Thus $g_n \to g$ pointwise where $g(x) = 1 + |x|$. (Note that there is nothing special about $[-1, 1]$ here, and indeed $g_n \to g$ pointwise on all of $\mathbb{R}$ as well. We are restricting the domain for the purpose of another property we will soon talk about.)

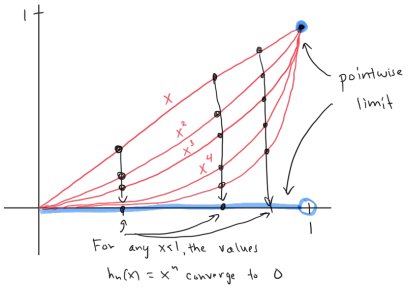
Consider now the sequence $h_n : [0, 1] \to \mathbb{R}$ where $h_n(x) = x^n$. So, our sequence looks like

$$x,\ x^2,\ x^3,\ x^4,\ \ldots,$$

each viewed as functions on $[0, 1]$. For $x = 1$, we have $h_n(1) = 1$ for all n, so $h_n(1) \to 1$ as $n \to \infty$. But for $0 \leq x < 1$, we have that $h_n(x) = x^n \to 0$ as $n \to \infty$, so we conclude that this sequence converges pointwise to the function $h : [0, 1] \to \mathbb{R}$ defined by

$$h(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x = 1. \end{cases}$$

Graphically this convergence looks like



Uniform convergence. The example $h_n(x) = x^n$ on $[0, 1]$ is a standard one which shows that continuity is not preserved under pointwise convergence: each h_n is continuous on $[0, 1]$, but the pointwise limit is not due to the "jump" at 1. Ideally, we want a notion of convergence where continuity *is* preserved, with the intuition being that if two functions f and g are "close" to one another in appropriate sense, continuity of one should transfer over to continuity of the other. Pointwise convergence is not a strong notion of convergence of guarantee this, which makes sense: pointwise convergence details only what happens point-by-point, with the convergence at one point having no bearing on the convergence elsewhere, whereas continuity depends not only on the behavior of a function at a single point but rather on its behavior *near* that point as well. We need a notion of convergence where the behavior of functions at all points matters.

This is provided by the notion of *uniform convergence*:

A sequence (f_n) of functions $f_n : X \to \mathbb{R}$ *converges uniformly* to $f : X \to \mathbb{R}$ if for all $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$|f_n(x) - f(x)| < \epsilon \text{ if } n \geq N \text{ for all } x \in X.$$

If so, we call f the *uniform limit* of (f_n).

The "for all $x \in X$" is the key point: to say that $f_n \to f$ pointwise means that for each x we can guarantee $f_n(x)$ is within ϵ of $f(x)$ for all n large enough, but the large enough n after which this starts to happen might change as x does. As we change x perhaps we have to go further in our sequence $(f_n(x))$ to guarantee that $f_n(x)$ is close to $f(x)$, with no restriction on how large n needs to be, whereas to say that $f_n \to f$ uniformly means that we can find one single N which guarantees $f_n(x)$ is close to $f(x)$ if $n \geq N$ *uniformly* for all x once. (The use of the term "uniform" here should remind you of how it is used in the notion of "uniform continuity", where in that case one δ works for all points at once. In general, "uniform" properties are ones which can be guaranteed to hold in a way which is independent of any one point we are looking at.)

Note in general that uniform convergence implies pointwise convergence, since fixing x in the definition of pointwise convergence immediately gives the definition of what it means for $(f_n(x))$ to converge to $f(x)$ in $\mathbb{R}$. Thus, if a sequence is going to converge uniformly, the only to which it could converge uniformly is its pointwise limit, assuming this pointwise limits exists.

Back to examples. We saw before that the functions $f_n(x) = \frac{1}{n}\sin x$ converged pointwise to $f = 0$ on $\mathbb{R}$. We claim that this convergence is actually uniform. This follows from

$$|\tfrac{1}{n}\sin x - 0| = \tfrac{1}{n}|\sin x| \leq \tfrac{1}{n} \text{ for all } x \in \mathbb{R}.$$

If $\epsilon > 0$, picking $N \in \mathbb{N}$ such that $\frac{1}{N} < \epsilon$ then gives $|\frac{1}{n}\sin x - 0| \leq \frac{1}{n} \leq \frac{1}{N} < \epsilon$ if $n \geq N$ for all $x \in \mathbb{R}$. Practically, what made this possible is that fact that we were able to find a *uniform* (meaning independent of x) bound $|\frac{1}{n}\sin x| \leq \frac{1}{n}$ on our function values that could be made arbitrarily small.

The sequence $g_n(x) = \cos(\frac{x}{n}) + \sqrt{x^2 + \frac{1}{n}}$ on $[-1, 1]$ also converges uniformly. We computed the pointwise limit before to be $g(x) = 1 + |x|$, which we now want to explicitly think of as being $g(x) = \cos 0 + \sqrt{x^2}$. In order to justify uniform convergence we consider

$$|g_n(x) - g(x)| \leq |\cos(\tfrac{x}{n}) - \cos(0)| + |\sqrt{x^2 + \tfrac{1}{n}} - \sqrt{x^2}|.$$

For fixed $\epsilon > 0$, we aim to make each term on the right uniformly smaller than $\frac{\epsilon}{2}$. For the first term, we can use the mean value theorem: $\cos(\frac{x}{n}) - \cos(0) = -\sin(c)(\frac{x}{n} - 0)$ for some c between 0 and $\frac{x}{n}$, so

$$|\cos(\tfrac{x}{n}) - \cos(0)| = |\sin(\tfrac{c}{n})||\tfrac{x}{n}| \leq \tfrac{|x|}{n} \leq \tfrac{1}{n},$$

where the last inequality follows from the fact that we are working on the interval $[-1, 1]$. Thus for n large enough we have the uniform estimate we want.

For the square root term, we use the fact that $|\sqrt{a} - \sqrt{b}| \leq \sqrt{|a - b|}$ for $a, b \geq 0$. (If you haven't seen this inequality before, it can be justified through some algebraic manipulations after squaring both sides.) This gives

$$|\sqrt{x^2 + \tfrac{1}{n}} - \sqrt{x^2}| \leq \sqrt{|(x^2 + \tfrac{1}{n}) - x^2} = \sqrt{\tfrac{1}{n}},$$

which again we make uniformly small. To put it all together, fix $\epsilon > 0$ and pick $N \in \mathbb{N}$ such that $\frac{2}{\sqrt{N}} < \frac{\epsilon}{2}$. Then for $n \geq N$ and any $x \in [-1, 1]$, we have

$$\begin{aligned} |g_n(x) - g(x)| &\leq |\cos(\tfrac{x}{n}) - \cos(0)| + |\sqrt{x^2 + \tfrac{1}{n}} - \sqrt{x^2}| \\ &\leq \frac{1}{n} + \frac{1}{\sqrt{n}} \\ &\leq \frac{2}{\sqrt{n}} \\ &\leq \frac{2}{\sqrt{N}} \\ &< \epsilon, \end{aligned}$$

so $g_n \to g$ uniformly on $[-1, 1]$ as claimed.

Note that we can replace $[-1, 1]$ here by any bounded interval and get uniform convergence on such domains as well. However, we cannot extend this to get uniform convergence $g_n \to g$ on *all* of $\mathbb{R}$. Certainly the argument above will not work as is since we would not be able to uniformly bound $\frac{x}{n}$ from the

$$|\cos(\tfrac{x}{n}) - \cos(0)| = |\sin(\tfrac{c}{n})||\tfrac{x}{n}| \leq \tfrac{|x|}{n}$$

estimate we used if we allowed $x \in \mathbb{R}$. Of course, just knowing that this argument no longer works is not enough to conclude that the convergence cannot be uniform on $\mathbb{R}$, but instead we can argue as follows. First, the square root part definitely convergences uniformly on $\mathbb{R}$ using

$$|\sqrt{x^2 + \tfrac{1}{n}} - \sqrt{x^2}| \leq \sqrt{|(x^2 + \tfrac{1}{n}) - x^2} = \sqrt{\tfrac{1}{n}},$$

and one can show that the sum of uniformly convergent sequences is uniformly convergent, so $g_n \to g$ uniformly on $\mathbb{R}$ if and only if $\cos(\frac{x}{n}) \to 1$ uniformly on $\mathbb{R}$. But for any $N \in \mathbb{N}$ we have for $x = \pi N/2$ that $|\cos(\frac{x}{N}) - 1| = |\cos(\pi/2) - 1| = 1$, so that $|\cos(x/n) - 1|$ cannot be made uniformly smaller than, say, $\frac{1}{2}$ on all of $\mathbb{R}$ no matter how large n is. The upshot is that the domain matters when discussing these notions of convergence.

Graphical interpretation. The sequence $h_n(x) = x^n$ on $[0, 1]$ does not converge uniformly to its pointwise limit

$$h(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x = 1. \end{cases}$$

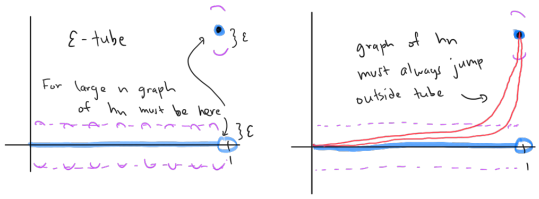
This we can derive as a consequence of the fact that uniform convergence preserves continuity, which we will prove next time, but it is also helpful to understand this graphically as follows. The condition for uniform convergence is that

$$|h_n(x) - h(x)| < \epsilon \text{ for all } x \text{ and large enough } n,$$

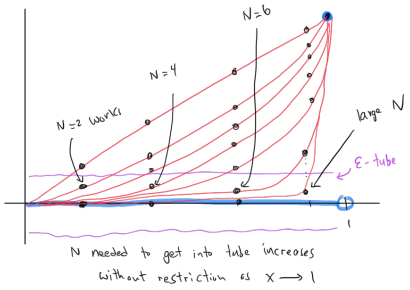
which we can write as

$$h_n(x) \in (h(x) - \epsilon, h(x) + \epsilon) \text{ for all } x \text{ and large enough } n.$$

This says that the entire graph of h_n must lie "within ϵ" of the graph of h once n is large enough; that is, if at every point x we move a distance ϵ away (vertically up and down) from $h(x)$, we get an "ϵ-tube" around the graph of h which must contain the graph of h_n for large n:

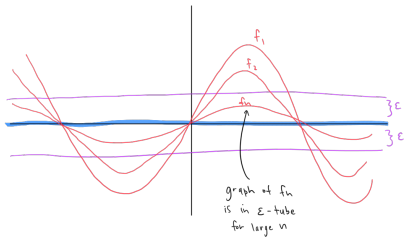


But in this case, no matter how large n is, the graph of h_n must eventually (as we get closer to $x = 1$) jump outside this tube, for $\epsilon = \frac{1}{4}$ for example, since the graph of h_n must hit $h_n(x) = 1$ at $x = 1$. Thus, visually, there is no N for which the graph of h_N lies within ϵ of the graph of h, so the convergence is not uniform. The issue is that, although for each $x \in [0, 1]$ there is an N for which $|h_n(x) - h(x)| < \epsilon$ can be guaranteed, the N that works increases as $x \to 1$, so there will be no single N that works for all $x \in [0, 1]$ at once:

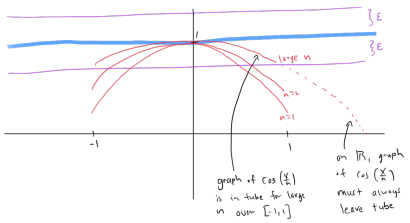


Note that even if we exclude $x = 1$ from our domain (thereby avoiding the issue that the limit is not continuous), the convergence is still not uniform for these graphical reasons. (To be more precise than just "graphicall", if the convergence were uniform on $[0, 1)$, it would also be uniform on $[0, 1]$ since $h_n(1) - h(1) = 0$ anyway, so that include this one extra point will not effect whether $|h_n(x) - h(x)| < \epsilon$ holds or not.) On any interval $[0, b)$ with $b < 1$ which is bounded away from 1, however, we do have $x^n \to 0$ uniformly.

For the example of $\frac{1}{n} \sin x \to 0$ uniformly on $\mathbb{R}$, the graphs look like



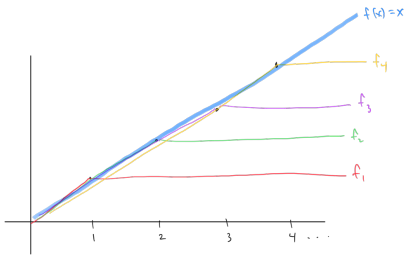
so that sure enough, once n is large enough, the entire graph of $f_n(x) = \frac{1}{n} \sin x$ lies within a fixed ϵ-tube around the graph of 0. For $\cos(\frac{x}{n}) \to 1$ uniformly on $[-1, 1]$ but not on $\mathbb{R}$, the picture is



.

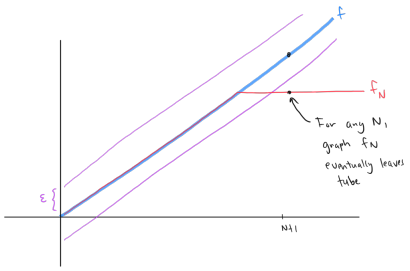
Lecture 7: More on Uniform Convergence

Warm-Up 1. Define $f_n : [0, \infty) \to \mathbb{R}$ by $f_n(x) = \max\{n, x\}$. We claim that f_n converges pointwise on $[0, \infty)$ but not uniformly. First, let us draw the graphs to make the intuition clear:



For fixed x, we have $f_n(x) = x$ once $n > x$, so $\lim_{n\to\infty} f_n(x) = x$ and thus (f_n) converges pointwise on $[0, \infty)$ to $f(x) = x$.

Graphically, the convergence is not uniform since given any "tube" around the graph of $f(x) = x$, the graph of any f_n eventually lies outside the tube as x increases. But to be more precise, for any $N \in \mathbb{N}$, we have $|f_N(N+1) - f(N+1)| = |N - (N+1)| = 1$, so $|f_n(x) - f(x)|$ cannot be made uniformly smaller than, say, $\frac{1}{2}$ no matter how large n is:



Warm-Up 2. We show that the uniform limit of bounded functions is bounded, so that boundedness is a property preserved by uniform convergence. Suppose $f_n : X \to \mathbb{R}$ are bounded functions converging uniformly to f. Then we have

$$|f_N(x) - f(x)| < 1 \text{ for all } x \in X$$

and some large N, which implies

$$|f(x)| < 1 + |f_N(x)| \text{ for all } x \in X.$$

If $M > 0$ is a bound on f_N, then $1 + M$ is a bound on f, so f is bounded as claimed.

Note that this gives another approach to the first Warm-Up: each f_n is bounded on $[0, \infty)$, but $f(x) = x$ is not, so the convergence is not uniform.

Supremum metric. To say that $f_n \to f$ uniformly X means that for all $\epsilon > 0$ we have

$$|f_n(x) - f(x)| < \epsilon \text{ for all } x \in X \text{ and large } n.$$

But since this is meant to hold for all x, we get that the supremum of the terms on the left is at most ϵ. Conversely, if this supremum is at most ϵ, then we get the inequality above required of uniform convergence. Even better: we can phrase this using $\sup < \epsilon$ and not just $\sup \leq \epsilon$, since

if we had the latter for every ϵ, picking smaller ϵ's would give the former for every ϵ. That is, the requirement of uniform convergence can be rephrased as saying that for all $\epsilon > 0$,

$$\sup_{x \in X} |f_n(x) - f(x)| < \epsilon \text{ for large enough } n,$$

where we turn the "uniform" requirement into a requirement on supremums instead.

The supremum above defines a metric on a space of functions, so that uniform convergence really takes its proper context in the setting of metric spaces. If we use $F(X)$ to denote the set of all functions $X \to \mathbb{R}$, then we will use $B(X)$ to denote the subset of those functions which are bounded:

$$B(X) := \{f : X \to \mathbb{R} \mid f \text{ is bounded}\}.$$

On this set we define the *supremum metric* by

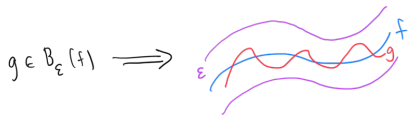
$$d(f, g) = \sup_{x \in X} |f(x) - g(x)|.$$

One can show that this is indeed a metric (which you should do if you didn't do it last quarter!), and the upshot is that uniform convergence in $B(X)$ is precisely convergence with respect to this particular metric. (Note that we work in $B(X)$ since boundedness guarantee that the supremum we are using always exists. Uniform convergence, of course, can also apply to unbounded functions, like $\sqrt{x^2 + \frac{1}{n}} \to \sqrt{x^2}$ uniformly on $\mathbb{R}$, it's just that in such a case we cannot easily interpret it as metric convergence in a concrete metric space.)

With respect to this metric, the pictures we've been drawing of "ϵ-tubes" are then just pictures of ϵ-balls, or more precisely the graphs of functions in ϵ-balls. Indeed, to say that $g \in B_\epsilon(f)$ means that

$$d(f, g) = \sup_{x \in X} |f(x) - g(x)| < \epsilon,$$

which means that $|f(x) - g(x)| < \epsilon$ for all x, and hence that the graph of g lies within the ϵ-tube around the graph of f:



(Actually, we get graphs within the ϵ-tube which do not come arbitrarily close to the edges of the tube, since if the graph of g did come are arbitrarily close to the edge we would have that $\sup |f(x) - g(x)|$ *equals* ϵ rather than being smaller than ϵ.)

Uniform preserves continuity. We now show that uniform convergence preserves continuity: if each $f_n : X \to \mathbb{R}$ is continuous and $f_n \to f$ uniformly, then $f : X \to \mathbb{R}$ is continuous. In the language of metric spaces, this implies that if

$$C(X) := \{f \in B(X) \mid f \text{ is continuous}\}$$

denotes the set of bounded continuous functions on X, then $C(X)$ is a *closed* subset of $B(X)$ since $C(X)$ will contain all of its limit points, where we are considering $B(X)$ equipped with the sup metric. This result makes intuitive sense: if f_n is continuous so that its values cannot differ too wildly from one another, and if the values of f are uniformly close to those of f_n, then the values of f should also not differ too wildly from one another.

To prove this, fix $y \in X$, let $\epsilon > 0$, and use uniform convergence to pick $N \in \mathbb{N}$ such that

$$|f_N(x) - f(x)| < \tfrac{\epsilon}{3} \text{ for all } x \in X.$$

Since f_N is continuous at y, there exists $\delta > 0$ such that

$$|f_N(x) - f(y)| < \tfrac{\epsilon}{3} \text{ when } |x - y| < \delta.$$

Thus if $|x - y| < \delta$, we have

$$|f(y) - f(x)| \leq |f(y) - f_N(y)| + |f_N(y) - f_N(x)| + |f_N(x) - f(x)| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.$$

(To be clear, we control the first and third terms using uniform convergence—note we are evaluating these at different points, which is why pointwise convergence is not enough—and we control the second term using continuity.) This shows that f is continuous at y, so f is continuous on X. If each f_n were uniformly continuous, the same argument would show that f is uniformly continuous.

Pointwise vs integration. The next claim is that uniform convergence preserves integration, but first we give examples showing that, again, pointwise convergence is not enough. This also makes sense: integration depends on the behavior of a function over an entire interval, so knowing what happens only point-by-point is not good enough. Enumerate the countably many rationals in $\mathbb{Q} \cap [0, 1]$ as

$$\mathbb{Q} \cap [0, 1] = \{r_1, r_2, r_3, \ldots\}$$

and define $f_n : [0, 1] \to \mathbb{R}$ to be 1 at $r_1, \ldots, r_n$ and 0 elsewhere. Then each f_n is integrable since it is discontinuous only at the finitely many $r_1, \ldots, r_n$, but the pointwise limit of the f_n is

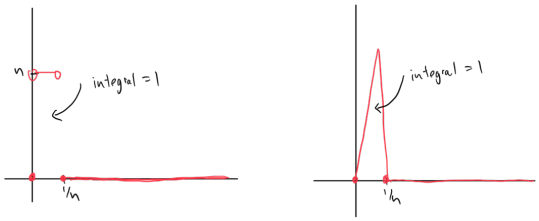
$$f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q}, \end{cases}$$

which is not integrable on $[0, 1]$. (In the limit we include more and more rationals at which the value is 1, which is why we get the function f above in the end.)

Even if the pointwise limit is integrable, it is also not true that values of the integrals themselves are preserved, which would be a nice property to have: if f and g are "close", $\int_a^b f$ and $\int_a^b g$ should ideally be "close" as well. For this take $g_n : [0, 1] \to \mathbb{R}$ to be

$$g_n(x) = \begin{cases} n & \text{if } 0 < x < \frac{1}{n} \\ 0 & \text{otherwise.} \end{cases}$$

Then $\int_0^1 g_n = 1$ for all n, but the g_n's converge pointwise to $g(x) = 0$ (since for each fixed $x \in (0, 1]$ we have $g_n(x) = 0$ once $\frac{1}{n} < x$), whose integral is 0. Hence $\int_0^1 g_n \not\to \int_0^1 g$ in this case. We can even cook up an example of this where each g_n is actually continuous by using some triangular shapes instead:



(Here g_n as defined above is on the left and on the right would be the continuous modification.)

Uniform preserves integration. But everything works fine with uniform convergence: integrability is preserved, as are the values of the integrals. To be clear, we claim that if the $f_n : [a,b] \to \mathbb{R}$ are integrable and $f_n \to f$ uniformly, then f is integrable on $[a,b]$ and

$$\lim_{n\to\infty} \int_a^b f_n = \int_a^b f.$$

(This equality says that the integration and limit operations can be exchanged under uniform convergence: $\lim_{n\to\infty} \int_a^b f_n = \int_a^b (\lim_{n\to\infty} f_n)$.) In the language of metric space, this says that the set $R([a,b])$ of Riemann integrable functions is closed in $B([a,b])$ with respect to the sup metric.

Here's a proof of integrability using Darboux sums. Let $\epsilon > 0$ and pick $N \in \mathbb{N}$ such that

$$|f(x) - f_N(x)| < \frac{\epsilon}{3(b-a)} \text{ for all } x \in [a,b].$$

Then $\sup(f - f_N) \leq \frac{\epsilon}{3(b-a)}$ and $\inf(f - f_N) \geq -\frac{\epsilon}{3(b-a)}$ on any subinterval of $[a,b]$, so that

$$U(f - f_N, P) = \sum \sup(f - f_N) \cdot \text{length} \leq \sum \frac{\epsilon}{3(b-a)} \cdot \text{length} = \frac{\epsilon}{3}$$

and similarly $L(f - f_N, P) \geq -\frac{\epsilon}{3}$. Thus if we pick a partition P of $[a,b]$ such that

$$U(f_N, P) - L(f_N, P) < \frac{\epsilon}{3},$$

using integrability of f_N, we have

$$\begin{aligned} U(f,P) - L(f,P) &\leq U(f - f_N, P) + U(f_N, P) - L(f_N, P) - L(f - f_N, P) \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \\ &= \epsilon, \end{aligned}$$

where in the first step we think of f and $f = (f - f_N) + f_N$ and use $U(g+h,P) \leq U(g,P) + U(h,P)$ and $L(g+h,P) \geq L(g,P) + L(h,P)$. This shows that f is integrable on $[a,b]$.

For a quicker proof making use of Riemann-Lebesgue, note that our proof of "uniform preserves continuity" actually showed that if each f_n is continuous at some y, then f is continuous at y as well. Thus, if f is not continuous at y, then some f_n is not continuous at y, so

$$D(f) \subseteq \bigcup_n D(f_n)$$

where the D's denote discontinuity sets. If each f_n is integrable, each $D(f_n)$ has measure zero, and hence so does their union, and thus so does $D(f)$. Hence f is integrable. (Note that we already know f is bounded since uniform convergence preserves boundedness by the second Warm-Up.)

To see that the values of the integrals are preserved, for $\epsilon > 0$ pick $N \in \mathbb{N}$ such that

$$|f_n(x) - f(x)| < \frac{\epsilon}{b-a} \text{ for } n \geq N \text{ and all } x \in [a,b].$$

Then if $n \geq N$, we have

$$\left| \int_a^b f_n - \int_a^b f \right| = \left| \int_a^b (f_n - f) \right| \leq \int_a^b |f - f_n| < \int_a^b \tfrac{\epsilon}{b-a} = \epsilon,$$

which shows that the numbers $\int_a^b f_n$ converge to the number $\int_a^b f$. (This what we mean by saying that the values of the integrals are "preserved".)

Lecture 8: Uniform Completeness

Warm-Up. We compute the limit

$$\lim_{n\to\infty} \int_0^2 e^{x^2/n}\, dx.$$

The point is that the given integral cannot be computed directly since $e^{x^2/n}$ has no antiderivative expressible in terms of the basic functions we all know and love, so trying to compute the integral and then limit leads to nowhere. Instead, we want to be able to exchange the limit and integration operations, and for this we need to know that the sequence $e^{x^2/n}$ converges uniformly on $[0, 2]$.

For fixed x, $e^{x^2/n} \to e^0 = 1$, so the pointwise limit of $e^{x^2/n}$ is the constant function 1. To establish uniform convergence we need to make

$$|e^{x^2/n} - 1|$$

uniformly small. Here we exploit the fact that the exponential function is increasing to say that $1 = e^0 \leq e^{x^2/n} \leq e^{4/n}$ for $x \in [0, 2]$, so

$$|e^{x^2/n} - 1| = e^{x^2/n - 1} \leq e^{4/n} - 1.$$

For $\epsilon > 0$, we can thus pick $N \in \mathbb{N}$ such that $e^{4/n} - 1 < \epsilon$—using the fact that the sequence $e^{4/n}$ converges to 1 in $\mathbb{R}$, which in turn uses continuity of the exponential function—and we get

$$|e^{x^2/n} - 1| \leq e^{4/n} - 1 < \epsilon \text{ for all } x \in [0, 2],$$

so the convergence $e^{x^2/n} \to 1$ is uniform on $[0, 2]$. Thus we have

$$\lim_{n\to\infty} \int_0^2 e^{x^2/n}\, dx = \int_0^2 \left(\lim_{n\to\infty} e^{x^2/n}\right) dx = \int_0^2 1\, dx = 2,$$

so 2 is the desired value.

What about derivatives? Uniform convergence preserves continuity and integration, so the next question to ask is whether differentiation is preserved as well? That is, if $f_n \to f$ uniformly and each f_n is differentiable, is f differentiable as well and do we have

$$\lim_{n\to\infty} f_n' = (\lim_{n\to\infty} f_n)' = f'?$$

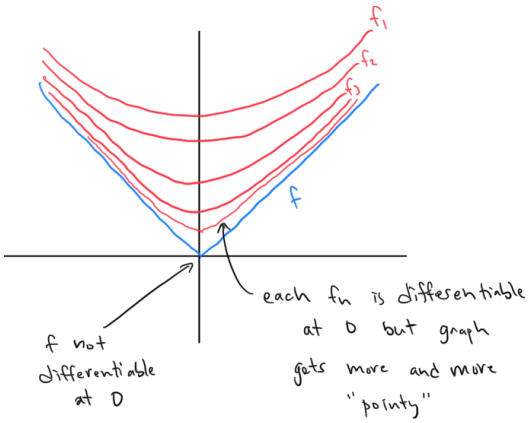
The answer, unfortunately is no. We give examples of what goes wrong below, but first we note that there is no reason why we should expect the answer to be "yes" since derivatives measure how rapidly a function changes, and two functions which are "close" to one another can still change at completely different rates:

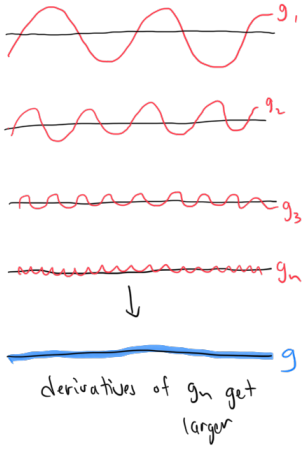
Take $f_n(x) = \sqrt{x^2 + \frac{1}{n}}$, which converges uniformly to $\sqrt{x^2} = |x|$ on an interval (even unbounded) around zero. Each f_n is differentiable everywhere, but the uniform limit $|x|$ is not differentiable at 0, so differentiability is not preserved by uniform convergence. Even if the limit is differentiable, we cannot guarantee that uniform convergence preserves the derivatives in the sense

that $\lim_{n\to\infty} f_n' = (\lim_{n\to\infty} f_n)'$: take $g_n(x) = \frac{\sin(nx)}{\sqrt{n}}$, which is always differentiable and converges uniformly to $g(x) = 0$, which is also differentiable, but

$$g_n'(x) = \sqrt{n}\cos(nx)$$

does not converge uniformly at all let alone to $g'(x) = 0$. Visually these examples look like





Uniform completeness. All is not lost, however, and we will see that under a mild assumption—essentially saying that the derivatives f_n' are not *too* wildly behaved, everything works in the way one might hope. But for this we first need to discuss a completeness property, upon which the correct "uniform preserves differentiation" result will depend.

To say that a sequence of functions (f_n) on X is Cauchy with respect to the sup metric means that for all $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$\sup_{x\in X} |f_n(x) - f_m(x)| < \epsilon \text{ for } n, m \geq N.$$

But, using the same reasoning as for why convergence with respect to the sup metric is equivalent to uniform convergence, this is equivalent to saying that for all $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$|f_n(x) - f_m(x)| < \epsilon \text{ for } n, m \geq N \text{ and all } x \in X.$$

We say that the sequence (f_n) is *uniformly Cauchy* on X in this setting.

The key fact about uniformly Cauchy sequences is that they are always uniformly convergent, just as Cauchy sequences in $\mathbb{R}$ are always convergent. This says that $B(X)$, equipped with the sup metric, is a *complete* metric space; as a consequence, the subspace $C(X)$ of continuous bounded functions is also complete as closed subsets of complete spaces always are. Suppose (f_n) is uniformly Cauchy. Then in particular for each $x \in X$, $(f_n(x))$ is a Cauchy sequence in $\mathbb{R}$, which we get by fixing x in the uniformly Cauchy definition. But $\mathbb{R}$ is complete, so $(f_n(x))$ converges for each $x \in \mathbb{R}$, and we denote the limit of this sequence by $f(x)$:

$$f(x) := \lim_{n\to\infty} f_n(x) \text{ for each } x \in X, \text{ one at a time.}$$

This defines a function $f : X \to \mathbb{R}$, and we claim that (f_n) converges uniformly to this f.

To see this, let $\epsilon > 0$ and pick $N \in \mathbb{N}$ such that

$$|f_n(x) - f_m(x)| < \frac{\epsilon}{2} \text{ for } n, m \geq N \text{ and all } x \in X.$$

For each $x \in X$, we have $f_n(x) \to f(x)$, so there exists $m_x \in \mathbb{N}$ such that $|f_{m_x}(x) - f(x)| < \frac{\epsilon}{2}$. (Note the dependence of m_x on x since at this point we only have pointwise convergence $f_n \to f$.) We can make m_x larger if need to guarantee that $m_x \geq N$ as well, and then

$$|f_n(x) - f(x)| \leq |f_n(x) - f_{m_x}(x)| + |f_{m_x}(x) - f(x)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Thus $|f_n(x) - f(x)| < \epsilon$ for $n \geq \mathbb{N}$ and all $x \in X$, so $f_n \to f$ uniformly.

Note the subtle argument here, where in order to justify uniform convergence we have to make use of non-uniform reasoning! The m_x which we use is chosen non-uniformly as it explicitly depends on x, and yet the actual inequality $|f_n(x) - f(x)| < \epsilon$ we derive in the end *is* uniform since N was chosen uniformly. This is a point which Rudin glosses over, but is essential to our reasoning.

Uniform differentiation. Now we can finally give our version of a sense in which uniform convergence does preserve differentiation. The claim is that if (f_n) is a sequence of differentiable functions on $[a, b]$ which converges uniformly to $f : [a, b] \to \mathbb{R}$, *and* the sequence of derivatives (f_n') converges uniformly as well, then f is differentiable and $f_n' \to f'$ uniformly. So, as long as the derivatives f_n' converge uniformly to *something*, so that they are not too wildly behaved, everything is OK in the world.

The proof uses a clever rephrasing of what differentiability means together with some of the key properties of uniform convergence we have already developed. Fix $x \in [a, b]$ and introduce the auxiliary functions

$$\phi_n(t) = \begin{cases} \frac{f_n(t) - f_n(x)}{t - x} & \text{if } t \neq x \\ f_n'(x) & \text{if } t = x. \end{cases}$$

Note that f_n being differentiable at x is equivalent to continuity of ϕ_n at x:

$$\lim_{t \to x} \frac{f_n(t) - f_n(x)}{t - x} = f_n'(x) \iff \lim_{t \to x} \phi_n(t) = \phi_n(x).$$

For now denote the uniform limit of the f_n by g, so that sequence ϕ_n above converges at least pointwise (recall $f_n \to f$) to

$$\phi(t) = \begin{cases} \frac{f(t) - f(x)}{t - x} & \text{if } t \neq x \\ g(x) & \text{if } t = x. \end{cases}$$

Now, here is the point: the statements that f is differentiable at x *and* $f'(x) = \lim_{n \to \infty} f_n(x)$ are together equivalent to the single statement that ϕ is continuous at x since

$$f'(x) := \lim_{t \to x} \frac{f(t) - f(x)}{t - x} \text{ exists and equals } g(x) \iff \lim_{t \to x} \phi_n(t) \text{ exists and equals } \phi(x).$$

Thus, if we can show that ϕ is continuous, we will be finished.

To show that ϕ is continuous we exploit continuity of the ϕ_n: if we can show that $\phi_n \to \phi$ uniformly, then we get continuity of ϕ automatically. (This is our first example of showing that a function is continuous *not* by verifying it satisfies the definition of continuity directly, but rather by showing that it is the uniform limit of functions which are already known to be continuous!) And to show that $\phi_n \to \phi$ uniformly we actually avoid any mention of ϕ altogether and show instead that the ϕ_n are uniformly Cauchy. If so, they ϕ_n will converge uniformly, and since ϕ is at least the pointwise limit of the ϕ_n, then the thing to which the ϕ_n converge uniformly must be ϕ itself. (This is overall an amazing argument which may take a few read throughs to grasp in full!)

For $t \neq x$, we have

$$\phi_n(t) - \phi_m(t) = \frac{f_n(t) - f_n(x)}{t - x} - \frac{f_m(t) - f_m(x)}{t - x} = \frac{[f_n(t) - f_m(t)] - [f_n(x) - f_m(x)]}{t - x}.$$

Applying the mean value theorem to the function $f_n - f_m$ gives

$$\phi_n(t) - \phi_m(t) = \frac{[f_n(t) - f_m(t)] - [f_n(x) - f_m(x)]}{t - x} = \frac{(f_n'(c_t) - f_m'(c_t)(t - x)}{t - x} = f_n'(c_t) - f_m'(c_t)$$

for some c_t between t and x. Since (f_n') converges uniformly, (f_n') is uniformly Cauchy, so we can make $|f_n'(t) - f_m'(t)|$ uniformly small. Thus for $\epsilon > 0$ we can pick $N \in \mathbb{N}$ such that

$$|\phi_n(t) - \phi_m(t)| = |f_n'(c_t) - f_m'(c_t)| < \epsilon \text{ for } n, m \geq N \text{ and all } t \neq x.$$

We can also include $t = x$ here since we know $|\phi_n(x) - \phi_m(x)| = |f_n'(x) - f_m'(x)| < \epsilon$ by the (f_n') being uniformly Cauchy condition. Thus (ϕ_n) is uniformly Cauchy, so it converges to its pointwise limit ϕ uniformly, so ϕ is continuous since each ϕ_n is continuous, so f is differentiable at x and $f'(x) = \lim_{n\to\infty} f_n'(x)$. (Phew!)

Lecture 9: Contractions

Warm-Up. In the differentiability result we proved last time, we assumed that $f_n \to f$ uniformly on $[a, b]$ with each f_n differentiable, and that (f_n') converged uniformly in order to get the result that f is then differentiable and $f' = \lim_{n\to\infty} f_n'$. Actually, we can get away with a bit less and assume only that $f_n(x_0)$ converges for *some* $x_0 \in [a, b]$, while still assuming that (f_n') converges uniformly. We prove that this weaker assumption alone already guarantees that (f_n) converges uniformly, so that we end up in the setting of the previous result. The point is that what matters is the control we have over (f_n'), and not so much the behavior of (f_n) originally.

To get a sense of why we should expect that knowing how (f_n') and $(f_n(x_0))$ behave alone is enough to know how (f_n) behaves, we appeal to the mean value theorem. We get that for each x there exists $c_{x,n}$ (dependent on both x and n) such that

$$f_n(x) = f_n(x_0) + f_n'(c_{x,n})(x - x_0).$$

The intuition is that by knowing that $(f_n(x_0))$ converges and that (f_n') converges, we can control the behavior of the right side, thereby controlling the left side. This alone does not give an honest proof, however, since we cannot use the uniform convergence of the f_n' as is because the points at which these derivatives are being evaluated change as n does. Instead, we show that (f_n) converges uniformly by forgetting about any mention of a potential limit and instead show that (f_n) is uniformly Cauchy.

For $x \in [a, b]$, applying the mean value theorem to the function $f_n - f_m$ gives

$$[f_n(x) - f_m(x)] - [f_n(x_0) - f_m(x_0)] = [f_n'(c) - f_m'(c)](x - x_0)$$

for some c between x and x_0. (This mean value application is the same idea we used in the uniform differentiation proof last time.) This then gives

$$\begin{aligned} |f_n(x) - f_m(x)| &\leq |[f_n(x) - f_m(x)] - [f_n(x_0) - f_m(x_0)]| + |f_n(x_0) - f_m(x_0)| \\ &\leq |f_n'(c) - f_m'(c)||x - x_0| + |f_n(x_0) - f_m(x_0)| \\ &\leq |f_n'(c) - f_m'(c)|(b - a) + |f_n(x_0) - f_m(x_0)| \end{aligned}$$

where in the first step we subtracted and added $f_n(x_0) - f_m(x_0)$ before using the triangle inequality and in the final step we use that $x, x_0 \in [a, b]$ to say that $|x - x_0| \leq b - a$. Everything now can be controlled: for $\epsilon > 0$, pick $N \in \mathbb{N}$ such that

$$|f_n'(x) - f_m'(x)| < \frac{\epsilon}{2(b-a)} \text{ for } n, m \geq N \text{ and all } x$$

using uniform convergence (which implies uniformly Cauchy) of the f_n', and pick $M \in \mathbb{N}$ such that

$$|f_n(x_0) - f_m(x_0)| < \frac{\epsilon}{2} \text{ for } n, m \geq N$$

(no uniformity needed here since x_0 is a single point), to get that for $n, m \geq \max\{N, M\}$,

$$|f_n(x) - f_m(x)| \leq |f_n'(c) - f_m'(c)|(b-a) + |f_n(x_0) - f_m(x_0)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

for all $x \in [a, b]$, so that (f_n) is uniformly Cauchy as desired.

A detour. We will now take a bit of a detour and consider a topic, namely *contractions*, which Rudin does not cover until later in the setting of multivariable differentiation. Contractions play a key role in the proof of the general inverse function theorem, which will be one of the last results we derive this quarter, but contractions have other uses beyond this alone and the type of result we will prove regarding them has much broader importance in mathematics. Indeed, the reason why we consider this topic now in these notes is to give one application of contractions to function spaces, which just so happens to be my favorite application of all time.

Before jumping into this, we recall (or prove if you did not see this last quarter) one metric space fact: if (p_n) is a sequence in a metric space X for which there exists $0 \leq r < 1$ satisfying

$$d(p_{n+1}, p_n) \leq r^n \text{ for all } n,$$

then (p_n) is Cauchy. The intuition is that as n increases r^n decreases since $0 \leq r < 1$, so the distance between successive terms in (p_n) is decreasing as well by at least a fixed factor r at each step. This should imply that all terms (not just successive ones) get closer and closer as we go in in the sequence, which is the Cauchy condition. Note that knowing that $d(p_{n+1}, p_n)$ decreases (but not necessarily by at least a fixed factor $0 \leq r < 1$) is not enough to guarantee being Cauchy as the sequence $x_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$ of partial sums of the harmonic series shows.

For any n and $k > 0$ we have

$$\begin{aligned} d(p_{n+k}, p_n) &\leq d(p_{n+k}, p_{n+k-1}) + d(p_{n+k-1}, p_{n+k-2}) + \cdots + d(p_{n+2}, p_{n+1}) + d(p_{n+1}, p_n) \\ &\leq r^{n+k-1} + r^{n+k-2} + \cdots + r^{n+1} + r^n, \end{aligned}$$

where the first step comes from repeated applications of the triangle inequality using all intermediate points between p_n and p_{n+k}. The resulting sum is a difference of partial sums of the geometric series $\sum_n r^n$, so since this series converges (this is where $0 \leq r < 1$ is used), the sequence of partial sums is Cauchy so we can make the sum above smaller than any $\epsilon > 0$ for large enough n and arbitrary $k > 0$. Hence (p_n) is Cauchy, and therefore if X is complete, (p_n) will converge. The same argument works if we have

$$d(p_{n+1}, p_n) \leq Cr^n$$

for some fixed constant $C \geq 0$ since this only multiplies all expressions used above by C (in particular we end up using convergence of the series $\sum_n Cr^n$) and does not affect the ability to make things small.

Banach contraction principle. A *contraction* on a metric space X is a function $f : X \to X$ such that there exists $0 \leq K < 1$ satisfying

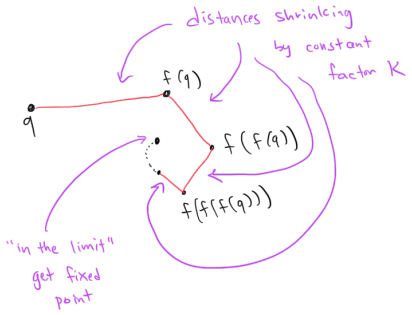
$$d(f(p), f(q)) \leq Kd(p, q) \text{ for all } p, q \in X.$$

As the name suggests, contractions "shrink" distances—the distance between two outputs is always smaller than the distance between the inputs—and do so by at least a fixed factor of $0 \leq K < 1$. The point is that things get "smaller" or "closer" when applying contractions in a way which we can control. Note that contractions are always continuous—in fact uniformly continuous—since $d(p, q) < \frac{\epsilon}{K}$ (in the $K \neq 0$ case) implies $d(f(p), f(q)) \leq Kd(p, q) < \epsilon$.

The fact we need about contractions is the *Banach contraction principle*, which also known as the *Banach fixed point theorem*:

> If X is a complete metric space and $f : X \to X$ is a contraction, then f has a unique fixed point, which is a point $p \in X$ such that $f(p) = p$.

So, contractions on complete spaces always leave one, and only one, point unchanged upon application. The intuition is that if we start with any random $q \in X$, applying f over and over and again gives points that are getting closer and closer to one another, thereby "clustering" near *something* which can no longer be made close to anything else upon applying f because it remains as is:



The proof essentially turns this intuition into a proper argument. Take any $q \in X$ and consider the sequence

$$q,\ f(q),\ f(f(q)),\ f(f(f(q)),\ \ldots$$

of iterates. We use $f^n(q)$ to denote the point obtained by applying f n times to q. Since f is a contraction, we get that

$$d(f^{n+1}(q), f^n(q)) \leq Kd(f^n(q), f^{n-1}(q)).$$

Using the contraction property again gives

$$d(f^{n+1}(q), f^n(q)) \leq Kd(f^n(q), f^{n-1}(q)) \leq KKd(f^{n-1}(q), f^{n-2}(q)),$$

and so on: each time we "unwind" an application of f using the contraction property, we introduce a new factor of K, so that in the end we get

$$d(f^{n+1}(q), f^n(q)) \leq K^n d(f(q), q).$$

But this then implies that the sequence $p_n := f^n(q)$ is Cauchy, so it converges, say to $p \in X$, since X is complete. Because $f^n(q) \to p$ and f is continuous, we get $f(f^n(q)) \to f(p)$. But $f(f^n(q)) = f^{n+1}(q)$ is just a subsequence of the original $f^n(q)$, so it must converge to the same thing—namely p—as the original sequence, and thus by uniqueness of limits we get $f(p) = p$,

so that p is the fixed point we want! Uniqueness is then an easy consequence of the contraction property: if $f(p) = p$ and $f(p') = p'$, then

$$d(p, p') = d(f(p), f(p')) \leq K d(p, p'),$$

which implies that $d(p, p') = 0$ since $0 \leq K < 1$, so the fixed point $p = p'$ is unique.

My favorite application. We will make use of the Banach contraction principle at the end of the course in proving the inverse function theorem, but, as promised, we use it here to give my favorite application in all of mathematics. Consider the differential equation

$$f'(x) = 3xf(x)^2 - \log(e^{\sin(\cos x)} + 1)$$

with initial value condition $f(1) = 1$. A solution of this initial value problem is a function $f(x)$ which satisfies both the differential equation and the initial value condition. For simpler equations, solutions can at times be found explicitly; for example, $f(x) = e^x$ satisfies $f'(x) = f(x), f(0) = 1$, and $f(x) = \frac{1}{1-x}$ satisfies $f'(x) = -f(x)^2, f(0) = 1$. But for general equations there is no hope of finding an explicit solution, so how can we know that one even exists? We claim that for the initial value problem above there does exist a solution, in fact a unique one, on a small enough interval $I = [1 - \delta, 1 + \delta]$ around 1.

The idea in proving this is to recast the problem in a different way so that other non-obvious tools suddenly become available. After integrating both sides of the given equation, we see that a function f satisfies

$$f'(x) = 3xf(x)^2 - \log(e^{\sin(\cos x)} + 1)$$

if and only if it satisfies

$$f(x) = c + \int_1^x [3tf(t)^2 - \log(e^{\sin(\cos t)} + 1)]\, dt$$

for some constant c. Indeed, if f is continuous, the fundamental theorem of calculus implies that the integral expression on the right is differentiable with respect to x and that its derivative is the integrand evaluated at $t = x$, so that taking derivatives of both sides indeed reproduces our original differential equation. (Note that if we assume only that f is continuous, it might not be clear that the derivative of the left side $f(x)$ even exists, but the point is that it will as a consequence of the fact that this left side equals the differentiable expression given on the right side.) The constant c is determined by the initial condition $f(1) = 1$: since an integral from 1 to 1 is always zero, we need $c = 1$ in order to have $f(1) = 1$. Thus, the upshot is that a function f satisfies

$$f'(x) = 3xf(x)^2 - \log(e^{\sin(\cos x)} + 1) \text{ with initial condition } f(1) = 1$$

if and only if it satisfies the single integral equation:

$$f(x) = 1 + \int_1^x [3tf(t)^2 - \log(e^{\sin(\cos t)} + 1)]\, dt.$$

So our goal is now to show that there is a function satisfying this integral equation. It might not seem that we've made much progress, but here is the amazing observation which makes everything work out: we can rephrase this integral equation as a fixed-point problem! Indeed, consider the metric space $C(I)$ of continuous functions on $I = [1 - \delta, 1 + \delta]$ for some to-be-determined constant

$\delta > 0$ equipped with the sup metric, and define the map $T : C(I) \to C(I)$ by setting, for each $f \in C(I)$, Tf to be the function on I whose value at $x \in I$ is:

$$(Tf)(x) = 1 + \int_1^x [3tf(t)^2 - \log(e^{\sin(\cos t)} + 1)]\, dt.$$

The function Tf is continuous again since the right side is differentiable by the fundamental theorem of calculus. Then, saying that f satisfies

$$f(x) = 1 + \int_1^x [3tf(t)^2 - \log(e^{\sin(\cos t)} + 1)]\, dt$$

is the same as saying that the function Tf equals f itself, so that what we want is to establish that T has a unique fixed point on some small I.

To show that T as defined above has a unique fixed point we will appeal to the Banach Contraction Principle: since $C(I)$ is complete with respect to the sup metric, showing that $T : C(I) \to C(I)$ is a contraction (guaranteeing that it is a contraction is where the choice of a small $\delta > 0$ will come in) implies that T has a unique fixed point, and my favorite application will be complete. We will work out the contraction details next time.

Lecture 10: Series of Functions

Warm-Up. We show that the map $T : C(I) \to C(I)$ defined by

$$(Tf)(x) = 1 + \int_1^x [3tf(t)^2 - \log(e^{\sin(\cos t)} + 1)]\, dt$$

is a contraction for a small enough $I = [1 - \delta, 1 + \delta]$. (Actually, we'll have to restrict the definition of $C(I)$ as well, as we'll see.) Recall the point of this is that, once we know we have a contraction, the completeness of $C(I)$ guarantees that T has a unique fixed point f, and this fixed point then satisfies the initial value problem

$$f'(x) = 3xf(x)^2 - \log(e^{\sin(\cos x)} + 1), \;\; f(1) = 1,$$

thereby showing that this initial value problem has a solution—in fact a unique one—on some small enough interval around 1.

The claim is that there exists $0 \leq K < 1$ such that $d(Tf, Tg) \leq Kd(f, g)$ for and all $f, g \in C(I)$, which if we spell out the details of the sup metric becomes

$$\sup_{x \in I} |(Tf)(x) - (Tg)(x)| \leq K \sup_{x \in I} |f(x) - g(x)|.$$

We have (noting that the constants 1 in the definitions of $(Tf)(x)$ and $(Tg)(x)$ subtract away)

$$\begin{aligned} |(Tf)(x) - (Tg)(x)| &= \left| \int_1^x [3tf(t)^2 - \log(e^{\sin(\cos t)} + 1)]\, dt - \int_1^x [3tf(t)^2 - \log(e^{\sin(\cos t)} + 1)]\, dt \right| \\ &= \left| \int_1^x 3t(f(t)^2 - g(t)^2)\, dt \right| \\ &\leq \int_{\min\{1,x\}}^{\max\{1,x\}} 3|t||f(t)^2 - g(t)^2|\, dt \end{aligned}$$

$$= \int_{\min\{1,x\}}^{\max\{1,x\}} 3|t||f(t)+g(t)||f(t)-g(t)|\,dt.$$

The $|f(t)-g(t)|$ term will be bounded by $d(f,g)$, which is what will end up giving the right side of the contraction property.

What we need now, however, is a way to bound $|f(t)+g(t)|$. Certainly for fixed f and g we can find a bound since f and g(continuous on compact) themselves are bounded, but we actually need a *uniform* bound on all functions under consideration since the contraction constant K we are deriving should be independent of whatever functions we use. The set of all continuous functions on I is not going to be uniformly bounded, so we must restrict the types of functions we consider and consider only those which, say, are bounded by 10:

$$C(I,[-10,10]) := \{f \in C(I) \mid f \text{ maps } I \text{ into } [-10,10]\}.$$

(The specific bound 10 we use here is irrelevant—we just need some bound larger than 1 to account for the eventual initial value requirement that $f(1)=1$.) After all, if the fixed point we desire is meant to satisfy $f(1)=1$ and be continuous, it's value near 1 should not differ too much from $f(1)=1$, so we do not really lose anything by making such a restriction. The restricted space of functions $C(I,[-10,10])$ which take values only in $[-10,10]$ is still complete (the proof of complements in the $I \to \mathbb{R}$ case depended only on the fact that the codomain $\mathbb{R}$ was complete, so since $[-10,10]$ is complete the same argument works), so the Banach contraction principle still applies. In order to ensure that the functions Tf we get as outputs are indeed in $C(I,[-10,10])$ (so that T maps this space into itself), we shrink our eventual $\delta > 0$ if needed. Let us also say that δ will be small enough so that $[1-\delta,1+\delta] \subseteq [0,2]$ so that we can use 2 as a bound on the variable of integration t.

With f and g in this restricted space of functions, and for the small enough δ, we have

$$\begin{aligned}
|(Tf)(x)-(Tg)(x)| &\leq \int_{\min\{1,x\}}^{\max\{1,x\}} 3|t||f(t)+g(t)||f(t)-g(t)|\,dt \\
&\leq \int_{\min\{1,x\}}^{\max\{1,x\}} 60|t||f(t)-g(t)|\,dt \\
&\leq \int_{\min\{1,x\}}^{\max\{1,x\}} 120\,d(f,g)\,dt \\
&\leq 120\delta\,d(f,g)
\end{aligned}$$

for $x \in I = [1-\delta,1+\delta]$. This gives that the supremum of the terms on the left is bounded by $120\delta\,d(f,g)$, so

$$d(Tf,Tg) \leq 120\delta\,d(f,g).$$

Thus by picking $\delta < \frac{1}{120}$ (which already guarantees $I \subseteq [0,2]$), we get that T is a contraction as desired. (As mentioned before, we also have to guarantee that Tf is within the same space $C(I,[-10,10])$ of functions with restricted codomain; since

$$\begin{aligned}
|Tf(x)| &\leq 1 + \int_{\min\{1,x\}}^{\max\{1,x\}} \left(3|t||f(t)|^2 + \log(e^{\sin(\cos t)}+1)\,dt\right) \\
&\leq 1 + \int_{\min\{1,x\}}^{\max\{1,x\}} (600+\log(4))\,dt \\
&\leq 1 + 700\delta,
\end{aligned}$$

where we use $|t| \leq 2$ and $|f(t)| \leq 10$, the $\delta < \frac{1}{120}$ we already have does guarantee that $|Tf(x)| \leq 10$. If it did not, we would just make δ smaller.)

Fun with contractions. Contractions are powerful tools in many areas of mathematics. The general version of this existence and uniqueness result for solutions of differential equations—called the *Picard-Lindelöf* theorem—is proved by the same type of fixed-point/contraction argument, and is something you will look at on the homework. Note that not only does this result guarantee the existence of a solution, it also gives a way to approximate it. Start with any continuous function f whatsoever, and form the sequence of iterates

$$f,\ Tf,\ T^2f,\ T^3f,\ \dots.$$

The proof of the contraction principle shows that this sequence of iterates converges to the fixed point of T, and hence to the solution of our differential equation. This gives the method known as *Picard iteration* for approximating solutions of differential equations. (For the equation $f'(x) = f(x), f(0) = 1$, the corresponding contraction is defined by $(Tf)(x) = 1 + \int_0^x f(t)\,dt$. For a fun thing to do on your own, take the constant function $f = 1$ as a starting point and compute the iterates above; you will get a well known sequence of polynomials which does indeed, as we will show next week, converge to the unique function satisfying $f'(x) = f(x), f(0) = 1$.)

As mentioned previously, we will also use contractions to prove the inverse function theorem at the end of this course, which is a completely different type of application. On the homework you will also see an example of obtaining well-known sets as fixed points of certain contractions, so the takeaway is that contractions are everywhere. This is my favorite application because it best exemplifies to me what the point of modern mathematics is, and why we spend so much effort dealing with abstraction: once we are able to prove things about abstract concepts, we are able to apply the results to a wide range of different scenarios all at once!

Series of functions. The next natural thing to do after considering sequences of functions is to add them together to get a series of functions like

$$\sum_n f_n = f_1 + f_2 + f_3 + \cdots.$$

Convergence of series is defined in terms of convergence of the sequence of partial sums, so we say that $\sum_n f_n$ converges ADJECTIVE if the sequence of partial sums

$$f_1 + \cdots + f_n$$

converges ADJECTIVE, where we put in place of ADJECTIVE whatever type of convergence we want, such as pointwise, uniformly, or absolutely. In the case of uniform convergence, properties of the terms f_n carry over to properties of the sum: if $\sum_n f_n$ converges uniformly on whatever domain we need, the function $f = \sum_n f_n$ is

- continuous if each f_n is continuous (since the partial sums $f_1 + \cdots + f_n$ are then continuous),
- integrable if each f_n is integrable (since the partial sums are then integrable) *and* we have

$$\int_a^b f = \int_a^b \sum_n f_n = \sum_n \int_a^b f_n$$

so that sums can be integrated term-by-term (since $\int (f_1 + \cdots + f_n) = \int f_1 + \cdots + \int f_n$),

- differentiable if each f_n is differentiable *and* $\sum_n f_n'$ also converges uniformly, in which case the sum can be differentiated term-by-term

$$f' = \left(\sum_n f_n\right)' = \sum_n f_n',$$

  which all comes from considering the analogous result for the sequence of partial sums and using $(f_1 + \cdots + f_n)' = f_1' + \cdots + f_n'$.

The point of doing all the work before relating uniform convergence to continuity, integrability, and differentiability is that we no longer have to check these properties by hand for functions defined via series—we get them for free as long as we can verify uniform convergence.

Weierstrass M-test. In order to make such results useful, we need a clean way of checking for uniform convergence. The standard approach is the following, which is typically called the *Weierstrass M-test* and exploits what we know about series of numbers already. The claim is that if (f_n) is a sequence of functions for which we can find bounds $|f_n| \leq M_n$ such that (the numerical series) $\sum_n M_n$ converges, then $\sum_n f_n$ converges uniformly on whatever domain these bounds hold. (In fact, we get absolute uniform convergence, which implies usual uniform convergence.)

Indeed, for $m \geq n$, we have

$$|f_n + \cdots + f_m| \leq |f_n| + \cdots + |f_m| \leq M_n + \cdots + M_m.$$

Thus if $\sum_n M_n$ converges, we can make the sum on the right however small we like (since the sequence of partial sums of $\sum_n M_n$ is Cauchy), and this implies that the sequence of partial sums of $\sum_n |f_n|$ is uniformly Cauchy, so that $\sum_n f_n$ converges absolutely and uniformly.

Example. We derive some properties of the function defined on $(0, \infty)$ by

$$f(x) = \sum_{n=0}^{\infty} e^{-nx} = 1 + e^{-x} + e^{-2x} + \cdots.$$

First we argue that this function is actually well-defined, meaning that the series on the right converges. For this we directly jump to verifying uniform convergence. We have

$$|e^{-nx}| = \frac{1}{e^{nx}} = \left(\frac{1}{e^x}\right)^n.$$

To bound this uniformly by some constant M_n requires that we make x as small as possible, but this we cannot do on all of $(0, \infty)$ at once. (Note that $\frac{1}{e^x} \leq \frac{1}{e^0} = 1$ will not help since $\sum_n 1$ does not converge.) So, we instead fix $a > 0$ and consider only convergence on $[a, \infty) \subseteq (0, \infty)$ for the time being. On this domain we have

$$|e^{-nx}| = \frac{1}{e^{nx}} \leq \frac{1}{e^{ax}} \leq \left(\frac{1}{e^a}\right)^n.$$

Since $a > 0$, $\frac{1}{e^a} < 1$, so $\sum_n (\frac{1}{e^a})^n$ converges, and hence $\sum_n e^{-nx}$ converges uniformly on $[a, \infty)$ by the M-test. Thus $f(x) = \sum_n e^{-nx}$ defines a function on $[a, \infty)$, which is in fact continuous since e^{-nx} is continuous and the convergence on $[a, \infty)$ is uniform.

Now by taking a to approach 0 we can extend the domain of f to be all of $(0, \infty)$, and still have continuity on all of $(0, \infty)$. To be clear, we are not claiming that $\sum_n e^{-nx}$ converges uniformly on

all of $(0, \infty)$, which is not true; rather, the argument is that for any $x \in (0, \infty)$, there is an interval $[a, \infty)$ for some $a < x$ that contains x, so that uniform convergence on $[a, \infty)$ implies convergence and continuity at x specifically. Since each e^{-nx} is integrable on any $[a, b] \subseteq (0, \infty)$, we also get integrability of $f(x) = \sum_n e^{-nx}$ for free (ok, in this case this can be obtained from continuity alone), and

$$\int_a^b f(x)\,dx = \sum_{n=0}^{\infty} \int_a^b e^{-nx}\,dx = \sum_{n=0}^{\infty} \frac{e^{na} - e^{nb}}{n}.$$

To show that $f(x) = \sum_n e^{-nx}$ is differentiable on $(0, \infty)$, we show it is differentiable on each $[a, \infty) \subseteq (0, \infty)$, and for this we need to know that $\sum_n f_n'$ converges uniformly. We have

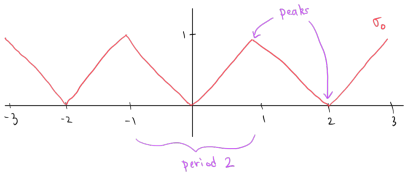
$$|f_n'(x)| = \left|-ne^{-nx}\right| \leq \frac{n}{e^{an}}$$

for $x \in [a, \infty)$. Since $\sum_n \frac{n}{e^{an}}$ converges (say, by the root test), we get that $\sum_n ne^{-nx}$ does converge uniformly on $[a, \infty)$, so that $f(x) = \sum_n e^{-nx}$ is differentiable on this domain. Again by taking a to approach zero, we get differentiability on all of $(0, \infty)$ and that

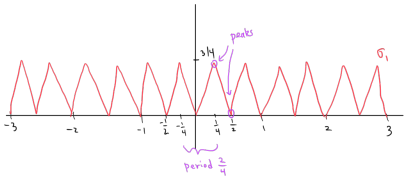
$$f'(x) = \sum_{n=0}^{\infty} (e^{-nx})' = \sum_{n=0}^{\infty} -ne^{-nx}.$$

Continuous nowhere differentiable. With these tools at hand we can now give a famous example in analysis, that of a function which is continuous on all of $\mathbb{R}$ yet nowhere differentiable. (I believe you showed at the end of last quarter that such functions exist—and in fact that, in a sense, "most" continuous functions are nowhere differentiable—but giving an explicit example of such a function was out of reach back then.) The function we want will be defined via a uniformly convergent series, so that continuity is not something we will have to check by hand.

First take σ_0 to be the function defined by $\sigma_0(x) = |x|$ for $-1 \leq x \leq 1$, and then extended elsewhere to have period 2:



Note that σ_0 is continuous everywhere, fails to be differentiable at the integers where the "peaks" (high or low) are, and has size $|\sigma_0| \leq 1$. Now shrink the period by a factor of 4 and the size by a factor of $\frac{3}{4}$ to get the function $\sigma_1(x) := \frac{3}{4}\sigma_0(4x)$:



This has period $\frac{2}{4}$, is still continuous, has size $|\sigma_1| \leq \frac{3}{4}$, and fails to be differentiable at quarters of integers. Do the same thing to σ_1 to get $\sigma_2(x) := \frac{3}{4}\sigma_1(4x) = (\frac{3}{4})^2\sigma_0(4^2x)$, and keeping going to define in general

$$\sigma_k(x) := (\tfrac{3}{4})^k \sigma_k(4^k x).$$

The σ_k's are all continuous on $\mathbb{R}$, have smaller and smaller sizes, smaller and smaller periods ($\frac{2}{4^k}$ in general), and the points at which they fail to be differentiable (countably many at each step) get lumped closer and closer together:

We claim that the function defined by adding all σ_k's together is the function we want:

$$\sigma(x) := \sum_{k=0}^{\infty} \sigma_k(x).$$

The intuition is that the "lumping" of non-differentiability points as k increases makes it harder and harder for the sum of σ_k's to be differentiable, so that indeed σ as defined above where will be nowhere differentiable! Continuity comes for free from uniform convergence: we have

$$|\sigma_k(x)| \leq \left(\frac{3}{4}\right)^k |\sigma_0(4x)| \leq \left(\frac{3}{4}\right)^k \text{ for all } x \in \mathbb{R},$$

so since $\sum_k (\frac{3}{4})^k$ converges, the sum defining σ converges uniformly on $\mathbb{R}$, so σ is defined for all x and is continuous everywhere.

To show that σ is differentiable requires some work and exploits the specific way in which the σ_k's were defined. Fix $x \in \mathbb{R}$. We want to show that

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

does not exist, so that σ is not differentiable at the arbitrary x. Define the sequence (h_n) by setting

$$h_n = \pm \frac{1}{2 \cdot 4^n},$$

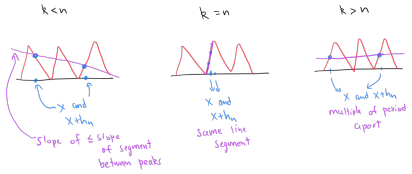
where the choice between plus and minus will be made in a bit. This sequence converges to 0 and we claim that the sequence

$$\frac{\sigma(x+h_n) - \sigma(x)}{h_n}$$

diverges to ∞, which is why the limit above defining $\sigma'(x)$ will not exist. We take this difference quotient for fixed n, and break it up into those terms occurring before $k = n$, the term at $k = n$, and the terms after $k = n$:

$$\begin{aligned} \frac{\sigma(x+h_n) - \sigma(x)}{h_n} &= \sum_{k=0}^{\infty} \frac{\sigma_k(x+h_n) - \sigma_k(x)}{h_n} \\ &= \sum_{k<n} \frac{\sigma_k(x+h_n) - \sigma_k(x)}{h_n} + \frac{\sigma_n(x+h_n) - \sigma_n(x)}{h_n} + \sum_{k>n} \frac{\sigma_k(x+h_n) - \sigma_k(x)}{h_n}. \end{aligned}$$

The pictures to have in mind for the behavior of each of these pieces are the following:



To start, for $k > n$ we have that $h_n = \pm\frac{1}{2 \cdot 4^n}$ is precisely an integer multiple of the period of σ_k, which is $\frac{2}{4^k}$, since

$$\pm\frac{1}{2 \cdot 4^n} = \pm 4^{k-(n+1)}\frac{2}{4^k} \text{ if } k \geq n+1.$$

This means that $x + h_n$ and x are a multiple of the period apart from one another when $k > n$, so $\sigma_k(x + h_n)$ and $\sigma_k(x)$ have the same value, and thus the sum

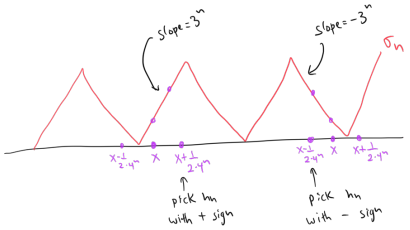
$$\sum_{k>n} \frac{\sigma_k(x + h_n) - \sigma_k(x)}{h_n}$$

vanishes since all numerators involved are zero.

Hence we are left with

$$\frac{\sigma(x + h_n) - \sigma(x)}{h_n} = \sum_{k<n} \frac{\sigma_k(x + h_n) - \sigma_k(x)}{h_n} + \frac{\sigma_n(x + h_n) - \sigma_n(x)}{h_n}.$$

For the $k = n$ term on the right, the point is that $h_n = \pm\frac{1}{2 \cdot 4^n}$ is small enough to guarantee $x + h_n$ and x lie on the same linear "segment" of σ_n, at least for the appropriate choice of sign:



This works because the distance between "peaks" (high or low) in the graph of σ_n is $\frac{1}{4^n}$—half the period—so one of $x - \frac{1}{2 \cdot 4^n}$ or $x + \frac{1}{2 \cdot 4^n}$ lies between the same peaks as x. Thus with these choices of signs we get that

$$\frac{\sigma_n(x + h_n) - \sigma_n(x)}{h_n}$$

is the slope of a linear segment in the graph of σ_n, which is $\pm 3^n$, so that

$$\left|\frac{\sigma_n(x + h_n) - \sigma_n(x)}{h_n}\right| = 3^n.$$

(The slope is 3^n because from $\sigma_n(x) = (\frac{3}{4})^n \sigma_0(4^n x)$ we see that the difference in two function values on the same segment is $(\frac{3}{4})^n$, while the difference in inputs is 4^n, so that the 4^n pieces cancel, leaving the "rise over run" as 3^n. Note that the slopes in the σ_0 case are just ± 1.) For the $k < n$ terms, we do not care to be so precise, and simply use the fact that the "slope" between any two points on the graph of σ_k for $k < n$ is never larger/smaller than the slope of one of the linear segments, which are $\pm 3^k$, so that

$$\left|\frac{\sigma_k(x + h_n) - \sigma_k(x)}{h_n}\right| \leq 3^k \text{ for } k < n.$$

Thus, a reverse triangle inequality gives

$$\left|\frac{\sigma(x + h_n) - \sigma(x)}{h_n}\right| = \left|\sum_{k<n} \frac{\sigma_k(x + h_n) - \sigma_k(x)}{h_n} + \frac{\sigma_n(x + h_n) - \sigma_n(x)}{h_n}\right|$$

$$\begin{aligned}
&\geq \left|\frac{\sigma_n(x+h_n)-\sigma_n(x)}{h_n}\right| - \left|\sum_{k<n}\frac{\sigma_k(x+h_n)-\sigma_k(x)}{h_n}\right| \\
&\geq 3^n - (1+3+3^2+\cdots+3^{n-1}) \\
&= 3^n - \frac{1-3^n}{1-3} \\
&= \frac{1}{2}(3^n+1).
\end{aligned}$$

As $n \to \infty$, this diverges to ∞, so $\frac{\sigma(x+h_n)-\sigma(x)}{h_n} \to \infty$ and thus

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

does not exist as claimed, so σ is not differentiable at any $x \in \mathbb{R}$. (Phew!)

Lecture 11: Arzela-Ascoli Theorem

Warm-Up. We show that

$$f(x) = \sum_{n=1}^{\infty}\left(1-\cos\frac{x}{n}\right)$$

defines a differentiable function on all of $\mathbb{R}$. On the closed interval $[-M, M]$, we have

$$\left|1-\cos\tfrac{x}{n}\right| = |-\sin c|\tfrac{|x|}{n} \leq |c|\tfrac{|x|}{n} \leq \tfrac{|x|^2}{n^2} \leq \tfrac{M^2}{n^2},$$

where we use one application of the mean value theorem to get $1-\cos\frac{x}{n} = (-\sin c)\frac{x}{n}$ for some c between 0 and $\frac{x}{n}$, and then another to get $|\sin c| = |\cos d||c| \leq |c|$ for some d between 0 and c. Since $\sum_n \frac{M^2}{n^2}$ converges, the M-test implies that the given series converges uniformly on $[-M, M]$. By taking $M \to \infty$, we thus get that

$$f(x) = \sum_{n=1}^{\infty}\left(1-\cos\frac{x}{n}\right)$$

is well-defined on all of $\mathbb{R}$. Note that f is also continuous on all of $\mathbb{R}$: it is continuous on each $[-M, M]$ by uniform convergence, and hence continuous on $\mathbb{R}$ by taking $M \to \infty$.

To check differentiability of f, we consider the term-by-term derivative series:

$$\sum_{n=1}^{\infty}\frac{1}{n}\sin\frac{x}{n}.$$

On any $[-M, M]$ we have

$$|\tfrac{1}{n}\sin\tfrac{x}{n}| \leq \tfrac{1}{n}\tfrac{|x|}{n} \leq \tfrac{M}{n^2},$$

where use the same $|\sin c| \leq |c|$ as before with $c = \frac{x}{n}$. Again $\sum_n \frac{M}{n^2}$ converges, so $\sum_{n=1}^{\infty}\frac{1}{n}\sin\frac{x}{n}$ converges uniformly on $[-M, M]$. Hence

$$f(x) = \sum_{n=1}^{\infty}\left(1-\cos\frac{x}{n}\right)$$

is differentiable on $[-M, M]$, and thus on all of $\mathbb{R}$ by taking $M \to \infty$.

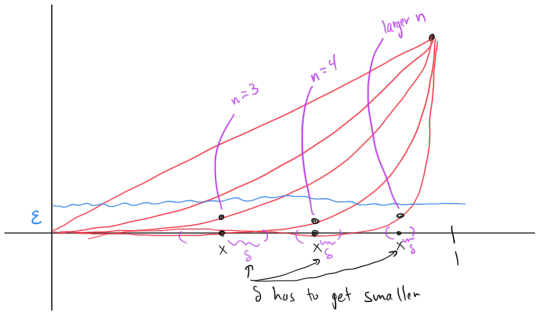
Compactness in function spaces. So far we have developed the notion of convergence (for sequences and series) in function spaces, and seen examples of continuous functions (such as the contractions in differential equation applications) from one function space to another. We have seen examples of closed and open sets in function spaces, and of complete function spaces. The next natural thing to wonder is what compactness looks like in the setting of function spaces. Of course, one answer is "sets of functions for which any open cover has a finite subcover", or "sets of functions where any sequence in that set has a convergent subsequence", but what we are really after is a way to characterize these properties in simpler terms.

In $\mathbb{R}^n$ we know that compactness is equivalent to being closed and bounded, so we first ask if this is true in function spaces as well. To be precise, from now on the function spaces we are talking about are $C([a, b])$—real-valued continuous functions on $[a, b]$ (or perhaps on a more general compact domain K)—equipped with the sup metric, so that convergence means the same thing as uniform convergence. Does closed and bounded in $C([a, b])$ imply compact? The answer is no: the closed ball

$$\overline{B_1(0)} = \{f \in C([0, 1]) \mid d(f, 0) \leq 1\}$$

of radius 1 around the constant zero function on $[0, 1]$ is closed and bounded, but we claim not compact. This is closed since it is the closure of the open ball $B_1(0) = \{f \in C([0, 1]) \mid d(f, 0) < 1\}$, and it is bounded in the metric space sense, which in our current language means that its elements are uniformly bounded, in fact by 1 in this case. But it is not compact since the sequence $f_n(x) = x^n$ in $\overline{B_1(0)}$ has no convergent subsequence: any potential uniformly convergent subsequence would have to converge to its pointwise limit, which is the same discontinuous function defined by $f(x) = 0$ for $0 \leq x < 1$ and $f(1) = 1$ we saw before.

Equicontinuity. But all is not lost, as we will see that closed and bounded *plus* one more property *is* enough to guarantee compactness in these function spaces. To get a sense of what this additional property might be, let us think about the functions x^n on $[0, 1]$ a bit more:



We previously argued (intuitively via the graph) that the reason why this sequence fails to converge uniformly is that given $\epsilon > 0$, to end up within ϵ away from the limit value 0 requires that the n for which this is true get larger and larger as $x \to 1$ from the left. This is due to the fact that we have steeper and steeper slopes of x^n close to $x = 1$ as n increases. But such steep slopes are also a reflection of the δ's needed in the definition of continuity: the $\delta > 0$ needed to end up within ϵ away from x^n has to get smaller and smaller as $x \to 1$ and $n \to \infty$. The reason why no one N is enough to satisfy the definition of uniform convergence is essentially the same for why there is no "minimal" δ that satisfies the definition of continuous for all x^n at once.

The additional property we claim we need to compare compactness to closed and is thus that of being *equicontinuous*, which was a term introduced on a recent homework problem. (In fact,

that homework problem will play a key role in the main result for today.) Recall the definition: a collection S of functions is equicontinuous if they are all continuous "in the same way", meaning that for all $\epsilon > 0$ there exists $\delta > 0$ such that

$$|f(x) - f(y)| < \epsilon \text{ whenever } |x - y| < \delta \text{ for all } f \in S.$$

That is, one δ satisfies the definition of (uniform) continuity for all functions in S at once. Having one single such δ at least avoids the type of bad behavior we saw in the x^n example.

We now show that uniform convergence in $C([a, b])$ (or more a general compact domain) implies equicontinuity, which suggests that equicontinuity is indeed a property we should expect to need if are looking for something that might imply the existence of uniformly convergent subsequences. Suppose (f_n) is an equicontinuous sequence of functions converging uniformly to f in $C([a, b])$. (The limit f must be in $C([a, b])$ as well if each f_n is.) Fix $\epsilon > 0$ and and pick $N \in \mathbb{N}$ such that

$$|f_n(x) - f(x)| < \frac{\epsilon}{3} \text{ for } n \geq \mathbb{N} \text{ and all } x \in [a, b].$$

Since f is continuous on $[a, b]$, it is uniformly continuous so we can pick $\delta > 0$ such that

$$|f(x) - f(y)| < \frac{\epsilon}{3} \text{ whenever } |x - y| < \delta.$$

Thus if $n \geq N$ and $|x - y| < \delta$, we get

$$|f_n(x) - f_n(y)| \leq |f_n(x) - f(x)| + |f(x) - f(y)| + |f(y) - f_n(y)| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon,$$

so that this δ satisfies the definition of equicontinuity for all f_n with $n \geq N$. To include the remaining functions $f_1, \ldots, f_{N-1}$ as well, we simply note that each of these is uniformly continuous on $[a, b]$, so we can pick δ_i for each $1 \leq i \leq N - 1$ to satisfy the definition of uniform continuity, and then $\min\{\delta, \delta_1, \ldots, \delta_n\} > 0$ satisfies the definition for all f_n.

Pointwise plus equicontinuity. On the homework you showed that equicontinuity is enough for pointwise convergence to actually imply uniform convergence on compact domains. In fact, we can get away with a bit less, and assume only pointwise convergence on a countable *dense* subset of the domain. The same argument works as, in the end, using compactness, it comes down to only working with a finite number of points anyway.

Indeed, in the case of domain $[a, b]$, take $[a, b] \cap \mathbb{Q}$ to be our countable dense set. If (f_n) is an equicontinuous sequence in $C([a, b])$ which converges pointwise on $[a, b] \cap \mathbb{Q}$, for fixed $\epsilon > 0$ we pick $\delta > 0$ as in the definition of equicontinuity to make

$$|f_n(x) - f_n(y)| < \frac{\epsilon}{3} \text{ for } |x - y| < \delta.$$

The open balls of radius δ around each point in $[a, b] \cap \mathbb{Q}$ cover all of $[a, b]$ since, by denseness, in any $(x - \delta, x + \delta)$ there exists some $p \in [a, b] \cap \mathbb{Q}$, which means that in turn x is in $(p - \delta, p + \delta)$. Thus we get a finite subcover by taking these open balls only at some $p_1, \ldots, p_n \in [a, b] \cap \mathbb{Q}$. If we pick $N \in \mathbb{N}$ such that

$$|f_n(p_i) - f_m(p_i)| < \frac{\epsilon}{3}$$

for $m, n \geq N$ (pointwise convergent implies pointwise Cauchy) and all $i = 1, \ldots, n$ (possible by taking a maximum of finitely many N's), then for any $x \in [a, b]$ and $m, n \in \mathbb{N}$ we have

$$|f_n(x) - f_m(x)| \leq |f_n(x) - f_n(p_i)| + |f_n(p_i) - f_m(p_i)| + |f_m(p_i) - f_m(x)| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon$$

where p_i is the specific $p_1, \ldots, p_n$ for which $x \in (p_i - \delta, p_i + \delta)$. Hence (f_n) is uniformly Cauchy on $[a, b]$, so it is uniformly convergent.

Arzela-Ascoli. Finally we come to our main compactness result, which gives a general way to guarantee the existence of uniformly convergent subsequences. View this as a function space analog of the Bolzano-Weierstrass theorem, and indeed Bolzano-Weierstrass is crucial to the proof. The *Arzela-Ascoli theorem* (note that Rudin does not use this name) states that if (f_n) is an equicontinuous uniformly bounded sequence in $C([a, b])$, then (f_n) has a uniformly convergent subsequence. Constructing this subsequence takes some real care (as we will see), but note that we need only construct a subsequence that converges pointwise, and indeed only on $[a, b] \cap \mathbb{Q}$ since the result above then gives uniform convergence on all of $[a, b]$.

Enumerate the elements of $[a, b] \cap \mathbb{Q}$ as $[a, b] \cap \mathbb{Q} = \{p_1, p_2, p_3, \ldots\}$. Since (f_n) is uniformly bounded, the sequence $(f_n(p_1))$ of values at p_1 specifically is bounded in $\mathbb{R}$, so it has a convergent subsequence by Bolzano-Weierstrass; we denote this convergent subsequence and its limit by

$$f_{1,1}(p_1),\ f_{1,2}(p_1),\ f_{1,3}(p_1),\ \ldots \longrightarrow y_1.$$

To be clear, the functions $f_{1,k}$ here form a subsequence of the original f_n. We use the double index notation to keep track of the point p_1 we are evaluating at as well as the location k a term occurs at moving horizontally above. Now evaluate the functions $f_{1,k}$ used here at the point p_2 to get a new bounded sequence $(f_{1,k}(p_2))$ in $\mathbb{R}$ which has its own convergent subsequence, say

$$f_{2,1}(p_2),\ f_{2,2}(p_2),\ f_{2,3}(p_2), \ldots \longrightarrow y_2.$$

The $f_{2,k}$'s here come from a subsequence of the previous $f_{1,k}$'s, and thus evaluating the $f_{2,k}$'s at the previous point p_1 maintains the convergence we had previously:

$$f_{2,1}(p_1),\ f_{2,2}(p_1),\ f_{2,3}(p_1),\ \ldots \longrightarrow y_1.$$

Now do the same with p_3: the sequence $(f_{2,k}(p_3))$ has a convergent subsequence

$$f_{3,1}(p_3),\ f_{3,2}(p_3),\ f_{3,3}(p_3),\ \ldots \longrightarrow y_3,$$

and since the $f_{3,k}$'s are a subsequence of the $f_{2,k}$'s (and hence also of the $f_{1,k}$'s) we still have

$$f_{3,1}(p_1),\ f_{3,2}(p_1),\ f_{3,3}(p_1),\ \ldots \longrightarrow y_1,$$

and

$$f_{3,1}(p_2),\ f_{3,2}(p_2),\ f_{3,3}(p_2),\ \ldots \longrightarrow y_2.$$

And so on and so on, we get at the m-th stage a sequence

$$f_{m,1}(p_m),\ f_{m,2}(p_m),\ f_{m,3}(p_m),\ \ldots \longrightarrow y_m$$

where the $f_{m,k}$'s maintain all previous convergences as well when evaluated at the previous p_i's. We end up with a big grid of convergences

$$\begin{array}{c}
f_{1,1}(p_1),\ f_{1,2}(p_1),\ f_{1,3}(p_1),\ \ldots \longrightarrow y_1 \\
f_{2,1}(p_2),\ f_{2,2}(p_2),\ f_{2,3}(p_2),\ \ldots \longrightarrow y_2 \\
f_{3,1}(p_3),\ f_{3,2}(p_3),\ f_{3,3}(p_3),\ \ldots \longrightarrow y_3 \\
\vdots \qquad\qquad\qquad\qquad \vdots
\end{array}$$

$$\begin{array}{ccc} f_{m,1}(p_m),\ f_{m,2}(p_m),\ f_{m,3}(p_m),\ \ldots \longrightarrow y_m \\ \vdots \qquad\qquad\qquad\qquad \vdots \end{array}$$

where each row maintains the previous ones in the sense we've described above.

To get our desired subsequence of the original f_n's, we first pick k_1 such that

$$|f_{1,k_1}(p_1) - y_1| < 1$$

from the first row, and so that this holds for all terms after f_{1,k_1} in the first row. Next we pick k_2 such that

$$|f_{2,k_2}(p_2) - y_2| < \frac{1}{2}$$

from the second row, far enough along so that this is still true for all terms after f_{2,k_2} in the second row, *and* far enough along the first row so that f_{2,k_2} occurs after f_{1,k_2}, which implies that

$$|f_{2,k_2}(p_1) - y_1| < 1$$

as well. Then pick k_3 such that

$$|f_{3,k_3}(p_3) - y_3| < \frac{1}{3}$$

far enough long the third row, and far enough along the first two rows so that f_{3,k_3} is after f_{1,k_1} and after f_{2,k_2}, so that

$$|f_{3,k_3}(p_1) - y_1| < 1 \quad \text{and} \quad |f_{3,k_3}(p_2) - y_2| < \frac{1}{2}.$$

Continue, where at the m-th stage we pick k_m such that

$$|f_{m,k_m}(p_m) - y_m| < \frac{1}{m}$$

and far enough to maintain the previously inequalities for $p_1, \ldots, p_{m-1}$. The resulting subsequence (f_{m,k_m})) of (f_n) then converges pointwise on $\mathbb{Q} \cap [a,b] = \{p_1, p_2, p_3, \ldots\}$ since

$$|f_{m,k_m}(p_i) - y_i| < \frac{1}{i} \text{ for } m \geq i \implies f_{m,k_m}(p_i) \to y_i \text{ for all } i.$$

Since (f_{m,k_m}) converges pointwise on the dense set $\mathbb{Q} \cap [a,b]$, we have that (f_{m,k_m}) converges uniformly, so this indeed is our desired uniformly convergent subsequence of (f_n). The same argument works for $C(K)$ with K being any compact space, where we replace $[a,b] \cap \mathbb{Q}$ by any countable dense subset, which exists in any compact metric space. (Phew again!)

Lecture 12: Weierstrass Approximation

Warm-Up 1. Suppose $f_n : [a,b] \to \mathbb{R}$ is a sequence of differentiable functions with uniformly bounded derivatives, and such that the sequence $(f_n(x_0))$ in $\mathbb{R}$ is bounded for at lest one $x_0 \in [a,b]$. We show that (f_n) has a uniformly convergent subsequence. Clearly (based on the fact that we want to produce a convergent subsequence) this meant to be an Arzela-Ascoli problem, so the goal is to verify that the assumptions of the Arzela-Ascoli theorem are satisfied.

Let $M > 0$ be a uniform bound on the f_n' over $[a, b]$. An application of the mean value theorem gives that for all $x \neq y \in [a, b]$, we have

$$|f_n(x) - f_n(y)| = |f_n'(c_n)||x - y| \leq M|x - y| \text{ for all } n,$$

where the c_n are some numbers between x and y. Thus for $\epsilon > 0$, any $0 < \delta < \frac{\epsilon}{M}$ satisfies the requirement of equicontinuity since

$$|x - y| < \delta \implies |f_n(x) - f_n(y)| \leq M|x - y| < M\frac{\epsilon}{M} = \epsilon \text{ for all } n.$$

Moreover, if we pick $y = x_0$ above, we get

$$|f_n(x)| \leq |f_n(x) - f_n(x_0)| + |f_n(x_0)| \leq M|x - x_0| + |f_n(x_0)| \leq M(b - a) + |f_n(x_0)|$$

for all n and $x \in [a, b]$, so picking a bound on the $f_n(x_0)$ gives a uniform bound on the f_n. Since (f_n) in $C([a, b])$ is thus uniformly bounded and equicontinuous, it has a uniformly convergent subsequence by the Arzela-Ascoli theorem.

Warm-Up 2. We prove the *Heine-Borel theorem for* $C([a, b])$, which is the claim that $K \subseteq C([a, b])$ is compact if and only if K is closed, bounded, and equicontinuous. (Recall the usual Heine-Borel theorem for $\mathbb{R}^n$ is the claim that compact means closed and bounded, so this is now the function space version of this result where equicontinuity is the only extra thing we need.) The backwards direction is just the Arzela-Ascoli theorem: if K is closed, bounded, and equicontinuous, then any sequence (f_n) is K is uniformly bounded and equicontinuous, so Arzela-Ascoli gives a uniformly convergent subsequence, and the fact that K is closed guarantees that the limit of this subsequence remains in K, so K is compact by the sequential characterization of compactness.

For the forward direction, if K is compact we get K being closed and bounded for free since compact implies closed and bounded in any metric space. To show that K is equicontinuous, we use the essentially the same argument as in the "uniform convergence implies equicontinuity" result from last time, only modified to work for all elements of K and not just a single sequence. Let $\epsilon > 0$ and consider the collection of all open balls $B_{\epsilon/3}(f)$ of radius $\epsilon/3$ centered at all elements f of K. This is an open cover, so we get a finite subcover

$$B_{\epsilon/3}(f_1), \ \ldots, \ B_{\epsilon/3}(f_n)$$

by a compactness. Each of these f_i is uniformly continuous, so we can pick a minimal $\delta > 0$ which satisfies the uniform continuity requirement for all of them. If $f \in K$ is any function, then f belongs to some $B_{\epsilon/3}(f_i)$ since these cover K, so we get that if $|x - y| < \delta$, then

$$|f(x) - f(y)| \leq |f(x) - f_i(x)| + |f_i(x) - f_i(y)| + |f_i(y) - f(y)| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon,$$

where the second $\epsilon/3$ comes from uniform continuity of f_i and the first and third $\epsilon/3$'s come from $f \in B_{\epsilon/3}(f_i)$, so that $\sup |f - f_i| \leq \epsilon/3$. Hence K is equicontinuous as claimed.

Another ODE application. To give a sense of how the Arzela-Ascoli theorem is used in practice, let us give another ODE (ordinary differential equation) application. We claim there is a function $f : [0, 1] \to [0, 10]$ which satisfies

$$f'(x) = \sqrt{f(x)} + 1, \ f(0) = 0$$

(On the homework you are asked to prove the Picard-Lindelöf theorem, which is the general version of the contraction ODE application we had before, but this theorem does not apply in this example, as we will explain shortly.) As in our previous ODE application, we rephrase this problem as a fixed point problem instead, where we are thus looking for f which satisfies

$$f(x) = \int_0^x (\sqrt{f(t)} + 1)\, dt.$$

If we take the same approach as before and try to show that

$$(Tf)(x) := \int_0^x (\sqrt{f(t)} + 1)\, dt$$

defines a contraction $T : C([0,1]) \to C([0,1])$ (where really we should consider only those functions with restricted codomain $[0, 10]$, and where we potentially have to shrink $[0, 1]$), we run into the issue that we end up with

$$|(Tf)(x) - (Tg)(x)| \leq \int_0^x (\sqrt{f(t)} - \sqrt{g(t)})\, dt \leq \int_0^x \sqrt{|f(t) - g(t)|}\, dt \leq \delta\sqrt{d(f,g)}$$

for a small δ, which gives

$$d(Tf, Tg) \leq \delta\sqrt{d(f,g)}$$

with d the sup metric. But this cannot be made into a contraction property anymore because of the presence of the square root on the right! Indeed, for small $d(f,g)$, $d(f,g)$ will actually be smaller than $\sqrt{d(f,g)}$, and so the inequality above cannot be used to bound $d(Tf, Tg)$ by a constant times $d(f,g)$ instead. So, the Picard-Lindelöf argument does not work here.

Instead, we approach this via Arzela-Ascoli. (The real point of the contraction argument is to obtain the desired fixed point as the uniform limit of some sequence of iterates, and we get around this by producing the fixed point as the uniform limit of some subsequence of a carefully constructed sequence instead. Note that in this approach there will be no uniqueness claim, since uniqueness in the previous approach came from having a contraction. All we are claiming now is the existence of at least one solution.) The idea is that if f were not a solution to our ODE, so that it did not already satisfy

$$f(x) = \int_0^x (\sqrt{f(t)} + 1)\, dt,$$

we can perhaps try to say something about how far off from being a solution f is by somehow controlling the "error" in

$$f(x) = \int_0^x (\sqrt{f(t)} + 1)\, dt + \text{error}.$$

If the "error" is small, f is close to a solution, so by controlling the error we can try to obtain a solution via a limiting process. We claim that we can construct functions $f_n : [0,1] \to [0,10]$ for which the error is expressible in a particularly nice way:

$$f_n(x) = \int_0^x (\sqrt{f_n(t)} + 1)\, dt + g_n(x)$$

where $g_n(x)$ is some function defined via some type of integral. (We will not go into the construction of these functions here as you will do this in a more general setting on the homework anyway; our goal for now is simply to illustrate how and why Arzela-Ascoli appears.) These error functions g_n will have the property that they converge uniformly to 0, and thus *if* the sequence (f_n) we

construct also converges uniformly, what it converges uniformly to would then be a solution of our ODE since taking limits in

$$f_n(x) = \int_0^x (\sqrt{f_n(t)} + 1)\, dt + g_n(x) \quad \text{gives} \quad f(x) = \int_0^x (\sqrt{f(t)} + 1)\, dt + 0$$

with f the (if it exists) uniform limit of the f_n.

The problem is that the functions f_n constructed here are not guaranteed to converge uniformly. But actually, we do not need the entire sequence to converge uniformly—having a uniformly convergent subsequence is enough! Indeed, the same reasoning as above shows that the uniform limit of a convergent subsequence would have to satisfy

$$f(x) = \int_0^x (\sqrt{f(t)} + 1)\, dt,$$

thereby giving our existence result. Thus to finish off our argument, we need only show that Arzela-Ascoli is application to be carefully constructed f_n mentioned above. In fact, the first Warm-Up gets us much of the way there: we need only show that the derivatives of the f_n are uniformly bounded (it will also be true that all f_n satisfy $f_n(0) = 0$ in this particular setup, so we definitely have boundedness of $(f_n(x_0))$ for at least one x_0) in order to get the existence of a uniformly convergent subsequence.

The fact that the f_n' will be uniformly bounded will come from the initial property they are meant to satisfy

$$f_n(x) = \int_0^x (\sqrt{f_n(t)} + 1)\, dt + g_n(x),$$

so that

$$f_n'(x) = (\sqrt{f_n(x)} + 1) + g_n'(x).$$

The g_n' will be shown to be uniformly bounded by the way in which they are defined, and the first term on the right will be uniformly bounded by the restriction that we only consider functions taking values in $[0, 10]$, and hence we will be finished. Again, rather than finishing this particular example, you will carry out a general version of this argument on the homework to prove what's called the *Peano existence theorem* (no uniqueness!) for solutions of ordinary differential equations.

Weierstrass approximation theorem. The last "topological" notion to discuss in the context of function spaces is that of *denseness*; namely, what do dense subsets of $C([a, b])$ look like? The first result in this direction is the *Weierstrass approximation theorem*, which is the statement that the set of polynomial functions is dense in $C([a, b])$. If you interpret "dense" in terms of sequences, the claim is that given any continuous function $f : [a, b] \to \mathbb{R}$, there is a sequence of polynomials $P_n(x)$ which converges to f uniformly on $[a, b]$, so that continuous functions can always be "uniformly approximated" by polynomials to whatever accuracy we want.

Here is one explicit example of such a convergence. You might recognize the polynomials

$$1,\ 1 + x,\ 1 + x + \tfrac{1}{2}x^2,\ 1 + x + \tfrac{1}{2}x^2 + \tfrac{1}{3!}x^3,\ \ldots$$

as the Taylor polynomials of e^x centered at 0. As we will show next week, this sequence of polynomials does in fact converge uniformly to e^x on any $[a, b]$, so the statement of Weierstrass approximation holds for e^x, at least. But the types of functions for which an analogous Taylor polynomial approach works is quite limited (these are what are called *analytic* functions, which we will study soon enough), whereas Weierstrass approximation is meant to hold for all continuous

functions. Constructing the polynomials that converge to a general continuous f is not an easy endeavor, and before doing so we look at one example of why having such a result can be useful.

Example. Suppose $f : [a, b] \to \mathbb{R}$ is a continuous function such that

$$\int_a^b f(x)\, x^n\, dx = 0 \text{ for all } n \in \mathbb{N}.$$

We claim that this then forces $f = 0$ to be the constant zero function. The intuition is that, by knowing only something about the behavior of polynomial functions, we should be able to deduce (via Weierstrass approximation) information about the behavior of f itself, essentially because in this case integration behaves well with respect to uniform convergence.

So, take a sequence $P_n \to f$ of polynomials converging to f uniformly on $[a, b]$. Then

$$\int_a^b f(x)\, P_n(x)\, dx = 0 \text{ for all } n$$

since we can break up the integral of $f(x)P_n(x)$ into sums of scalar multiples of the $\int_a^b f(x)\, x^n\, dx$, which will all be zero. Since $P_n \to f$ uniformly and f is bounded, we have that $P_n f \to ff$ uniformly as well. (We need f to be bounded here in order to bound the $|f(x)|$ term on the right side of $|f(x)P_n(x) - f(x)f(x)| = |f(x)||P_n(x) - f(x)|$.) Since $P_n f \to f^2$ uniformly on $[a, b]$, we thus get

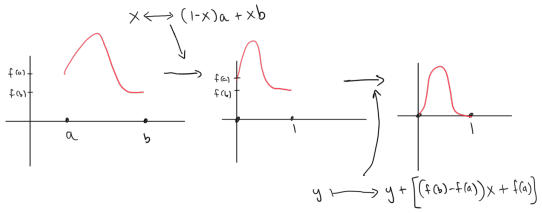
$$\int_a^b f(x)P_n(x)\, dx \to \int_a^b f(x)^2\, dx$$

as $n \to \infty$. The terms on the left are zero, so we get

$$\int_a^b f(x)^2\, dx = 0.$$

Since $f(x)^2$ is nonnegative and continuous, we must have $f(x)^2 = 0$ (see Homework 2), so $f(x) = 0$ for all $x \in [a, b]$ as claimed.

Constructing the polynomials. To prove Weierstrass approximation, we first make a simplification. Our given continuous f is defined on $[a, b]$, but after applying a linear change of variables to x we can assume f is defined on $[0, 1]$ instead:

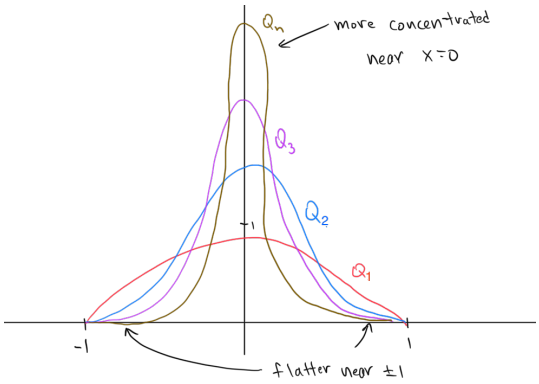


A sequence of polynomials converging uniformly to his new function can be modified to produce one converging uniformly to the original f since making a linear change of variables in a polynomial still gives a polynomial. Moreover, after modifying the output by a linear function we can also make it so that the values at 0 and 1 are both 0, which is the third picture above. Again, such a linear change does not affect the ability to be uniformly approximated by polynomials.

Thus we reduce down to the case where $f : [0,1] \to \mathbb{R}$ is continuous and $f(0) = f(1) = 0$. We will need f to be defined elsewhere as well, so we extend the definition of f to be 0 outside $[0,1]$. Introduce the polymoials

$$Q_n(x) = c_n(1-x^2)^n \text{ on } [-1,1], \text{ where } c_n = \frac{1}{\int_{-1}^{1}(1-x^2)^n\,dx}.$$

(Note that c_n is just the constant we need in order to ensure that $\int_{-1}^{1} Q_n(x)\,dx = 1$ for all n.) Here are the graphs of these polynomials:



Each $(1-x^2)^n$ is nonegative and has values 0 at ± 1 and 1 at 0, so the scaling factor c_n increases (decreases in the Q_1 case) the height as needed to make the area underneath 1. As n increases, the graphs get more and more concentrated near $x = 0$, and the portions near ± 1 get flatter and flatter, so that Q_n becomes more and more "neglible" near these endpoints—this is perhaps the most crucial property we will need!

The polynomials we claim will converge uniformly to f are then defined by

$$P_n(x) := \int_{-1}^{1} f(x+t)\,Q_n(t)\,dt.$$

Note that $x \in [0,1]$ here is fixed one-at-a-time and t is the variable of integration. From this expression it is not at all clear that $P_n(x)$ is actually a polynomial, but this can be deduced by making the change of variables $u = x + t$:

$$P_n(x) = \int_{-1}^{1} f(x+t)\,Q_n(t)\,dt = \int_{x-1}^{x+1} f(u)\,Q_n(u-x)\,du = \int_{0}^{1} f(u)\,Q_n(u-x)\,dx,$$

where in the last step we use that f is 0 outside $[0,1]$ to reduce the integral over $[x-1, x+1] \supseteq [0,1]$ (recall $x \in [0,1]$) to one over $[0,1]$ alone. We see after expanding that

$$Q_n(u-x) = c_n(1-(u-x)^2)^n$$

is a polynomial in terms of x, and all the instances of u integrate away when computing

$$\int_{0}^{1} f(u)\,Q_n(u-x)\,dx,$$

so that this integral does in the end result in constants and powers of x alone, so that $P_n(x)$ thus defined is indeed a polynomial.

We will prove next time that these P_n do converge uniformly to f on $[0,1]$, which will make use of key properties of the Q_n's. This type of convergence strategy (phrasing convergence in terms of integrals) is a general one we will see show up again when discussing Fourier series, so we will also give a bit more context behind this idea next time.

Lecture 13: Stone-Weierstrass Theorem

Warm-Up. Suppose $f : [a, b] \to \mathbb{R}$ is continuous. We show that there exists a sequence of polynomials p_n such that $p_n \to f$ and $p_n' \to f'$ uniformly on $[a, b]$. The point here is that we get convergence not only of $p_n \to f$ but also of the derivatives simultaneously. Certainly, since f is continuous by Weierstrass approximation we get polynomials $p_n \to f$, but for such a sequence we cannot guarantee that $p_n' \to f'$ as well unless we knew that (p_n') also converged uniformly, which is not a given unless we construct the p_n more carefully.

Instead, we apply Weierstrass approximation not to f but to f', which is also continuous, to get a sequence of polynomials q_n converging to f' uniformly on $[a, b]$. Then set

$$p_n(x) = f(a) + \int_a^x q_n(t)\, dt.$$

Each of these is a polynomial in x since we can compute the integral on the right explicitly by using an antiderivative of q_n, and antiderivatives of polynomials are themselves polynomials. Since $q_n \to f'$ uniformly on $[a, b]$, $q_n \to f'$ uniformly on any $[a, x]$ as well, and hence

$$\int_a^x q_n(t)\, dt \to \int_a^x f'(t)\, dt$$

uniformly on $[a, b]$. (To be clear, we can use

$$\left| \int_a^x q_n(t)\, dt - \int_a^x f'(t)\, dt \right| \leq \int_a^x |q_n(t) - f'(t)|\, dt \leq \int_a^b |q_n(t) - f'(t)|\, dt$$

to get uniform convergence with respect to x of the $\int_a^x$ integrals.) Thus

$$p_n(x) = f(a) + \int_a^x q_n(t)\, dt \longrightarrow f(a) + \int_a^x f'(t)\, dt = f(a) + [f(x) - f(a)] = f(x)$$

uniformly, where we use the fundamental theorem of calculus at the end, and

$$p_n'(x) = q_n(x) \to f'(x)$$

uniformly by the choice of q_n. Hence these p_n satisfy our requirements.

Landau kernels. Recall that in order to prove the Weierstrass approximation theorem for a continuous f on $[0, 1]$, extended to be 0 outside $[0, 1]$, we constructed the polynomials

$$P_n(x) = \int_{-1}^1 f(x + t)\, Q_n(t)\, dt$$

where $Q_n(t) = c_n \int_{-1}^1 (1 - t^2)^n\, dt$ with $c_n = 1/(\int_{-1}^1 (1 - t^2)^n\, dt$. (Check the change of variables argument from last time to verify that these are indeed polynomials.) The claim is that these polynomials do converge uniformly to f, which proves Weierstrass approximation.

But before finishing this argument, let us give some context as this same type of idea occurs more broadly in analysis, and indeed is what we will use soon enough to prove convergence of Fourier series. The polynomial functions $Q_n(x) = c_n(1 - x^2)^n$ used here are known as the *Landau kernels*, and are an example of what are called *integral kernels*. An integral kernel is simply a

function we use to turn one function into another via an integration process: a "kernel" function $g(x)$ can be used to transform $f(x)$ into the function of x defined by

$$\int_a^b f(x+t)\,g(t)\,dt.$$

(The map that sends f to the function above is called an *integral transform*. Note that the use of the word "kernel" here is different from the typical use of this word in algebra, where kernel usually means things that are sent to zero. In general, some care needs to be taken to make sure the integral above makes sense by being clear about the domains of f and g. Such an integral is in fact an infinite-dimensional analog of matrix multiplication, but we will not take the space here to explain why. Ask in office hours if you want to know more!) The integral above is what's called the *convolution* of f and g and is typically denoted by $f \star g$. Integral kernels are thus used to prove convergence statements like

$$f \star g_n \to f,$$

where by picking appropriate kernel functions g_n we get a desired type of function $f \star g_n$ on the left, such as a polynomial in the case of the Landau kernels. (For the convergence of Fourier series we will soon discuss, the kernels we will use are what are known as the *Dirichlet kernels*.)

In this language, the claim is that the Landau kernels (where convolutions give polynomials) satisfy the requirement that

$$P_n := f \star Q_n \to f$$

uniformly on $[0,1]$. The key properties we will need in order to show this are:

$$\int_{-1}^{1} |Q_n(t)|\,dt = 1 \quad \text{and} \quad Q_n \to 0 \text{ uniformly on } [-1,-\delta] \cup [\delta,1] \text{ for any } 0 < \delta < 1.$$

The first property is just from the choice of the c_n in the definition of $|Q_n(x)| = Q_n(x) = c_n(1-x^2)^n$. The second property is the one we pointed out informally last time when mentioning that the graphs of the Q_n become "flatter and flatter" near the endpoints of $[-1,1]$ as n increases; the precise claim now is that, as long we remain bounded away from $x = 0$, no matter by how small an amount, the Q_n can be made uniformly small. (Integral kernels are typically assumed to have similar types of properties in general, which is what makes convergences like $f \star g_n \to f$ possible.)

To justify the second property, we note first that

$$1 = \int_{-1}^{1} c_n(1-x^2)^n\,dx \geq 2c_n \int_0^{1/\sqrt{n}} (1-x^2)^n\,dx \geq 2c_n \int_0^{1/\sqrt{n}} (1-nx^2)\,dx > \frac{c_n}{\sqrt{n}},$$

where the first inequality comes from noting that $(1-x^2)^n$ is even and nonnegative, so that $\int_{-1}^{1} = 2\int_0^1 \geq 2\int_0^{1/\sqrt{n}}$, the next inequality comes from showing that $(1-x^2)^n - (1-nx^2)$ is nonnegative by computing its derivative to show that it is increasing, and the final inequality comes from explicitly computing $\int_0^{1/\sqrt{n}}(1-nx^2)\,dx$. This gives $c_n < \sqrt{n}$ for all n. (Note that $c_n = Q_n(0)$ is the maximum value on the graph of Q_n, and the pictures of these graphs from last time do suggest that these values increase, which now we know happens at a rate we can control.) Thus for fixed $0 < \delta < 1$, we have

$$Q_n(x) = c_n(1-x^2)^n < \sqrt{n}(1-x^2)^n \leq \sqrt{n}(1-\delta^2)^n \text{ for } \delta \leq |x| \leq 1.$$

Since $|1-\delta^2| < 1$, $\sqrt{n}(1-\delta^2)^n \to 0$ since the left is bounded by $n(1-\delta^2)^n$, which converges to 0. The inequality above then implies that

$$Q_n(x) \to 0 \text{ uniformly for } \delta \leq |x| \leq 1.$$

Proof of uniform convergence. To now show that $P_n \to f$ uniformly on $[0,1]$, use the fact that $\int_{-1}^{1} Q_n(t)\,dt = 1$ to write

$$P_n(x) - f(x) = \int_{-1}^{1} f(x+t)Q_n(t)\,dt - \int_{-1}^{1} f(x)Q_n(t)\,dt = \int_{-1}^{1} (f(x+t) - f(x))Q_n(t)\,dt.$$

Our desired convergence claim then amounts to controlling the behavior of this integral. Since f is (uniformly) continuous, we can hope to control the

$$f(x+t) - f(x)$$

term when t is small so that $x+t$ is close to x. This will only work for $|t| \leq \delta$ with δ chosen by continuity, so we break up the domain of our integral as

$$\begin{aligned} |P_n(x) - f(x)| &= \left| \int_{-1}^{1} (f(x+t) - f(x))Q_n(t)\,dt \right| \\ &\leq \int_{-1}^{1} |f(x+t) - f(x)|Q_n(t)\,dt \\ &= \int_{|t|\leq\delta} |f(x+t) - f(x)|Q_n(t)\,dt + \int_{\delta\leq|t|\leq 1} |f(x+t) - f(x)|Q_n(t)\,dt. \end{aligned}$$

(Note that the final integral is actually a sum of two integrals, one over $[-1,-\delta]$ and the other over $[\delta,1]$. Also, we use the fact that $Q_n \geq 0$ to avoid absolute values on the Q_n terms.)

Now we are set: we first use continuity to make $|f(x+t) - f(x)|$ small and then the fact that

$$\int_{|t|\leq\delta} Q_n(t)\,dt \leq \int_{-1}^{1} Q_n(t)\,dt = 1$$

to make the first integral uniformly small, and then we bound $|f(x+t) - f(x)|$ by whatever and use uniform convergence $Q_n \to 0$ on $\delta \leq |t| \leq 1$ to make the second integral uniformly small as well. Et voilà! Magic. We leave writing out this argument more formally to you, and in fact you will show on the homework that the same type of reasoning works for more general appropriately "nice" integral kernel functions.

Function algebras. Polynomials are dense in $C([a,b])$, so one might wonder just what it is about the structure of the set of polynomials that allows for this to happen, and whether other such subsets of $C([a,b])$ work just as well. The *Stone-Weierstrass theorem* shows that, indeed, other subsets of $C([a,b])$ can be shown to be dense too as long as they satisfy some key algebraic properties. (This applies more generally to continuous real-valued functions on *any* compact space.)

We say that a set $\mathcal{A} \subseteq C([a,b])$ of continuous functions is an *algebra* if it is closed under addition, multiplication, and scalar multplication; i.e., if

$$\text{for all } f, g \in \mathcal{A} \text{ and } c \in \mathbb{R}, \text{ the functions } f+g, fg, \text{ and } cf \text{ are all in } \mathcal{A} \text{ as well.}$$

We say that $\mathcal{A}$ *vanishes nowhere* if for all $p \in [a,b]$ there exists $f \in \mathcal{A}$ such that $f(p) \neq 0$, so that there is no point at which all elements of $\mathcal{A}$ vanish. And we say that $\mathcal{A}$ *separates points* if for all distinct $p_1, p_2 \in [a,b]$ there exists $f \in \mathcal{A}$ such that $f(p_1) \neq f(p_2)$, so that distinct points can be "separated" by an element of $\mathcal{A}$. Essentially, these properties ensure that $\mathcal{A}$ has "enough" elements. The set of constant functions, for example, is an algebra but does not separate points, while the set of positive powers of $x - p$ (not an algebra) vanishes at p. In both of these cases,

there are not enough elements in that set to have any hope of being able to uniformly approximate all continuous functions.

The set of polynomials is an algebra (sums, products, and scalar multiples of polynomials are polynomials), vanishes nowhere (non-zero constant polynomials do not vanish), and separates points (x gives different values at different points), so Stone-Weierstrass applies to say that the set of polynomials is dense. One might ask, then, why we went to the trouble of establishing this fact first via Weierstrass approximation if it was just going to be a consequence of Stone-Weierstrass anyway? The answer is that, as we will see, the proof of Stone-Weierstrass relies on Weierstrass approximation applied to the absolute value function, so we had to prove the latter first. (Proving Weierstrass approximation independently of Stone-Weierstrass is also useful for the sake of illustrating the idea of an integral kernel!)

For a newer example, we claim that the set of *trigonometric polynomials*, which are functions built out of sines and cosines and expressible as

$$\sum_{n=0}^{k}(a_n \cos(nx) + b_n \sin(nx)),$$

on the unit circle is also dense in $C(\text{unit circle})$. (Here, we think of the unit circle as the interval $[-\pi, \pi]$ with the endpoints $-\pi$ and π glued to one another; in other words, we identify a point on the circle with the angle at which it occurs in the standard $(\cos t, \sin t)$ parametrization, where $\pm\pi$ give the same point. A continuous function on the circle is then a continuous function on $[-\pi, \pi]$ such that $f(-\pi) = f(\pi)$.) Sums and scalar multiples of such things are certainly of this same form. For products we can use various trigonometric identities, such as

$$\sin^2 x = \tfrac{1}{2}(1 - \cos 2x), \ \sin^3 x = \tfrac{1}{4}(3\sin x - \sin 3x), \text{ and } \sin x \cos x = \tfrac{1}{2}\sin 2x$$

for example, to show that the set of trigonometric polynomials is closed under multiplication as well, so it is an algebra. Constants are trigonometric polynomials (the $n = 0$ case in the sum above), so they show that this set vanishes nowhere. At points in $[-\pi, \pi]$ where $\sin x$ has the same value, $\cos x$ does not (look at the unit circle), so these two show that the set of trigonometric polynomials on $[-\pi, \pi]$ separates points. Hence this set is dense, so any continuous function on $[-\pi, \pi]$ with the same values at $\pm\pi$ can be uniformly approximated by sine and cosine expressions alone.

The key property that being a nowhere vanishing and point-separating function algebra guarantees is the following: for any $p_1, p_2 \in [a, b]$ and $c_1, c_2 \in \mathbb{R}$, there exists $f \in \mathcal{A}$ such that

$$f(p_1) = c_1 \quad \text{and} \quad f(p_2) = c_2.$$

This is the main property we need in order to prove Stone-Weierstrass, so why do we not take this as our assumption instead of nowhere vanishing and point separating? Because verifying nowhere vanishing and point-separating by hand is typically easier than verifying this consequence! To prove this consequence, pick $g \in \mathcal{A}$ such that $g(p_1) \neq g(p_2)$ (point separation), and $h_1, h_2 \in \mathcal{A}$ such that $h_1(p_1) \neq 0$, and $h_2(p_2) \neq 0$ (nowhere vanishing). Then

$$f(x) = c_1 \frac{[g(x) - g(p_2)]h_1(x)}{[g(p_1) - g(p_2)]h_1(p_1)} + c_2 \frac{[g(x) - g(p_1)]h_2(x)}{[g(p_2) - g(p_1)]h_2(p_2)}$$

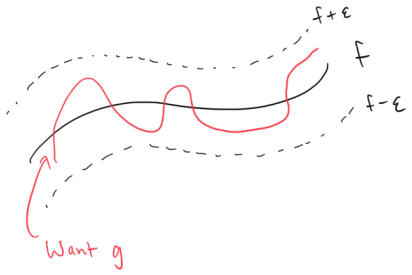
is the function we want. (This is essentially a "it works because it works" construction!) The denominators are nonzero and the entire thing is a sum of scalar multiples and products of things in $\mathcal{A}$, so $f \in \mathcal{A}$, and you can check that $f(p_i) = c_i$.

Stone-Weierstrass theorem. So, suppose $\mathcal{A} \subseteq C([a,b])$ is a nowhere vanishing point-separating function algebra. Given $f \in C([a,b])$ and $\epsilon > 0$, our goal is to find $g \in A$ such that

$$|g(x) - f(x)| < \epsilon \text{ for all } x \in [a,b],$$

so that for all $\epsilon > 0$ we have $B_\epsilon(f) \cap \mathcal{A} \neq \emptyset$, meaning that f is a limit point of $\mathcal{A}$ and hence that $\mathcal{A}$ is dense in $C([a,b])$. We will structure the argument by leaving the more tedious details to the end and focusing on the construction of a candidate g first, so we will point out along the way which properties we use still need to be verified. This is opposite to how Rudin and every other source approaches this, where the argument builds from the ground up, but I feel the "big picture" essence is simpler to grasp by postponing some details.

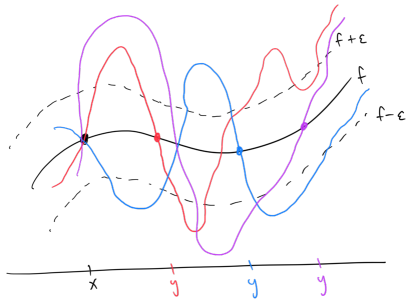
Here's the picture, where we take an ϵ-tube around the graph of f:



Our goal is to construct $g \in \mathcal{A}$ whose graph is fully within this tube. Fix $x \in [a,b]$. For each $y \in [a,b]$, we can find $g_{x,y} \in \mathcal{A}$ such that which agrees with f at x and y:

$$g_{x,y}(x) = f(x) \quad \text{and} \quad g_{x,y}(y) = f(y).$$

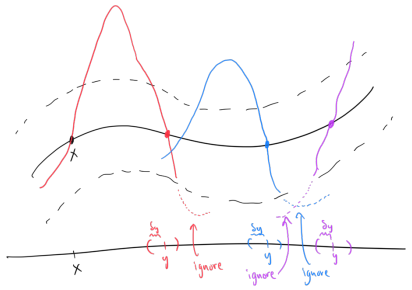
This comes from the key consequence we showed nowhere vanishing and point-separating gives us. As y varies, we get a whole collection of such functions:



The main point is that near x and y at least, $g_{x,y}$ is not too far off from f. Indeed, for each y by continuity (of f at y and of $g_{x,y}$ at y) we can find $\delta_y > 0$ such that

$$g_{x,y}(t) > f(t) - \epsilon \text{ for } t \in (y - \delta_y, y + \delta_y).$$

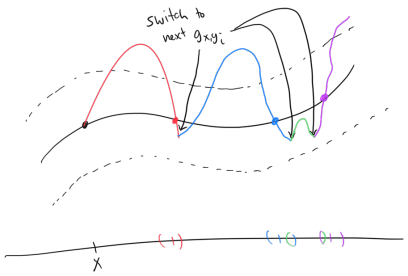
(Apply continuity to get $|g_{x,y}(t) - g_{x,y}(y)| < \frac{\epsilon}{2}$ and $|f(t) - f(y)| < \frac{\epsilon}{2}$ for $|t - y| < \delta_y$, then hit $|g_{x,y}(t) - f(t)|$ with a triangle inequality and use the fact that $g_{x,y}(y) = f(y)$.) The picture is



As long as we remain within these intervals, we have functions $g_{x,y}$ which do not fall below the ϵ-tube around f. But these intervals cover $[a, b]$ as y varies, so compactness gives a finite number of such intervals and corresponding functions

$$g_{x,y_1}, \ldots, g_{x,y_n}.$$

Taking the maximum $g_x := \max\{g_{x,y_1}, \ldots, g_{x,y_n}\}$ of these functions thus gives a single function whose graph never dips below this tube:

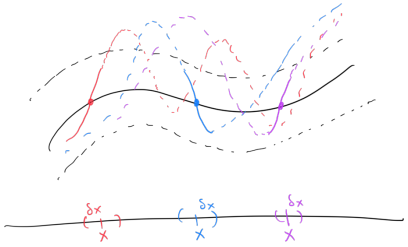


On some $(y_i - \delta_{y_i}, y_i + \delta y_i)$, g_{x,y_i} stays above the bottom of the tube, and as it begins to dip below once we leave this interval, we switch to a different g_{x,y_j} on a different interval which is the new maximum on this new interval. Now, here is one assumption we will have to come back to later: the maximum g_x constructed in this way is in fact in the *closure* (in the metric space sense) of $\mathcal{A}$. We cannot guarantee that $g_x \in \mathcal{A}$, but being in the closure will be good enough, as we'll see.

Now vary x. The functions g_x constructed above all lie above $f - \epsilon$, but we have no control so far over how large they can be. To achieve this control, we do the same thing we did before, only now restricting how large our functions are instead of how small in order to get a function that does not rise above the tube. For each x use continuity to get $\delta_x > 0$ such that

$$g_x(t) < f(t) + \epsilon \text{ for } t \in (x - \delta_x, x + \delta_x),$$

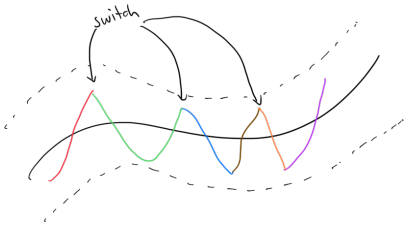
which uses the fact that $g_x(x) = f(x)$:



Each such g_x now lies within the ϵ-tube around f near x. The intervals $(x - \delta_x, x + \delta_x)$ cover $[a, b]$, so we get finitely many such intervals and functions

$$g_{x_1}, \ldots, g_{x_m}.$$

The minimum $g := \min\{g_{x_1}, \ldots, g_{x_m}\}$ of these is then continuous and has graph that lies fully within the tube $f(x) - \epsilon < g(x) < f(x) + \epsilon$ uniformly, since as we move from one interval $(x_i - \delta_{x_i}, x_i + \delta x_i)$ to the next we switch from one g_{x_i} which begins to head above the tube to the next where we move back in:



And there you have it! Almost. Just as with the construction of g_x, we can only guarantee that the minimum g above lies in the closure of $\mathcal{A}$, so this is not quite the final $g \in \mathcal{A}$ we want. But what this argument does show is that our original f is actually a *limit point* of $\overline{\mathcal{A}}$ since any open ball around it contains an element of $\mathcal{A}$, so f is in the closure of $\overline{\mathcal{A}}$, which is the same thing as $\overline{\mathcal{A}}$ since closures are closed. Thus $f \in \overline{\mathcal{A}}$, which is what it means for $\mathcal{A}$ to be dense in $C([a, b])$. If you want to get an actual $g \in \mathcal{A}$ (instead of in the closure of $\mathcal{A}$) which is uniformly close to f, apply the argument above to get $h \in \overline{\mathcal{A}}$ which is within $\frac{\epsilon}{2}$ from f, and then pick $g \in \mathcal{A}$ which is within $\frac{\epsilon}{2}$ from h (h is a limit point of $\mathcal{A}$), so that $g \in \mathcal{A}$ is then within ϵ of f as desired. (Phew!)

Closures of function algebras. We now justify the remaining technical details in the proof of the Stone-Weierstrass theorem. The key functions in the proof are constructed as maxima or minima of functions first in $\mathcal{A}$ (in the maximum case) and then in the closure $\overline{\mathcal{A}}$ (in the minimum) case, so we have to know that such maxima/maxima remain at least in $\overline{\mathcal{A}}$. This will come from the fact that we can express the maximum and minimum of two functions f, g as

$$\max\{f, g\} = \tfrac{1}{2}(f + g + |f - g|) \quad \text{and} \quad \min\{f, g\} = \tfrac{1}{2}(f + g - |f - g|).$$

(This we can simply check point-by-point using the fact that $|f - g| = f - g$ at points where $f > g$ and the opposite at points where $f < g$.) If we start with f and g in $\overline{\mathcal{A}}$, *and* we know that $\overline{\mathcal{A}}$ is itself closed under addition and scalar multiplication *and* closed under taking absolute values, then all terms on the right sides above are in $\overline{\mathcal{A}}$, and so the max and min will be as well. From two functions we can extend to a finite number inductively, so the max's and min's used in the proof of Stone-Weierstrass will indeed be in $\overline{\mathcal{A}}$.

So, we are left showing that $\overline{\mathcal{A}}$ is closed under addition, scalar multiplication, and taking absolute values. (The closure of $\mathcal{A}$ is also closed under multiplication, so that it is a function algebra itself, but we do not need this final property here.) If $f, g \in \overline{\mathcal{A}}$ and $c \in \mathbb{R}$, we can pick sequences $f_n \to f$ and $g_n \to g$ with $f_n, g_n \in \mathcal{A}$, so that we get

$$f_n + g_n \to f + g \quad \text{and} \quad cf_n \to cf,$$

which means that $f + g, cf \in \overline{\mathcal{A}}$.

The statement about being closed under absolute values is the claim that if $f \in \overline{\mathcal{A}}$, then $|f| \in \overline{\mathcal{A}}$. (This is the point where working in the closure of $\mathcal{A}$ instead of just $\mathcal{A}$ is required, and why we can only guarantee that max's and min's are in the closure: if $f \in \mathcal{A}$, it is not true that $|f|$ must be in $\mathcal{A}$ well—we can only guarantee that $|f|$ is in the closure of $\mathcal{A}$. For example, the absolute value of the polynomial x is not a polynomial.) To prove this, we must make use of Weierstrass approximation by polynomials applied to $|x|$, which is why we needed that result first. Given $\epsilon > 0$, pick a polynomial such that

$$|(a_0 + a_1 x + \cdots + a_n x^n) - |x|| < \epsilon \text{ for } x \in [-M, M].$$

where $M > 0$ is a bound on f. Note that taking $x = 0$ in particular gives $|a_0| < \epsilon$. In fact, we can assume that $a_0 = 0$ since otherwise we can replace the polynomial above with it minus the constant

polynomial a_0 and still maintain the uniform approximation: i.e.

$$|(a_1x + \cdots + a_nx_n) - |x|| \leq |(a_0 + a_1x + \cdots + a_nx^n) - |x|| + |a_0|$$

can be made uniformly smaller than ϵ if the original $|(a_0 + a_1x + \cdots + a_nx^n) - |x||$ can be made uniformly smaller than $\frac{\epsilon}{2}$. So, our uniform approximation looks like

$$|(a_1x + \cdots + a_nx^n) - |x|| < \epsilon \text{ for } x \in [-M, M].$$

Hence for $f(x) \in [-M, M]$ with $x \in [a, b]$ (now we are changing what we denote by x), we get

$$|(a_1f(x) + \cdots + a_nf(x)^n) - |f(x)|| < \epsilon$$

uniformly. This says that the function $a_1f + a_2f^2 + \cdots + a_nf^n$ is ϵ-uniformly close to f, so since $a_1f + a_2f^2 + \cdots + a_nf^n \in \overline{\mathcal{A}}$ (closed under addition and scalar multiplication), this says that f is a limit point of $\overline{\mathcal{A}}$. But $\overline{\mathcal{A}}$ contains all of its limit points, so $f \in \overline{\mathcal{A}}$ as desired. (Phew again! Note that we needed to have $a_0 = 0$ since otherwise we would get the function $a_0 + a_1f + \cdots + a_nf^n$, which is not guaranteed to be in $\overline{\mathcal{A}}$ unless $\overline{\mathcal{A}}$ contains all constant functions, which is not necessarily true.)

Big time analysis. Let us take a pause here and reflect on what we have done the past five days. Within these five days, we have developed some absolutely monster results of analysis: Picard-Lindelöf (contraction application to ODEs), explicit example of a continuous nowhere differentiable function, Arzela-Ascoli, Weierstrass approximation (including a brief introduction to integral kernels), and Stone-Weierstrass. Each of these took some serious effort to make sense of, as exemplified by all the times the word "phew" appeared at the end of a proof!

The point is that you should not expect that all of this will come easily on a first, or even second or third or INSERT LARGER NUMBER HERE read through as many huge ideas are involved. (Indeed, in class we only outlined the main ideas behind the proof of Stone-Weierstrass, as going through the details in full would take a good amount of time that, in my opinion, is not obviously worth it.) We have covered 100 years worth of analysis (from Weierstrass around 1830 to Stone around 1930) in the span of five days, so it is good to keep this perspective in mind as you work to make sense of it all.

Lecture 14: Power Series

Warm-Up. Suppose f is a continuous real-valued function on the unit circle, which as we explained last time can be taken to be a continuous function on $[-\pi, \pi]$ such that $f(-\pi) = f(\pi)$. We show that if

$$\int_{-\pi}^{\pi} f(x)\cos(nx)\,dx = 0 = \int_{-\pi}^{\pi} f(x)\sin(nx)\,dx$$

for all $n = 0, 1, 2, 3, \ldots$, then $f = 0$ everywhere. We saw a simpler type of result previously where a continuous $f : [a, b] \to \mathbb{R}$ satisfying

$$\int_a^b f(x)\,x^n\,dx = 0$$

for all $n \geq 0$ is necessarily zero, and in fact the argument here is exactly the same: the given assumptions imply that

$$\int_{-\pi}^{\pi} f(x)(\text{trigonometric polynomial})\,dx = 0$$

for any trigonometric polynomial, so take a sequence of such things converging uniformly to f (using denseness of trig polynomials in $C(\text{circle})$ by Stone-Weierstrass) shows that $\int_a^b f(x)^2\,dx = 0$, from which $f = 0$ follows.

But rather than repeat the details of this same argument again, let us now recast both of these examples as arising from how continuous functions behave on dense sets in general. The point is that for fixed f, the mapping

$$T : C([a,b]) \to \mathbb{R} \text{ defined by } T(g) := \int_a^b f(x)g(x)\,dx$$

is continuous when $C([a,b])$ is equipped with the sup metric. (Same thing for $C(\text{circle})$ where "circle" is $[-\pi,\pi]$ as above.) Indeed, we have

$$|T(g_1) - T(g_2)| \leq \int_a^b |f(x)||g_1(x) - g_2(x)|\,dx \leq \int_a^b |f(x)|d(g_1,g_2)\,dx$$

where $d(g_1,g_2)$ is the sup distance between g_1 and g_2. If $M > 0$ is a bound on f over $[a,b]$, this gives

$$|T(g_1) - T(g_2)| \leq \int_a^b M\,d(g_1,g_2)\,dx = M(b-a)\,d(g_1,g_2),$$

which implies continuity of T by taking $\delta = \frac{\epsilon}{M(b-a)}$ for a given $\epsilon > 0$.

Since $T : C([a,b] \to \mathbb{R}$ is continuous, its behavior is completely characterized by its values on a dense subset of the domain. In the previous example, this dense subset is the set of polynomials on $[a,b]$, and we get that T sends any such polynomial to 0, so it must send everything to 0, in particular f itself so that $T(f) = \int_a^b f(x)^2\,dx = 0$. In the current case, the dense subset is the set of trigonometric polynomials in $C(\text{circle})$, so T being zero on this dense subset implies it is zero everywhere, so in particular again $T(f) = \int_{-\pi}^{\pi} f(x)^2\,dx = 0$, which implies $f = 0$. Taking more general dense subsets of $C([a,b])$ as the ones against which integrating f gives 0 would give precisely the same outcome.

A nicer form of denseness. We know that the set of ordinary polynomials is dense in $C([a,b])$, and we know that the set of trigonometric polynomials is dense in $C(\text{circle})$. But both of these results are in a sense non-explicit, in that we prove there exists ordinary/trig polynomials converging uniformly to a given f without being able to say much about what these ordinary/trig polynomials look like. Of course, we do have concrete expressions for polynomials that work in the Weierstrass approximation case via

$$P_n(x) := \int_{-1}^{1} f(x+t)Q_n(t)\,dt$$

where Q_n are the Landau kernels, but we have no idea what these polynomial actually looks like in general without doing a big computation. The situation is even worse in the trig polynomial case, where the existence comes from the max/min construction in the proof of Stone-Weierstras, and who knows what type of thing this gives explicitly.

The question now is whether there are "nicer" versions of these results, where we can say something more explicit about the ordinary/trig polynomials we need. For example, we previously claimed that

$$1,\ 1+x,\ 1+x+\tfrac{1}{2}x^2,\ 1+x+\tfrac{1}{2}x^2+\tfrac{1}{3!}x^3,\ \ldots$$

converges uniformly to e^x on any $[a,b]$, where here we do have incredibly nice descriptions of the polynomials that work. Are there analogs of this in general? What about the trig polynomial case?

The answer is that there *are* other such "nice" analogs, although we will have restrict the types of functions we care about, and they are provided by the notions of *power series* in the ordinary polynomial case and *Fourier series* in the trig polynomial case. We will study power series first, and Fourier series later.

Power series convergence. A power series is a series of the form

$$\sum_{n=0}^{\infty} c_n(x-a)^n$$

with variable x. (We say in this case that the power series is *centered at* a.) Recall (earlier in Rudin) that the issue of *pointwise* convergence of such a series is settled by the root test: the series converges—in fact absolutely—if

$$\limsup \sqrt[n]{|c_n(x-a)|^n} = |x-a| \limsup \sqrt[n]{|c_n|} < 1$$

and diverges if this lim sup is larger than 1. The root test is inconclusive when the lim sup equals 1, but we actually will not about these edge cases here. We thus have convergence when

$$|x-a| < \frac{1}{\limsup \sqrt[n]{|c_n|}},$$

where we interpret the right side as infinite (so no restriction on x in that case) when $\limsup \sqrt[n]{|c_n|}$ is 0. We call

$$R := \frac{1}{\limsup \sqrt[n]{|c_n|}}$$

the *radius of convergence* of the power series, and the corresponding interval $(a-R, a+R)$ centered at a the *interval of convergence*, or simply the "domain" of the power series. (Again, we might or might not have convergence at the endpoints $a \pm R$, but we will ignore such behavior here. By "interval of convergence" or "domain" we will always mean *open* interval or domain.)

So, power series converge pointwise on their intervals of convergence. However, we cannot guarantee that this convergence is in fact *uniform* on the entire open domain, as you essentially showed in a previous homework problem: the partial sums of $\sum_n x^n$ are all bounded on $(-1,1)$, but the pointwise sum $\frac{1}{1-x}$ of this series is unbounded on $(-1,1)$, so the convergence cannot be uniform on all of $(-1,1)$. But, the observation now is that, as long as we do not allow ourselves to come arbitrarily close to the endpoints, we *can* guarantee uniform convergence, and in fact this is a property of power series in general.

The claim is that if $R > 0$ is the radius of convergence of $\sum_n c_n(x-a)^n$, then the convergence is uniform on any closed interval of the domain $(a-R, a+R)$. (The modern lingo is that power series "converge uniformly on compact subsets" of their domain. When $R = 0$, then the series converges pointwise only at a, so asking for uniform convergence in this case is moot. Also, if we do have convergence at an endpoint, it turns out that the uniform convergence does in fact extend to that endpoint as well; this is what's known as *Abel's theorem*, which we will not prove as we will not need it.) It is enough to show this for closed intervals of the form $[a-r, a+r]$ with $0 < r < R$ since all closed intervals are subsets of these types alone. On $[a-r, a+r]$, so that $|x-a| < r$, we have

$$|c_n(x-a)^n| \leq |c_n| r^n = |c_n|((a+r)-a)^n,$$

where the point is that we have removed the dependence on The number $x = a + r$ lies in the interval of convergence $(a - R, a + R)$ of the power series, so $\sum_n |c_n|((a + r) - a)^n$ converges (this uses the fact that convergence for power series is always absolute), so

$$\sum_{n=0}^{\infty} c_n(x - a)^n$$

converges uniformly on $[a - r, a + r] \subseteq (a - R, a + R)$ by the Weierstrass M-test.

Since each $c_n(x - a)^n$ is continuous, we thus get that

$$f(x) = \sum_{n=0}^{\infty} c_n(x - a)^n$$

is continuous on any $[a-r, a+r] \subseteq (a-R, a+R)$, and since any point in $(a-R, a+R)$ lies in such a closed interval, we get continuity of f on its entire domain. (Again, if we also have convergence at and endpoint, continuity in fact extends to that endpoint as well as a consequence of Abel's theorem as alluded to earlier.) We also get integrability on any $[c, d] \subseteq (a - R, a + R)$ for free, and the fact that integration can be carried out term by term.

Smoothness. For differentiability, we have to also know that the series of term-by-term derivatives

$$\sum_{n=0}^{\infty} nc_n(x - a)^{n-1} = \sum_{n=1}^{\infty}(n + 1)c_{n+1}(x - a)^n$$

also converges uniformly. (Note that this term-by-term derivative is yet another power series!) But since $n^{1/n} \to 1$ as $n \to \infty$, we have that

$$\limsup \sqrt[n]{n|c_n|} = \limsup \sqrt[n]{|c_n|},$$

so a power series and its term-by-term derivative always have the same radius of convergence. Thus by the general machinery of power series derived above, we do have that

$$\sum_{n=0}^{\infty} nc_n(x - a)^{n-1} = \sum_{n=1}^{\infty}(n + 1)c_{n+1}(x - a)^n$$

converges uniformly on any $[a - r, a + r] \subseteq (a - R, a + R)$, and hence that

$$f(x) = \sum_{n=0}^{\infty} c_n(x - a)^n$$

is differentiable on any $[a - r, a + r]$ with derivative equal to the term-by-term derivative; taking r to approach R then gives differentiability on all of $(a - R, a + R)$.

Applying the same general power series machinery to

$$f'(x) = \sum_{n=1}^{\infty}(n + 1)c_{n+1}(x - a)^n$$

shows that f' is differentiable on its domain, so f is twice-differentiable, and then applying the machinery to f'', then f''', etc shows that f is in fact infinitely-differentiable on $(a-R, a+R)$. We will use the term *smooth* to refer to the property of being infinitely-differentiable, and thus smoothness

is a first restriction we need to impose if we are wanting to express functions as convergent power series: if it were possible to express f in this way, f would have to be smooth since power series are always smooth. It is not true however that smoothness is enough as there are smooth functions which are not expressible as power series; we will clarify the distinction between "smooth" and "expressible as a power series" next time.

Example. We have

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n \text{ on } (-1,1).$$

Differentiating gives

$$\frac{1}{(1-x)^2} = \sum_{n=1}^{\infty} (n+1)x^n \text{ on } (-1,1),$$

and differentiating again gives

$$\frac{2}{(1-x)^3} = \sum_{n=2}^{\infty} (n+2)(n+1)x^n \text{ on } (-1,1).$$

And so on we can keep going, all of which we obtain for free from the general machinery of power series, no longer having to do any hard work ourselves. The same works with integration/anti-differentiation, so that

$$-\ln(1-x) = \sum_{n=0}^{\infty} \frac{x^{n+1}}{n+1} \text{ on } (-1,1),$$

and then

$$-\int_0^x \ln(1-t)\,dt = \sum_{n=0}^{\infty} \frac{x^{n+2}}{(n+2)(n+1)} \text{ on } (-1,1),$$

and so on. (As a side remark, the series above expressing $-\ln(1-x)$ as a power series does also make sense at the endpoint -1, so this gives the identity

$$-\ln 2 = \sum_{n=0}^{\infty} \frac{(-1)^{n+1}}{n+1}$$

for example. The original series $\frac{1}{1-x} = \sum_n x^n$ from which this was derived, however, does not converge at $x = -1$, so this illustrates that in general convergence at endpoints for a power series has nothing to do with convergence of endpoints of its derivative, which is one reason why we are excluding the behavior at endpoints from this entire discussion.)

Sums and products. Our goal now is to better thus understand the types of functions which convergent power series define. If

$$f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n \quad \text{and} \quad g(x) = \sum_{n=0}^{\infty} d_n(x-a)^n$$

with radii of convergence R_1 and R_2 respectively, then it is true that $f+g$ and fg are also expressible as convergent power series, at least within whichever of the radii R_i is smaller. (We will see later that composing power series also works, as does dividing as long as we assume nonzero

denominator.) Indeed, these come from general properties of absolutely convergent series covered earlier in Rudin: for $|x-a| < \min\{R_1, R_2\}$, we have that

$$\sum_{n=0}^{\infty}(c_n + d_n)(x-a)^n \text{ converges to } f(x) + g(x)$$

and that

$$\sum_{n=0}^{\infty}\left(\sum_{k=0}^{n}\binom{n}{k}c_k d_{n-k}\right)(x-a)^n \text{ converges to } f(x)g(x).$$

(The first expression comes from combining like—meaning same degree—terms in

$$(c_0 + c_1(x-a) + c_2(x-a)^2 + \cdots) + (d_0 + d_1(x-a) + d_2(x-a)^2 + \cdots)$$

and the second from doing the same in

$$(c_0 + c_1(x-a) + c_2(x-a)^2 + \cdots)(d_0 + d_1(x-a) + d_2(x-a)^2 + \cdots) = c_0 d_0 + (c_0 d_1 + c_1 d_0)(x-a) + \cdots$$

Note that being able to rearrange the terms—which is what absolute convergence gives us—is crucial to making this work.)

Lecture 15: Analytic Functions

Warm-Up. Let $a \neq 1$. We show that $f(x) = \frac{1}{1-x}$ can be expressed as a convergent power series centered at a on some interval around a. Rewrite $\frac{1}{1-x}$ as follows:

$$\frac{1}{1-x} = \frac{1}{(1-a)-(x-a)} = \frac{1}{1-a}\left(\frac{1}{1-\frac{x-a}{1-a}}\right).$$

Using $\frac{1}{1-y} = \sum_{n=0}^{\infty} y^n$ for $|y| < 1$ with $y = \frac{x-a}{1-a}$, we have:

$$\frac{1}{1-x} = \frac{1}{1-a}\sum_{n=0}^{\infty}\left(\frac{x-a}{1-a}\right)^n = \sum_{n=0}^{\infty}\frac{1}{(1-a)^{n+1}}(x-a)^n \text{ when } \left|\frac{x-a}{1-a}\right| < 1.$$

This gives the desired expression with radius of convergence $|1-a| > 0$. Specifically, for $a > 1$ this expresses $\frac{1}{1-x}$ as a convergent power series on $(1, 2a-1)$, while for $a < 1$ this expresses $\frac{1}{1-x}$ as a convergent power series on $(2a-1, 1)$.

Analytic functions. The types of functions which can be expressed as power series are special enough that we give them their own name. We say that f is *analytic* on an open set $U \subseteq \mathbb{R}$ if U can be covered by open intervals on each of which f can be expressed as a convergent power series centered at some point in that interval; i.e., each element of U belongs to an open interval I such that for some $a \in I$ there exist c_n such that

$$f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n \text{ on } I.$$

The power series we need in order to express f might change as we move from one interval to another, but such a series expansion always exists *locally*.

For example, $f(x) = \frac{1}{1-x}$ is analytic on $\mathbb{R}\backslash\{1\}$, the set of real numbers excluding 1, as the Warm-Up shows: the intervals $(1, 2a-1)$ for $a > 1$ and $(2a-1, 1)$ for $a < 1$ cover all of $\mathbb{R}\backslash\{1\}$, and on each f is expressible as a convergent power series. We cannot find one power series that works on all of $\mathbb{R}\backslash\{1\}$ at once, but that's OK since we only require that f be locally expressible as a power series throughout our domain.

Note that analytic functions will always be smooth since power series are always infinitely differentiable within their domains of convergence. It is not true, however, that all smooth functions are analytic, as we'll see.

Shifting the center. The definition of analytic only requires that express our function as a power series centered at *some* point in an interval, whereas in the example of $\frac{1}{1-x}$ we are able to find such an expression centered at *any* $a \neq 1$ we want. So we are led to wonder what the role the center of the power series plays in general. The answer is that the center is insignificant, as once we have a power series expression with some center, we can always come up with one with a different candidate center. For simplicity of notation, let us consider only the case of a power series

$$\sum_{n=0}^{\infty} c_n x^n$$

centered at 0, since an arbitrary $\sum_n c_n(x-a)^n$ can always be though of as $\sum_n c_n y^n$ with $y = x - a$.

The claim is that if the given series converges on $(-R, R)$, then for any $a \in (-R, R)$ we can rewrite this series as one centered at a instead on some $(a-r, a+r) \subseteq (-R, R)$. Indeed, write x as $x = a + (x-a)$ and use the binomial theorem

$$(c+d)^n = \sum_{k=0}^{n} \binom{n}{k} c^{n-k} d^k$$

to get

$$\sum_{n=0}^{\infty} c_n x^n = \sum_{n=0}^{\infty} c_n (a + (x-a))^n = \sum_{n=0}^{\infty} c_n \left(\sum_{k=0}^{n} \binom{n}{k} a^{n-k} (x-a)^k \right).$$

The specific form of the coefficients used in the binomial expansion do not really matter—what matters is that we get an expression in terms of power of $x - a$. In order to turn the resulting expression into what we want, namely a power series centered at a, we now need to swap the two summations; indeed, we are left with something involving $(x-a)^k$, so for this to be an appropriate power series we need to sum over k's in the outer sum:

$$\sum_{n=0}^{\infty} c_n \left(\sum_{k=0}^{n} \binom{n}{k} a^{n-k} (x-a)^k \right) = \sum_{k=0}^{\infty} \left(\sum_{n=k}^{\infty} c_n \binom{n}{k} a^{n-k} \right) (x-a)^k.$$

A first clarification needed here is in having the inner sum run from $n = k$ to ∞, which comes from the fact that a given exponent k in the sum on the left only occurs once the outer index n is large enough, so that for example we do not get a $(x-a)^5$ term in the sum on the left until n is at least 5 since we get no such term when $k = 0, 1, 2, 3, 4$.

The bigger issue is in knowing that such a sum swap makes sense and does not affect the convergence. This is a general problem about double summations

$$\sum_{n=0}^{\infty} \sum_{k=0}^{\infty} a_{nk} \stackrel{?}{=} \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} a_{nk},$$

with the key fact being that as long as the left side converges absolutely, then such a swap is indeed valid. (To apply this in our case, where in

$$\sum_{n=0}^{\infty} c_n \left(\sum_{k=0}^{n} \binom{n}{k} a^{n-k} (x-a)^k \right) = \sum_{k=0}^{\infty} \left(\sum_{n=k}^{\infty} c_n \binom{n}{k} a^{n-k} \right) (x-a)^k$$

we have only finitely many terms in the inner sum on the left and $n = k$ to ∞ instead of $n = 0$ to ∞ on the right, we simply take the general version with a bunch of the a_{nk}'s—namely those where $k > n$—equal to 0.) Taking this general fact for granted for now, we must thus know that

$$\sum_{n=0}^{\infty} \sum_{k=0}^{n} |c_n| \binom{n}{k} |a|^{n-k} |x-a|^k$$

converges. But after unwinding the binomial expansion, this sum is the same as

$$\sum_{n=0}^{\infty} \sum_{k=0}^{n} |c_n| \binom{n}{k} |a|^{n-k} |x-a|^k = \sum_{n=0}^{\infty} |c_n| (|a| + |x-a|)^n,$$

which converges as long as $|a| + |x-a| < R$ is within the radius of convergence of the original series, which thus requires that $|x-a| < R - |a|$. Hence,

$$\sum_{n=0}^{\infty} c_n x^n = \sum_{k=0}^{\infty} \left(\sum_{n=k}^{\infty} c_n \binom{n}{k} a^{n-k} \right) (x-a)^k$$

is valid on $(a-r, a+r) \subseteq (-R, R)$ for $r := R - |a| > 0$, so we have expressed our original power series as one centered at a instead as desired.

Double summations. The fact that

$$\sum_{n=0}^{\infty} \sum_{k=0}^{\infty} a_{nk} = \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} a_{nk}$$

as long as the left side converges absolutely is proved in Rudin using uniform convergence of functions, but we do not need all of that machinery to derive this. Instead, the same proof as the one used in showing that rearrangements of the terms of an absolutely convergent series do not affect the convergence works here as well, as long as we define what it means for a *double series* $\sum_{n,k} a_{nk}$ to converge.

First, we say that a doubly-indexed sequence s_{ij} converges to S if for all $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$|s_{ij} - S| < \epsilon \text{ for } i, j \geq N.$$

(So, same definition as for singly-indexed sequences, only where we take both indices to be large enough in a common way.) If so, one can show that the limit S can be computed "index-by-index" in the sense that

$$\text{if } \lim_{j \to \infty} a_{ij} \text{ exists for each } i, \text{ then } S = \lim_{i \to \infty} \left(\lim_{j \to \infty} a_{ij} \right).$$

We then say that a double series $\sum_{n,k} a_{nk}$ converges if the doubly-indexed sequence of partial sums

$$s_{ij} := \sum_{n=0}^{i} \sum_{k=0}^{j} a_{nk}$$

converges in the sense above. If so, then we get that the sum can be computed index-by-index in the sense that

$$\text{if } \sum_{k=0}^{\infty} a_{nk} \text{ exists for each } n, \text{ then } \sum_{n,k} a_{nk} = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} a_{nk}.$$

We then get a Cauchy criterion for convergence of double sums, a notion of absolute convergence, and so on. In particular, absolute convergence guarantees that rearrangements are possible (with essentially the same proof as in the usual series case as mentioned before), which gives our desired sum swap. We will not give further details of double sums here as the proofs really are analogous to things you saw last quarter.

Taylor series. If f is analytic and hence expressible (locally) as a power series, then there is in fact only one series that can do the job. Suppose that

$$f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n = c_0 + c_1(x-a) + c_2(x-a)^2 + \cdots .$$

Plugging in a on the right side gives only c_0 since all other terms will contain a power of $a - a = 0$. Thus $f(a) = c_0$ is the constant term. Next, taking derivatives gives

$$f'(x) = c_1 + 2c_2(x-a) + \cdots ,$$

so $f'(a) = c_1$. In general, taking k-th derivatives gives

$$f^{(k)}(x) = k!c_k + (\text{terms with } (x-a)), \text{ so } f^{(k)}(a) = k!c_k.$$

Thus the coefficients of the given power series *must* be given by $c_n = f^{(n)}(a)/n!$, meaning that the power series in question must be the *Taylor series* of f centered at a:

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n.$$

Hence, in the definition of analytic we can omit any reference to the existence of *some* convergent power series and replace this specifically by the Taylor series centered at a point. Analyticity of a smooth function then comes down to arguing that its Taylor series have positive radii of convergence *and* the thing to which they converge on their domains of convergence is the given function itself.

Examples. We claim that $f(x) = e^x$ is analytic. In fact, in this case we claim that we can express f *globally* as a power series, specifically its Taylor series centered at 0:

$$e^x = \sum_{n=0}^{\infty} \frac{1}{n!}x^n \text{ for all } x \in \mathbb{R}.$$

(Note that $f^{(n)}(0) = 1$ for all n in this case.) We can compute the radius of convergence of this series to see that it is infinite using the $\limsup \sqrt[n]{|c_n|}$ formula, but instead we will just verify convergence directly at any $x \in \mathbb{R}$. (Even with the $\limsup$ computation of the radius, we would still need to argue that the thing to which this series converges is indeed e^x.)

The key is Taylor's theorem, which expresses the difference between $f(x)$ and one of its Taylor polynomials. For fixed $x \in \mathbb{R}$, there thus exists c between 0 and x such that

$$\left| e^x - \sum_{n=0}^{N} \frac{1}{n!}x^n \right| = \left| \frac{f^{(N+1)}(c)}{(N+1)!}x^{N+1} \right| = \frac{e^c}{(N+1)!}|x|^{N+1}.$$

Since c is between 0 and x, we have $e^c \leq \max\{1, e^x\}$ (the 1 takes care of the case where $x < 0$), so

$$\left| e^x - \sum_{n=0}^{N} \frac{1}{n!} x^n \right| \leq \frac{\max\{1, e^x\}}{(N+1)!} x^{N+1}.$$

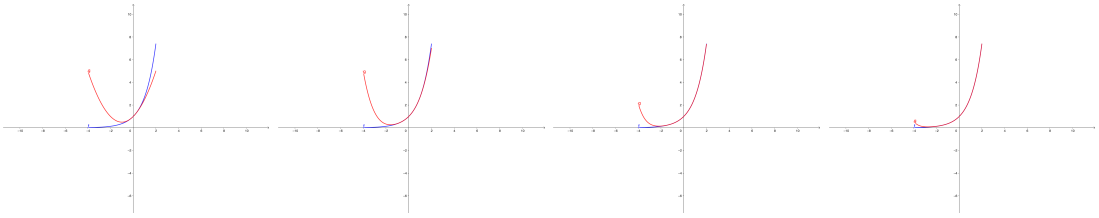
(The point here is that we have found a uniform bound on the $f^{(N+1)}$ term, independent of N.) As $N \to \infty$, we have $x^{N+1}/(N+1)! \to 0$, which shows that

$$e^x = \lim_{N \to \infty} \sum_{n=0}^{N} \frac{x^n}{n!}$$

as desired. (The same type of argument shows that $\sin x$ and $\cos x$ ae analytic on $\mathbb{R}$ as well, where we are able to find uniform bounds on the derivatives. In general, uniform bounds are not strictly required, but what we do need is some type of control over how quickly the derivatives can grow so that expressions like $\frac{f^{(N+1)}(c)}{(N+1)!} x^{N+1}$ can be forced to go to 0 as N increases. A problem in discussion section will explore this idea further.) As a consequence (general power series stuff), we then get that

$$1 + x + \tfrac{x^2}{2} + \tfrac{x^3}{3!} + \cdots + \tfrac{x^n}{n!} \to e^x$$

uniformly on any closed interval. Here's what this looks like on, say, $[-4, 2]$ with the second, fourth, sixth, and eighth order Taylor polynomials:



(The convergence is slower for negative values, due to the fact that the odd-order polynomials—not drawn—have negative leading term x^n for x negative.)

Smooth but not analytic. The standard example of a non-analytic smooth function is the following. Define f on $\mathbb{R}$ by

$$f(x) = \begin{cases} e^{-1/x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0. \end{cases}$$

This function is indeed smooth, in particular it is infinitely-differentiable at 0, and in fact $f^{(n)}(0) = 0$ for all n. (This was an example done last quarter, but if it's unfamiliar try to prove it yourself!) If this function were analytic on $\mathbb{R}$, it would be expressible as a power series on some interval containing 0, and hence would have to equal its own Taylor series centered at 0 on some $(-R, R)$. (This is why being able to shift the center is important, since the failure to be analytic will come from the behavior when 0 is the center.) But the Taylor series centered at 0 is explicitly

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n = \sum_{n=0}^{\infty} 0 = 0,$$

and f does not equal 0 on any $(0, R)$, so we conclude that f is not analytic on $\mathbb{R}$.

The function f is certainly analytic on $(0, \infty)$ since it is constant there, and it is also analytic on $(0, \infty)$ since, as you will show on the homework, compositions of analytic functions like $e^{-1/x}$ are analytic too. The problem only arises when we try to include 0, illustrating that analyticity is very much dependent on the domain under consideration.

Lecture 16: More on Special Series

Warm-Up. Suppose f is analytic on $U \subseteq \mathbb{R}$ and not the constant zero function. We claim that the zeroes of f are then isolated, meaning that if $f(a) = 0$ for some $a \in U$, there exists $\delta > 0$ such that $(a - \delta, a + \delta)$ contains no other zeroes of f apart from a, so that zeroes of f occur at some positive distance from one another. This gives credence to the idea that an analytic function can thought of as an "infinitely long polynomial" via a power series expansion, since zeroes of nonzero polynomials are for sure isolated since there are only finitely many; in the "infinite" version of this, we might have infinitely many zeroes, but at least they are still guaranteed to be isolated. (Think about the zeroes of $\sin x$, for example.)

We can write f as a convergent power series on some interval around a as

$$f(x) = \sum_{n=0}^{\infty} c_n (x - a)^n.$$

If all coefficients c_n here are zero, then this series is the constant zero function and has infinite radius of convergence, meaning that this equality would be true everywhere, but this is not true since we are assuming that f is not the constant zero function. Thus this series must have some nonzero coefficients. Since $f(a) = 0 = c_0$ is the constant term in the series, we know that the first nonzero coefficient must hence be some c_k where $k \geq 1$.

With this k, we can then write our series as

$$f(x) = c_k (x - a)^k + c_{k+1}(x - a)^{k+1} + \cdots,$$

where we've ignored all the zero terms occurring before we hit $c_k(x - a)^k$. Factoring out $(x - a)^k$ gives

$$f(x) = (x - a)^k \underbrace{[c_k + c_{k+1}(x - a) + c_{k+2}(x - a)^2 + \cdots]}_{g(x)},$$

where we denote the function to which the power series in brackets converges to by g. Since $c_k \neq 0$, $h(a) = c_k \neq 0$, and since g is continuous (as is any convergent power series), we have that g is nonzero on some $(a - \delta, a + \delta) \subseteq U$. Hence on this interval the only way in which f can be zero is for the first factor $(x - a)^k$ above to be zero, which only occurs at $x = a$, so the zero at a is isolated.

Identity theorem. The fact about having isolated zeroes places severe restrictions on the behavior of analytic functions. The most important of which is the *identity theorem*, which states that if two analytic functions f and g agree on a set E which contains a limit point, then f and g must be the same everywhere. This is quite restrictive, since it says that knowing only how f behaves on the elements of E determine how it behaves *everywhere* else. For example, open intervals contain limits points, so if an analytic function is equal to 0 on some interval—no matter how small!!!—it must be zero everywhere. This is in stark contrast to what happens for even smooth functions (let alone ones which are only continuous), where having a function be zero, say, on a small interval says absolutely nothing about how it must behave anywhere else. This again rules out the smooth function which is $e^{-1/x}$ for $x > 0$ and 0 elsewhere from being analytic, since this is zero on all of $(-\infty, 0]$ but not on $(0, \infty)$.

To see this, suppose $f = g$ on E which contains a limit point p. Then $f - g$, which is still analytic, is zero at all elements of E, including p. But since p is a limit point, the zero at p is not isolated since any interval around it will contain an element of E and hence another zero of

$f - g$, so the Warm-Up shows that $f - g$ must in fact be the constant zero function, so that $f = g$ everywhere. If an analytic function f has, for example, value

$$f(\tfrac{1}{n}) = e^{1/n} \text{ for all } n \in \mathbb{N},$$

then by continuity we must have $f(0) = e^0$ as well, so that $f(x)$ and e^x agree on the limit-point-containing set $E = \{0\} \cup \{\frac{1}{n} \mid n \in \mathbb{N}\}$, and hence the identity theorem guarantees that $f(x) = e^x$ everywhere, solely from knowledge of the values $f(\frac{1}{n})$. Quite restrictive indeed.

Intro to complex analysis. Most functions you've ever written down in your life—at least unless you were purposefully trying to create some discontinuities—were analytic. Indeed, things made out of polynomials, exponentials, sines, and cosines are analytic, and then taking sums, products, reciprocals, and compositions (see the next homework) gives more analytic things. So,

$$f(x) = \frac{1}{1 + x^2}$$

is certainly analytic on $\mathbb{R}$. And yet, finding explicit power series expressions, let alone their radii of convergence, is not always so easy. In the example above, this is pretty easy when we take the center to be zero using the standard geometric series $\frac{1}{1-y} = \sum_n y^n$, but finding a series expansion around 1, say, is challenging. This is due to the fact that the Taylor coefficients are not easy to compute directly since there will be no "nice" discernible pattern which arises. So, how do we actually work with such series expansions in general, and why do we actually care about analytic functions at all?

The answer comes from the realm of *complex analysis*, which is where the notion of analyticity finds its true home. We give a crash course introduction to this subject here, if only to put our discussion into the proper context. Let C denote the set of continuous functions on some domain, and C^k the set of functions which are k-times continuously differentiable. We then have the containments

$$C \supseteq C^1 \supseteq C^2 \supseteq C^3 \supseteq \cdots C^\infty \supseteq C^\omega := \{\text{analytic functions}\},$$

where the idea is that as we move to the right functions get "nicer": differentiable functions are nicer to work with than continuous functions, twice differentiable ones nicer than differentiable ones, and so on. Note that each of these containments is *strict*, since we can find examples of functions belonging to one set but not the next; in particular, we saw last time an example of a function which is infinitely-differentiable but not analytic. In a sense, real analytic functions are the "end of the line" for how nice a real function can get.

But, real analytic functions are only the starting point in the subject of complex analysis. If real analysis studies functions defined on the set of real numbers $\mathbb{R}$, complex analysis studies functions defined on the set of complex numbers $\mathbb{C}$. The key definition is what it means for a complex function to be *complex differentiable*, which is based on the same type of limit which defines real differentiability:

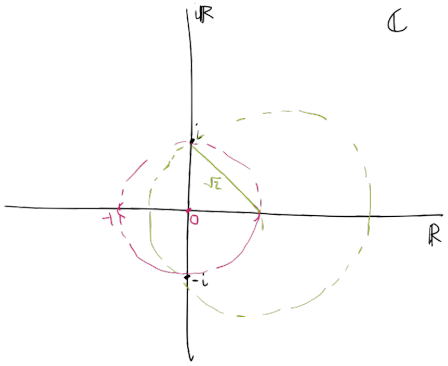
$$f'(z) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}.$$

Here, f denotes a complex function $f : \mathbb{C} \to \mathbb{C}$ (or with domain only some subset of $\mathbb{C}$), z is a complex variable, and z_0 a complex number. This limit makes perfect sense in this settings as well, and all the same derivative rules we had for real functions work in this setting too: the derivative of z^n is nz^{n-1}, the derivative of $\sin z$ is $\cos z$ (once you define what it means to take sine or cosine of a complex number), and so on.

The punchline is that when it comes to complex differentiability, the distinctions given in the containments above for real functions disappear: if a complex function is complex differentiable, it is *automatically* complex twice differentiable, three times differentiable, complex infinitely differentiable, and complex analytic!!! Thus, "analytic" in the complex setting means the same thing as "differentiable". This is perhaps the ultimate reason why real analytic functions are "nice": they are precisely the types of real functions which can be "extended" to complex differentiable ones.

For example, consider the complex function $f(z) = \frac{1}{1+z^2}$, where z is a complex variable. This function is complex differentiable at all z except $\pm i$ where the denominator is zero. Thus, it is complex analytic on $\mathbb{C}\backslash\{\pm i\}$. In particular, its restriction to the real axis is real analytic on $\mathbb{R}$, which is one way of showing that $\frac{1}{1+x^2}$ is analytic on $\mathbb{R}$. Now, here is another useful fact about complex analytic functions: if f is complex analytic, the Taylor series centered at some z_0 has radius of convergence equal to the radius of the largest possible disk which can be drawn centered at z_0 to avoid points at which f is not differentiable. (We visualize the set of complex numbers $\mathbb{C}$ as a plane with the set of real numbers the x-axis and the set of complex numbers of the form iy with $y \in \mathbb{R}$ the y-axis; the word "disk" here means a disk in this plane.)

Thus, radii of convergence in complex analysis are incredibly simple to compute and no lim sup computations are necessary. In the case of $\frac{1}{1+z^2}$, the largest disk (magenta in the picture below) centered at 0 which can be drawn without hitting a point where f is not differentiable is of radius 1 since f is not differentiable at $\pm i$:



Intersecting this disk with the real axis gives the interval $(-1, 1)$, which is indeed the interval of convergence of the Taylor series of $\frac{1}{1+x^2}$ centered at 0. The Taylor series of $\frac{1}{1+z^2}$ centered at 1 has radius of convergence equal to the distance from 1 to i, and intersecting the disk of convergence (in green) with the real axis shows that the radius of convergence of the Taylor series of $\frac{1}{1+x^2}$ centered at 1 is equal to $\sqrt{2}$. Complex analysis, if nothing else, gives us an easier way to answer questions about real analytic functions, but of course has numerous other uses in its own right. Take a course in complex analysis to learn more.

Trigonometric series. We now shift to our final topic in the study of function spaces, *Fourier series*. The motivation is the same as the one we initially took for studying power series: polynomials are dense among all continuous functions, and power series give a particular "nice" way to try to express a function as a limit of polynomials (Taylor polynomials to be precise, so now we know this only works for functions which are analytic), and now we seek to develop a similar story for trigonometric polynomials, which, as we have seen, are dense among all continuous functions on a circle thought of us $[-\pi, \pi]$ with the endpoints glued.

The role of power series is now played by what is known as a *trigonometric series*, which is a series of the form

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(nx) + b_n \sin(nx)].$$

(We will clarify why we write the constant term as $\frac{a_0}{2}$ instead of just a_0 soon enough.) The partial sums of such a series are just trigonometric polynomials as we defined them before. If we are trying to understand which functions can expressed in such a way, the first thing to note is that such a function would have to be 2π-periodic, since the sine and cosine terms used above are themselves 2π-periodic. The behavior of a 2π-periodic function is determined fully from its behavior on $[-\pi, \pi]$ alone, where we now have the additional constraint that the values at $\pm\pi$ should be the same due to the periodicity. The upshot is that a 2π-periodic function can just be thought of as a function on a circle in the sense we've described previously.

By modifying the period of the sine and cosine terms used above we can account for any possible period. For example,

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(n\pi x) + b_n \sin(n\pi x)]$$

would be used to 2-periodic functions, or equivalently functions on $[-1, 1]$ with the same values at ± 1, and more generally

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(\tfrac{2\pi n x}{L}) + b_n \sin(\tfrac{2\pi n x}{L})]$$

would work for $2L$-periodic functions, or functions on $[-L, L]$ with the values at $\pm L$. We will stick with the 2π-periodic version for simplicity of notation, but everything we do easily carries over the case of a general period after a small change of variables.

Fourier coefficients. Hence suppose that f is a function on $[-\pi, \pi]$ with $f(-\pi) = f(\pi)$. Assume for the time being that we can in fact express f as a uniformly convergent trigonometric series

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(nx) + b_n \sin(nx)].$$

For power series we saw that the coefficients needed must be those given by the coefficients of a Taylor series, so we ask whether we can similarly determine what the coefficients a_0, a_n, b_n above would have to be if such a uniform convergence were to hold. The fact is that as long as f is integrable, such coefficients can be determined explicitly.

Take the expression above and multiply through by some $\cos(mx)$:

$$f(x)\cos(mx) = \tfrac{a_0}{2}\cos(mx) + \sum_{n=1}^{\infty} [a_n \cos(nx)\cos(mx) + b_n \sin(nx)\cos(mx)].$$

If f is integrable, we can integrate both sides and use the uniform convergence to exchange integration and summation to get

$$\begin{aligned}\int_{-\pi}^{\pi} f(x)\cos(mx)\,dx = \frac{a_0}{2}\int_{-\pi}^{\pi} \cos(0x)\cos(mx)\,dx \\ + \sum_{n=1}^{\infty}\left[a_n \int_{-\pi}^{\pi} \cos(nx)\cos(mx)\,dx + b_n \int_{-\pi}^{\pi} \sin(nx)\cos(mx)\,dx\right].\end{aligned}$$

(Note that we wrote the first integral on the right as $\int \cos(0x)\cos(mx)$ instead of $\int \cos(mx)$ using $\cos(0x) = 1$, for reasons that will soon be clear.) To further simplify we need to know the values of all the resulting integrals, but these all have easy values given by what are called the *orthogonality relations* for sine and cosine:

$$\int_{-\pi}^{\pi} \cos(nx)\cos(mx)\,dx = \begin{cases} 0 & m \neq n \\ \pi & m = n \neq 0 \\ 2\pi & m = n = 0 \end{cases}$$

$$\int_{-\pi}^{\pi} \cos(nx)\sin(mx)\,dx = 0 \text{ for all } m, n$$

$$\int_{-\pi}^{\pi} \sin(nx)\sin(mx)\,dx = \begin{cases} 0 & m \neq n \text{ or } m = n = 0 \\ \pi & m = n \neq 0 \end{cases}$$

These can all be justified by direct computations, using various trig identities where appropriate. We will skip the details here. (We will talk about why we use the term "orthogonality" when describing these identities next time.)

If $m = 0$, all the terms in the summation on the right of

$$\int_{-\pi}^{\pi} f(x)\cos(mx)\,dx = \frac{a_0}{2}\int_{-\pi}^{\pi} \cos(0x)\cos(mx)\,dx + \sum_{n=1}^{\infty}\left[a_n \int_{-\pi}^{\pi} \cos(nx)\cos(mx)\,dx + b_n \int_{-\pi}^{\pi} \sin(nx)\cos(mx)\,dx\right].$$

are thus zero and $\int_{-\pi}^{\pi} \cos(0x)\cos(mx)\,dx = 2\pi$, so we are left with

$$\int_{-\pi}^{\pi} f(x)\cos(0x)\,dx = \frac{a_0}{2}(2\pi), \text{ and thus } a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(0x)\,dx = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\,dx.$$

For $m \neq 0$, $\int_{-\pi}^{\pi} \cos(0x)\cos(mx)\,dx = 0$ and the only nonzero term in the infinite summation above is the $a_n \int_{-\pi}^{\pi} \cos(nx)\cos(mx)\,dx$ term in the case where $n = m$, so we get

$$\int_{-\pi}^{\pi} f(x)\cos(mx)\,dx = a_m \int_{-\pi}^{\pi} \cos(mx)\cos(mx)\,dx = a_m\pi, \text{ so } a_m = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(mx)\,dx.$$

Thus for all n, even $n = 0$, we have $a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)\,dx$. (This is the reason for writing the constant term in the trigonometric series as $\frac{a_0}{2}$ instead of as a_0: it allows us to use the same integral formula for a_n for all n at once. Otherwise, because $\int_{-\pi}^{\pi} \cos(0x)\cos(0x)\,dx = 2\pi$ instead of π, we would have to use $a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\cos(0x)\,dx$ as the constant term. This just comes down to a matter of preference, and we prefer to absorb the extra factor of 2 into $\frac{a_0}{2}$ instead of in the integral formula for a_0.)

A similar argument where we multiply both sides of

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty}[a_n\cos(nx) + b_n\sin(nx)]$$

by $\sin(mx)$ and then integrate will yield

$$\int_{-\pi}^{\pi} f(x)\sin(mx)\,dx = a_m \int_{-\pi}^{\pi} \sin(mx)\,dx\sin(m)\,dx = a_m\pi, \text{ so } a_m = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(mx)\,dx.$$

The conclusion is that if

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(nx) + b_n \sin(nx)]$$

is to converge uniformly, the coefficients must be given by

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(nx)\, dx \quad \text{and} \quad b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(nx)\, dx.$$

So, if we are looking for a trigonometric series which we hope will converge to a given integrable f (on $[-\pi, \pi]$ with same values at $\pm\pi$) uniformly, the trigonometric series with these specific coefficients is the one we would need. This is called the *Fourier series* of f, and we will investigate its convergence properties over the next two days.

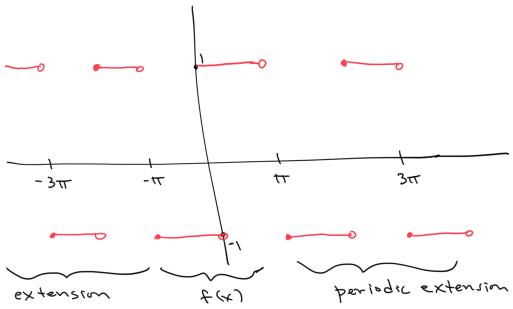
Note that we are considering only Fourier series for real-valued functions here. Rudin discusses the slightly more general case of complex-valued functions, where the benefit is that we can use complex exponentials instead of sines and cosines to express the Fourier series in a more compact way. The theory works exactly the same, so we will stick with the real case so as to be consistent with everything else we have done.

Lecture 17: Fourier Analysis

Warm-Up. We compute the Fourier series of the "square wave" function defined over $[-\pi, \pi]$ by

$$f(x) = \begin{cases} -1 & -\pi \leq x < 0 \\ 1 & 0 \leq x < \pi. \end{cases}$$

To be more precise, we extend this function to be periodic over the rest of $\mathbb{R}$ with period 2π, and compute the Fourier series of the resulting function. The name "square wave" comes from the picture of its graph:



The Fourier coefficients of f are:

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx\, dx = 0 \text{ for } n \geq 0$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx\, dx = \frac{1}{\pi} \left(\int_{-\pi}^{0} -\sin nx\, dx + \int_{0}^{\pi} \sin nx\, dx \right) = \frac{2(1 - \cos n\pi)}{n\pi} \text{ for } n > 0.$$

A quick way of seeing the value of a_n is to note that $f(x)$ is an odd function, so $f(x) \cos nx$ is odd and hence should integrate to 0 over a symmetric interval. The point is that the Fourier series of

an odd function should have no cosine terms in it at all since such terms would prevent the series from being odd; similarly, the Fourier series of an even function should have no sine terms in it. The Fourier series of f is thus

$$\sum_{n=1}^{\infty} \frac{2(1-\cos n\pi)}{n\pi} \sin nx = \sum_{k=0}^{\infty} \frac{4}{(2k+1)\pi} \sin(2k+1)x,$$

where the second expression comes from rewriting the first to only consider odd integers $n = 2k+1$, which we can do because $1 - \cos n\pi = 0$ for even n.

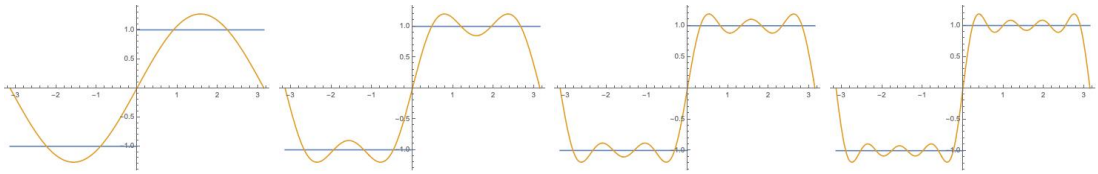
Let us get a sense of the convergence of this Fourier series by considering the partial sums

$$(S_N f)(x) = \sum_{k=1}^{N} \frac{2(1-\cos k\pi)}{k\pi} \sin kx.$$

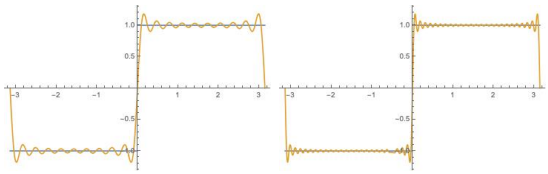
In particular, using the simplified form where we only consider odd indices, we have

$$\begin{aligned}
(S_1 f)(x) &= \frac{4}{\pi} \sin x \\
(S_3 f)(x) &= \frac{4}{\pi} \sin x + \frac{4}{3\pi} \sin 3x \\
(S_5 f)(x) &= \frac{4}{\pi} \sin x + \frac{4}{3\pi} \sin 3x + \frac{4}{5\pi} \sin 5x \\
(S_7 f)(x) &= \frac{4}{\pi} \sin x + \frac{4}{3\pi} \sin 3x + \frac{4}{5\pi} \sin 5x + \frac{4}{7\pi} \sin 7x
\end{aligned}$$

The graphs of these (in yellow) vs $f(x)$ (in blue) look like



Here are $S_{19} f$ and $S_{49} f$:

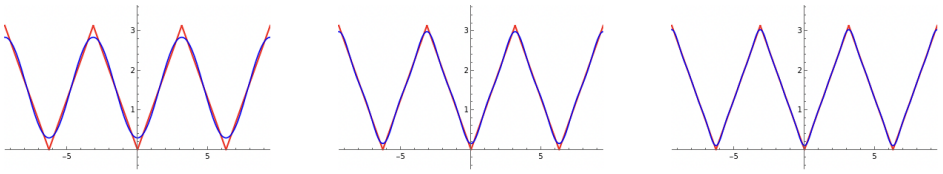


The takeaway is that the Fourier partial sums do seem to be visually approaching f, at least at points where f is continuous! (So, not at 0 or $\pm\pi$.) This is the type of result we aim to justify, under appropriate assumptions. (In fact, continuity alone will not be enough as there are examples of continuous functions whose Fourier series do not converge pointwise! But, everything works fine with just some slightly stronger forms of continuity.)

For another visual example, the "triangular wave" function which is $f(x) = |x|$ for $-\pi \le x \le \pi$ and then extended 2π-periodically (this is similar to the function we used previously to construct a continuous but nowhere differentiable function, only with different period) has Fourier series

$$\frac{\pi}{2} + \sum_{n=1}^{\infty} \frac{2(\cos n\pi - 1)}{\pi n^2} \cos(nx),$$

whose first few partial sums look like



Again, these do appear to be approaching the triangular wave, in this case everywhere since the triangular wave is continuous (in the slightly stronger way we'll describe next time) everywhere.

Fourier partial sums. In order to approach the problem of determining convergence of Fourier series we must investigate their partial sums

$$(S_N f)(x) := \frac{a_0}{2} + \sum_{n=1}^{N} [a_n \cos(nx) + b_n \sin(nx)].$$

It will be useful to have an alternative way of writing this partial sum expression. To get this, we substitute in the integral formulas for all the coefficients:

$$(S_N f)(x) = \frac{\frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\,dt}{2} + \sum_{n=1}^{N} \left[\frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos(nt)\cos(nx)\,dt + \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin(nt)\sin(nx)\,dt \right].$$

Note that we use t as the variable of integration in the Fourier coefficient formulas to distinguish it from x, which is the value at which the partial sum $S_N f$ is being evaluated.

We can now manipulate by exchanging integrations and (finite) summations, and then factoring a common $f(t)$ out of every term to get

$$(S_N f)(x) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t) \left(\frac{1}{2} + \sum_{n=1}^{N} [\cos(nt)\cos(nx) + \sin(nt)\sin(mx)] \right) dt.$$

Using the angle sum formula $\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b)$ for cosine, we can further write the partial sum above as

$$(S_N f)(x) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t) \left(\frac{1}{2} + \sum_{n=1}^{N} \cos[n(x-t)] \right) dt.$$

Controlling the behavior of these partial sums then comes down to controlling the behavior of this integral, and in particular the behavior of the function being multiplied by $f(t)$.

Dirichlet kernel. Set D_N to be the function defined by

$$D_N(x) = \frac{1}{2} + \sum_{n=1}^{N} \cos(nx),$$

so that we can write the Fourier partial sums as

$$(S_N f)(x) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t) D_N(x-t)\,dt.$$

After a change of variables $u = x - t$, we get

$$(S_N f)(x) = \frac{1}{\pi}\int_{x-\pi}^{x+\pi} f(x-u) D_N(u)\,du = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x-u) D_N(u)\,du,$$

where in the second integral we use the fact that f and D_N are 2π-periodic to write the integral over $[x-\pi, x+\pi]$ as that over $[-\pi, \pi]$ instead. (Integrating a periodic function over any interval whose length matches the period always gives the same value.)

The function D_N is called the N-th order *Dirichlet kernel*, and understanding its properties is crucial to understanding convergence of Fourier series. The integral expression above for $S_N f$ is called the *convolution* of f with D_N, and is typically denoted by $f \star D_N$:

$$(f \star D_N)(x) := (S_N f)(x) = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x-u) D_N(u)\, du.$$

We previously saw the terms "kernel" and "convolution" in the context of the Landau kernels used to prove Weierstrass approximation, and they were also used on the previous homework. (The homework problem this refers to deals with so-called "good" kernels, where the result is some uniform convergence property of convolutions with good kernels. The Dirichlet kernels are not "good" in that sense, however, which is why convergence of Fourier series can be quite a delicate problem. We'll say more next time.) In both the homework and the previous Landau case the convolution was written using $f(x+t)$ instead of $f(x-t)$ as given here, but the two expressions give the same result in this case since D_N is an even function and the interval of integration is symmetric about 0, so that the change of variables $u \mapsto -u$ does not change the integral.

Back to linear algebra. Before moving on, it will be useful to describe the linear-algebraic context behind the study of Fourier series. Recall that given an *orthogonal* basis $\mathbf{b}_1, \dots, \mathbf{b}_n$ for $\mathbb{R}^n$—where the dot product between any two vectors is zero—all $\mathbf{x} \in \mathbb{R}^n$ can be decomposed in terms of this basis as

$$\mathbf{x} = \left(\frac{\mathbf{x} \cdot \mathbf{b}_1}{\mathbf{b}_1 \cdot \mathbf{b}_1}\right) \mathbf{b}_1 + \cdots + \left(\frac{\mathbf{x} \cdot \mathbf{b}_n}{\mathbf{b}_n \cdot \mathbf{b}_n}\right) \mathbf{b}_n.$$

Moreover, the part of this sum that goes up to $\mathbf{e}_k$ gives the *orthogonal projection* of $\mathbf{x}$ onto the subspace V spanned by $\mathbf{b}_1, \dots, \mathbf{b}_k$:

$$\operatorname{proj}_V \mathbf{x} = \left(\frac{\mathbf{x} \cdot \mathbf{b}_1}{\mathbf{b}_1 \cdot \mathbf{b}_1}\right) \mathbf{b}_1 + \cdots + \left(\frac{\mathbf{x} \cdot \mathbf{b}_k}{\mathbf{b}_k \cdot \mathbf{b}_k}\right) \mathbf{b}_k.$$

This orthogonal projection is characterized by the property that $\mathbf{x} - \operatorname{proj}_V \mathbf{x}$ is orthogonal to all elements of V, or equivalently that $\operatorname{proj}_V \mathbf{x}$ is the element of V that is closest to $\mathbf{x}$ in the sense that it minimizes the length of $\mathbf{x} - \mathbf{v}$ among all $\mathbf{v} \in \operatorname{span}(\mathbf{b}_1, \dots, \mathbf{b}_k)$.

The point is that this is precisely what is going on with Fourier series as well. Define an *inner product* (generalization of dot product) on functions by

$$\langle f, g \rangle := \int_{-\pi}^{\pi} f(x) g(x)\, dx.$$

(The right side is an "uncountable sum" analog of the usual dot product sum $\mathbf{x} \cdot \mathbf{y} = \sum_i x_i y_i$.) The "orthogonality relations"

$$\langle \cos(nx), \cos(mx) \rangle = \int_{-\pi}^{\pi} \cos(nx)\cos(mx)\, dx = \begin{cases} 0 & m \neq n \\ \pi & m = n \neq 0 \\ 2\pi & m = n = 0 \end{cases}$$

$$\langle \cos(nx), \sin(mx) \rangle = \int_{-\pi}^{\pi} \cos(nx)\sin(mx)\, dx = 0 \text{ for all } m, n$$

$$\langle \sin(nx), \sin(mx) \rangle = \int_{-\pi}^{\pi} \sin(nx)\sin(mx)\,dx = \begin{cases} 0 & m \neq n \text{ or } m = n = 0 \\ \pi & m = n \neq 0 \end{cases}$$

we introduced last time then say precisely that the functions $1, \cos(nx), \sin(nx)$ for $n > 0$ (note $1 = \cos(0x)$) are orthogonal with respect to this inner product. The Fourier coefficients are then

$$a_n = \frac{\int_{-\pi}^{\pi} f(x)\cos(nx)\,dx}{\pi} = \frac{\langle f(x), \cos(nx)\rangle}{\langle \cos(nx), \cos(nx)\rangle} \text{ and } b_n = \frac{\int_{-\pi}^{\pi} f(x)\sin(nx)\,dx}{\pi} = \frac{\langle f(x), \sin(nx)\rangle}{\langle \sin(nx), \sin(nx)\rangle}$$

when $n > 0$, and

$$\frac{a_0}{2} = \frac{\int_{-\pi}^{\pi} f(x)\cos(0x)\,dx}{2\pi} = \frac{\langle f(x), \cos(0x)\rangle}{\langle \cos(0x), \cos(0x)\rangle},$$

so the Fourier series decomposition (when valid)

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(nx) + b_n \sin(nx)]$$

is thus nothing but an infinite sum analog of

$$\mathbf{x} = \left(\frac{\mathbf{x}\cdot\mathbf{b}_1}{\mathbf{b}_1\cdot\mathbf{b}_1}\right)\mathbf{b}_1 + \cdots + \left(\frac{\mathbf{x}\cdot\mathbf{b}_n}{\mathbf{b}_n\cdot\mathbf{b}_n}\right)\mathbf{b}_n.$$

In the $\mathbb{R}^n$ case we can turn an orthogonal basis into an *orthonormal* basis $\mathbf{u}_i = \frac{\mathbf{b}_i}{\|\mathbf{b}_i\|}$ (where each basis vector now has norm 1) by dividing by lengths $\|\mathbf{b}_i\| = \sqrt{\mathbf{b}_i \cdot \mathbf{b}_i}$, and then the decomposition of $\mathbf{x}$ looks like

$$\mathbf{x} = (\mathbf{x}\cdot\mathbf{u}_1)\mathbf{u}_1 + \cdots + (\mathbf{x}\cdot\mathbf{u}_n)\mathbf{u}_n.$$

Similarly, we get "orthonormal" functions by dividing each of $1, \cos(nx), \sin(nx)$ by its "length", which is the square root of the inner product of it with itself:

$$\frac{1}{\sqrt{2\pi}},\ \frac{\cos(x)}{\sqrt{\pi}},\ \frac{\sin(x)}{\sqrt{\pi}},\ \frac{\cos(2x)}{\sqrt{\pi}},\ \frac{\sin(2x)}{\sqrt{\pi}},\ \ldots.$$

For ease of notation, let us denote these orthonormal functions by $\phi_0, \phi_1, \phi_2, \ldots$ as they appear, so

$$\phi_0 = \frac{1}{\sqrt{2\pi}},\ \phi_1 = \frac{\cos(x)}{\sqrt{\pi}},\ \phi_2 = \frac{\sin(x)}{\sqrt{\pi}},\ \phi_3 = \frac{\cos(2x)}{\sqrt{\pi}},\ \phi_4 = \frac{\sin(2x)}{\sqrt{\pi}},\ \ldots.$$

(So, odd indices for the cosine ones and even indices for the sine ones, except for ϕ_0 which uses $\cos(0x) = 1$.) Then

$$a_n \cos(nx) = \frac{\langle f(x), \cos(nx)\rangle}{\langle \cos(nx), \cos(nx)\rangle}\cos(nx) = \left\langle f(x), \frac{\cos(nx)}{\sqrt{\pi}}\right\rangle \frac{\cos(nx)}{\sqrt{\pi}} = \langle f(x), \phi_{2n-1}(x)\rangle\, \phi_{2n-1}(x)$$

for $n \geq 1$, and similarly

$$\frac{a_0}{2}\cos(0x) = \langle f(x), \phi_0(x)\rangle\, \phi_0(x) \quad \text{and} \quad b_n \sin(nx) = \langle f(x), \phi_{2n}(x)\rangle\, \phi_{2n}(x).$$

The upshot is that the Fourier series expression

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(nx) + b_n \sin(nx)] = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(nx) + \sum_{n=1}^{\infty} b_n \sin(nx)$$

then looks like

$$\sum_{n=0}^{\infty} \langle f, \phi_n \rangle \, \phi_n,$$

which is an infinite sum analog of $(\mathbf{x} \cdot \mathbf{u}_1)\mathbf{u}_1 + \cdots + (\mathbf{x} \cdot \mathbf{u}_n)\mathbf{u}_n$. The coefficients $c_n := \langle f, \phi_n \rangle$ here are related to the previous coefficients (before we normalized) by $c_0 = \sqrt{2\pi}\,\frac{a_0}{2}$ and

$$c_{2n-1} = \sqrt{\pi}\, a_n \quad \text{and} \quad c_{2n} = \sqrt{\pi}\, b_n \text{ for } n \geq 1.$$

Integral minimizers. If the linear-algebraic interpretation above is to be the correct one, we would expect a Fourier partial sum

$$\sum_{n=0}^{N} \langle f, \phi_n \rangle \, \phi_n$$

to be the "orthogonal projection" of f onto the space spanned by $\phi_0, \ldots, \phi_N$, which we take as meaning that this partial sum should be the element in this span that minimizes "distance" to f. Here we interpret "distance" as the uncountable analog of $\sqrt{\sum_i (x_i - y_i)^2}$ given by

$$\sqrt{\int_{-\pi}^{\pi} (f(x) - g(x))^2 \, dx}.$$

This square root is minimized when the integral of which the square root is taken is minimized, so the claim is that

$$\sum_{n=0}^{N} \langle f, \phi_n \rangle \, \phi_n$$

minimizes the integral

$$\int_{-\pi}^{\pi} (f - T_N)^2$$

among all trigonometric polynomials T_N spanned by $\phi_0, \ldots, \phi_N$.

To see this, write the Fourier series of f as $\sum_{n=0}^{\infty} c_n \phi_n$ with $c_n = \langle f, \phi_n \rangle$, and write an arbitrary $T_N = \sum_{m=0}^{N} d_m \phi_m$ with $d_m \in \mathbb{R}$. The expression we want to minimize is

$$\int_{-\pi}^{\pi} (f - T_N)^2 = \int_{-\pi}^{\pi} f^2 - 2 \int_{-\pi}^{\pi} f T_N + \int_{-\pi}^{\pi} T_N^2.$$

Using the fact that the ϕ_i are orthonormal with respect to $\langle f, g \rangle = \int_{-\pi}^{\pi} fg$, we have

$$\int_{-\pi}^{\pi} T_N^2 = \int_{-\pi}^{\pi} \left(\sum_{n=0}^{N} d_n \phi_n \right) \left(\sum_{m=0}^{N} d_m \phi_m \right) = \sum_{n,m=0}^{N} d_n d_m \underbrace{\int_{-\pi}^{\pi} \phi_n \phi_m}_{\langle \phi_n, \phi_m \rangle = 0 \text{ or } 1} = \sum_{n=0}^{N} d_n^2.$$

By the definition $c_n = \langle f, \phi_n \rangle = \int_{-\pi}^{\pi} f \phi_n$, we have

$$\int_{-\pi}^{\pi} f T_N == \int_{-\pi}^{\pi} f \sum_{m=0}^{N} d_m \phi_m = \sum_{m=0}^{N} d_m \int_{-\pi}^{\pi} f \phi_m = \sum_{m=0}^{N} d_m c_m.$$

Thus

$$\int_{-\pi}^{\pi} (f - T_N)^2 = \int_{-\pi}^{\pi} f^2 - 2 \int_{-\pi}^{\pi} f T_N + \int_{-\pi}^{\pi} T_N^2$$

$$= \int_{-\pi}^{\pi} f^2 - 2\sum_{n=0}^{N} d_n c_n + \sum_{n=0}^{N} d_n^2$$
$$= \int_{-\pi}^{\pi} f^2 - \sum_{n=0}^{N} c_n^2 + \sum_{n=0}^{N} (d_n - c_n)^2,$$

where in the last step we use $(d_n - c_n)^2 - c_n^2 = -2d_n c_n + d_n^2$. This final line is the expression we wish to minimize, and since the final term $\sum_{n=0}^{N}(c_n - d_n)^2$ is nonnegative, the expression is minimized when this final term is zero, which occurs if and only if each $(d_n - c_n)^2$ is zero, so if and only if $d_n = c_n$ for $n = 0, \ldots, N$. Thus, $\int_{-\pi}^{\pi}(f - T_N)^2$ is minimized when $T_N = \sum_{n=0}^{N} c_n \phi_n$ is indeed the N-th partial sum of the Fourier series $\sum_{n=0}^{\infty} c_n \phi_n$. (The same proof works with respect to any inner product, and in particular gives the fact that orthogonal projections in $\mathbb{R}^n$ minimize distance.)

Bessel and Riemann-Lebesgue. As a consequence of the work above, when T_N *is* a partial sum of the Fourier series of f we have

$$\int_{-\pi}^{\pi} (f - T_N)^2 = \int_{-\pi}^{\pi} f^2 - \sum_{n=0}^{N} c_n^2$$

with c_n the Fourier coefficients. The left side is nonnegative, so we get that

$$\sum_{n=0}^{N} c_n^2 \leq \int_{-\pi}^{\pi} f^2.$$

Since this holds for all N, we then get

$$\sum_{n=0}^{\infty} c_n^2 \leq \int_{-\pi}^{\pi} f^2,$$

a result known as *Bessel's inequality*. The point here is that, as a consequence, the series on the left actually *converges*. Thus, the c_n^2, and hence the c_n must converge to 0, a result known as the *Riemann-Lebesgue lemma*. (Not to be confused with the Riemann-Lebesgue theorem we saw earlier in the context of integrability via measure zero. The phrase "Riemann-Lebesgue theorem" for the integrability result is common but not super widespread, whereas everyone on Earth does use "Riemann-Lebesgue lemma" for the Fourier coefficient result.)

If we translate the results above in terms of c_n, ϕ_n into corresponding statements in terms of $a_n, \cos(nx), b_n, \sin(nx)$ using

$$c_0 = \sqrt{2\pi}\,\frac{a_0}{2},\ c_{2n-1} = \sqrt{\pi}\,a_n,\ c_{2n} = \sqrt{\pi}\,b_n \text{ for } n \geq 1$$

as described earlier, we get that Bessel's inequality is

$$\frac{a_0^2}{2} + \sum_{n=0}^{\infty}(a_n^2 + b_n^2) \leq \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)^2\,dx$$

and that the Riemann-Lebesgue lemma becomes

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)\,dx \to 0 \quad \text{and} \quad b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)\,dx \to 0$$

as $n \to \infty$. The fact that Fourier coefficients decay to 0 is crucial to understanding the convergence of Fourier series. (It can be shown that Bessel's inequality is actually an equality, a result known as *Parseval's theorem*. Rudin proves this for a general integrable f, but we will not give the proof here as we will not need this result going forward. We will, however, give a proof in the case where f is continuous as an application of another result we'll see. Parseval's theorem is nothing but an infnite-dimensional analog of the Pythagorean theorem—convince yourself why!)

Lecture 18: Convergence of Fourier Series

Warm-Up 1. We show that if continuous functions f and g have the same Fourier coefficients (and hence the same Fourier series), then $f = g$. This a type of uniqueness result in that an integrable function is uniquely determined by its Fourier coefficients. Note that if Fourier series always converged uniformly (or even pointwise) this would be immediate, as then f can be recovered directly from

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(nx) + b_n \sin(nx)].$$

But the claim here is more general assuming only continuity of our functions, which is not enough to guarantee any type of convergence of the Fourier series.

A key observation is that Fourier coefficients are linear in the sense that the Fourier coefficients of $f + g$ are those of f plus those of g:

$$a_n(f+g) = \int_{-\pi}^{\pi} (f(x)+g(x)) \cos(nx)\, dx = \int_{-\pi}^{\pi} f(x) \cos(nx)\, dx + \int_{-\pi}^{\pi} g(x) \cos(nx)\, dx = a_n(f)+a_n(g)$$

and similar for the sine coefficients b_n. (Here we include the function of which we are computing the Fourier coefficients as part of the notation, so $a_n(f)$ versus a_n alone.) If f and g thus have the same Fourier coefficients, then $f - g$ has all zero Fourier coefficients:

$$a_n(f-g) = a_n(f) - a_n(g) = 0 \quad \text{and} \quad b_n(f-g) = b_n(f) - b_n(g) = 0.$$

But actually, we have already studied functions satisfying $a_n(f - g) = 0 = b_n(f - g)$ for all n before, namely as the Warm-Up in Lecture 14. There we showed, using denseness of trigonometric polynomials, that

$$a_n(h) = \int_{-\pi}^{\pi} h(x) \cos(nx)\, dx = 0 = \int_{-\pi}^{\pi} h(x) \sin(nx)\, dx = b_n(h)$$

for all n forces $h = 0$ (assuming h is continuous), which is precisely the setup we are looking at now. We thus get immediately that $f - g = 0$, so $f = g$. The takeaway is that that previous denseness result is really a result about Fourier coefficients uniquely determining a function.

Warm-Up 2. We show that if f is C^1, then $na_n \to 0$ and $nb_n \to 0$ as $n \to \infty$. We know already from the Riemann-Lebesgue lemma that a_n and b_n both decay to zero, so what this now gives us is a faster type of decay, at least for C^1 functions. (More generally, the same argument works even if f' is just integrable, but C^1 is the usual way to guarantee that.) We have

$$a_n(f) = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(nx)\, dx.$$

The point is that we can rewrite this by integrating by parts:

$$\begin{aligned} a_n(f) &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)\,dx \\ &= \frac{1}{n\pi} f(x)\sin(nx)\Big|_{-\pi}^{\pi} - \frac{1}{n\pi}\int_{-\pi}^{\pi} f'(x)\sin(nx)\,dx. \end{aligned}$$

The first term is zero, and $\frac{1}{\pi}\int_{-\pi}^{\pi} f'(x)\sin(nx)\,dx$ in the second term is precisely the Fourier coefficient $b_n(f')$ of f', so

$$a_n(f) = -\frac{1}{n}\left(\frac{1}{\pi}\int_{-\pi}^{\pi} f'(x)\sin(nx)\,dx\right) = -\frac{1}{n}b_n(f').$$

A similar integration by parts computation (which will now use the fact that f is periodic to deal with the first term) gives

$$b_n(f) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)\,dx = \frac{1}{n\pi}\int_{-\pi}^{\pi} f'(x)\cos(nx)\,dx = \frac{1}{n}a_n(f').$$

(So, taking derivatives exchanges the roles of a_n and b_n, up to some easy factor.) Thus we get

$$na_n(f) = -b_n(f') \quad \text{and} \quad nb_n(f) = a_n(f'),$$

so since the Fourier coefficients of f' approach zero by Riemann-Lebesgue, so do $na_n(f)$ and $nb_n(f')$.

More integration by parts applications will show that if f is C^2 then $n^2a_n, n^2b_n \to 0$, if f is C^3 then $n^3a_n, n^3b_n \to 0$, and so on, so that higher-orders of differentiability give better rates of decay of Fourier coefficients.

Back to convergence. Last time we derived the expression

$$(S_N f)(x) = (f \star D_N)(x) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x-t)D_N(t)\,dt$$

for the partial sums of a Fourier series, where D_N is the N-th order Dirichlet kernel and $\star$ is convolution. To write $(f \star D_N)(x) - f(x)$ in a useful way, we use the fact that

$$\int_{-\pi}^{\pi} D_0(t)\,dt = \int_{-\pi}^{\pi}\frac{1}{2}\,dt = \pi$$

and

$$\int_{-\pi}^{\pi} D_N(t)\,dt = \int_{-\pi}^{\pi}\left(\frac{1}{2} + \sum_{n=1}^{N}\cos(nt)\right)dt = \pi + \sum_{n=1}^{N}\int_{-\pi}^{\pi}\cos(nt)\,dt = \pi + \sum_{n=1}^{N} 0 = \pi$$

for $N \geq 1$. (Thus, the Dirichlet kernels satisfy the first property required of a "good" kernel, appropriately modified for the interval $[-\pi, \pi]$ as opposed to $[-1, 1]$ as on the last homework.) This gives

$$\frac{1}{\pi}\int_{-\pi}^{\pi} D_N(t)\,dt = 1 \text{ for all } N \geq 0,$$

so

$$(f \star D_N)(x) - f(x) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x-t)D_N(t)\,dt - f(x)\underbrace{\frac{1}{\pi}\int_{-\pi}^{\pi} D_N(t)\,dt}_{1}$$

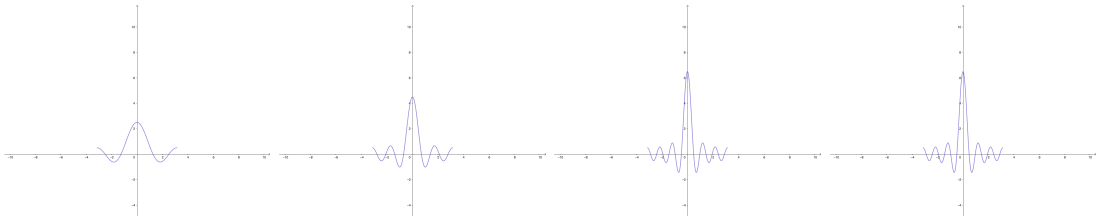
$$= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x-t)D_N(t)\,dt - \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)D_N(t)\,dt$$
$$= \frac{1}{\pi}\int_{-\pi}^{\pi} [f(x-t)-f(x)]D_N(t)\,dt.$$

Showing convergence of Fourier series thus comes down to make this final expression small.

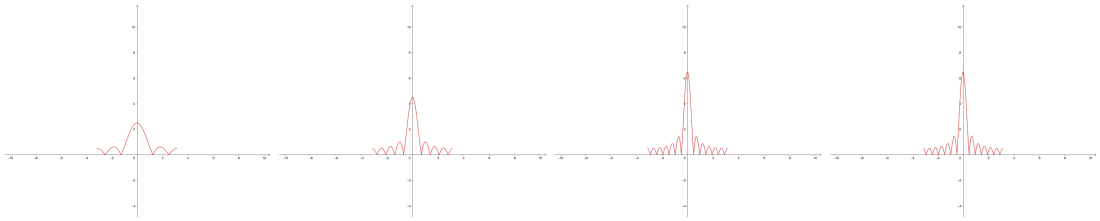
If the D_N did form a good kernel, we would be finished (assuming f is continuous), as we've stated previously. But the D_N do not form a good kernel, specifically because the property which asks that the integrals

$$\int_{-\pi}^{\pi} |D_N(t)|\,dt$$

be bounded by some common $M > 0$ fails. (This was needed in the good kernel problem on the homework since at some point you have to use continuity to make $f(x-t)-f(x)$ small and need something like $\int_{|t|\leq\delta} |D_N(t)|\,dt \leq M$ to deal with the rest of the integral near 0.) The Dirichlet kernels look like (these are D_2, D_4, D_6, D_8)



Each net area is π, and in fact we do have convergence of the integral of $|D_N|$ to 0 near the endpoints away from $t = 0$, so the third property of "good kernel" will actually hold. The problem is that D_N can take on negative values (for $N \geq 1$), which is what allows for "cancellations" to occur so that the net area remains 1 always, but after we take absolute values, all negative values get flipped above, and in fact this causes the areas to explode to ∞ as N increases:



So, no possible bound on the integrals of all the absolute values, so no good kernel. Sad days.

Lipschitz is enough. Instead of relying a general good kernel argument, we instead use the decay of the Fourier coefficients as given by Riemann-Lebesgue to derive a convergence result, specifically one for functions which are *Lipschitz* continuous. To say that f is Lipschitz means that there exists $C > 0$ such that

$$|f(x) - f(y)| \leq C|x-y| \text{ for all } x, y.$$

This implies ordinary continuity (in fact uniform continuity), and is actually a property we've seen a few times throughout the quarter: contractions, for example, are the case where $C < 1$, and if f is C^1 on some $[a, b]$ we can bound the derivative in a mean value application

$$|f(x) - f(y)| = |f'(c)||x-y|$$

to get the Lipschitz condition. So, a fairly mild restriction in the end.

We will prove that if f is Lipschitz, then the Fourier series of f converges to f pointwise. (It is in fact true that the Fourier series will converge uniformly as well. We'll point out at the end what would be needed in order to get uniform convergence.) We have

$$(f \star D_N)(x) - f(x) = \frac{1}{\pi} \int_{-\pi}^{\pi} [f(x-t) - f(x)] D_N(t)\, dt.$$

After the use of some various trig identities (that we will skip), it turns out that the Dirchlet kernel can be written as

$$D_N(t) = \frac{\sin[(n+\frac{1}{2})t]}{2\sin\frac{t}{2}} \text{ for } t \neq 0, \text{ and } D_N(0) = \tfrac{1}{2} + N.$$

(The value at 0, or more generally any even multiple of π where $\sin\frac{t}{2}$ is zero, can be found from the original definition of D_N as $\frac{1}{2}$ plus a sum of cosines. We'll ignore the value at 0 though since it will not affect the integral anyway. Note that this Dirchlet kernel differs slightly from the version Rudin gives because of the 2 in the denominator, which stems from the fact that Rudin is doing all this for complex-valued functions and not just real-valued. Not a big deal.) With this, we have

$$(f \star D_N)(x) - f(x) = \frac{1}{\pi} \int_{-\pi}^{\pi} [f(x-t) - f(x)] \frac{\sin(Nt + \frac{t}{2})}{2\sin\frac{t}{2}}\, dt$$

Using the angle addition formula for $\sin(A+B)$, we can write $\sin(Nt + \frac{t}{2})$ as a product of two cosines plus a product of two signs, and after doing so and breaking up the integral (and moving the location of the $2\sin\frac{t}{2}$ in the denominator) we get that the expression above is

$$\frac{1}{\pi} \int_{-\pi}^{\pi} \left(\frac{f(x-t) - f(t)}{2\sin\frac{t}{2}} \right) \cos(\tfrac{t}{2}) \cos(Nt)\, dt + \frac{1}{\pi} \int_{-\pi}^{\pi} \left(\frac{f(x-t) - f(t)}{2\sin\frac{t}{2}} \right) \sin(\tfrac{t}{2}) \sin(Nt)\, dt.$$

For simplicity of notation, let us do as Rudin does and set

$$g(t) = \frac{f(x-t) - f(x)}{2\sin\frac{t}{2}},$$

(technically only valid for $t \neq 0$, but no matter, just set $g(0) = 0$) so that

$$(f \star D_N)(x) - f(x) = \frac{1}{\pi} \int_{-\pi}^{\pi} g(t) \cos(\tfrac{t}{2}) \cos(Nt)\, dt + \frac{1}{\pi} \int_{-\pi}^{\pi} g(t) \sin(\tfrac{t}{2}) \sin(Nt)\, dt.$$

The key point is that the two integrals on the right are precisely Fourier coefficients, but for the functions $g(t)\cos(\frac{t}{2})$ and $g(t)\sin(\frac{t}{2})$:

$$(f \star D_N)(x) - f(x) = a_N(g(t)\cos\tfrac{t}{2}) + b_N(g(t)\sin\tfrac{t}{2}).$$

Riemann-Lebesgue implies that these coefficients go to 0, and hence $(f \star D_N)(x) \to f(x)$ as desired, almost: this works as long as we know that g is integrable, since we need $g(t)\cos\frac{t}{2}$ and $g(t)\sin\frac{t}{2}$ to be integrable in order to talk about Fourier coefficients and be able to apply Riemann-Lebesgue. In terms of the Riemann-Lebesgue theorem (not lemma!), we can see for sure that

$$g(t) = \frac{f(x-t) - f(x)}{2\sin\frac{t}{2}}$$

fails to be continuous only possibly at the points where f (in the numerator) fails to be continuous or maybe also at $t = 0$, which is the only point where the denominator is zero in $[-\pi, \pi]$. Thus, if

f is integrable, it has a measure zero discontinuity set, so throwing one more point will not change the measure zero property, and hence g has a measure zero discontinuity set.

But we also need to know that g is bounded, and this is where the Lipschitz condition is needed. (Rudin just states that g is bounded without making any attempt to actually justify it—it is not obvious!) The numerator is bounded, but the issue is that the denominator will approach 0 as t approaches 0, so saying that the fraction is still bounded anyway takes some work. For this we use the fact that

$$\frac{\sin y}{y} \to 1 \text{ as } y \to 0.$$

This gives that

$$\left|\frac{\sin \frac{t}{2}}{\frac{t}{2}}\right| \geq \frac{1}{2} \text{ when } |t| < \delta$$

for some $\delta > 0$, so on $(-\delta, \delta)$ we can bound the denominator of $g(t)$ from below by

$$|2 \sin \tfrac{t}{2}| \geq |\tfrac{t}{2}|.$$

Hence for $|t| < \delta$ we have

$$|g(t)| = \frac{|f(x-t) - f(x)|}{|2 \sin \frac{t}{2}|} \leq \frac{|f(x-t) - f(t)|}{\frac{|t|}{2}} \leq \frac{C|t|}{\frac{|t|}{2}} = 2C,$$

where we use the Lipschitz property just before the end. Thus g is bounded on a small $(-\delta, \delta)$, and it is bounded on the rest of $[-\pi, \pi]$ because the denominator does not come close to 0, so g is bounded on all of $[-\pi, \pi]$. Therefore g is integrable, and our argument is complete, meaning that

$$S_N f = f \star D_N \to f \text{ pointwise if } f \text{ is Lipschitz.}$$

Actually, we do not need Lipschitz over all of $[-\pi, \pi]$ to get a convergence result since our argument applies to one x at a time: if we only assume a Lipschitz condition at some x (meaning fix x in the Lipschitz definition and vary the other point), we will get Fourier convergence at least at that x.

In order to turn this into uniform convergence instead, we would need a uniform version of the Riemann-Lebesgue lemma. The argument above used Riemann-Lebesgue with the function

$$g(t) = \frac{f(x-t) - f(x)}{2 \sin \frac{t}{2}}$$

where x is fixed, so for uniform convergence we would instead consider

$$g(x,t) = \frac{f(x-t) - f(x)}{2 \sin \frac{t}{2}}$$

as a function of both x and t. We would then need a version of the Riemann-Lebesgue lemma that guaranteed decay of such "uniform" Fourier coefficients:

$$a_n(g(x,t)), b_n(g(x,t)) \to 0 \text{ uniformly.}$$

You will justify such a version on the homework, and then the pointwise convergence argument here can indeed be turned into uniform convergence.

Example. For the square wave we saw last time with Fourier series

$$\sum_{k=0}^{\infty} \frac{4}{(2k+1)\pi} \sin(2k+1)x,$$

we can now get some concrete values of the sum. The square wave is constant 1 near, say, $x = \frac{\pi}{2}$, so it is Lipschitz at $\frac{\pi}{2}$ and thus the series above converges to 1 at $\frac{\pi}{2}$:

$$1 = \sum_{k=0}^{\infty} \frac{4}{(2k+1)\pi} \sin\left(\frac{[2k+1]\pi}{2}\right) = \sum_{k=0}^{\infty} \frac{4}{(2k+1)\pi}(-1)^k = \frac{4}{\pi}\sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1}.$$

Thus the alternating sum of reciprocals of odd positive integers is

$$\frac{\pi}{4} = \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \cdots.$$

(This sum can also be derived via a Taylor series argument for $\arctan x$, which is analytic on $\mathbb{R}$, by integrating $\frac{1}{1+x^2} = \sum_{n=0}^{\infty}(-1)^n x^{2n}$.)

The triangular wave function is Lipschitz continuous everywhere, so its Fourier series

$$\frac{\pi}{2} - \sum_{n=0}^{\infty} \frac{4}{\pi(2n+1)^2} \cos([2n+1]x)$$

converges to the value of the triangular wave at all x. At $x = \pi$, for example, we thus get

$$\pi = \frac{\pi}{2} - \frac{4}{\pi}\sum_{n=0}^{\infty} \frac{1}{(2n+1)^2} \cos([2n+1]\pi) = \frac{\pi}{2} + \frac{4}{\pi}\sum_{n=0}^{\infty} \frac{1}{(2n+1)^2}.$$

After rearranging, we have

$$\frac{\pi^2}{8} = \sum_{n=0}^{\infty} \frac{1}{(2n+1)^2} = 1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots$$

as the sum of squares of reciprocals of odd positive integers. (This can also be derived using a Taylor series for $\arcsin x$, but I think the Fourier approach is simpler in this case.) Since

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \sum_{n=0}^{\infty} \frac{1}{(2n+1)^2} + \sum_{n=1}^{\infty} \frac{1}{(2n)^2} = \frac{\pi^2}{8} + \frac{1}{4}\sum_{n=1}^{\infty} \frac{1}{n^2},$$

after rearranging we get

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

The series $\sum_{n=1}^{\infty} \frac{1}{n^2}$ has appeared a few times before (in particular on the homework), and now we know the definite value.

Cesàro summability. Instead of pursuing full convergence of Fourier series from the get-go, we can aim for some middle ground via an averaging procedure. If a sequence (a_n) in $\mathbb{R}$ converges to L, then the sequence of averages of terms in (a_n) converges to L as well:

$$a_n \to L \quad \text{implies} \quad \frac{a_1 + \cdots + a_n}{n} \to L.$$

(Proved last quarter, I believe!) The point is that instead of asking about convergence of (a_n) at the start, we can first ask about convergence of these averages, which are generally better behaved

than the a_n since, in a sense, taking averages has the effect of "smoothing out" the effect of some poorly behaved a_n's. If the sequence of averages diverges, then the original sequence diverges as well, whereas if the sequence of averages converges, we might not know that the original sequence converges too, but we at least know what candidate limit it would have to converge to (the same as the averages) if it were to converge. The averages thus give an initial step towards understanding convergence.

We can then apply this idea to the sequence of partial sums of a series. We say that $\sum_{n=0}^{\infty} a_n$ is *Cesàro summable* with *Cesàro sum* L if the sequence of averages

$$\frac{s_0 + \cdots + s_n}{n+1}$$

of the partial sums $s_k = a_0 + \cdots + a_k$—these averages are called the *Cesàro means* of the series—converges to L. If a series converges in the usual sense, it will be Cesàro summable with Cesàro sum equal to the usual sum. Again, Cesàro summable does not imply convergent, but it is a first step towards convergence.

Fejér kernels. Applying this to a Fourier series gives the following. The partial sums of the Fourier series of f are

$$(S_N f)(x) = \frac{a_0}{2} + \sum_{n=1}^{N} [a_n \cos(nx) + b_n \sin(nx)] = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x-t) D_N(t)\, dt.$$

The Cesàro means are thus

$$\begin{aligned} \frac{(S_0 f)(x) + \cdots + (S_N f)(x)}{N+1} &= \frac{\frac{1}{\pi}\int_{-\pi}^{\pi} f(x-t)[D_0(t) + \cdots + D_N(t)]\, dt}{N+1} \\ &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x-t) \left[\frac{D_0(t) + \cdots + D_N(t)}{N+1} \right] dt. \end{aligned}$$

If we set

$$K_N(t) = \frac{D_0(t) + \cdots + D_N(t)}{N+1}$$

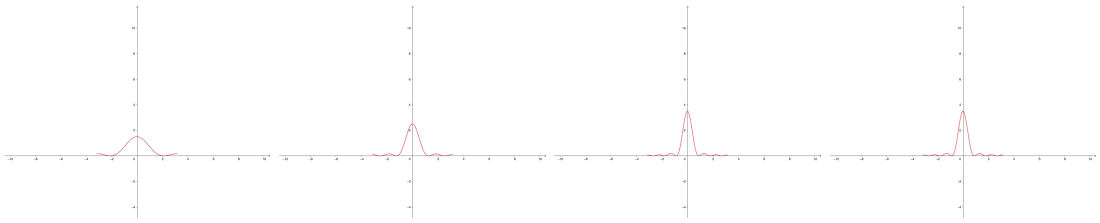
to be the averages of the Dirichlet kernels, we get that the Cesàro means of the Fourier series are given by convolution with K_N:

$$\frac{(S_0 f)(x) + \cdots + (S_N f)(x)}{N+1} = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x-t) K_N(t)\, dt =: (f \star K_N)(x).$$

The function K_N is called the *N-th order Fejér kernel*. As you showed on the last homework, the Fejér kernels *are* a good kernel, and hence, as you showed, if f is continuous we have that

$$f \star K_N \to f$$

uniformly. The upshot is that the Fourier series of a continuous functions is always uniformly Cesàro summable to that function, a result which is known as *Fejér's theorem*. Here are pictures of the graphs of some Fejér kernels (K_2, K_4, K_6, K_8), which illustrate their good kernel properties: their integral over $[-\pi, \pi]$ is always π, which since the Fejér kernels are nonnegative (hence they "smooth out" the non-good kernel properties of the Dirichlet kernels), gives a bound on the integrals of their absolute values, and they flatten out away from 0:



Lecture 19: Limits and Linearity

Warm-Up 1. We prove that polynomials are dense in $C([a,b])$ for $[a,b] \subseteq (-\pi,\pi)$. Of course, we already know this is true due to the Weierstrass approximation theorem, but the point here is to give an alternative proof based on the theory of Fourier series. The restriction $[a,b] \subseteq (-\pi,\pi)$ we use here is only needed because polynomials (more generally continuous functions) are not necessarily 2π-periodic, but this restriction can be removed using Fourier series with periods other than 2π with different sine and cosine expressions: given any $[a,b]$, use Fourier series with a period larger than $b-a$, and the argument we give here will still work. This and the next Warm-Up also demonstrate that Cesàro summability has applications in its own right regardless of whether or not the Fourier converges.

The strategy is as follows. Given $f \in C([a,b])$, Fejérs theorem guarantees that we can uniformly approximate f by the Cesàro means of its Fourier series, which are trigonometric polynomials. (So, we also avoid having to use Stone-Weierstrass.) Each of these trigonometric polynomials involves sines and cosines, but sine and cosine are both analytic with global power series expansions, so we can uniformly approximate them on all of $[a,b]$ using their Taylor polynomials. Putting all of these approximations together gives a uniform approximation of f by polynomials as desired.

To be precise, let $\epsilon > 0$ and pick N such that

$$|f - f \star K_N| < \frac{\epsilon}{2}$$

where K_N is the N-th order Fejér kernel, which is possible because $f \star K_N \to f$ uniformly by Fejér's theorem. Since $f \star K_N$ is an average of the trigonometric polynomials $S_N f$ (partial sums of the Fourier series of f), $f \star K_N$ is also a trigonometric polynomial, say

$$(f \star K_N)(x) = c_0 + \sum_{n=1}^{N} (c_n \cos(nx) + d_n \sin(nx)).$$

Since the Taylor polynomials of cosine and sine centered at 0 converge to cosine and sine on all of $\mathbb{R}$, they do so uniformly on $[a,b]$; call these Taylor polynomials

$$P_k(x) = \sum_{n=0}^{k} \frac{(-1)^k x^{2k}}{(2k)!} \text{ for } \cos x \text{ and } Q_k(x) = \sum_{n=0}^{k} \frac{(-1)^k x^{2k+1}}{(2k+1)!} \text{ for } \sin x.$$

Pick K_1 and K_2 such that

$$|\cos - P_{K_1}| < \frac{\epsilon}{4N(\max\{|c_1|,\ldots,|c_N|\}+1)} \quad \text{and} \quad |\sin - Q_{K_2}| < \frac{\epsilon}{4N(\max\{|d_1|,\ldots,|d_N|\}+1)}.$$

(The $+1$ is there just to avoid the dumb case where everything is zero.) Then

$$R(x) := c_0 + \sum_{n=1}^{N} (c_n P_{K_1}(nx) + d_n Q_{K_2}(nx))$$

is a polynomial and

$$\begin{aligned}
|f - R| &\le |f - f \star K_N| + |f \star K_n - R| \\
&< \frac{\epsilon}{2} + \left| \sum_{n=1}^{N} [c_n(\cos - P_{K_1}) + d_n(\sin - Q_{K_2})] \right| \\
&\le \frac{\epsilon}{2} + \sum_{n=1}^{N} [|c_n||\cos - P_{K_1})| + |d_n||\sin - Q_{K_2}|] \\
&< \frac{\epsilon}{2} + \sum_{n=1}^{N} \left(\frac{\epsilon}{4N} + \frac{\epsilon}{4N} \right) \\
&= \epsilon,
\end{aligned}$$

so f can be uniformly approximated by polynomials within any $\epsilon > 0$.

Warm-Up 2. We prove *Parseval's theorem*

$$\frac{a_0^2}{2} + \sum_{n=1}^{\infty} (a_n^2 + b_n^2) = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)^2 \, dx$$

for f continuous on $[-\pi, \pi]$ (with same values at $\pm\pi$) where a_n, b_n are the Fourier coefficients of f. In fact, Parseval's theorem holds more generally for f integrable, but this requires a little more work to justify, which we will leave to discussion section. (Parseval's theorem is just an infinite-dimensional analog of the Pythagorean theorem. In the form $f = \sum_{n=0}^{\infty} \langle f, \phi_n \rangle \, \phi_n$ of the Fourier series of f where we use orthonormal functions ϕ_n—so $\frac{1}{\sqrt{2\pi}}, \frac{\cos x}{\sqrt{\pi}}, \frac{\sin x}{\sqrt{\pi}}$, etc—Parseval's theorem looks like

$$\sum_{n=0}^{\infty} \langle f, \phi_n \rangle^2 = \langle f, f \rangle^2 .$$

The usual Pythagorean theorem in $\mathbb{R}^n$ says that if

$$(\mathbf{x} \cdot \mathbf{u}_1)\mathbf{u}_1 + \cdots + (\mathbf{x} \cdot \mathbf{u}_n)\mathbf{u}_n = \mathbf{x}$$

with respect to an orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_n$, then

$$(\mathbf{x} \cdot \mathbf{u}_1)^2 + \cdots + (\mathbf{x} \cdot \mathbf{u}_n)^2 = \mathbf{x} \cdot \mathbf{x} = \|\mathbf{x}\|^2 ,$$

so indeed we are looking at an infinite version of this.)

Bessel's inequality already gives

$$\frac{a_0^2}{2} + \sum_{n=1}^{\infty} (a_n^2 + b_n^2) \le \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)^2 \, dx,$$

so we need only justify the opposite inequality. From the derivation of Bessel's inequality, we have

$$\frac{1}{\pi} \int_{-\pi}^{\pi} (f - S_N f)^2 = \frac{1}{\pi} \int_{-\pi}^{\pi} f^2 - \left(\frac{a_0^2}{2} + \sum_{n=1}^{N} (a_n^2 + b_n^2) \right)$$

for each N, where $S_N f$ is the N-th order Fourier partial sum. The Cesáro mean $f \star K_N$ is a trigonometric polynomial of order N, so since the $S_N f$ minimize integrals of squares, we have

$$\frac{1}{\pi}\int_{-\pi}^{\pi} f^2 - \left(\frac{a_0^2}{2} + \sum_{n=1}^{N}(a_n^2 + b_n^2)\right) = \frac{1}{\pi}\int_{-\pi}^{\pi}(f - S_N f)^2 \leq \frac{1}{\pi}\int_{-\pi}^{\pi}(f - f \star K_N)^2.$$

The integrands on the right converge to 0 as $N \to \infty$ since $f \star K_n \to f$ uniformly, so taking limits above gives

$$\frac{1}{\pi}\int_{-\pi}^{\pi} f^2 - \left(\frac{a_0^2}{2} + \sum_{n=1}^{\infty}(a_n^2 + b_n^2)\right) \leq 0,$$

which is the desired remaining inequality, so Parseval's theorem holds. (To get Parseval for integrable f, we have to know that integrals of arbitrary integrable functions can be approximated by integrals of continuous functions. Again, we'll leave the details to discussion section.)

Componentwise limits and continuity. We now move to studying differentiability in $\mathbb{R}^n$, which is our final topic of the quarter. The notion of a limit in $\mathbb{R}^n$ plays a key role, so we begin by revisiting the main ideas. (Limits of functions between arbitrary metric spaces was covered at the start of Chapter 4 in Rudin, but it is worth being clear about what this looks like in Euclidean space.)

Given a function $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ defined on an open subset U of $\mathbb{R}^n$ (when $m > 1$ we often say that f is *vector-valued*), for $a \in U$ we say that $\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x}) = \mathbf{L} \in \mathbb{R}^m$ if for all $\epsilon > 0$ there exists $\delta > 0$ such that

$$\|f(\mathbf{x}) - \mathbf{L}\| < \epsilon \text{ when } 0 < \|\mathbf{x} - \mathbf{a}\| < \delta.$$

Here, the first $\|\cdot\|$ is the Euclidean distance in $\mathbb{R}^m$ and the second $\|\cdot\|$ is the Euclidean distance in $\mathbb{R}^n$. Note first that, from this, the "approach via different paths" technique you would have seen in a multivariable calculus course is valid: if $\gamma(t)$ is any continuous path passing through $\mathbf{a} = \gamma(t_0)$, the values of $\gamma(t)$ eventually (for t close enough to t_0) fall within the range $0 < \|\mathbf{x} - \mathbf{a}\| < \delta$, so $\|f(\gamma(t)) - \mathbf{L}\| < \epsilon$ for those values and hence the limit of f when we approach $\mathbf{a}$ only along γ is $\mathbf{L}$. As a consequence, if different choices for γ give different limits, then $\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x})$ does not exist.

The main point for our purposes is that such limits can be determined component-by-component in the sense that if we denote the components of f by $f = (f_1, \ldots, f_m)$, where each f_m is scalar-valued, then

$$\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x}) = \mathbf{L} \text{ if and only if } \lim_{\mathbf{x}\to\mathbf{a}} f_i(\mathbf{x}) = L_i$$

where the L_i are the components of $\mathbf{L} = (L_1, \ldots, L_m)$. Thus, considering limits into $\mathbb{R}^m$ reduces to considering limits into $\mathbb{R}$, which are usually simpler to get a handle on. This fact follows from the inequalities

$$|x_i - y_i| = \sqrt{(x_i - y_i)^2} \leq \sqrt{(x_1 - y_1)^2 + \cdots + (x_m - y_m)^2} \leq \sqrt{m} \cdot \max\{|x_i - y_i|\}$$

for $(x_1, \ldots, x_m), (y_1, \ldots, y_m) \in \mathbb{R}^m$, so Euclidean distance in $\mathbb{R}^m$ can be made small if and only if componentwise distances in $\mathbb{R}$ can be made small. From this we then get that continuity occurs componentwise as well, meaning that $f = (f_1, \ldots, f_m) : U \to \mathbb{R}^m$ is continuous if and only if each $f_i : U \to \mathbb{R}$ is continuous.

Here is an example. Set $f : \mathbb{R}^2 \to \mathbb{R}^2$ to be

$$f(x, y) = \begin{cases} \left(\frac{xy(x^2 - y^2)}{x^2 + y^2}, e^{x^3 + y^3}\right) & \text{if } (x, y) \neq (0, 0) \\ (0, 1) & \text{if } (x, y) = (0, 0). \end{cases}$$

We claim that

$$\lim_{(x,y)\to(0,0)} f(x,y) = (0,1) = f(0,0),$$

so that f is continuous at $(0,0)$. Since limits can be computed componentwise, this is comes down to knowing that

$$\lim_{(x,y)\to(0,0)} \frac{xy(x^2-y^2)}{x^2+y^2} = 0 \quad \text{and} \quad \lim_{(x,y)\to(0,0)} e^{x^3+y^3} = 1.$$

The second just follows from continuity of the exponential function and of x^3+y^3. For the first, we use

$$|x| = \sqrt{x^2} \leq \sqrt{x^2+y^2} \text{ and similarly } |y| \leq \sqrt{x^2+y^2}.$$

Then

$$\left|\frac{xy(x^2-y^2)}{x^2+y^2}\right| \leq \left|\frac{(x^2+y^2)(x^2-y^2)}{x^2+y^2}\right| = |x^2-y^2|,$$

which goes to 0 as $(x,y)\to(0,0)$, so $\frac{xy(x^2-y^2)}{x^2+y^2}$ does as well.

Linear transformations. Differentiability is all about approximating nonlinear things by linear ones, so before saying what differentiability means in $\mathbb{R}^n$, we must know what we mean by "linear" here. The answer comes from linear algebra: a function $T : \mathbb{R}^n \to \mathbb{R}^m$ is *linear* (or is a *linear transformation* if it preserves addition and scalar multiplication in the sense that

$$T(\mathbf{x}+\mathbf{y}) = T(\mathbf{x}) + T(\mathbf{y}) \text{ and } T(c\mathbf{x}) = cT(\mathbf{x}) \text{ for all } \mathbf{x}, y \in \mathbb{R}^n, c \in \mathbb{R}.$$

As you would have seen in a linear algebra course, being linear is equivalent to the existence of an $m \times n$ matrix A such that $T(\mathbf{x}) = A\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^n$.

From this we can easily see that linear transformations are always continuous. Indeed, if we denote the entries of A by a_{ij} and those of $\mathbf{x}$ by x_i, then $A\mathbf{x}$ looks like

$$A\mathbf{x} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{bmatrix}.$$

The components of T, being polynomials, are thus all continuous, so T is continuous. As we'll see, linear transformations play the role of "derivatives" in the higher-dimensional setting.

Operator norms. We will need to be able to control the size of expressions involving linear transformations. We might first ask whether a given linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ is bounded, but this can never be the case if $T \neq 0$ using our usual definition of "bounded": for $\mathbf{x}$ such that $T(\mathbf{x}) \neq 0$, we get

$$\|T(c\mathbf{x})\| = \|cT(\mathbf{x})\| = |c|\,\|T(\mathbf{x})\| \to \infty \text{ as } |c| \to \infty,$$

so the image of $T \neq 0$ in $\mathbb{R}^m$ is never bounded. To get a meaningful bound we must restrict the types of vectors $\mathbf{x}$ we consider, and thus we define the *operator norm* of $T(\mathbf{x}) = A\mathbf{x}$ (also called the operator norm of A) as

$$\|T\| := \sup_{\|\mathbf{x}\|=1} \|T(\mathbf{x})\|.$$

That is, we ask for a bound on T but only among things of norm 1. Since the unit sphere $\{\|\mathbf{x}\| = 1\}$ is compact in $\mathbb{R}^n$ and T is continuous, $\|T\|$ achieves a maximum among such points by the extreme value theorem, and this maximum is the supremum above, which is thus finite.

The precise value of $\|T\|$ will not be important to know, but you will show in discussion section that it is actually the square root of the largest eigenvalue of the matrix A^TA. What *is* important is the following inequality it satisfies, which is what will allow us to bound expressions involving linear transformations. If $\mathbf{x} \neq 0$, then $\frac{\mathbf{x}}{\|\mathbf{x}\|}$ has norm 1, so we get

$$\left\| T(\tfrac{\mathbf{x}}{\|\mathbf{x}\|}) \right\| \leq \|T\| .$$

On the other hand,

$$\left\| T(\tfrac{\mathbf{x}}{\|\mathbf{x}\|}) \right\| = \left\| \tfrac{1}{\|\mathbf{x}\|} T(\mathbf{x}) \right\| = \tfrac{1}{\|\mathbf{x}\|} \|T(\mathbf{x})\|$$

by lineraity, so

$$\frac{1}{\|\mathbf{x}\|} \|T(\mathbf{x})\| \leq \|T\| , \text{ and hence } \|T(\mathbf{x})\| \leq \|T\| \|\mathbf{x}\| .$$

This also holds for $\mathbf{x} = 0$ since $T(\mathbf{0}) = \mathbf{0}$, so $\|T(\mathbf{x})\| \leq \|T\| \|\mathbf{x}\|$ for all $\mathbf{x} \in \mathbb{R}^n$. This is sometimes called the *Cauchy-Schwarz inequality* for the operator norm due so its similarities to the usual Cauchy-Schwarz inequality: $|\mathbf{x} \cdot \mathbf{y}| \leq \|\mathbf{x}\| \|\mathbf{y}\|$. (In fact, the two inequalities say the same thing in the case where $m = 1$, so that $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix}$ has only one row, as a consequence of the last problem on the next homework.) The Cauchy-Schwarz inequality thus says that, geometrically, $\|T\|$ is the largest factor by which T can *scale* vectors.

Note one immediate consequence of all this. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have

$$\|T(\mathbf{x}) - T(\mathbf{y})\| = \|T(\mathbf{x} - \mathbf{y})\| \leq \|T\| \|\mathbf{x} - \mathbf{y}\| ,$$

which says that linear transformations are Lipschitz and hence uniformly continuous.

Lecture 20: Differentiability in $\mathbb{R}^n$

Warm-Up 1. We justify some basic properties of the operator norm, namely that

$$\|AB\| \leq \|A\| \|B\| , \ \|cA\| = |c| \|A\| , \text{ and } \|A + B\| \leq \|A\| + \|B\|$$

for all matrices A and B (of appropriate sizes so that the sums and products make sense) and $c \in \mathbb{R}$. For $\|\mathbf{x}\| = 1$, we have

$$\|(AB)\mathbf{x}\| \leq \|A\| \|B\mathbf{x}\| \leq \|A\| \|B\| \|\mathbf{x}\| = \|A\| \|B\| ,$$

so the maximum of the things on the left is bounded by the thing on the right, which gives the first property. Next,

$$\|cA(\mathbf{x})\| = |c| \|A\mathbf{x}\| ,$$

so the maximum of the things on the left for $\|\mathbf{x}\| = 1$, which is $\|cA\|$, equals the maximum of the things on the right, which is $|c| \|A\|$, and this is the second property. Finally, if $\|\mathbf{x}\| = 1$,

$$\|(A + B)\mathbf{x}\| = \|A\mathbf{x} + B\mathbf{x}\| \leq \|A\mathbf{x}\| + \|B\mathbf{x}\| \leq \|A\| + \|B\| ,$$

which implies $\|A + B\| \leq \|A\| + \|B\|$.

Warm-Up 2. The final property above implies that $\|A - B\|$ satisfies triangle inequality:

$$\|A - B\| = \|(A - C) + (C - B)\| \leq \|A - C\| + \|C - B\| .$$

Since the only matrix of norm 0 is the zero matrix (if the norm is 0, by scaling you can show that the matrix sends every vector to the zero vector), this shows that the space of $m \times n$ matrices is a metric space with respect to the distance $\|A - B\|$. We will work with this metric space in just a few examples dealing with differentiability later, so it will not play a major role for us, but nonetheless we now justify some topological properties of this space—or rather of some of its subsets—just to get a feel for what it looks like.

For example, the set of invertible $n \times n$ matrices is open in the metric space of all $n \times n$ matrices equipped with the operator norm. Rudin gives a proof of this, and then a proof of the fact that the map sending such a matrix to its inverse is continuous, using the operator norm directly. This requires a careful study of the operator norm, but instead we take the approach stated in the last problem on the current homework that topological questions dealing with the operator norm are equivalent to those dealing with the Euclidean norm instead. The point is that we can think if an $n \times n$ matrix as a vector in $\mathbb{R}^{n^2}$ by arranging its entries as one single long column, and we can then instead consider the usual Euclidean norm on this vector. The Euclidean norm obtained in this way is *not* the same as the operator norm (except for in very special cases, see the homework), but they are related by some inequalities that imply that "open" with repsect to the operator norm is equivalent to "open" with respect to the Euclidean norm. As a consequence, "continuous" means the same thing in both contexts as well.

A square matrix is invertible if and only if its determinant is nonzero, so the set of invertible $n \times n$ matrices—typically denoted by $GL_n(\mathbb{R})$—is the preimage of $\mathbb{R}\backslash\{0\} = (-\infty, 0) \cup (0, \infty)$ under the function

$$\det : \{n \times n \text{ matrices}\} \to \mathbb{R}$$

that sends a matrix to its determinant. This determinant function is continuous since determinants can be written as polynomial expressions in the entries of the matrix, and polynomials are always continuous. (For example, $\det \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] = ad - bc$ is a quadratic polynomial in the entries a, b, c, d, and the determinant of a 3×3 matrix will be a cubic polynomial in the entries.) The preimage of an open set under a continuous function is open, so we get that the set U of invertible matrices is open among all matrices of the same size. The function $U \to U$ that sends a matrix to its inverse is continuous because its components are continuous since the entries of A^{-1} can be written as fractions of polynomials whose the denominators are the (nonzero) determinant; for example,

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} d/(ad - bc) & -b/(ad - bc) \\ -c/(ad - bc) & a/(ad - bc) \end{bmatrix},$$

and similar expressions for larger matrices.

Here are some other examples. The set of $n \times n$ matrices of determinant 1—typically denoted by $SL_n(\mathbb{R})$—is closed (in the metric space sense) in the space of all matrices of the same size. This uses the same continuous determinant argument: this set is the preimage of $\{1\}$ under the determinant function, and preimages of closed sets are closed. The set of *orthogonal* $n \times n$ matrices—denoted by $O_n(\mathbb{R})$—which are matrices with orthonormal columns, or equivalently satisfying $Q^T Q = 1$, is, we claim, compact. Indeed, this set is bounded since the Euclidean norm of the corresponding vector in $\mathbb{R}^{n^2}$ is bounded because all columns are meant to have length 1. Moreover, this set is compact because it is the preimage of a closed set under a continuous function: the entries of $Q^T Q$ are all polynomials in the entries of Q, so the map $Q \mapsto Q^T Q$ is continuous as a map from the space of $n \times n$ matrices to itself, and the set of orthogonal matrices is the preimage of the closed set $\{I\}$. Since the set of orthogonal matrices is closed and bounded (in $\mathbb{R}^{n^2}$), it is compact, and hence still compact with respect to the operator norm. (Note, as a constrast, that the set of matrices of determinant 1 is not compact since it is not bounded: in the 2×2 case, for example, we can

ensure $ad - bc = 1$ holds for a arbitrarily large and hence unbounded as long as we in turn make d appropriately small.)

The set of orthogonal matrices is not connected (again in the metric space sense), however, since the subset of orthogonal matrices of determinant 1 and the subset of orthogonal matrices of determinant -1 (any orthogonal matrix has determinant ± 1) are each open, nonempty, and disjoint from one another. If we restrict to the set of orthogonal matrices of determinant 1—denoted $SO_n(\mathbb{R})$—however, this is now connected and compact!

Differentiability. To motivate the definition of differentiability in higher dimensions, let us start with the single-variable case, where derivatives are meant to give linear approximations. This means that something like

$$f(x) \approx f(a) + f'(a)(x - a) \text{ for } x \approx a$$

should be true. The precise way of saying this is that the difference (or "error")

$$f(x) - [f(a) + f'(a)(x - a)]$$

should go to 0 as $x \to a$ "more rapidly" than $x - a$ does in the sense that

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

Manipulating this expression yields

$$\lim_{x \to a} \frac{f(x) - [f(a) + f'(a)(x - a)]}{x - a} = 0 \iff \lim_{x \to a} \frac{f(x) - f(a)}{x - a} = f'(a),$$

so we can take the former as the definition of what it means for f to be differentiable at a, or more precisely the statement that there exists $f'(a) \in \mathbb{R}$—which we call the *derivative* of f at a—for which first limit is zero as the definition.

For a function $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$, we then expect a similar "linear approximation" property to hold if f is to be differentiable:

$$f(\mathbf{x}) \approx f(\mathbf{a}) + f'(\mathbf{a})(\mathbf{x} - \mathbf{a}) \text{ for } \mathbf{x} \approx \mathbf{a}.$$

But what type of object should the "derivative $f'(\mathbf{a})$" here be? The vectors $\mathbf{x}$ and $\mathbf{a}$ are in $U \subseteq \mathbb{R}^n$, while $f(\mathbf{x}), f(\mathbf{a}) \mathbf{i} \mathbb{R}^m$, so $f'(\mathbf{a})$ should be something that will transform vectors in $\mathbb{R}^n$ into vectors in $\mathbb{R}^m$ in a "linear" way, so that $f'(\mathbf{a})$ should actually be a linear transformation $\mathbb{R}^n \to \mathbb{R}^m$, or in other words an $m \times n$ matrix! We thus say that f is *differentiable* at $\mathbf{a} \in U$ if there exists an $m \times n$ matrix (equivalently linear transformation $B : \mathbb{R}^n \to \mathbb{R}^m$) such that

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{a} + \mathbf{h}) - [f(\mathbf{a}) + B\mathbf{h}]}{\|\mathbf{h}\|} = \mathbf{0}.$$

Here, $\mathbf{x} := \mathbf{a} + \mathbf{h}$ is a point that will be "close" to $\mathbf{a}$ as $\mathbf{h} = \mathbf{x} - \mathbf{a}$ gets small, and the numerator is the error in approximating the value $f(\mathbf{a} + \mathbf{h})$ at this nearby point by the linear expression $f(\mathbf{a}) + B\mathbf{h} = f(\mathbf{a}) + B(\mathbf{x} - \mathbf{a})$; to say that f is differentiable is to say that this error decays more quickly than does $\|\mathbf{h}\| = \|\mathbf{x} - \mathbf{a}\|$. If this is true, we call the matrix B the *derivative* of f at $\mathbf{a}$, and denote it by $f'(\mathbf{a}) = B$. (We will show next time that this B, if it exists, is unique, and can be determined explicitly from the partial derivatives of f.)

Note that this definition gives the expected result when $m = n = 1$: $B = (b)$ is then a 1×1 matrix, and this new definition of differentiable becomes

$$\lim_{x \to a} \frac{f(x) - [f(a) + b(x-a)]}{x - a} = 0 \iff \lim_{x \to a} \frac{f(x) - f(a)}{x - a} = b,$$

so we recover the previous definition in the single-variable case. In the single-variable case the word "matrix" never appears because you only ever deal with 1×1 matrices anyway, but they're there!

Example. Define $f : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$f(x, y) = (x^2 + y^2, xy + y).$$

We show that f is differentiable at $(0, 1)$ with derivative $f'(0, 1) = \left[\begin{smallmatrix} 0 & 2 \\ 1 & 1 \end{smallmatrix}\right]$. Set $\mathbf{a} = (0, 1)$, $\mathbf{h} = (h, k)$, and $\mathbf{x} = (x, y)$, so that

$$\begin{aligned} f(\mathbf{a} + \mathbf{h}) - f(\mathbf{a}) - f'(\mathbf{a})\mathbf{h} &= f(h, k+1) - f(0, 1) - \left[\begin{smallmatrix} 0 & 2 \\ 1 & 1 \end{smallmatrix}\right] \left[\begin{smallmatrix} h \\ k \end{smallmatrix}\right] \\ &= (h^2 + [k+1]^2, h(k+1) + [k+1]) - (1, 1) - (2k, h + k) \\ &= (h^2 + k^2 + 2k + 1, hk + k + 1) - (1, 1) - (2k, h + k) \\ &= (h^2 + k^2, hk). \end{aligned}$$

(Technically we should not use the notation $f'(\mathbf{a})$ above until we know that this derivative exists, and we should instead use $B = \left[\begin{smallmatrix} 0 & 2 \\ 1 & 1 \end{smallmatrix}\right]$, but whatever.)

Thus in order to establish differentiability, we must know that

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{a} + \mathbf{h}) - f(\mathbf{a}) - f'(\mathbf{a})\mathbf{h}}{\|\mathbf{h}\|} = \lim_{(h,k) \to (0,0)} \frac{(h^2 + k^2, hk)}{\sqrt{h^2 + k^2}} = (0, 0).$$

But this can be determined componentwise, so we must know that

$$\lim_{(h,k) \to (0,0)} \frac{h^2 + k^2}{\sqrt{h^2 + k^2}} = 0 \quad \text{and} \quad \lim_{(h,k) \to (0,0)} \frac{hk}{\sqrt{h^2 + k^2}} = 0.$$

The first follows from $\frac{h^2+k^2}{\sqrt{h^2+k^2}} = \sqrt{h^2 + k^2}$, and the second from the bound $|h||k| \leq \sqrt{h^2 + k^2}\sqrt{h^2 + k^2}$, so f is differentiable at $(0, 1)$ as claimed with derivative $f'(0, 1) = \left[\begin{smallmatrix} 0 & 2 \\ 1 & 1 \end{smallmatrix}\right]$. (Why was this the right matrix to consider in the first place? We'll see next time.)

Another example. Linear transformations $T(\mathbf{x}) = A\mathbf{x}$ are always differentiable with constant derivative A. Indeed, the numerator in the limit defining differentiability is

$$T(\mathbf{x} + \mathbf{h}) - T(\mathbf{x}) - A\mathbf{h} = T(\mathbf{x}) + T(\mathbf{h}) - T(\mathbf{x}) - A\mathbf{h} = \mathbf{0}$$

by linearity, so

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{T(\mathbf{x} + \mathbf{h}) - T(\mathbf{x}) - A\mathbf{h}}{\|\mathbf{h}\|} = \lim_{\mathbf{h} \to \mathbf{0}} \frac{\mathbf{0}}{\|\mathbf{h}\|} = \mathbf{0}.$$

Hence T is differentiable at any $\mathbf{x} \in \mathbb{R}^n$ and $T'(\mathbf{x}) = A$. (This is just the higher-dimensional analog of the fact that $f(x) = ax$ is differentiable and $f'(x) = a$ is constant. The same is true for something like $T(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, so linear plus a constant—what are typically referred to as *affine* transformations—which is analogous to what happens for $f(x) = ax + b$.)

Differentiable implies continuous. We will build up more on differentiability next time, but as a start we should know that differentiable implies continuous, just as in the single-variable case. Suppose $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\mathbf{x} \in U$ with derivative $f'(\mathbf{x})$, an $m \times n$ matrix. Then

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

so in particular there exists $\delta > 0$ such that

$$\left\| \frac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - f'(\mathbf{x})\mathbf{h}}{\|\mathbf{h}\|} \right\| < 1 \text{ when } \|\mathbf{h}\| < \delta.$$

This gives

$$\left\| f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - f'(\mathbf{x})\mathbf{h} \right\| < \|\mathbf{h}\| \text{ for small } \|\mathbf{h}\|,$$

and thus

$$\|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x})\| \leq \|\mathbf{h}\| + \left\| f'(\mathbf{x})\mathbf{h} \right\| \text{ for small } \|\mathbf{h}\|$$

after a reverse triangle inequality. Using Cauchy-Schwarz on $\|f'(\mathbf{x})\mathbf{h}\| \leq \|f'(\mathbf{x})\| \, \|\mathbf{h}\|$, where $\|f'(\mathbf{x})\|$ is an operator norm, we get

$$\|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x})\| \leq \|\mathbf{h}\| + \left\| f'(\mathbf{x}) \right\| \|\mathbf{h}\| = (1 + \left\| f'(\mathbf{x}) \right\|) \|\mathbf{h}\|$$

for small $\|\mathbf{h}\|$. The right side now goes to 0 as $\mathbf{h} \to \mathbf{0}$ (note that $1 + \|f'(\mathbf{x})\|$ is a fixed constant), so the left side does as well, which means that

$$\lim_{\mathbf{h} \to \mathbf{0}} f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}),$$

so that f is continuous at $\mathbf{x}$ as claimed.

Lecture 21: Jacobian Matrices

Warm-Up 1. We show that if f is differentiable at $\mathbf{x}$, the derivative matrix $f'(\mathbf{x})$ is unique in the sense that there can only be one choice which satisfies the definition of differentiability. Now, shortly we will derive an explicit description of what the entries in $f'(\mathbf{x})$ must look anyway, which also guarantees uniqueness, but the goal here is to give an standalone argument which does not depend on knowing what $f'(\mathbf{x})$ is.

Suppose B, B' are two matrices that satisfy the differentiability definition, so

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - B\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0} \quad \text{and} \quad \lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - B'\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0}.$$

Subtracting gives

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{B\mathbf{h} - B'\mathbf{h}}{\|\mathbf{h}\|} - \mathbf{0}$$

since everything else cancels out. Rewriting and using linearity of $B - B'$ gives

$$\lim_{\mathbf{h} \to \mathbf{0}} (B - B') \tfrac{\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0}.$$

This means for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\left\| (B - B') \tfrac{\mathbf{h}}{\|\mathbf{h}\|} \right\| < \epsilon \text{ for } 0 < \|\mathbf{h}\| < \delta.$$

Now, take $\|\mathbf{x}\| = 1$. Then $\mathbf{h} = \frac{\delta}{2}\mathbf{x}$ satisfies $0 < \|\mathbf{h}\| < \delta$, so we get

$$\left\|(B - B')\mathbf{x}\right\| = \left\|(B - B')\tfrac{\mathbf{h}}{\|\mathbf{h}\|}\right\| < \epsilon.$$

(If $\mathbf{h} = \frac{\delta}{2}\mathbf{x}$, dividing $\mathbf{h}$ by its norm just recovers $\mathbf{x}$ since $\|\mathbf{x}\| = 1$.) The operator norm of $B - B'$ is the maximum of the things on the left, so this gives

$$\left\|B - B'\right\| < \epsilon.$$

Since this holds for all $\epsilon > 0$, $\|B - B'\| = 0$, so $B - B' = 0$ and hence the derivative $f'(\mathbf{x})$ is unique.

Warm-Up 2. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x, y) = \begin{cases} \frac{-2x^3+3y^4}{x^2+y^2} & (x, y) \neq (0, 0) \\ 0 & (x, y) = (0, 0). \end{cases}$$

We claim that the 1×2 matrix $B = \begin{bmatrix} -2 & 0 \end{bmatrix}$ does not satisfy the definition of differentiability for f at $(0, 0)$. Now, at this point this is not enough to show that f is not differentiable at $(0, 0)$ since it does not rule out the possibility that some *other* 1×2 matrix could satisfy the definition, but in fact we will see afterwards that if f were going to be differentiable at $(0, 0)$, this specific matrix is the only one that could work. So, this Warm-Up actually is, in the end, showing that f is not differentiable at $(0, 0)$

For $\mathbf{h} = (h, k) \neq \mathbf{0}$, we have

$$\begin{aligned} f(\mathbf{0} + \mathbf{h}) - f(\mathbf{0}) - B\mathbf{h} &= f(h, k) - f(0, 0) - \begin{bmatrix} -2 & 0 \end{bmatrix}\begin{bmatrix} h \\ k \end{bmatrix} \\ &= \frac{-2h^3 + 3k^4}{h^2 + k^2} - 0 - (-2h + 0) \\ &= \frac{-2h^3 + 3k^4}{h^2 + k^2} + 2h \\ &= \frac{3k^4 + 2hk^2}{h^2 + k^2}. \end{aligned}$$

Thus, the limit in the definition of differentiability looks like

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{0} + \mathbf{h}) - f(\mathbf{0}) - B\mathbf{h}}{\|\mathbf{h}\|} = \lim_{(h,k)\to(0,0)} \frac{\frac{3k^4+2hk^2}{h^2+k^2}}{\sqrt{h^2 + k^2}} = \lim_{(h,k)\to(0,0)} \frac{3k^4 + 2hk^2}{(h^2 + k^2)\sqrt{h^2 + k^2}}.$$

If this limit were going to exist and equal $\mathbf{0}$, we would get a limit value of $\mathbf{0}$ along any curve we choose to approach $(0, 0)$ along. But along $h = k$, this limit restricts to

$$\lim_{h\to 0} \frac{3h^4 + 2h^3}{2h^2\sqrt{2h^2}} = \lim_{h\to 0} \left(\frac{3h^2}{2\sqrt{2}|h|} + \frac{2h}{\sqrt{2}|h|} \right).$$

The first term on the right does approach 0 as $h \to 0$, but the limit of the second term does not exist since $\frac{h}{|h|} = \pm 1$ depending on whether h is positive or negative. Hence the limit above does not exist, and it if does not exist it certainly does not equal 0, so

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{0} + \mathbf{h}) - f(\mathbf{0}) - B\mathbf{h}}{\|\mathbf{h}\|} \neq \mathbf{0}$$

and hence the differentiability definition fails with candidate matrix $B = \begin{bmatrix} -2 & 0 \end{bmatrix}$.

Deriving the derivative. If f is differentiable at $\mathbf{x}$, we now derive the entries of the (unique) matrix B satisfying

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - B\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0}.$$

First, since this limit exists, we get the same limit value for the restriction of f to any path passing through $\mathbf{0}$, so let us take specifically the path formed by the points $\mathbf{h} = t\mathbf{e}_j$ as t varies with $j = 1, \ldots, n$ fixed, where $\mathbf{e}_j$ is the vector which has 1 in the j-th entry and 0 elsewhere. (A "standard basis vector" in the language of linear algebra.) Along this path, our limit becomes

$$\lim_{t\to 0} \frac{f(x+t\mathbf{e}_j) - f(\mathbf{x}) - B(t\mathbf{e}_j)}{|t|} = \mathbf{0}$$

where we use $\|t\mathbf{e}_j\| = |t|\,\|\mathbf{e}_j\| = |t|$ in the denominator.

Now, $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ here is vector-valued (scalar-valued when $m = 1$), so let us denote its components by $f = (f_1, \ldots, f_m)$. Limits can be determined componentwise, so we get that

$$\lim_{t\to 0} \frac{f_i(\mathbf{x}+t\mathbf{e}_j) - f_o(\mathbf{x}) - (B\mathbf{e}_j)_i t}{t} = \mathbf{0}$$

for each $i = 1, \ldots, m$ where $(B\mathbf{e}_j)_i$ denotes the i-th component of the vector $B\mathbf{e}_j$, and where we use linearity to say that $B(t\mathbf{e}_j) = (B\mathbf{e}_j)t$. (Also, we drop the absolute value on t because the absolute value does not affect whether a limit does or doesn't equal 0.) This is now a single-variable (in terms of t) scalar-valued expression, and so can be manipulated algebraically to get

$$\lim_{t\to 0} \frac{f_i(\mathbf{x}+t\mathbf{e}_j) - f_i(\mathbf{x})}{t} = (B\mathbf{e}_j)_i.$$

So, we get two conclusions here: not only does the limit

$$\lim_{t\to 0} \frac{f_i(\mathbf{x}+t\mathbf{e}_j) - f_i(\mathbf{x})}{t}$$

exist for each i, j, but its value is given by $(B\mathbf{e}_j)_i$. This limit defines what's called the *partial derivative* $\frac{\partial f_i}{\partial x_j}(\mathbf{x})$ of f_i with respect to x_j at $\mathbf{x}$, where the point is that it is the single-derivative of the function we get by varying only x_j in $f_i(\mathbf{x})$ and keeping the other coordinates fixed. (We vary only x_j because $\mathbf{x} + t\mathbf{e}_j$ and $\mathbf{x}$ agree in the other coordinates.) Thus we have

$$\frac{\partial f_i}{\partial x_j}(\mathbf{x}) = (B\mathbf{e}_j)_i.$$

But the product $B\mathbf{e}_j$ is precisely the j-th column of B, and hence

$$\frac{\partial f_i}{\partial x_j}(\mathbf{x}) = (B\mathbf{e}_j)_i = \text{ the entry in the } i\text{-th row and } j\text{-th column of } B.$$

Thus, to summarize, if f is differentiable at $\mathbf{x}$, then all partial derivatives of all components of f exist at $\mathbf{x}$, *and* $f'(\mathbf{x})$ is the matrix which has these partial derivatives as its entries.

Jacobian matrices. We define the *Jacobian matrix* of f at $\mathbf{x}$ to be this matrix of partial derivatives (assuming they exist), arranged as

$$Df(\mathbf{x}) := \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\mathbf{x}) \end{bmatrix}.$$

The i-th row contains all partial derivatives of the i-th component f_i of f, and the j-th column contains the partial derivatives of all components with respect to x_j. (So, change the variable as we move along the row, and change the component as we move down a column. Note that the resulting matrix then does have size $m \times n$.) With this we can thus rephrase the definition of differentiability at $\mathbf{x}$ as saying that the Jaocbian matrix $Df(\mathbf{x})$ exists (in other words all partial derivatives exist at $\mathbf{x}$), and satisfies

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{x}+\mathbf{h}) - [f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h}]}{\|\mathbf{h}\|} = \mathbf{0}.$$

(Note that Rudin just continues to use the notation $f'(\mathbf{x})$ for this derivative, but we prefer to reserve the prime notation for single-variable derivatives and use $Df(\mathbf{x})$ for the derivative matrix in order to emphasize that it is an entire matrix, not just a number.)

In the function in the second Warm-Up, $\frac{\partial f}{\partial x}(0,0)$ is the derivative of the single-variable function obtained by varying x and fixing $y = 0$ at $x = 0$, so this is the derivative of

$$f(x,0) = \frac{-2x^3 + 0}{x^2 + 0} = -2x$$

at $x = 0$, which is indeed -2. Similarly, $\frac{\partial f}{\partial y}(0,0)$ is the derivative of

$$f(0,y) = \frac{0 + 3y^4}{0 + y^2} = 3y^2$$

at $y = 0$, which is 0. Hence $Df(0,0)$ exists and equals $Df(0,0) = \begin{bmatrix} -2 & 0 \end{bmatrix}$, which is precisely the matrix we showed in the Warm-Up does not satisfy the definition of differentiability. Thus, as claimed there, f is not differentiable at $(0,0)$ since this is the matrix that would have to work if it was going to be differentiable. (This example shows also that existence of partial derivatives does not guarantee differentiability!)

Example. Set

$$f(x,y) = \begin{cases} \frac{x^3+y^3}{\sqrt{x^2+y^2}} & (x,y) \neq (0,0) \\ 0 & (x,y) = (0,0). \end{cases}$$

We show that f is differentiable at $(0,0)$. The partial derivatives at $(0,0)$ are

$$\frac{\partial f}{\partial x}(0,0) = \frac{d}{dx}\bigg|_{x=0} f(x,0) = \frac{d}{dx}\bigg|_{x=0} \frac{x^3}{\sqrt{x^2}} = \frac{d}{dx}\bigg|_{x=0} x|x| = 0$$
$$\frac{\partial f}{\partial y}(0,0) = \frac{d}{dy}\bigg|_{y=0} f(0,y) = \frac{d}{dy}\bigg|_{y=0} \frac{y^3}{\sqrt{y^2}} = \frac{d}{dy}\bigg|_{y=0} y|y| = 0.$$

Thus $Df(0,0) = \begin{bmatrix} 0 & 0 \end{bmatrix}$, so

$$\frac{f(\mathbf{0}+\mathbf{h}) - f(\mathbf{0}) - Df(\mathbf{0})\mathbf{h}}{\|\mathbf{h}\|} = \frac{f(h,k) - f(0,0) - \begin{bmatrix} 0 & 0 \end{bmatrix}\begin{bmatrix} h \\ k \end{bmatrix}}{\sqrt{h^2+k^2}}$$

$$= \frac{\frac{h^3+k^3}{\sqrt{h^2+k^2}} - 0 - 0}{\sqrt{h^2+k^2}}$$
$$= \frac{h^3+k^3}{h^2+k^2}.$$

Using $|h| \leq \sqrt{h^2+k^2}$ and $|k| \leq \sqrt{h^2+k^2}$, we can show that the limit of the expression above as $(h,k) \to (0,0)$ is 0 as a consequence of the squeeze theorem, so f is indeed differentiable at $(0,0)$.

Continuous differentiability. We can next try to check differentiability of the function above at points other than the origin. The partial derivatives at such points are straightforward to compute using ordinary quotient and change rules, so Jacobian matrices can be found explicitly. The problem, however, is that checking the limit definition of differentiability at non-origin points is anything but simple, since the expressesions involved will be fairly messy.

But, there is another approach to differentiability here that avoids having to check the definition directly, based on the fact that the partial derivatives, once you compute them, can be seen to be continuous at non-origin points. We say that a function $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is C^1, or continuously differentiable, if all partial derivatives of all components of f exist and are continuous at all points of U. Note that this is a statement about the behaviour of partial derivatives, which a priori do not guarantee anything about differentiability of f, but in fact the key result here is that C^1 *does* imply differentiability. So, existence of partial derivatives alone does not lead to f being differentiable, but continuity of partial derivatives does. Note that this only works in the

$$C^1 \implies \text{differentiable}$$

direction, as differentiability in general only guarantees that partial derivatives exist but not that they are continuous. (We'll give an example of this next time.)

We will give the proof of this result next time, as it will lead us down a bit of a rabbit hole into the mean value theorem and then the chain rule in $\mathbb{R}^n$. For now, let us give perhaps the "proper" way of thinking about the C^1 condition. If $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ has partial derivatives throughout U, so that the Jacobian matrix $Df(\mathbf{x})$ exists at all $x \in U$, we can consider the map Df that sends $\mathbf{x} \in U$ to the $m \times n$ matrix $Df(\mathbf{x})$; if we denote the space of all $m \times n$ matrices as $L(\mathbb{R}^n, \mathbb{R}^m)$ (here L stands for "linear transformation"), then this is a map

$$Df : U \subseteq \mathbb{R}^n \to L(\mathbb{R}^n, \mathbb{R}^m).$$

If we equip $L(\mathbb{R}^n, \mathbb{R}^m)$ with the metric induced by the operator norm, we can then ask whether Df is continuous. But as we've stated before, based on a problem on the homework this is the same as asking whether the same map but viewed as

$$Df : U \subseteq \mathbb{R}^n \to L(\mathbb{R}^n, \mathbb{R}^m) = \mathbb{R}^{mn},$$

where we identify matrices with long vectors, is continuous when we consider the usual Euclidean metric on the right. This is true if and only each component of Df is continuous, but the components of Df are just the partial derivatives of the components of f, and so we get that the C^1 condition just means the same as saying that Df is continuous. Thus, C^1 does mean that f has a "continuous derivative" when we interpret "derivative" and "continuous" correctly. This perspective will be useful since it says, in particular, that we can make things like

$$\|Df(\mathbf{x}) - Df(\mathbf{y})\| \quad \text{(this is an operator norm)}$$

small if $\|\mathbf{x} - \mathbf{y}\|$ is small, which will play a role in various results we will see later.

Lecture 22: Mean Value Theorem

Warm-Up. Suppose $f, g : \mathbb{R}^n \to \mathbb{R}^m$ are differentiable at $\mathbf{x} \in \mathbb{R}^n$. Define the "dot product" function $f \cdot g : \mathbb{R}^n \to \mathbb{R}$ by

$$(f \cdot g)(\mathbf{x}) := f(\mathbf{x}) \cdot g(\mathbf{x}).$$

(Note that $f(\mathbf{x})$ and $g(\mathbf{x})$ are each vectors in $\mathbb{R}^m$, so that taking their dot product makes sense and gives a scalar value as a result.) We prove the higher-dimensional *product rule*, which says that $f \cdot g$ is differentiable at $\mathbf{x}$ and has Jacobian derivative given by

$$D(f \cdot g)(\mathbf{x}) = g(\mathbf{x})^T Df(\mathbf{x}) + f(\mathbf{x})^T Dg(\mathbf{x}).$$

Here, $g(\mathbf{x})^T$ and $f(\mathbf{x})^T$ are $1 \times n$ row vectors, and $Df(\mathbf{x})$ and $Dg(\mathbf{x})$ are $m \times n$ matrices so the right side above is defined and results in a $1 \times n$ matrix, which is precisely the type of object $D(f \cdot g)(\mathbf{x})$ should be. We will derive this expression for $D(f \cdot g)(\mathbf{x})$ in the course of proving this product rule, but note that it makes sense as an analog of the usual product rule since it does look like "g times derivative of f plus f times derivative of g". Indeed, if we compute partial derivatives of $f(\mathbf{x}) \cdot g(\mathbf{x})$ using the usual product rule, we get some terms involving partial derivatives of the components of f times the components of g, and other terms involving partial derivatives of the components of g times the components of f, so $D(f \cdot g)(\mathbf{x})$ *should* consist of such terms.

We approach this using the notion of linear *errors*, which is also what we will use in proving the chain rule later. Take a linear expansion of f at $\mathbf{x}$ with error ϵ:

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h} + \epsilon(\mathbf{h}).$$

The error $\epsilon(\mathbf{h})$ is *defined* to be the difference of the thing on the left with the first two things on the right, so it is the error/remainder obtained when approximating $f(\mathbf{x} + \mathbf{h})$ with the linear expression $f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h}$. In other words, $\epsilon(\mathbf{h})$ is the numerator that appears in the limit definition of differentiability, so to say that f is differentiable at $\mathbf{x}$ means precisely that

$$\frac{\epsilon(\mathbf{h})}{\|\mathbf{h}\|} \to \mathbf{0} \quad \text{as} \quad \mathbf{h} \to \mathbf{0}.$$

Similarly, we expand g linearly as

$$g(\mathbf{x} + \mathbf{h}) = g(\mathbf{x}) + Dg(\mathbf{x})\mathbf{h} + \delta(\mathbf{h})$$

with error δ which satisfies $\frac{\delta(\mathbf{h})}{\|\mathbf{h}\|} \to \mathbf{0}$ since g is differentiable at $\mathbf{x}$.

To determine differentiability of $f \cdot g$ we thus consider a linear expansion of $f \cdot g$. First we use the expansion for f to write

$$\begin{aligned}(f \cdot g)(\mathbf{x} + \mathbf{h}) = f(\mathbf{x} + \mathbf{h}) \cdot g(\mathbf{x} + \mathbf{h}) &= [f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h} + \epsilon(\mathbf{h}) \cdot g(\mathbf{x} + \mathbf{h}) \\ &= [f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h}] \cdot g(\mathbf{x} + \mathbf{h}) + \epsilon(\mathbf{h}) \cdot g(\mathbf{x} + \mathbf{h}),\end{aligned}$$

were in the last step we use the distributive property $(\mathbf{a} + \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot \mathbf{c} + \mathbf{b} \cdot \mathbf{c}$ of dot products. For the first term in what remains we use the expansion for g to get

$$\begin{aligned}f(\mathbf{x} + \mathbf{h}) \cdot g(\mathbf{x} + \mathbf{h}) &= [f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h}] \cdot g(\mathbf{x} + \mathbf{h}) + \epsilon(\mathbf{h}) \cdot g(\mathbf{x} + \mathbf{h}) \\ &= [f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h}] \cdot [g(\mathbf{x}) + Dg(\mathbf{x})\mathbf{h} + \delta(\mathbf{h})] + \epsilon(\mathbf{h}) \cdot g(\mathbf{x} + \mathbf{h}) \\ &= f(\mathbf{x}) \cdot g(\mathbf{x}) + [Df(\mathbf{x})\mathbf{h} \cdot g(\mathbf{x}) + f(\mathbf{x}) \cdot Dg(\mathbf{x})\mathbf{h}] + \text{higher-order stuff}\end{aligned}$$

where the higher-order stuff is

$$[f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h}] \cdot \delta(\mathbf{h}) + Df(\mathbf{x})\mathbf{h} \cdot Dg(\mathbf{x})\mathbf{h} + \epsilon(\mathbf{h}) \cdot g(\mathbf{x} + \mathbf{h}).$$

The term $Df(\mathbf{x})\mathbf{h} \cdot g(\mathbf{x}) + f(\mathbf{x}) \cdot Dg(\mathbf{x})\mathbf{h}$ before "higher-order stuff" above comes from taking those things which are *linear* in $\mathbf{h}$, which essentially means the terms where $\mathbf{h}$ appears only once in a product; the "higher-order stuff' thus consists of all the "non-linear" things in $\mathbf{h}$. In general, the linear terms in an expansion like this are the ones that should describe the value of the derivative, as is true here since we can write these linear terms as

$$\begin{aligned} Df(\mathbf{x})\mathbf{h} \cdot g(\mathbf{x}) + f(\mathbf{x}) \cdot Dg(\mathbf{x})\mathbf{h} &= g(\mathbf{x}) \cdot Df(\mathbf{x})\mathbf{h} + f(\mathbf{x}) \cdot Dg(\mathbf{x})\mathbf{h} \\ &= g(\mathbf{x})^T Df(\mathbf{x})\mathbf{h} + f(\mathbf{x})^T Dg(\mathbf{x})\mathbf{h} \\ &= [g(\mathbf{x})^T Df(\mathbf{x}) + f(\mathbf{x})^T Dg(\mathbf{x})]\mathbf{h} \end{aligned}$$

where we use $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$ in the first step and $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T\mathbf{b}$ in the second. Thus, we see that linear term in this expansion indeed uses the $1 \times n$ matrix $g(\mathbf{x})^T Df(\mathbf{x}) + f(\mathbf{x})^T Dg(\mathbf{x})$ we are claiming is the correct derivative $D(f \cdot g)(\mathbf{x})$ for $f \cdot g$. (The point is that if we did not have a guess ahead of time for what this derivative should be, we can derive it by considering the linear terms in such linear expansions with error.) Thus, we this as what we claim is the correct value of the derivative, we have that

$$(f \cdot g)(\mathbf{x} + \mathbf{h}) - (f \cdot g)(\mathbf{x}) - (\text{what we claim is the derivative})\mathbf{h}$$

equals the "higher-order stuff" above, which means that this higher-order stuff is just the linear error for the function $f \cdot g$:

$$\text{error for } f \cdot g = [f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h}] \cdot \delta(\mathbf{h}) + Df(\mathbf{x})\mathbf{h} \cdot Dg(\mathbf{x})\mathbf{h} + \epsilon(\mathbf{h}) \cdot g(\mathbf{x} + \mathbf{h}).$$

To say that $f \cdot g$ is differentiable at $\mathbf{x}$ with derivative $D(f \cdot g)(\mathbf{x}) = g(\mathbf{x})^T Df(\mathbf{x}) + f(\mathbf{x})^T Dg(\mathbf{x})$ is then to say that this linear error for $f \cdot g$ goes to $\mathbf{0}$ when divided by $\|\mathbf{h}\|$:

$$\frac{[f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h}] \cdot \delta(\mathbf{h}) + Df(\mathbf{x})\mathbf{h} \cdot Dg(\mathbf{x})\mathbf{h} + \epsilon(\mathbf{h}) \cdot g(\mathbf{x} + \mathbf{h})}{\|\mathbf{h}\|} \to \mathbf{0} \quad \text{as} \quad \mathbf{h} \to \mathbf{0}.$$

Proving *this* is what our argument thus comes down to! Everything up to this point was just about performing algebraic manipulations in order to obtain an expression for the linear error of $f \cdot g$ at $\mathbf{x}$ with which we can now do some work.

Let us consider each piece of this linear error divided by norm separately:

$$\frac{[f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h}] \cdot \delta(\mathbf{h})}{\|\mathbf{h}\|}, \quad \frac{Df(\mathbf{x})\mathbf{h} \cdot Dg(\mathbf{x})\mathbf{h}}{\|\mathbf{h}\|}, \quad \frac{\epsilon(\mathbf{h}) \cdot g(\mathbf{x} + \mathbf{h})}{\|\mathbf{h}\|}.$$

We show that each of these goes to $\mathbf{0}$ as $\mathbf{h} \to \mathbf{0}$ one at a time, which will complete our proof of the product rule. For the third term, we use the Cauchy-Schwarz inequality (for vectors) to bound as

$$\frac{|\epsilon(\mathbf{h}) \cdot g(\mathbf{x} + \mathbf{h})|}{\|\mathbf{h}\|} \leq \frac{\|\epsilon(\mathbf{h})\|}{\|\mathbf{h}\|} \|g(\mathbf{x} + \mathbf{h})\|.$$

Since f is differentiable at $\mathbf{x}$, $\frac{\epsilon(\mathbf{h})}{\|\mathbf{h}\|} \to \mathbf{0}$, so the product above will also go to $\mathbf{0}$ as long as the $\|g(\mathbf{x} + \mathbf{h})\|$ remains bounded, but since differentiability (of g at $\mathbf{x}$) implies continuity we have

$$g(\mathbf{x} + \mathbf{h}) \to g(\mathbf{x}) \quad \text{as} \quad \mathbf{h} \to \mathbf{0}$$

and hence

$$\frac{\|\epsilon(\mathbf{h})\|}{\|\mathbf{h}\|}\|g(\mathbf{x}+\mathbf{h})\| \to 0\,\|g(\mathbf{x})\| = 0$$

so this term works out. For the first term in the linear error divided by norm, we have

$$\frac{|[f(\mathbf{x})+Df(\mathbf{x})\mathbf{h}]\cdot\delta(\mathbf{h})|}{\|\mathbf{h}\|} \leq \|f(\mathbf{x})+Df(\mathbf{x})\mathbf{h}\|\,\frac{\|\delta(\mathbf{h})\|}{\|\mathbf{h}\|}.$$

Linear/matrix transformations are continuous, so the first factor on the right goes to

$$\|f(\mathbf{x})+Df(\mathbf{x})\mathbf{0}\| = \|f(\mathbf{x})\|,$$

and hence the entire product goes to 0 since $\frac{\delta(\mathbf{h})}{\|\mathbf{h}\|} \to \mathbf{0}$ because g is differentiable at $\mathbf{x}$. This takes care of the first term in the linear error.

Finally, using Cauchy-Schwarz for vectors and for matrices, for the "quadratic factor" in the linear error we have

$$\frac{|Df(\mathbf{x})\mathbf{h}\cdot Dg(\mathbf{x})\mathbf{h}|}{\|\mathbf{h}\|} \leq \frac{\|Df(\mathbf{x})\mathbf{h}\|\,\|Dg(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|} \leq \frac{\|Df(\mathbf{x})\|\,\|\mathbf{h}\|\,\|Dg(\mathbf{x})\|\,\|\mathbf{h}\|}{\|\mathbf{h}\|} = \|Df(\mathbf{x})\|\,\|Dg(\mathbf{x})\|\,\|\mathbf{h}\|,$$

which goes to 0 as $\mathbf{h} \to \mathbf{0}$. Thus we do have

$$\frac{\text{linear error for } f\cdot g}{\|\mathbf{h}\|} = \frac{[f(\mathbf{x})+Df(\mathbf{x})\mathbf{h}]\cdot\delta(\mathbf{h}) + Df(\mathbf{x})\mathbf{h}\cdot Dg(\mathbf{x})\mathbf{h} + \epsilon(\mathbf{h})\cdot g(\mathbf{x}+\mathbf{h})}{\|\mathbf{h}\|} \to \mathbf{0}$$

as $\mathbf{h} \to \mathbf{0}$, so $f\cdot g$ is differentiable at $\mathbf{x}$ with derivative as we claimed above.

(This was an elaborate argument! But it is indicative of many types of arguments in this subject where the goal is to control linear errors. Indeed, we will see the same idea appear in the proof of the chain rule, and you will use it in proving an analog of the quotient rule on the homework.)

C^1 implies differentiable. We now prove that if $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is C^1, then f is differentiable. Since f is C^1, the Jacobian matrix $Df(\mathbf{x})$ at $\mathbf{x} \in U$ exists and has continuous entries. We must show that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0}.$$

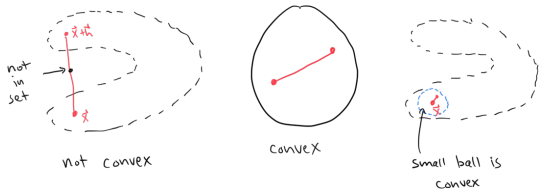
It is enough to consider each component of f one at a time, so we show that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f_i(\mathbf{x}+\mathbf{h}) - f_i(\mathbf{x}) - Df_i(\mathbf{x})\mathbf{h}}{\|\mathbf{h}\|}.$$

What we need is a way to relate $f_i(\mathbf{x}+\mathbf{h}) - f_i(\mathbf{x})$ to $Df_i(\mathbf{x})\mathbf{h}$, so this requires some type of mean value theorem application. In fact, the mean value theorem in this case (assuming differentiability) works just as it did in the single-variable case: there exists $\mathbf{c}$ between $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$ such that

$$f_i(\mathbf{x}+\mathbf{h}) - f_i(\mathbf{x}) = Df_i(\mathbf{c})\mathbf{h}.$$

Some remarks are in order. First, it is important that f_i is scalar-valued here (i.e., maps into $\mathbb{R}$) instead of vector-valued, as the mean value theorem does *not* work in this same way for vector-valued functions, as we will soon see. This is why we consider each component of f one at a time. Second, what does saying $\mathbf{c}$ is "between" $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$ mean if all vectors here are in $\mathbb{R}^n$? The answer is that by "between" here we mean that $\mathbf{c}$ is on the *line segment* between $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$, or in other words that $\mathbf{c} = \mathbf{x} + c\mathbf{h}$ for some $c \in [0,1]$. But this brings up the issue that, for an arbitrary open set U, it is not true that all points on the line segment between $\mathbf{x}, \mathbf{x}+\mathbf{h} \in U$ themselves belong to U:



This is true for *convex* sets (convexity is precisely the property that the set fully contains line segments between points in that set), but a general U might not be convex. However, this is easy to deal with: since U is open, we can find a ball around $\mathbf{x} \in U$ that remains in U, and open balls *are* always convex. When taking the limit $\mathbf{h} \to \mathbf{0}$, it is enough to consider points $\mathbf{x}+\mathbf{h}$ that belong this open ball anyway, so the line segment between $\mathbf{x}$ and such $\mathbf{x}+\mathbf{h}$ is guaranteed to be in U.

More importantly, this version of the mean value theorem depends on knowing that f is already differentiable, so it cannot actually be used for this specific C^1 result since the goal is to prove differentiability itself! So, for this specific result we have to be more clever about how we make use of "mean value thinking". Instead, we apply the single-variable mean value theorem "coordinate by coordinate": if we fix all coordinates in $f_i(x_1, \dots, x_n)$ except for one, we have

$$f_i(\dots, x_j + h_j, \dots) - f_i(\dots, x_j, \dots) = \frac{\partial f_i}{\partial x_j}(\dots, c_j, \dots)h_j$$

for some c_j between x_j and $x_j + h_j$. (Set $\mathbf{c}_j$ to be the input on the right, which has c_j in the j-th entry and the same entries elsewhere as the inputs on the left.) To get

$$\text{from } f_i(\mathbf{x}+\mathbf{h}) = f_i(x_1 + h_1, \dots, x_n + h_n) \text{ to } f_i(\mathbf{x}) = f_i(x_1, \dots, x_n)$$

in a way which only changes one coordinate at-a-time, we subtract and add terms which move from $x_1 + h_1$ to x_1 while leaving every other coordinate the same, then subtract and terms which move from $x_2 + h_2$ to x_2 while leaving the rest the same, then the terms moving from $x_3 + h_3$ to x_3, and so on, subtracting and adding intermediate terms along the way:

$$f_i(\mathbf{x}+\mathbf{h}) - f_i(\mathbf{x}) = \sum_j [f_i(\dots, x_j + h_j, \dots) - f_i(\dots, x_j, \dots)]$$

where in both function evaluations on the right all inputs before $x_j + h_j$ are just x_k for $k < j$ and the inputs after are $x_k + h_k$ for $k > j$. (In the three-variable case, this looks like

$$\begin{aligned} f_i(x_1 + h_1, x_2 + h_2, x_3 + h_3) - f_i(x_1, x_2, x_3) = {} & f_i(x_1 + h_1, x_2 + h_2, x_3 + h_3) - f_i(x_1, x_2 + h_2, x_3 + h_3) \\ & + f_i(x_1, x_2 + h_2, x_3 + h_3) - f_i(x_1, x_2, x_3 + h_3) \\ & + f_i(x_1, x_2, x_3 + h_3) - f_i(x_1, x_2, x_3) \end{aligned}$$

where the terms in each difference on the right only differ in one input.) Applying the single-variable mean value theorem in each coordinate then gives

$$\sum_j [f_i(\dots, x_j + h_j, \dots) - f_i(\dots, x_j, \dots)] = \sum_j \frac{\partial f_i}{\partial x_j}(\mathbf{c}_j)h_j = \begin{bmatrix} \frac{\partial f_i}{\partial x_1}(\mathbf{c}_1) & \dots & \frac{\partial f_i}{\partial x_n}(\mathbf{c}_n) \end{bmatrix} \mathbf{h}$$

where $\mathbf{h}$ at the end is a column vector.

Thus we have

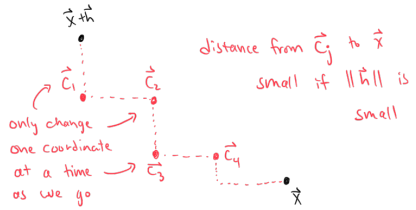
$$f_i(\mathbf{x}+\mathbf{h}) - f_i(\mathbf{x}) - Df_i(\mathbf{x})\mathbf{h} = \begin{bmatrix} \frac{\partial f_i}{\partial x_1}(\mathbf{c}_1) & \dots & \frac{\partial f_i}{\partial x_n}(\mathbf{c}_n) \end{bmatrix} \mathbf{h} - \begin{bmatrix} \frac{\partial f_i}{\partial x_1}(\mathbf{x}) & \dots & \frac{\partial f_i}{\partial x_n}(\mathbf{x}) \end{bmatrix} \mathbf{h}$$

$$= \left[\frac{\partial f_i}{\partial x_1}(\mathbf{c}_1) - \frac{\partial f_i}{\partial x_1}(\mathbf{x}) \quad \dots \quad \frac{\partial f_i}{\partial x_n}(\mathbf{c}_n) - \frac{\partial f_i}{\partial x_1}(\mathbf{x}) \right] \mathbf{h},$$

so taking norms and dividing by $\|\mathbf{h}\|$ gives

$$\begin{aligned} \frac{\|f_i(\mathbf{x}+\mathbf{h}) - f_i(\mathbf{x}) - Df_i(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|} &= \frac{\left\| \left[\frac{\partial f_i}{\partial x_1}(\mathbf{c}_1) - \frac{\partial f_i}{\partial x_1}(\mathbf{x}) \quad \dots \quad \frac{\partial f_i}{\partial x_n}(\mathbf{c}_n) - \frac{\partial f_i}{\partial x_1}(\mathbf{x}) \right] \mathbf{h} \right\|}{\|\mathbf{h}\|} \\ &\leq \frac{\left\| \left[\frac{\partial f_i}{\partial x_1}(\mathbf{c}_1) - \frac{\partial f_i}{\partial x_1}(\mathbf{x}) \quad \dots \quad \frac{\partial f_i}{\partial x_n}(\mathbf{c}_n) - \frac{\partial f_i}{\partial x_1}(\mathbf{x}) \right] \right\| \|\mathbf{h}\|}{\|\mathbf{h}\|} \\ &= \left\| \left[\frac{\partial f_i}{\partial x_1}(\mathbf{c}_1) - \frac{\partial f_i}{\partial x_1}(\mathbf{x}) \quad \dots \quad \frac{\partial f_i}{\partial x_n}(\mathbf{c}_n) - \frac{\partial f_i}{\partial x_1}(\mathbf{x}) \right] \right\|. \end{aligned}$$

By the way in which the $\mathbf{c}_j$ were defined, where for each we modify only one coordinate compared to what came before and what comes after using some c_j between x_j and $x_j + h_j$, each $\mathbf{c}_j$ will be close to $\mathbf{x}$ if $\mathbf{x} + \mathbf{h}$ is close to $\mathbf{x}$:



Thus by continuity of the $\frac{\partial f_i}{\partial x_j}$, each entry in

$$\left[\frac{\partial f_i}{\partial x_1}(\mathbf{c}_1) - \frac{\partial f_i}{\partial x_1}(\mathbf{x}) \quad \dots \quad \frac{\partial f_i}{\partial x_n}(\mathbf{c}_n) - \frac{\partial f_i}{\partial x_1}(\mathbf{x}) \right]$$

will approach 0 as $\mathbf{h}$ (and hence each $\mathbf{c}_j$) approaches $\mathbf{0}$, so we conclude that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f_i(\mathbf{x}+\mathbf{h}) - f_i(\mathbf{x}) - Df_i(\mathbf{x})\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0},$$

meaning that f is differentiable at $\mathbf{x}$ as claimed.

Example. The fact that C^1 implies differentiable is often a quick way to guarantee differentiability, but not the only way since this is not an equivalence. For example, the function $f : \mathbb{R}^2 to \mathbb{R}$ defined by

$$f(x,y) = \begin{cases} (x^2+y^2)\sin(\frac{1}{\sqrt{x^2+y^2}}) & \text{if } (x,y) \neq (0,0) \\ 0 & \text{if } (x,y) = (0,0) \end{cases}$$

is differentiable at $(0,0)$ but its partial derivatives are not continuous at $(0,0)$, so it is not C^1. Indeed, this function has partial derivatives at the origin which are both 0, so $Df(0,0) = 0$, and then

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{h}) - f(\mathbf{0}) - Df(\mathbf{0})\mathbf{h}}{\|\mathbf{h}\|} = \lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{h}\|^2 \sin(\frac{1}{\|\mathbf{h}\|})}{\|\mathbf{h}\|} = \lim_{\mathbf{h}\to\mathbf{0}} \|\mathbf{h}\| \sin(\frac{1}{\|\mathbf{h}\|}) = 0,$$

which gives differentiability. The partial derivatives at non-origin points can be computed using usual product, chain, and quotient rules, and after doing so it will be clear that they are not continuous at $(0,0)$.

Mean value theorem, scalar-valued version. We now prove the mean value theorem for scalar-valued functions. The claim is that if f is a scalar-valued differentiable function on an open convex set $E \subseteq \mathbb{R}^n$, then for any $\mathbf{x}, \mathbf{x}+\mathbf{h} \in E$ there exists $\mathbf{c}$ between $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$ (meaning on the line segment between) such that

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) = Df(\mathbf{c})\mathbf{h}.$$

Note, at least, that all of the sizes match up: $Df(\mathbf{c})$ is a $1 \times n$ matrix, and $\mathbf{h} \in \mathbb{R}^n$ is a $n \times 1$ column vector, so the right side above is a 1×1 scalar, just as the left side should be.

To prove this we make use of the single-variable mean value theorem by turning our multivariable function into a single-variable one by restricting to the line segment between $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$. That is, consider the function $g : [0,1] \to \mathbb{R}$ defined by

$$g(t) := f(\mathbf{x}+t\mathbf{h}).$$

(Convexity of E guarantees that $\mathbf{x}+t\mathbf{h} \in E$ for all $0 \leq t \leq 1$, so the right side above makes sense.) As t varies from 0 to 1, $\mathbf{x}+t\mathbf{h}$ fills out the desired segment with $g(0) = \mathbf{x}$ and $g(1) = \mathbf{x}+\mathbf{h}$. The function g is in fact differentiable by the *chain rule*, which we will prove next time. (The point is that g is the composition of the differentiable functions f and $t \mapsto \mathbf{x}+t\mathbf{h}$.) Taking this for granted now, the single-variable mean value theorem thus gives $c \in (0,1)$ such that

$$g(1) - g(0) = g'(c)(1-0).$$

The left side is $f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x})$. To compute the derivative on the right, we again use the still-to-be-proven chain rule: the derivative if $f(\mathbf{x}+t\mathbf{h})$ is the derivative of f evaluated at $\mathbf{x}+t\mathbf{h}$ times the derivative of $t \mapsto \mathbf{x}+t\mathbf{h}$, which is just $\mathbf{h}$. (Note that $\mathbf{x}+t\mathbf{h}$ is linear in t.) With this we have

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) = g(1) - g(0) = g'(c) = Df(\mathbf{x}+c\mathbf{h})\mathbf{h}.$$

Since $\mathbf{c} := \mathbf{x}+c\mathbf{h}$ is on the line segment between $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$, we have our desired claim.

Generalizing mean value. In an ideal world, the mean value theorem would work for differentiable $f : E \subseteq \mathbb{R}^n \to \mathbb{R}^m$ with $m > 1$ as well (so, f is vector-valued; E is still convex), but alas the world is not so nice. Certainly the equation one might expect the mean value theorem to give

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) = Df(\mathbf{c})\mathbf{h}$$

has all of the correct sizes: the left side is $m \times 1$ (thinking of vectors as column vectors), $Df(\mathbf{c})$ is $m \times n$ and $\mathbf{h}$ is $n \times 1$, so that $Df(\mathbf{c})\mathbf{h}$ is indeed $m \times 1$ as well. So the issue is not a technical one due to dimensions, but rather a fatal flaw. The problem is that, although we can apply the mean value theorem to each component of f and get equations like

$$f_i(\mathbf{x}+\mathbf{h}) - f_i(\mathbf{x}) = Df_i(\mathbf{c}_i)\mathbf{h},$$

the $\mathbf{c}_i$ at which the derivatives are evaluated can change as we move from component to component, whereas $Df(\mathbf{c})$—whose rows consist of the Df_i—would required a single $\mathbf{c}$ in all rows. (This is similar to what happens in the proof that C^1 implies differentiable, where we get partial derivatives at evaluated at different $\mathbf{c}_j$.)

A standard example of where things do not work is $f : [0, 2\pi] \to \mathbb{R}^2$ defined by

$$f(t) = (\cos t, \sin t).$$

Then $Df(t) = \left[\begin{smallmatrix} -\sin t \\ \cos t \end{smallmatrix}\right]$ and $f(0) = f(2\pi)$, so if the mean value equation did hold we would get

$$\mathbf{0} = f(2\pi) - f(0) = Df(t)(2\pi - 0) = 2\pi \begin{bmatrix} -\sin t \\ \cos t \end{bmatrix}.$$

But there is no t at which $\sin t$ and $\cos t$ are both zero, so this cannot be. We can get the $\mathbf{0}$ on the left if we allow different points at which to evaluate the derivatives in the component of Df, but not with one point for both.

Mean value theorem, vector-valued version. The best we can do in the vector-valued case is an *inequality* rather than an equality in the mean value statement. The claim is that if f is vector-valued and differentiable on an open and convex E, and if there exists $M \geq 0$ which bounds the operator norm of all $Df(\mathbf{x})$ for $\mathbf{x} \in U$, then for any $\mathbf{x}, \mathbf{a} \in E$ we have

$$\|f(\mathbf{x}) - f(\mathbf{a})\| \leq M \|\mathbf{x} - \mathbf{a}\|.$$

(In particular, f is Lipschitz.) If you look back at previous mean value applications, you will notice that in most cases the main takeaway is the inequality obtained by bounding derivative terms anyway, so having this inequality (as opposed to equality) in the vector-valued case will actually be good enough for our needs. The assumption that there be $M > 0$ such that $\|Df(\mathbf{x})\| \leq M$ for all $\mathbf{x}$ is not too difficult to justify in practice; for example, if f is C^1, then $\mathbf{x} \mapsto \|Df(\mathbf{x})\|$ is continuous, so we get such a maximum bound M at least on any *compact* and convex set in the domain E.

To prove this general mean value theorem, we make use of the scalar-valued version. Fix $\mathbf{x}, \mathbf{a} \in E$ and define $g : E \to \mathbb{R}$ as a function of $\mathbf{y} \in E$ by

$$g(\mathbf{y}) = (f(\mathbf{x}) - f(\mathbf{a}) \cdot f(\mathbf{y}).$$

By the scalar mean value theorem, there exists $\mathbf{c} \in E$ between $\mathbf{x}$ and $\mathbf{a}$ such that

$$g(\mathbf{x}) - g(\mathbf{a}) = Dg(\mathbf{c})(\mathbf{x} - \mathbf{a}).$$

The left side is

$$\underbrace{(f(\mathbf{x}) - f(\mathbf{a})) \cdot f(\mathbf{x})}_{g(\mathbf{x})} - \underbrace{(f(\mathbf{x}) - f(\mathbf{a})) \cdot f(\mathbf{a})}_{g(\mathbf{a})} = (f(\mathbf{x}) - f(\mathbf{a})) \cdot (f(\mathbf{x}) - f(\mathbf{a})) = \|f(\mathbf{x}) - f(\mathbf{a})\|^2.$$

By the product rule in the Warm-Up, the derivative of

$$g(\mathbf{y}) = (f(\mathbf{x}) - f(\mathbf{a}) \cdot f(\mathbf{y})$$

is "the derivative of $f(\mathbf{x}) - f(\mathbf{a})$ dot $f(\mathbf{y})$ plus $f(\mathbf{x}) - f(\mathbf{a})$ dot the derivative of $Df(\mathbf{y})$", but since these are derivatives taken with respect to $\mathbf{y}$, the derivative of $f(\mathbf{x}) - f(\mathbf{a})$ is zero, so only the second term survives and

$$Dg(\mathbf{c})(\mathbf{x} - \mathbf{a}) = (f(\mathbf{x}) - f(\mathbf{a})) \cdot Df(\mathbf{c})(\mathbf{x} - \mathbf{a}).$$

The scalar mean value result above thus becomes

$$\|f(\mathbf{x}) - f(\mathbf{a})\|^2 = (f(\mathbf{x}) - f(\mathbf{a})) \cdot Df(\mathbf{c})(\mathbf{x} - \mathbf{a}).$$

Taking norms and using Cauchy-Schwarz (for both vectors and matrices) gives

$$\|f(\mathbf{x}) - f(\mathbf{a})\|^2 = (f(\mathbf{x}) - f(\mathbf{a})) \cdot Df(\mathbf{c})(\mathbf{x} - \mathbf{a}) \leq \|f(\mathbf{x}) - f(\mathbf{a})\| \, \|Df(\mathbf{c})\| \, \|\mathbf{x} - \mathbf{a}\|,$$

and after bounding $\|Df(\mathbf{c})\| \leq M$ and dividing by $\|f(\mathbf{x}) - f(\mathbf{a})\|$ we get the desired

$$\|f(\mathbf{x}) - f(\mathbf{a})\| \leq M \|\mathbf{x} - \mathbf{a}\|.$$

Lecture 23: Chain Rule and More

Warm-Up. We show that if $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable with $Df(\mathbf{x}) = 0$ at all $\mathbf{x}$, then f is constant. Here is a first approach. Each component of $f = (f_1, \ldots, f_m)$ satisfies $Df_i(\mathbf{x}) = \mathbf{0}$ for all $\mathbf{x}$, so by the scalar mean value theorem we have

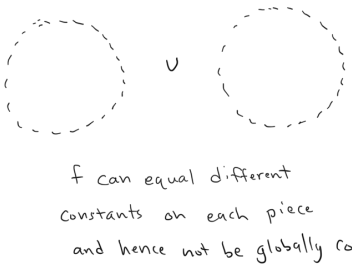
$$f(\mathbf{x}) - f(\mathbf{a}) = Df_i(\mathbf{c})(\mathbf{x} - \mathbf{a}) = 0$$

for $\mathbf{c}$ between $\mathbf{x}$ and $\mathbf{a}$, so $f_i(\mathbf{x}) = f_i(\mathbf{a})$ for all $\mathbf{a}, \mathbf{x}$. This means that each component f_i is constant, so f is constant as well. For a second approach, we use the vector mean value theorem. Since $\|Df(x)\| = \|0\| = 0$ at all points, we have

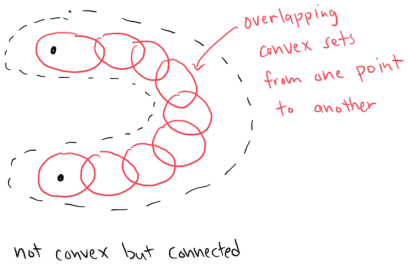
$$\|f(\mathbf{x}) - f(\mathbf{a})\| \leq 0 \, \|\mathbf{x} - \mathbf{a}\| = 0$$

for all $\mathbf{x}$ and $\mathbf{a}$, so $f(\mathbf{x}) - f(\mathbf{a}) = \mathbf{0}$ for all points and hence f is constant.

Now, the same reasoning applies in either approach if we replace the domain $\mathbb{R}^n$ with any open and convex set. However, we have to be careful with non-convex domains. Consider for example a domain consisting of the union of two disjoint open balls:



A function which has derivative zero at all points in such a domain does *not* have to be constant because the "constant" we get over each piece could be different; the derivative a point only depends on the behavior near that point, so one happens in one piece has no bearing on the derivative on the other piece. However, if we have a *connected* open domain, then we can extend the result of this Warm-Up to get that f must be constant. The proof above does not apply directly if the domain is not convex, but the intuition is that we can get from one point to any other point via a collection of convex subsets:



If these convex subsets overlap, the constant that we get as we move from one to another will stay the same, so f will be constant on the entire domain. You will implement this idea more formally on the next homework.

Chain rule. We now prove the multivariable chain rule, which we already used last time when proving the scalar version of the mean value theorem. The claim is that if f is differentiable at $\mathbf{x}$ and g is differentiable at $f(\mathbf{x})$ (all on appropriate domains), then $g \circ f$ is differentiable at $\mathbf{x}$ and

$$D(g \circ f)(\mathbf{x}) = Dg(f(\mathbf{x}))Df(\mathbf{x}).$$

So, the derivative of a composition is indeed the product of individual derivatives (in the same order as that occurring in the composition) interpreted as a product of matrices.

To prove this, we use the same notion of linear error as in the proof of the product rule from last time. We expand f near $\mathbf{x}$ as

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h} + \epsilon(\mathbf{h})$$

and g near $f(\mathbf{x})$ as

$$g(f(\mathbf{x})+\mathbf{k}) = g(f(\mathbf{x})) + Dg(f(\mathbf{x}))\mathbf{k} + \delta(\mathbf{k})$$

where the linear errors $\epsilon(\mathbf{h})$ and $\delta(\mathbf{k})$ satisfy

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\epsilon(\mathbf{h})}{\|\mathbf{h}\|} = \mathbf{0} \quad \text{and} \quad \lim_{\mathbf{k}\to\mathbf{0}} \frac{\delta(\mathbf{k})}{\|\mathbf{k}\|} = \mathbf{0}$$

since f and g are differentiable at $\mathbf{x}$ and $f(\mathbf{x})$ respectively. With this, we expand $g \circ f$ near $\mathbf{x}$ as

$$\begin{aligned} g(f(\mathbf{x}+\mathbf{h})) &= g(f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h} + \epsilon(\mathbf{h}) \\ &= g(f(\mathbf{x})) + Dg(f(\mathbf{x}))[Df(\mathbf{x})\mathbf{h} + \epsilon(\mathbf{h})] + \delta(\mathbf{k}) \\ &= g(f(\mathbf{x})) + Dg(f(\mathbf{x}))Df(\mathbf{x})\mathbf{h} + Dg(f(\mathbf{x}))\epsilon(\mathbf{h}) + \delta(\mathbf{k}) \end{aligned}$$

where in the first step we plugged in the expansion for $f(\mathbf{x}+\mathbf{h})$ and in the second we used the expansion for $g(f(\mathbf{x})+\mathbf{k})$ with $\mathbf{k} = Df(\mathbf{x})\mathbf{h} + \epsilon(\mathbf{h})$. The term in this expansion for $g(f(\mathbf{x}+\mathbf{h}))$ which is linear in $\mathbf{h}$ is $Dg(f(\mathbf{x}))Df(\mathbf{x})\mathbf{h}$, which is precisely what we are claiming the derivative transformation for $g \circ f$ at $\mathbf{x}$ should be, which makes sense since derivatives should linear terms in expansions in general. To show that this is the correct derivative, we must thus show that the "linear error" in this expansion, which is

$$Dg(f(\mathbf{x}))\epsilon(\mathbf{h}) + \delta(\mathbf{k}),$$

will go to zero as $\mathbf{h} \to \mathbf{0}$ upon being divided by $\|\mathbf{h}\|$.

To do this, we consider the two resulting terms

$$\frac{Dg(f(\mathbf{x}))\epsilon(\mathbf{h})}{\|\mathbf{h}\|} \quad \text{and} \quad \frac{\delta(\mathbf{k})}{\|\mathbf{h}\|}$$

separately. The first is easy to manage: taking norms gives

$$\frac{\|Dg(f(\mathbf{x}))\epsilon(\mathbf{h})\|}{\|\mathbf{h}\|} \leq \frac{\|Dg(f(\mathbf{x}))\|\,\|\epsilon(\mathbf{h})\|}{\|\mathbf{h}\|},$$

which goes to zero since $\frac{\epsilon(\mathbf{h})}{\|\mathbf{h}\|} \to \mathbf{0}$ and $\|Dg(f(\mathbf{x}))\|$ is fixed. For the second term, we would like to be able to rewrite as

$$\frac{\delta(\mathbf{k})}{\|\mathbf{h}\|} = \frac{\delta(\mathbf{k})}{\|\mathbf{k}\|}\frac{\|\mathbf{k}\|}{\|\mathbf{h}\|}$$

so that we can use $\frac{\delta(k)}{\|\mathbf{k}\|} \to \mathbf{0}$. There are a few issues here. First, we only know that $\frac{\delta(k)}{\|\mathbf{k}\|} \to \mathbf{0}$ as $\mathbf{k} \to \mathbf{0}$ whereas we are now taking the limit as $\mathbf{h} \to \mathbf{0}$, but this is not a problem: our value of $\mathbf{k} = Df(\mathbf{x})\mathbf{h} + \epsilon(\mathbf{h})$ does indeed go to $\mathbf{0}$ as $\mathbf{h} \to \mathbf{0}$, so $\frac{\delta(k)}{\|\mathbf{k}\|} \to \mathbf{0}$ as $\mathbf{h} \to \mathbf{0}$. The second issue is that the rewritten form above only makes sense if $\mathbf{k} \neq \mathbf{0}$ since otherwise the first fraction on the right does not exist. (When taking $\mathbf{h} \to \mathbf{0}$, we are certainly only considering nonzero $\mathbf{h}$, but

$\mathbf{k} = Df(\mathbf{x})\mathbf{h} + \epsilon(\mathbf{h})$ might in fact be zero even if $\mathbf{h}$ is not.) However, when $\mathbf{k} = \mathbf{0}$, the linear expansion for g near $f(\mathbf{x})$ from which $\delta(\mathbf{k})$ was obtained looks like

$$g(f(\mathbf{x}) + \mathbf{0}) = g(f(\mathbf{x})) + Dg(f(\mathbf{x}))\mathbf{h} + \delta(\mathbf{0}),$$

so that $\delta(\mathbf{0}) = \mathbf{0}$ because there *is* no error in this case as $g(f(\mathbf{x})) = g(f(\mathbf{x}))$ is true on the nose. Thus, $\delta(\mathbf{k}) = \mathbf{0}$ when $\mathbf{k} = \mathbf{0}$, so that we can write

$$\frac{\delta(\mathbf{k})}{\|\mathbf{h}\|} = \begin{cases} \frac{\delta(\mathbf{k})}{\|\mathbf{k}\|}\frac{\|\mathbf{k}\|}{\|\mathbf{h}\|} & \text{if } \mathbf{k} \neq \mathbf{0} \\ 0 & \text{if } \mathbf{k} = \mathbf{0} \end{cases}.$$

We can now use $\frac{\delta(\mathbf{k})}{\|\mathbf{k}\|} \to \mathbf{0}$ as $\mathbf{h}$ (hence $\mathbf{k}$) goes to $\mathbf{0}$, as long as we know that the remaining $\frac{\|\mathbf{k}\|}{\|\mathbf{h}\|}$ term remains bounded. But since $\mathbf{k} = Df(\mathbf{x})\mathbf{h} + \epsilon(\mathbf{h})$, we have

$$\frac{\|\mathbf{k}\|}{\|\mathbf{h}\|} = \frac{\|Df(\mathbf{x})\mathbf{h} + \epsilon(\mathbf{h})\|}{\|\mathbf{h}\|} \leq \frac{\|Df(\mathbf{x})\|\,\|\mathbf{h}\| + \|\epsilon(\mathbf{h})\|}{\|\mathbf{h}\|} = \|Df(\mathbf{x})\| + \frac{\|\epsilon(\mathbf{h})\|}{\|\mathbf{h}\|},$$

which indeed remains bounded since $\frac{\epsilon(\mathbf{h})}{\|\mathbf{h}\|} \to \mathbf{0}$ and $\|Df(\mathbf{x})\|$ is just some number. We thus conclude that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{g(f(\mathbf{x}+\mathbf{h})) - g(f(\mathbf{x})) - Dg(f(\mathbf{x}))Df(\mathbf{x})\mathbf{h}}{\|\mathbf{h}\|} = \lim_{\mathbf{h}\to\mathbf{0}} \frac{Dg(f(\mathbf{x}))\epsilon(\mathbf{h}) + \delta(\mathbf{k})}{\|\mathbf{h}\|} = \mathbf{0},$$

so $g \circ f$ is differentiable at $\mathbf{x}$ with derivative $D(g \circ f)(\mathbf{x}) = Dg(f(\mathbf{x}))Df(\mathbf{x})$ as claimed.

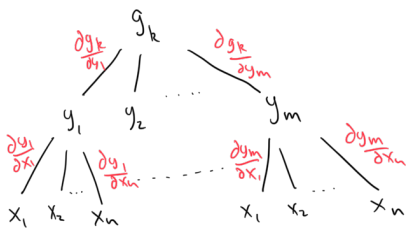
Partial derivative version. If we write out the entries of the Jacobian matrices $Dg(\mathbf{y})$ and $Df(\mathbf{x})$ (we'll suppress the points at which these are evaluated for the sake of clean notation), the product $Dg(\mathbf{y})Df(\mathbf{x})$ looks like

$$Dg(\mathbf{y})Df(\mathbf{x}) = \begin{bmatrix} \vdots & & \vdots \\ \frac{\partial g_k}{\partial y_1} & \cdots & \frac{\partial g_k}{\partial y_m} \\ \vdots & & \vdots \end{bmatrix} \begin{bmatrix} \cdots & \frac{\partial f_1}{\partial x_i} & \cdots \\ & \vdots & \\ \cdots & \frac{\partial f_n}{\partial x_i} & \cdots \end{bmatrix} = \text{matrix with } \sum_j \frac{\partial g_k}{\partial y_j}\frac{\partial f_j}{\partial x_i} \text{ in row } k, \text{ column } i.$$

Thus, as a consequence of the chain rule equality $D(g \circ f)(\mathbf{x}) = Dg(f(\mathbf{x}))Df(\mathbf{x})$, we get

$$\frac{\partial (g \circ f)_k}{\partial x_i} = \frac{\partial g_k}{\partial y_1}\frac{\partial f_1}{\partial x_i} + \cdots + \frac{\partial g_k}{\partial y_m}\frac{\partial f_m}{\partial x_i}.$$

This is the version of the multivariable chain rule one often sees in a multivariable calculus course, where the matrix version might not be covered. Indeed, one often draws "dependency tree diagrams" like



where $g_k(\mathbf{y})$ depends on $\mathbf{x}$ via $\mathbf{y} = f(\mathbf{x})$, and the partial derivatives of g_k with respect to the final x_i variables are obtained by following the branches down this tree. The upshot is that this all follows simply by considering the entries in the matrix (i.e., the true) version of the chain rule.

Derivatives beyond $\mathbb{R}^n$. The notion of differentiability we have developed extends beyond the setting of $\mathbb{R}^n$ alone; any setting in which we have a notion of "linearity" and norm gives rise to "differentiability" as well. We will not push this very far in this course, but let us consider at least the following example. With $M_n(\mathbb{R})$ denoting the space of $n \times n$ matrices, consider the function $f : M_n(\mathbb{R}) \to M_n(\mathbb{R})$ defined by $f(X) = X^2$. We claim that this is differentiable at any X. (Naively, we might expect that the derivative is $f'(X) = 2X$ based on what we know in the $M_1(\mathbb{R}) = \mathbb{R}$ case, but this is not quite right, as we will see.)

To be differentiable at X should mean that there is some type of "derivative" $Df(X)$ satisfying

$$\lim_{H \to 0} \frac{f(X+H) - f(X) - Df(X)(H)}{\|H\|} = 0,$$

where the limit is taken with respect to operator norms. But what type of "thing" should $Df(X)$ be? In the $\mathbb{R}^n$ case, $Df(\mathbf{x})$ is a matrix, but the point is that this matrix gives a linear transformation $\mathbb{R}^n \to \mathbb{R}^n$ and the $Df(\mathbf{x})\mathbf{h}$ term appearing in the differentiability definition is the result of applying this transformation to $\mathbf{h}$. The $Df(X)(H)$ term above should thus be the result of applying some linear transformation $Df(X)$ to H. Since $Df(X)(H)$ is a little cumbersome to read, we will instead denote the derivative at X by Df_X, so that to be differentiable at X means that there exists a linear transformation $Df_X : M_n(\mathbb{R}) \to M_n(\mathbb{R})$ such that

$$\lim_{H \to 0} \frac{f(X+H) - f(X) - Df_X(H)}{\|H\|} = 0,$$

again using operator norms. (If you have not seen the notion of a linear transformation between spaces of matrices before, the definition is the same as it is in the $\mathbb{R}^n$ case: Df_X is linear if

$$Df_X(H_1 + H_2) = Df_X(H_1) + Df_X(H_2) \quad \text{and} \quad Df_X(cH) = cDf_X(H)$$

for $H_1, H_2, H \in M_n(\mathbb{R})$ and $c \in \mathbb{R}$.)

To find the correct derivative Df_X, we expand $f(X+H)$ and extract the terms that are "linear" in H, just as we have done before. We have

$$(X+H)^2 = (X+H)(X+H) = X^2 + XH + HX + H^2.$$

(Note that matrix multiplication is not commutative, so we cannot necessarily combine XH and HX.) The "linear" terms here are XH and HX, so we expect that $Df_X(H) = XH + HX$; that is, $Df_X : M_n(\mathbb{R}) \to M_n(\mathbb{R})$ should be the linear transformation which sends H to $XH + HX$. We can check that this is indeed linear using properties of matrix multiplication:

$$\begin{aligned} Df_X(H_1 + H_2) &= X(H_1 + H_2) + (H_1 + H_2)X \\ &= XH_1 + XH_2 + H_1X + H_2X \\ &= XH_1 + H_1X + XH_2 + H_2X \\ &= Df_X(H_1) + Df_X(H_2) \\ Df_X(cH) &= X(cH) + (cH)X \\ &= cXH + cHX \end{aligned}$$

$$= c(XH + HX)$$
$$= cDf_X(H).$$

The "nonlinear" term H^2 in $(X+H)^2$ is thus the "linear error", and differentiability should mean that this go to zero as $H \to 0$. Indeed, we have

$$\frac{f(X+H) - f(X) - Df_X(H)}{\|H\|} = \frac{(X+H)^2 - X^2 - (XH+HX)}{\|H\|} = \frac{H^2}{\|H\|},$$

and after taking norms and using Cauchy-Schwarz (in the form $\left\|H^2\right\| = \|HH\| \leq \|H\|\,\|H\|$), we see that the limit of the final expression as $H \to 0$ is in fact zero. Thus $f(X) = X^2$ is indeed differentiable at X with derivative Df_X defined by $Df_X(H) = XH + HX$. (You will do this all for the functions $f(X) = X^3$ and $f(X) = X^{-1}$ on the homework!)

The naive guess "$f'(X) = 2X$" is incorrect, but it is in a sense "correct" if we interpret $2X = X + X$ in the right way, namely where multiply the two terms in $X + X$ by H, once on the right and once on the left to get $XH + HX$, which is the correct derivative value $Df_X(H)$. ($XH + HX$ *is* somehow the "infinitesimal version" of X^2, but making sense of this is best left to a course on what are called *manifolds*.) However, note that in the $n = 1$ case, so that $M_1(\mathbb{R}) = \mathbb{R}$ is the space of 1×1 matrices—i.e., numbers—we get

$$Df_x(h) = xh + hx = 2xh$$

since multiplication of 1×1 matrices *is* commutative. Lo-and-behold we see the usual single-variable derivative $f'(x) = 2x$ of $f(x) = x^2$ pop up, interpreted here as the linear transformation $\mathbb{R} \to \mathbb{R}$ which multiples h by $2x$. The upshot is that the setting $n \times n$ matrices subsumes what you know about $f(x) = x^2$, but this latter case is somehow too simplistic (since we are dealing with 1×1 matrices) to shed light on what is *really* going on behind the scene.

Lecture 24: Inverse Function Theorem

Warm-Up. We will not do much with higher-order derivatives in the $\mathbb{R}^n$ setting, apart from possibly a problem on the final homework. So, the point of this Warm-Up is just to give one example of a higher-order derivative result, namely the one fact about second-order derivatives you would no doubt have seen in a multivariable calculus course, namely the equality of *mixed* second-order partial derivatives. This goes by the name of *Clairaut's theorem*, which states that if f is a C^2 function (meaning all second-order partial derivatives are continuous), then

$$\frac{\partial^2 f}{\partial x_j\,\partial x_i} = \frac{\partial^2 f}{\partial x_i\,\partial x_j}.$$

(The second-order partial derivatives of $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$ are the first-order partial derivatives of $Df : U \to \mathbb{R}^n$ sending $\mathbf{x}$ to $Df(\mathbf{x})$ viewed as a vector in $\mathbb{R}^n$.) To simplify the notation, we consider only the case of a function f of two variables, which is enough since in the second-order derivatives above we vary only the two variables x_i and x_j anyway. (In other words, apply what we are about to do to the function $(x, y) \mapsto f(\ldots, x, \ldots, y, \ldots)$, with x and y in the i-th and j-th locations and all other variables are fixed.)

For fixed x, y, consider the expression

$$\frac{f(x+h, y+k) - f(x+h, y) - f(x, y+k) + f(x, y)}{hk},$$

of which will take the limit as $(h,k) \to (0,0)$. (The numerator evaluates f at the corners of a rectangle, and the idea is that we will shrink this rectangle when taking the limit. Clairaut's theorem essentially comes from first shrinking in the x direction and then y, and then vice versa.) If we introduce the function

$$g(t) = f(t, y+k) - f(t,y)$$

with y, k fixed, the numerator above is $g(x+h) - g(x)$, so a mean value application gives

$$f(x+h, y+k) - f(x+h, y) - f(x, y+k) + f(x,y) = [f_x(c, y+k) - f_x(c,y)]h$$

for some c between x and $x+h$. (Applying the mean value theorem first in the x-coordinate corresponds to first "shrinking in the x-direction".) This gives

$$\frac{f(x+h, y+k) - f(x+h, y) - f(x, y+k) + f(x,y)}{hk} = \frac{[f_x(c, y+k) - f_x(c,y)]h}{hk}.$$

Now we apply the mean value theorem in the y-coordinate (c fixed) to get

$$f_x(c, y+k) - f_x(c,y) = f_{xy}(c,d)k$$

for some d between y and $y+k$. (Here and above we are using subscript notation for the partial derivatives.) With this we get

$$\frac{[f_x(c, y+k) - f_x(c,y)]h}{hk} = \frac{f_{xy}(c,d)hk}{hk} = f_{xy}(c,d).$$

Since $f_{xy} = \frac{\partial^2 f}{\partial y \, \partial x}$ is continuous and $(c,d) \to (x,y)$ as $(h,k) \to (0,0)$ (because c is between x and $x+h$ and d is between y and $y+k$), we get

$$\lim_{(h,k)\to(0,0)} \frac{f(x+h, y+k) - f(x+h, y) - f(x, y+k) + f(x,y)}{hk} = \lim_{(h,k)\to(0,0)} f_{xy}(c,d) = f_{xy}(x,y).$$

Now go back and shrink in the opposite order, or more precisely apply mean value first in y and then in x. With the function

$$\ell(t) = f(x+h, t) - f(x,t),$$

the numerator in our limit is $\ell(y+k) - \ell(y)$, so applying the mean value theorem to this and then to the resulting $f_y(x+h, d) - f_y(x,d)$ will give

$$\frac{f(x+h, y+k) - f(x+h, y) - f(x, y+k) + f(x,y)}{hk} = f_{yx}(c,d)$$

for c between x and $x+h$ and d between y and $y+k$. By continuity of f_{yx}, we now get

$$\lim_{(h,k)\to(0,0)} \frac{f(x+h, y+k) - f(x+h, y) - f(x, y+k) + f(x,y)}{hk} = f_{yx}(x,y).$$

But the limit here is the same limit as before, so we must have $f_{xy}(x,y) = f_{yx}(x,y)$ as claimed.

The assumption that f be C^2 is crucial as Clairaut's theorem does not hold otherwise. (Actually, it holds assuming that at *at least* one of f_{xy} or f_{yx} is continuous, but not if neither are continuous.) The standard example is

$$f(x,y) = \begin{cases} \frac{xy(x^2-y^2)}{x^2+y^2} & \text{if } (x,y) \neq (0,0) \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

A direct computation shows that both f_{xy} and f_{yx} exist at the origin, but $f_{xy}(0,0) = -1$ whereas $f_{yx}(0,0) = 1$, so that the mixed second-order partials are not equal. The issue is that neither f_{xy} nor f_{yx} will be continuous at $(0,0)$.

Locally, calculus is linear algebra. Before moving on to the final two "big" theorems of differential analysis, let us setup the context by thinking about the point of calculus in general. I claim that all of differential calculus is borne out of taking linear-algebraic "infinitesimal" results that occur point by point and turning them into "local" results about behaviors of non-linear functions. Indeed, the entire point of viewing a higher-dimensional derivative as a matrix or linear transformation is to make this perspective precise.

Here is a table of concepts we have seen, or will see, that highlights which linear-algebraic notion each is meant to be the non-linear version of:

calculus	linear algebra
differentiable function	linear transformation
chain rule	compositions of linear transformations
mean value theorem	Cauchy-Schwarz inequality
inverse function theorem	solving $A\mathbf{x} = \mathbf{b}$ when A is invertible
implicit function theorem	solving $A\mathbf{x} = \mathbf{b}$ in general

First off, to say that a function is differentiable is exactly to say that it can be well-approximated by a linear transformation, or more precisely by a linear transformation plus a constant:

$$f(\mathbf{x}+\mathbf{h}) \approx \underbrace{f(\mathbf{x})}_{\text{constant}} + \underbrace{Df(\mathbf{x})}_{\text{linear}}\mathbf{h}.$$

Because of this, linear-algebraic properties of the matrix $Df(\mathbf{x})$ (i.e., "infinitesimal" properties of f at $\mathbf{x}$) should translate to local properties of f near $\mathbf{x}$. The chain rule says that this works for compositions, where local behavior of a composition $g \circ f$ comes from infinitesimal behavior of $Dg(f(\mathbf{x}))Df(\mathbf{x})$, which is the composition of the infinitesimal linear transformations $Dg(f(\mathbf{x}))$ and $Df(\mathbf{x})$. The mean value theorem in the inequality form

$$\|f(\mathbf{x}) - f(\mathbf{a})\| \leq M \|\mathbf{x} - \mathbf{a}\|$$

is the local version of the Cauchy-Schwarz ineqaulity

$$\|Df(\mathbf{x})\mathbf{h}_1 - Df(\mathbf{x})\mathbf{h}_2\| = \|Df(\mathbf{x})(\mathbf{h}_1 - \mathbf{h}_2)\| \leq \|Df(\mathbf{x})\| \|\mathbf{h}_1 - \mathbf{h}_2\|.$$

And so it is with our final theorems. I claim that the *inverse function theorem* is just about turning the fact that an $n \times n$ system of linear equations $A\mathbf{x} = \mathbf{b}$ always has a unique solution if and only if A is invertible into local behavior of f (where the matrix A will be $Df(\mathbf{x})$), and the *implicit function theorem* is the analog of this for $m \times n$ systems of linear equations with $m > n$. In general, whenever you have some linear-algebraic result about a derivative, you can expect there to be some calculus/analytic result that it gives rise to!

Solving nonlinear equations. The inverse function theorem is all about solving (systems of) equations. Consider for example the non-linear equations

$$\begin{aligned} xe^{xy} - \sin y &= a \\ x^9 y^{10} + 3\cos(xy) &= b. \end{aligned}$$

The question is whether this has a solution for x, y given a, b. Certainly it is not possible to solve for x, y *explicitly* here in terms of a and b, but all we are interested in is knowing whether such a solution exists. I claim that a unique solution is guaranteed to exist at least for

$$(a, b) \text{ close enough to } (e - \sin 1, 1 + 3\cos 1).$$

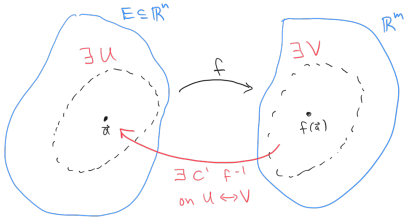
And even better: the functions $x(a, b), y(a, b)$ which describe the unique solutions (x, y) given (a, b) close enough to $(e - \sin 1, 1 + 3\cos 1)$ will be C^1. So, not only can the equations be solved (at least locally), but they can be solved in a continuously differentiable manner!

The reason why has to do with the derivative of the function

$$f(x, y) = (xe^{xy} - \sin y, x^9 y^{10} + 3\cos(xy)),$$

for which $(e - \sin 1, 1 + 3\cos 1)$ is actually just the value $f(1, 1)$ at $(1, 1)$. The inverse function theorem says that if $Df(1, 1)$ is invertible (which it is, as we will check next time), then the equation $f(x, y) = (a, b)$ can be "inverted" to express $(x, y) = g(a, b)$ in terms of a, b near $(1, 1)$. Thus, if an $n \times n$ equation is "infinitesimally solvable", then it is locally solvable.

Inverse function theorem. Here is the precise claim. Suppose $f : E \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is C^1 and that $Df(\mathbf{a})$ is invertible at some $\mathbf{a} \in E$. Then f is locally invertible near $\mathbf{a}$ with C^1 inverse, meaning that there exist open sets $U \subseteq E$ containing $\mathbf{a}$ and $V \subseteq \mathbb{R}^n$ containing $f(\mathbf{a})$ such that the restriction $f : U \to V$ is invertible with C^1 inverse $f^{-1} : V \to U$:



Moreover, the derivatives of the inverse are given by

$$D(f^{-1})(f(\mathbf{x})) = Df(\mathbf{x})^{-1}.$$

(So, although we will not be able to describe the inverse f^{-1} in any explicit way, we *will* be able to describe its derivatives explicitly—the derivative of the inverse is the inverse of the derivative—which is good enough at least to get some linear approximations to the inverse!) Intuitively, this should all indeed be true since for points $\mathbf{a} + \mathbf{h}$ near $\mathbf{a}$ the function f is approximated by

$$f(\mathbf{a}) + Df(\mathbf{a})\mathbf{h},$$

and if $Df(\mathbf{a})$ is an invertible matrix the right side is an invertible function of $\mathbf{h}$ (the inverse comes from solving $\mathbf{y} = f(\mathbf{a}) + Df(\mathbf{a})\mathbf{h}$ for $\mathbf{h}$ in terms of $\mathbf{y}$ using matrix operations), so f should be invertible near $\mathbf{a}$ as well.

We will give the full proof next time, but for now we focus on the key observation that makes it all work. Fix $\mathbf{y} \in \mathbb{R}^n$. To invert f we need to be able to find a unique $\mathbf{x}$ satisfying $f(\mathbf{x}) = \mathbf{y}$, as the candidate inverse will then send $\mathbf{y}$ to this $\mathbf{x}$. The goal is to phrase the problem of finding such $\mathbf{x}$ in a way that makes other tools we have developed applicable. Introduce the function

$$g(\mathbf{x}) = \mathbf{x} - Df(\mathbf{a})^{-1}(f(\mathbf{x}) - \mathbf{y}),$$

where $Df(\mathbf{a})^{-1}$ is the inverse matrix we are assuming exists. Then $f(\mathbf{x}) = \mathbf{y}$ if and only if $f(\mathbf{x}) - \mathbf{y} = \mathbf{0}$, which since $Df(\mathbf{a})^{-1}$ is invertible is true if and only if $Df(\mathbf{a})^{-1}(f(\mathbf{x}) - \mathbf{y}) = \mathbf{0}$, which is true if and only if $g(\mathbf{x}) = \mathbf{x}$, and hence (drumroll) the problem of finding $\mathbf{x}$, and hence inverting f, becomes a problem about fixed points (!!!) of g:

$$\mathbf{x} \text{ satisfies } f(\mathbf{x}) = \mathbf{y} \iff \mathbf{x} \text{ is a fixed point of } g.$$

This will be the reason why proving the inverse function theorem is possible.

In order to understand fixed points of g—in particular their existence and uniqueness—we should thus (because of the Banach contraction principle) be asking whether g is a contraction. The mean value theorem gives an inequality like

$$\left\|g(\mathbf{x}) - g(\mathbf{x}')\right\| \leq M \left\|\mathbf{x} - \mathbf{x}'\right\|,$$

which says that g indeed has the contraction property when the bound M on the derivatives of g is smaller than 1. The derivative of $g(\mathbf{x}) = \mathbf{x} - Df(\mathbf{a})^{-1}(f(\mathbf{x}) - \mathbf{y})$ is

$$Dg(\mathbf{x}) = I - Df(\mathbf{a})^{-1}Df(\mathbf{x}),$$

where the derivative of the first term in g—the identity function of $\mathbf{x}$—has derivative equal to the identity matrix I, and the derivative of the second term in g comes from the chain rule and the fact that the linear transformation $\mathbf{h} \mapsto Df(\mathbf{a})^{-1}\mathbf{h}$ has "constant" derivative $Df(\mathbf{a})^{-1}$, as we showed in a Warm-Up a few lectures back. We rewrite this derivative by factoring out $Df(\mathbf{a})^{-1}$ to get

$$Dg(\mathbf{x}) = Df(\mathbf{a})^{-1}[Df(\mathbf{a}) - Df(\mathbf{x})].$$

(The desire to end up with $Df(\mathbf{a}) - Df(\mathbf{x})$ is why we included the *inverse* of $Df(\mathbf{a})$ in the definition of g.) Taking norms gives

$$\|Dg(\mathbf{x})\| \leq \left\|Df(\mathbf{a})^{-1}\right\| \|Df(\mathbf{a}) - Df(\mathbf{x})\|.$$

Since f is C^1, we can make $\|Df(\mathbf{a}) - Df(\mathbf{x})\|$ as small as we want; in particular, there exists $r > 0$ such that

$$\|Df(\mathbf{a}) - Df(\mathbf{x})\| < \frac{1}{2\left\|Df(\mathbf{a})\right\|^{-1}} \text{ for } \|\mathbf{x} - \mathbf{a}\| < r.$$

Thus, on the open ball $B_r(\mathbf{a})$ we have

$$\|Dg(\mathbf{x})\| \leq \left\|Df(\mathbf{a})^{-1}\right\| \|Df(\mathbf{a}) - Df(\mathbf{x})\| < \frac{1}{2},$$

which thus implies the contraction property of g. (Boom!) There is still much work to be done to clarify exactly how we will use this property to invert f, but we will indeed exploit properties of contractions throughout the proof we will give next time.

Lecture 25: More on Inverses

Warm-Up 1. We show that there exist C^1 functions $x(a, b)$ and $y(a, b)$ on some open set in $\mathbb{R}^2$ containing $(1 - \sin 1, 1 + 3\cos 1)$ such that $x = x(a, b)$ and $y = y(a, b)$ satisfy

$$\begin{aligned} xe^{xy} - \sin y &= a \\ x^9y^{10} + 3\cos(xy) &= b. \end{aligned}$$

This finishes the example we introduced last time when motivating the inverse function theorem, since what we claim here says precisely that the equations above can be solved for (x, y) in terms of (a, b) in a continuously differentiable manner for (a, b) close enough to $(1 - \sin 1, 1 + 3\cos 1)$. We will not be able to say what the solution functions $x(a, b), y(a, b)$ look like explicitly, but we will know that they exist.

Define $f : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$f(x, y) = (xe^{xy} - \sin y, x^9y^{10} + 3\cos(xy))$$

and note that $f(1, 1) = (1 - \sin 1, 1 + 3\cos 1)$. We have

$$Df(x, y) = \begin{bmatrix} e^{xy} + xye^{xy} & x^2e^{xy} - \cos y \\ 9x^8y^{10} - 3y\sin(xy) & 10x^9y^9 - 3x\sin(xy) \end{bmatrix},$$

and thus

$$Df(1, 1) = \begin{bmatrix} 2e & e - \cos 1 \\ 9 - 3\sin 1 & 10 - 3\sin 1 \end{bmatrix}.$$

Since f is C^1 (the components of Df computed above are continuous) and $Df(1, 1)$ is invertible (for example by computing its determinant and seeing that it is nonzero, the inverse function theorem says that there are open sets $U \ni (1, 1)$ and $V \ni f(1, 1)$ such that

$$f : U \to V \text{ is invertible with } C^1 \text{ inverse.}$$

Setting $x(a, b)$ and $y(a, b)$ to be the components of f^{-1} give our desired functions:

$$(x(a, b), y(a, b)) := f^{-1}(a, b) \text{ for } (a, b) \in V.$$

These are C^1 since f^{-1} is C^1, and they satisfy

$$f(x(a, b), y(a, b)) = f(f^{-1}(a, b)) = (a, b),$$

which is just our original system of 2×2 nonlinear equations.

Warm-Up 2. Suppose f is invertible, differentiable at $\mathbf{x}$, and that f^{-1} is differentiable at $f(\mathbf{x})$. We show that

$$Df^{-1}(f(\mathbf{x})) = Df(\mathbf{x})^{-1},$$

so that "the derivative of the inverse is the inverse of the derivative" when evaluated at the appropriate points. This the general version of

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

from single-variable calculus, and guarantees that even if we do not know what f^{-1} is exactly, we for sure know what the derivative of f^{-1} is.

This just comes from the chain rule. We have

$$f^{-1}(f(\mathbf{x})) = \mathbf{x}$$

by the definition of an inverse. Since f is differentiable at $\mathbf{x}$ and f^{-1} is differentiable at $f(\mathbf{x})$, the chain rule says that the left side is differentiable at $\mathbf{x}$ with derivative given by the product

$$\underbrace{Df^{-1}(f\mathbf{x}))}_{\text{matrix}} \underbrace{Df(\mathbf{x})}_{\text{matrix}}.$$

The right side of $f^{-1}(f(\mathbf{x})) = \mathbf{x}$ above is the identity function $\mathbf{x} \mapsto \mathbf{x}$, and so has derivative equal to the identity matrix; thus after taking derivatives of both sides we get

$$Df^{-1}(f\mathbf{x}))Df(\mathbf{x}) = I.$$

Since both matrices are square matrices, this shows that they are inverses of one another, so

$$Df^{-1}(f(\mathbf{x})) = Df(\mathbf{x})^{-1}$$

as claimed.

For example, in the first Warm-Up, we have

$$\begin{aligned} Df^{-1}(f(1,1)) &= Df(1,1)^{-1} \\ &= \begin{bmatrix} 2e & e-\cos 1 \\ 9-3\sin 1 & 10-3\sin 1 \end{bmatrix}^{-1} \\ &= \frac{1}{11e-3e\sin 1+9\cos 1-3\sin 1\cos 1} \begin{bmatrix} 10-3\sin 1 & -e+\cos 1 \\ -9+3\sin 1 & 2e \end{bmatrix}. \end{aligned}$$

We do not know the functions $(x(a,b)), y(a,b)) = f^{-1}(a,b)$ we showed exist, but we at least know how to linear approximate them near $f(1,1) = (1-\sin 1, 1+3\cos 1)$:

$$\begin{aligned} \begin{bmatrix} x(a,b) \\ y(a,b) \end{bmatrix} &\approx \begin{bmatrix} x(f(1,1)) \\ y(f(1,1)) \end{bmatrix} + Df^{-1}(f(1,1)) \begin{bmatrix} a-[1-\sin 1] \\ b-[1+3\cos 1] \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \frac{1}{K} \begin{bmatrix} 10-3\sin 1 & -e+\cos 1 \\ -9+3\sin 1 & 2e \end{bmatrix} \begin{bmatrix} a-[1-\sin 1] \\ b-[1+3\cos 1] \end{bmatrix}. \end{aligned}$$

where $K = 11e - 3e\sin 1 + 9\cos 1 - 3\sin 1\cos 1$. In most practical applications, having a linear approximation to the inverse is good enough!

Proof of the inverse function theorem. We will now prove the inverse function theorem. This is going to be the most elaborate proof we have seen in the entire course, the difficulty of which is warranted since the implications are so strong! Indeed, the inverse function theorem is truly the only way we have of even knowing that $n \times n$ systems of nonlinear equations *have* solutions given that writing down explicit solutions is nearly always impossible. (But, as we saw in the Warm-Up, it will at least be possible to approximate the solutions, however.) Such a monster result is likely to have a monster proof! In fact, all key topics we have seen when it comes to differentiation—the definition of differentiable, the mean value theorem, and the chain rule—will play a role.

The key observation is the one we finished with last time, so let us recall it here. Given $\mathbf{y}$, to say that $\mathbf{x}$ satisfies $f(\mathbf{x}) = \mathbf{y}$ is to say that $\mathbf{x}$ is a fixed point of $g(\mathbf{x}) := \mathbf{x} - Df(\mathbf{a})^{-1}(f(\mathbf{x}) - \mathbf{y})$. Moreover, it is possible to find $r > 0$ such that $\|Dg(\mathbf{x})\| \leq \frac{1}{2}$, which uses the chain, for $\|\mathbf{x} - \mathbf{a}\| < r$, and hence

$$\left\|g(\mathbf{x}) - g(\mathbf{x}')\right\| \leq \frac{1}{2}\left\|\mathbf{x} - \mathbf{x}'\right\| \text{ for } \mathbf{x}, \mathbf{x}' \in B_r(\mathbf{a}),$$

which uses the mean value theorem. This hints at some contraction property for g, which is what makes understanding fixed points of g, and hence solutions of $f(\mathbf{x}) = \mathbf{y}$, possible.

Let us outline the strategy before jumping into the proof. First, we will construct a local inverse for f on some $U \ni \mathbf{a}$; this will use the fact that fixed points of contractions, when they exist, are unique. Second, we will show that $f(U)$ is open, and as a byproduct that $f^{-1} : f(U) \to U$ is continuous; this will use the full force of the Banach contraction principle on fixed points of

contractions on complete spaces. Then, we show that the local inverse f^{-1} thus constructed is differentiable, which will use some bounds developed from the function g above, and finally we show that the derivative of the local inverse f^{-1} is continuous, which will be easy. As a consequence of the third step, we will also derive the expression for Df^{-1} found in the second Warm-Up, only here without assuming differentiability of the inverse beforehand. Buckle up!

Step 1: constructing the local inverse. For fixed $\mathbf{y}$, we have

$$\left\|g(\mathbf{x}) - g(\mathbf{x}')\right\| \leq \frac{1}{2}\left\|\mathbf{x} - \mathbf{x}'\right\| \text{ on some } B_r(\mathbf{a})$$

where $g(\mathbf{x}) := \mathbf{x} - Df(\mathbf{a})^{-1}(f(\mathbf{x}) - \mathbf{y})$. So, g is a contraction on $B_r(\mathbf{a})$, but we cannot use the Banach contraction principle here to claim that g has a fixed point for two reasons: 1. $B_r(\mathbf{a})$ is not complete, but this is easy to fix by using the closure instead ($\mathbb{R}^n$ is complete and closed subsets of complete spaces are complete); but the real issue is 2. g does not necessarily map $B_r(\mathbf{a})$ (or its closure) into *itself*, which was the necessary setup of the Banach contraction principle. (The iteration process used in the proof of Banach contraction does not work if the contraction does not map the space into itself!)

So, here we will not use anything about existence of fixed points, but rather the fact that *if* a fixed point of g exists, then it must be unique. For this we do not need the space to be mapped to itself, as the contraction inequality above alone is enough: if $g(\mathbf{x}) = \mathbf{x}$ and $g(\mathbf{x}') = \mathbf{x}'$, then the inequality above becomes

$$\left\|\mathbf{x} - \mathbf{x}'\right\| \leq \frac{1}{2}\left\|\mathbf{x} - \mathbf{x}'\right\|,$$

which means $\|\mathbf{x} - \mathbf{x}'\| = 0$ and hence $\mathbf{x} = \mathbf{x}'$. So, the conclusion is that if $\mathbf{y}$ is a point for which $f(\mathbf{x}) = \mathbf{y}$ *does* have a solution $\mathbf{x}$, then it only has one; in other words, f is *injective* on $B_r(\mathbf{a})$.

Thus, set $U := B_r(\mathbf{a})$. Then $f : U \to f(U)$ is injective and surjective (we cut down the codomain in order to make it surjective), so it is bijective and hence invertible. This gives us our desired local inverse $f^{-1} : f(U) \to U$, with $\mathbf{a} \in U$ and $f(\mathbf{a}) \in f(U)$.

Step 2: continuity of the local inverse. Next we show that $f(U)$ is open in $\mathbb{R}^n$. This will then be the set $V := f(U)$ in the statement of the inverse function theorem, so that $f : U \to V$ and $f^{-1} : V \to U$. Let $\mathbf{y}_0 \in f(U)$ and pick $\mathbf{x}_0 \in U$ such that $f(\mathbf{x}_0) = \mathbf{y}_0$. We must find a ball around $\mathbf{y}_0$ that is contained in $f(U)$. Let $\mathbf{y} \in B_?(\mathbf{y}_0)$, where ? denotes some to-be-determined radius. We want $\mathbf{y} \in f(U)$, or in other words that there exists $\mathbf{x} \in U$ such that $f(\mathbf{x}) = \mathbf{y}$, and we know from our fixed-point rephrasing that this is equivalent to showing that $g(\mathbf{x}) := \mathbf{x} - Df(\mathbf{a})^{-1}(f(\mathbf{x}) - \mathbf{y})$ has a fixed point. Thus, we now care about the *existence* of a fixed point, whereas in the first step we only cared about the uniqueness.

Pick some $B_s(\mathbf{x}_0) \subseteq U$ (recall U is open), and by shrinking the radius we can ensure $\overline{B_s(\mathbf{x}_0)} \subseteq U$. We know that $g : \overline{B_s(\mathbf{x}_0)} \to f(U)$ is a contraction, but in order to guarantee the existence of a fixed point for g, we now need g to map the complete space $\overline{B_s(\mathbf{x}_0)}$ into itself. Thus, we must know that

$$\text{if } \mathbf{x} \in \overline{B_s(\mathbf{x}_0)}, \text{ then } g(\mathbf{x}) \in \overline{B_s(\mathbf{x}_0)}, \text{ or in other words } \|g(\mathbf{x}) - \mathbf{x}_0\| \leq s.$$

This will achieve via a triangle inequality comparison with $g(\mathbf{x}_0)$. From the definition of g, we have

$$\|g(\mathbf{x}_0) - \mathbf{x}_0\| = \left\|Df(\mathbf{a})^{-1}(f(\mathbf{x}_0) - \mathbf{y})\right\| = \left\|Df(\mathbf{a})^{-1}(\mathbf{y}_0 - \mathbf{y})\right\| \leq \left\|Df(\mathbf{a})^{-1}\right\| \|\mathbf{y} - \mathbf{y}_0\|.$$

Thus if we go back and rick the ? radius small enough, we can ensure this final expression is at most $\frac{s}{2}$. So, for $\mathbf{y} \in B_{\frac{1}{2\|Df(\mathbf{a})^{-1}\|}}(\mathbf{y}_0)$, we have

$$\|g(\mathbf{x}_0) - \mathbf{x}_0\| \leq \left\|Df(\mathbf{a})^{-1}\right\| \|\mathbf{y} - \mathbf{y}_0\| \leq \frac{s}{2}.$$

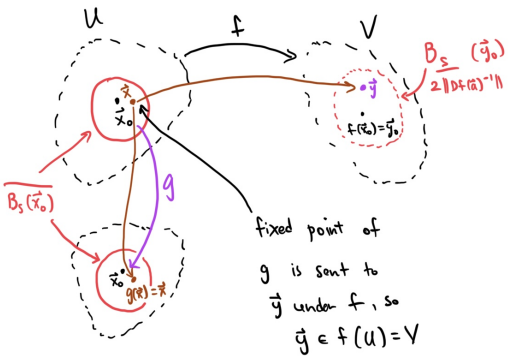
By the contraction property g satisfies, for $\mathbf{x} \in \overline{B_s(\mathbf{x}_0)}$ we have

$$\|g(\mathbf{x}) - g(\mathbf{x}_0)\| \leq \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\| \leq \frac{s}{2}.$$

Hence for such $\mathbf{x}$ we get

$$\|g(\mathbf{x}) - \mathbf{x}_0\| \leq \|g(\mathbf{x}) - g(\mathbf{x}_0)\| + \|g(\mathbf{x}_0) - \mathbf{x}_0\| \leq \frac{s}{2} + \frac{s}{2} = s.$$

Thus, g maps $\overline{B_s(\mathbf{x}_0)}$ into itself, and $\overline{B_s(\mathbf{x}_0)}$ is complete, so the Banach contraction principle guarantees that g has a fixed point, so there exists $\mathbf{x} \in \overline{B_s(\mathbf{x}_0)} \subseteq U$ such that $g(\mathbf{x}) = \mathbf{x}$, which is equivalent to $f(\mathbf{x}) = \mathbf{y}$ for the $\mathbf{y}$ we fixed in $B_{\frac{1}{2\|Df(\mathbf{a})^{-1}\|}}(\mathbf{y}_0)$. Thus $\mathbf{y} \in f(U)$, so $B_{\frac{1}{2\|Df(\mathbf{a})^{-1}\|}}(\mathbf{y}_0)$ is contained in $f(U)$ and hence $f(U)$ is open in $\mathbb{R}^n$. Here is the picture of what's going on:



So, we have our $f : U \to V$, with $U \ni \mathbf{a}$ open and $V = f(U) \ni f(\mathbf{a})$ open, and we have the inverse $f^{-1} : V \to U$. The same reasoning as above with U replaced by a smaller open subset of U works just as well, and shows that the image of any open set in U is open in V. This says that $f : U \to V$ is what's called an *open map*, which just means a function sending open sets to open sets. (This gives a proof of the fact that C^1 functions with invertible derivatives are always open.) But to say that f is open means the same thing as saying that $f^{-1} : V \to U$ is continuous, since the preimage of $Z \subseteq U$ under f^{-1} is just the image of Z under f.

Step 3: differentiability of the inverse. At this point we only know that $f : U \to V$ has invertible derivative at $\mathbf{a} \in U$ (by the setup of the inverse function theorem), but we can ensure invertible derivative everywhere by shrinking U if need be. The set of invertible $n \times n$ matrices is open in the set of all $n \times n$ matrices (with respect to the operator norm), so since $Df(\mathbf{a})$ is invertible, all nearby $Df(\mathbf{x})$ will be invertible too. Since f is C^1, we can make $Df(\mathbf{x})$ close to $Df(\mathbf{a})$ by making $\mathbf{x}$ close to $\mathbf{a}$, so if we shrink U, so that $\mathbf{x}$ is close enough to $\mathbf{a}$, we will indeed have that $Df(\mathbf{x})$ is invertible for all $\mathbf{x} \in U$.

Now we show that $f^{-1} : V \to U$ is differentiable. Let $\mathbf{y} \in V$ and let $\mathbf{y} + \mathbf{k} \in V$. ($V$ is open, so $\mathbf{y} + \mathbf{k}$ is in V for small enough $\mathbf{k}$.) Since $V = f(U)$, we can pick $\mathbf{x} \in U$ and $\mathbf{x} + \mathbf{h} \in U$ such that

$$f(\mathbf{x}) = \mathbf{y} \quad \text{and} \quad f(\mathbf{x} + \mathbf{h}) = \mathbf{y} + \mathbf{k}.$$

We claim that

$$\lim_{\mathbf{k} to \mathbf{0}} \frac{f^{-1}(\mathbf{y} + \mathbf{k}) - f^{-1}(\mathbf{y}) - Df(\mathbf{x})^{-1}\mathbf{k}}{\|\mathbf{k}\|} = \mathbf{0},$$

which shows that f^{-1} is differentiable at $\mathbf{y}$ with derivative $Df(\mathbf{x})^{-1}$. (Thus, as suggested by the Warm-Up, we get that $Df^{-1}(\mathbf{y}) = Df(\mathbf{x})^{-1} = Df(f^{-1}(\mathbf{y}))^{-1}$ as a consequence of this part.) As

with other differentiability arguments we have seen, this will come from writing the numerator above in a way which will make taking the limit possible.

We have $f^{-1}(\mathbf{y}+\mathbf{k}) - f^{-1}(\mathbf{y}) = (\mathbf{x}+\mathbf{h}) - \mathbf{x} = \mathbf{h}$, so

$$f^{-1}(\mathbf{y}+\mathbf{k}) - f^{-1}(\mathbf{y}) - Df(\mathbf{x})^{-1}\mathbf{k} = \mathbf{h} - Df(\mathbf{x})^{-1}\mathbf{k} = Df(\mathbf{x})^{-1}[Df(\mathbf{x})\mathbf{h} - \mathbf{k}],$$

where in the last step we use $\mathbf{h} = Df(\mathbf{x})^{-1}Df(\mathbf{x})\mathbf{h}$. But $\mathbf{k} = (\mathbf{y}+\mathbf{k}) - \mathbf{y} = f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x})$, so

$$Df(\mathbf{x})^{-1}[Df(\mathbf{x})\mathbf{h} - \mathbf{k}] = Df(\mathbf{x})^{-1}[Df(\mathbf{x})\mathbf{h} - f(\mathbf{x}+\mathbf{h}) + f(\mathbf{x})].$$

Thus, the limit we want to equal zero looks like

$$\lim_{\mathbf{k} to \mathbf{0}} \frac{f^{-1}(\mathbf{y}+\mathbf{k}) - f^{-1}(\mathbf{y}) - Df(\mathbf{x})^{-1}\mathbf{k}}{\|\mathbf{k}\|} = \lim_{\mathbf{k}\to\mathbf{0}} \frac{Df(\mathbf{x})^{-1}[Df(\mathbf{x})\mathbf{h} - f(\mathbf{x}+\mathbf{h}) + f(\mathbf{x})]}{\|\mathbf{k}\|}.$$

The key point is that $Df(\mathbf{x})\mathbf{h} - f(\mathbf{x}+\mathbf{h}) + f(\mathbf{x})$, or rather its negative, is precisely the linear error in expanding f around $\mathbf{x}$, so we know already that $\frac{Df(\mathbf{x})\mathbf{h} - f(\mathbf{x}+\mathbf{h}) + f(\mathbf{x})}{\|\mathbf{h}\|} \to \mathbf{0}$ as $\mathbf{h} \to \mathbf{0}$. Hence taking the limit above will be possible once we can relate the denominator $\|\mathbf{k}\|$ to $\|\mathbf{h}\|$.

We make use of $g(\mathbf{x}) = \mathbf{x} - Df(\mathbf{a})^{-1}(f(\mathbf{x}) - \mathbf{y})$ again to write

$$\begin{aligned} g(\mathbf{x}+\mathbf{h}) - g(\mathbf{x}) &= [(\mathbf{x}+\mathbf{h}) - Df(\mathbf{a})^{-1}(f(\mathbf{x}+\mathbf{h}) - \mathbf{y})] - [\mathbf{x} - Df(\mathbf{a})^{-1}(f(\mathbf{x}) - \mathbf{y})] \\ &= \mathbf{h} - Df(\mathbf{a})^{-1}(f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x})) \\ &= \mathbf{h} - Df(\mathbf{a})^{-1}\mathbf{k}, \end{aligned}$$

so $\left\|\mathbf{h} - Df(\mathbf{a})^{-1}\mathbf{k}\right\| = \|g(\mathbf{x}+h\mathbf{h}) - g(\mathbf{x})\| \leq \frac{1}{2}\|\mathbf{h}\|$, where we once again use the contraction inequality for g. The reverse triangle inequality gives

$$\|\mathbf{h}\| - \left\|Df(\mathbf{a})^{-1}\mathbf{k}\right\| \leq \frac{1}{2}\|\mathbf{h}\|, \text{ so } \frac{1}{2}\|\mathbf{h}\| \leq \left\|Df(\mathbf{a})^{-1}\mathbf{k}\right\| \leq \left\|Df(\mathbf{a})^{-1}\right\| \|\mathbf{k}\|.$$

Thus $\|\mathbf{k}\| \geq \frac{\|\mathbf{h}\|}{2\|Df(\mathbf{a})^{-1}\|}$, so (almost at the finish line!) we have

$$\begin{aligned} \frac{\left\|f^{-1}(\mathbf{y}+\mathbf{k}) - f^{-1}(\mathbf{y}) - Df(\mathbf{x})^{-1}\mathbf{k}\right\|}{\|\mathbf{k}\|} &= \frac{\left\|Df(\mathbf{x})^{-1}[Df(\mathbf{x})\mathbf{h} - f(\mathbf{x}+\mathbf{h}) + f(\mathbf{x})]\right\|}{\|\mathbf{k}\|} \\ &\leq \frac{\left\|Df(\mathbf{x})^{-1}\right\|}{1/(2\|Df(\mathbf{a})^{-1}\|)} \frac{\|Df(\mathbf{x})\mathbf{h} - f(\mathbf{x}+\mathbf{h}) + f(\mathbf{x})\|}{\|\mathbf{h}\|}. \end{aligned}$$

As $\mathbf{k} \to \mathbf{0}$, $\mathbf{h} = f^{-1}(\mathbf{y}+\mathbf{k}) - f^{-1}(\mathbf{y}) \to \mathbf{0}$ as well since f^{-1} is continuous, so the final expression above goes to $\mathbf{0}$ by differentiability of f at $\mathbf{x}$, and thus f^{-1} is differentiable at $\mathbf{y}$ with $Df^{-1}(\mathbf{y}) = Df(\mathbf{x})^{-1}$. Simply. Astonishing.

Step 4: continuity of the derivative. So, $f : U \to V$ is invertible and $f^{-1} : V \to U$ is differentiable. The remaining claim is that f^{-1} is C^1, meaning that Df^{-1} is continuous. But

$$Df^{-1}(\mathbf{y}) = Df(f^{-1}(\mathbf{y}))^{-1}$$

is a composition of continuous things (f^{-1} is continuous by Step 2, Df is continuous since f is C^1, and matrix inversion is a continuous operation), so we get that Df is continuous as well, and so we... are... done! (And the crowds rejoiced! Simply a tour de force argument of analysis.)

Lecture 26: Implicit Function Theorem

Warm-Up. The key idea in the proof of the inverse function theorem is that of rephrasing the problem of finding $\mathbf{x}$ satisfying $f(\mathbf{x}) = \mathbf{y}$ (for $\mathbf{x}$ close to $\mathbf{a}$ and $\mathbf{y}$ close to $f(\mathbf{a})$) as finding a fixed point of $g(\mathbf{x}) = \mathbf{x} - Df(\mathbf{a})^{-1}(f(\mathbf{x}) - \mathbf{y})$ instead. The existence of this fixed point—from the proof of the Banach contraction principle—comes from an iterative argument where for any starting $\mathbf{p}$ near $\mathbf{a}$, the sequence

$$\mathbf{p},\ g(\mathbf{p}),\ g(g(\mathbf{p})),\ g(g(g(\mathbf{p}))),\ \dots$$

will converge to the fixed point of g, and hence to the $\mathbf{x} = f^{-1}(\mathbf{y})$ we want to find. If nothing else, this can be used to give a sense as to why $g(\mathbf{x}) = \mathbf{x} - Df(\mathbf{a})^{-1}(f(\mathbf{x}) - \mathbf{y})$ was the right thing to look at. Recall the initial motivation we gave for the inverse function theorem: if

$$f(\mathbf{x}) \approx f(\mathbf{a}) + Df(\mathbf{a})(\mathbf{x} - \mathbf{a})$$

is to hold for $\mathbf{x}$ near $\mathbf{a}$, invertibility of the function on the right should correspond to invertibility of the function on the left. Thus given $\mathbf{y} = f(\mathbf{x})$, we have

$$\mathbf{y} \approx f(\mathbf{a}) + Df(\mathbf{a})(\mathbf{x} - \mathbf{a}),\ \text{so}\ \mathbf{x} \approx \mathbf{a} - Df(\mathbf{a})^{-1}(f(\mathbf{a}) - \mathbf{y})$$

where the second expression comes from solving for $\mathbf{x}$ in the first using the inverse of $Df(\mathbf{a})$. Hence $g(\mathbf{a}) = \mathbf{a} - Df(\mathbf{a})^{-1}(f(\mathbf{a}) - \mathbf{y})$ is a first approximation to $\mathbf{x} = f^{-1}(\mathbf{y})$, and we then iterate to get better approximations using precisely this "infinitesimal inverse" $g(\mathbf{x}) = \mathbf{x} - Df(\mathbf{a})^{-1}(f(\mathbf{x}) - \mathbf{y})$.

Our goal here is to see what this all looks like in the simple case of a single-variable function $f : U \subseteq \mathbb{R} \to \mathbb{R}$. In particular, we take the function $f(x) = \log(x)$ on $(0, \infty)$ defined by

$$\log(x) = \int_1^x \frac{1}{t}\, dt.$$

Now, we know what the inverse should be in this case, namely $f^{-1}(y) = e^y$, but imagine that we did not know anything about log apart from the definition above, and in particular did not know that it was the inverse of the exponential function. The goal is to get a sense for why this should be the case solely from the iterative argument described above. That is, for fixed y (near $f(1) = \log(1) = \int_1^1 \frac{1}{t}\, dt = 0$, we work out some iterates

$$1,\ g(1),\ g(g(1)),\ g(g(g(1))),\ \dots$$

and see (at least numerically) that they do appear to be converging to $f^{-1}(y) = e^y$. By the fundamental theorem of calculus, we have

$$f'(1) = \frac{1}{1} > 0,$$

so $f'(1)$ is invertible (as a 1×1 matrix), and hence the inverse function theorem and its proof apply. The contraction $g(\mathbf{x}) = \mathbf{x} - Df(\mathbf{a})^{-1}(f(\mathbf{x}) - \mathbf{y})$ in this case looks like

$$g(x) = x - 1(\log x - y) = x + y - \log x,$$

so our iterates starting at 1 look like

$$1,\ 1 + y,\ 1 + 2y - \log(1 + y),\ 1 + 3y - \log(1 + y) - \log(1 + 2y - \log(1 + y)),\ \dots.$$

Here are some numerical values of these iterates at some specific values of y near $f(1) = 0$:

y	3-rd iterate	10-th iterate	100-th iterate
1	2.47096...	2.70898...	2.71828...
2	4.54005	6.51645...	7.38905...
0.5	1.62795...	1.64869...	1.64872...
-0.5	0.55966...	0.60892...	0.60653...

Sure enough, these values appear to be getting closer and closer to e^y, just as the proof of the inverse function theorem would lead you to believe! (So, hidden within the proof of the inverse function theorem is a method for approximating the inverse, different from the linear approximation idea using $Df^{-1} = (Df)^{-1}$ we mentioned last time.)

Here is another example. Take $f(x) = x^2$ near $x = 2$. We have $f'(2) = 4$, which is invertible, so again the inverse function theorem applies. The contraction g in this case is

$$g(x) = x - \tfrac{1}{4}(x^2 - y) = x + \tfrac{1}{4}y - \tfrac{1}{4}x^2,$$

so the iterates starting at 2 look like

$$2,\ 1 + \tfrac{1}{4}y,\ 1 + \tfrac{1}{2}y - \tfrac{1}{4}(1 + \tfrac{1}{4}y)^2,\ 1 + \tfrac{3}{4}y - \tfrac{1}{4}(1 + \tfrac{1}{4}y)^2 - \tfrac{1}{4}(1 + \tfrac{1}{2}y - \tfrac{1}{4}(1 + \tfrac{1}{4}y)^2),\ \ldots$$

Computing some of these numerically near $f(2) = 4$ gives

y	3-rd iterate	10-th iterate	100-th iterate
3	1.73236...	1.73205...	1.73205...
2	1.42089...	1.41421...	1.41421...
5	2.23626...	2.23606...	2.23606...
4.5	2.12133...	2.12132...	212132...

which indeed appear to be converging to $\sqrt{y} = f^{-1}(y)$. The math works out! (If you are familiar with *Newton's method* for approximating solutions of equations, what we are doing here is just a modification of that. Note that the converges appears to happen more quickly for $\sqrt{y}$ than it did for e^y, as evidence by the fact that the final columns in the second table are identical.)

Solving general systems. The inverse function theorem dealt with solving systems of (nonlinear) equations $f(\mathbf{x}) = \mathbf{y}$ with as many equations as variables. For our final topic, we consider more general "underdetermined" systems with at least as many variables as equations. Take for example the system

$$\begin{aligned} xu^2 + yv^2 + xy &= 11 \\ xv^2 + yu^2 - xy &= -1 \end{aligned}$$

for $(x, y, u, v) \in \mathbb{R}^4$. We can check that $(2, 3, 1, 1)$ is one solution, and we want to know if there are other solutions and how they can be obtained.

It is customary to rewrite our equations so that each is set to equal 0:

$$\begin{aligned} xu^2 + yv^2 + xy - 11 &= 0 \\ xv^2 + yu^2 - xy + 1 &= 0. \end{aligned}$$

Define $F : \mathbb{R}^4 \to \mathbb{R}^2$ to be the function

$$F(x, y, u, v) = (xu^2 + yv^2 + xy - 11, xv^2 + yu^2 - xy + 1),$$

so our system of equations is $F(x, y, u, v) = (0, 0)$. The key observation is that the *partial Jacobian matrix* $DF_{(u,v)}$, where we take the part of DF obtained by differentiating with respect to u and v only, is invertible at our solution $(2, 3, 1, 1)$:

$$DF_{(u,v)}(x, y, u, v) = \begin{bmatrix} 2xu & 2yv \\ 2yu & 2xv \end{bmatrix} \rightsquigarrow DF_{(u,v)}(2, 3, 1, 1) = \begin{bmatrix} 4 & 6 \\ 6 & 4 \end{bmatrix} \text{ is invertible.}$$

The *implicit function theorem* (to be stated shortly) then guarantees that near $(2, 3, 1, 1)$, there exist C^1 functions $u = u(x, y)$ and $v = v(x, y)$ such that $(x, y, u(x, y), v(x, y))$ satisfies

$$F(x, y, u(x, y), v(x, y)) = (0, 0),$$

thus expressing (u, v) (i.e., solving for u and v) in terms of (x, y) in the system $F(x, y, u, v) = (0, 0)$.

So, not only do we have more solutions (infinitely many more near $(2, 3, 1, 1)$), they can be written in terms of x and y alone. The equations $u(x, y), v(x, y)$, together with x, y, define "parametric equations" for points in the set of solutions of our system, at least for those points near $(2, 3, 1, 1)$. This says that, geometrically, the set of solutions of our system is a *surface* in $\mathbb{R}^4$, or what is modern language called a 2-dimensional *submanifold* of $\mathbb{R}^4$. (k-dimensional manifolds are objects described by C^1, say, parametric equations with k independent parameters.) If we were to introduce a third constraining equation into our system, we would instead be able to describe three of our variables in terms of the fourth, and the set of solutions in that case would be a curve (i.e., 1-dimensional manifold) since points can be parametrized by one parameter alone.

Implicit function theorem. Here, then, is the statement of the implicit function theorem. Suppose $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ is C^1 (more generally the domain can just be an open subset of $\mathbb{R}^m \times \mathbb{R}^n$) and that $(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^m \times \mathbb{R}^n$ satisfies $F(\mathbf{a}, \mathbf{b}) = \mathbf{0}$. Denote points in the domain by $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^m \times \mathbb{R}^n$. If the partial Jacobian matrix $DF_{\mathbf{y}}(\mathbf{a}, \mathbf{b})$ is invertible, then there exists an open set $W \subseteq \mathbb{R}^m$ containing $\mathbf{a}$ and a C^1 function $g : W \to \mathbb{R}^n$ such that $g(\mathbf{a}) = \mathbf{b}$ and

$$F(\mathbf{x}, g(\mathbf{x})) = \mathbf{0} \text{ for all } \mathbf{x} \in W.$$

Moreover, for each $\mathbf{x} \in W$, the point $\mathbf{y} = g(\mathbf{x})$ satisfying $F(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ is unique, so g is unique.

Let us digest this. The domain $\mathbb{R}^m \times \mathbb{R}^n$ of F is $\mathbb{R}^{m+n}$ (or an open subset thereof), but we write it as $\mathbb{R}^m \times \mathbb{R}^n$ in order to separate our variables as $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$ since these play different roles in the statement. The equation

$$F(\mathbf{x}, \mathbf{y}) = \mathbf{0}$$

is then a system of n equations (the number of components of F) in $m + n$ variables (so, more variables than equations) satisfied by at least $(\mathbf{a}, \mathbf{b})$. The claim of the implicit function theorem is that in $F(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ we can essentially (under an invertibility assumption) solve for $\mathbf{y}$ in terms $\mathbf{x}$ near $(\mathbf{a}, \mathbf{b})$, which is what $\mathbf{y} = g(\mathbf{x})$ for $\mathbf{x} \in W$ gives us. The equation $F(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ thus "implicitly defines" $\mathbf{y}$ as a function $\mathbf{x}$. The partial Jacobian matrix $DF_{\mathbf{y}}$ is taken with respect to the variables which we are trying to solve *for* in terms of the remaining variables. This is an $n \times n$ matrix, so the number of equations in our setup determines the number of variables we can expect to be able to express in terms of the others.

We will prove this next time, where the key step is to find a way to apply the inverse function theorem. This latter theorem cannot be applied directly since the dimensions in $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ do not match up, so we will need a function $\widetilde{F} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m \times \mathbb{R}^n$ where the dimensions do match up and which encodes F in an appropriate way.

Implicit derivatives. As is the case with the inverse function theorem, the implicitly defined function $g : W \to \mathbb{R}^n$ we get from the implicit function theorem cannot be described explicitly in general, but in fact we *can* give its derivative explicitly, just as what happened with the inverse function theorem. With the notation above, the claim is that

$$Dg(\mathbf{x}) = -DF_{\mathbf{y}}(\mathbf{x}, g(\mathbf{x}))^{-1} DF_{\mathbf{x}}(\mathbf{x}, g(\mathbf{x})) \text{ for all } \mathbf{x} \in W.$$

The two terms on the right are partial Jacobian matrices, the first $n \times n$ and the second $n \times m$, so the result is $n \times m$, which is the correct size for $Dg(\mathbf{x})$. The right side is explicit since we know what F is in the setup. A first clarification is that, while in the setup we only know that $DF_{\mathbf{y}}$ is invertible at $(\mathbf{a}, g(\mathbf{a})) = (\mathbf{a}, \mathbf{b})$, the fact that F is C^1 and that invertibility of matrices is an open condition (also used in the proof of the inverse function theorem) guarantees that $DF_{\mathbf{y}}(\mathbf{x}, g(\mathbf{x}))$ is also invertible for $\mathbf{x}$ close to $\mathbf{a}$, so we restrict W if need be so that this holds.

To prove this we differentiate both sides of the equation

$$F(\mathbf{x}, g(\mathbf{x})) = \mathbf{0}$$

characterizing the implicit function g. The right side differentiates to the zero matrix. For the left side, we consider the composition of

$$\mathbf{x} \mapsto (\mathbf{x}, g(\mathbf{x})) \quad \text{and} \quad (\mathbf{x}, \mathbf{y}) \mapsto F(\mathbf{x}, \mathbf{y}).$$

The derivatives of each of these are

$$\begin{bmatrix} I \\ Dg(\mathbf{x}) \end{bmatrix} \quad \text{and} \quad DF(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} DF_{\mathbf{x}}(\mathbf{x}, \mathbf{y}) & DF_{\mathbf{y}}(\mathbf{x}, \mathbf{y}) \end{bmatrix},$$

where in the first I denotes an $m \times m$ identity matrix (this comes from differentiating the first component of $(\mathbf{x}, g(\mathbf{x}))$ with respect to $\mathbf{x}$) and $Dg(\mathbf{x})$ is $n \times m$, and in the second we break DF up into the $n \times m$ piece $DF_{\mathbf{x}}$ where we differentiate with respect to $\mathbf{x}$ and the $n \times n$ piece $DF_{\mathbf{y}}$ where we differentiate with respect to $\mathbf{y}$. Thus by the chain rule, the derivative of $F(\mathbf{x}, g(\mathbf{x}))$ is

$$\begin{bmatrix} DF_{\mathbf{x}}(\mathbf{x}, g(\mathbf{x})) & DF_{\mathbf{y}}(\mathbf{x}, g(\mathbf{x})) \end{bmatrix} \begin{bmatrix} I \\ Dg(\mathbf{x}) \end{bmatrix} = DF_{\mathbf{x}}(\mathbf{x}, g(\mathbf{x})) + DF_{\mathbf{y}}(\mathbf{x}, g(\mathbf{x})) Dg(\mathbf{x}).$$

This should equal the zero matrix since $F(\mathbf{x}, g(\mathbf{x})) = \mathbf{0}$ for all $\mathbf{x}$, so we get

$$DF_{\mathbf{x}}(\mathbf{x}, g(\mathbf{x})) + DF_{\mathbf{y}}(\mathbf{x}, g(\mathbf{x})) Dg(\mathbf{x}) = 0,$$

and solving for $Dg(\mathbf{x})$ gives $Dg(\mathbf{x}) = -DF_{\mathbf{y}}(\mathbf{x}, g(\mathbf{x}))^{-1} DF_{\mathbf{x}}(\mathbf{x}, g(\mathbf{x}))$ as claimed. By considering entries of both sides here, we then get the partial derivatives of g described in terms of those of F.

Relation to linear algebra. We will do more with all this next time, but we finish now with the linear-algebraic fact of which the implicit function theorem is meant to be the local analog. Consider the usual method for solving an under-determined system of linear equations, and let us use the example

$$\begin{aligned} x_1 + 2x_2 + \ \ x_3 + \ \ 4x_4 - \ \ x_5 &= 5 \\ 3x_1 + 6x_2 + 5x_3 + 10x_4 - 4x_5 &= 14 \\ -x_1 - 2x_2 + \ \ x_3 - \ \ 2x_4 - 4x_5 &= -2. \end{aligned}$$

as a guide. To solve this you would use row operations on augmented matrix to reduce down to "echelon form", with the end result being that the solutions of the system above are the same as those of

$$\begin{aligned} x_1 + 2x_2 \qquad + \tfrac{9}{2}x_5 &= \tfrac{1}{2} \\ x_3 \quad - \tfrac{3}{2}x_5 &= \tfrac{1}{2} \\ x_4 - \quad x_5 &= 2. \end{aligned}$$

Then you express the "pivot" variables x_1, x_3, x_4 in terms of the "free" variables x_2, x_5 to get your general solution

$$x_1 = -2x_2 - \tfrac{9}{2}x_5 + \tfrac{1}{2},\ x_3 = \tfrac{3}{2}x_5 + \tfrac{1}{2},\ x_4 = x_5 + 1 \rightsquigarrow \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -2x_2 - \frac{9}{2}x_5 + \frac{1}{2} \\ x_2 \\ \frac{3}{2}x_5 + \frac{1}{2} \\ x_5 + 1 \\ x_5 \end{bmatrix}.$$

In retrospect, what is happening here is that the original equations "implicitly define" x_1, x_3, x_4 as functions of x_2, x_5, and the process of solving gives the explicit expressions for x_1, x_3, x_4 in terms of x_2, x_5. Let us rearrange the variables so that the pivot variables are all grouped together at the start and the free variables at the end; our original system then looks like

$$\begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix} = \mathbf{b}$$

where $\mathbf{y}$ are the pivot variables, A the coefficients of the pivot variables, $\mathbf{x}$ the free variables, B the coefficients of the free variables, and $\mathbf{b}$ the vector encoding the right sides. After expanding the left our system looks like

$$A\mathbf{y} + B\mathbf{x} = \mathbf{b}.$$

Since the $\mathbf{y}$'s are pivot variables, the "partial Jacobian matrix" of $F(\mathbf{x}, \mathbf{y}) = A\mathbf{y} + B\mathbf{x}$ with respect to $\mathbf{y}$, which is just A, is invertible, so we can solve for $\mathbf{y}$ above to get

$$\mathbf{y} = -A^{-1}B\mathbf{x} + A^{-1}\mathbf{b}.$$

This gives $\mathbf{y} = g(\mathbf{x})$ as a function of $\mathbf{x}$, which in the example above looks like

$$\begin{bmatrix} x_1 \\ x_3 \\ x_5 \end{bmatrix} = g\left(\begin{bmatrix} x_2 \\ x_5 \end{bmatrix}\right) = \begin{bmatrix} -2x_2 - \frac{9}{2}x_+4 + \frac{1}{2} \\ \frac{3}{2}x_4 + \frac{1}{2} \\ x_4 + 1 \end{bmatrix}.$$

This is precisely the content of the implicit function theorem, which seeks to do this same thing with non-linear equations instead. (In the linear case we can always get an explicit expression for the implicit g.) Note that the matrix $-A^{-1}B$ describing the linear term in the solution for $\mathbf{y}$ above is just $-(DF_{\mathbf{y}})^{-1}(DF_{\mathbf{x}})$ in this particular example where $DF_{\mathbf{y}} = A$ and $DF_{\mathbf{x}} = B$, so the expression derived for the implicit derivative before is also a reflection of what happens in the linear case. The upshot is that solving $F(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ "infinitesimally" using linear algebra gives rise a local solution.

Lecture 27: More on Implicit Functions

Warm-Up. Suppose $F : \mathbb{R}^3 \to \mathbb{R}$ is C^1 and that $DF(x, y, z) \neq 0$ for all $(x, y, z) \in \mathbb{R}^3$. We show that $F(x, y, z) = 0$ then defines a *surface* in $\mathbb{R}^3$, and will derive an equation for the tangent

plane to this surface at any point. (Part of the goal here is to give precise meaning to the terms "surface" and "tangent plane".) The end result is something you would have seen in a multivariable calculus course, so the intent here is to see how these old results are formally derived from the implicit function theorem. As we go, we will also work out the details in the example of a sphere $x^2 + y^2 + z^2 = 1$ to see what the abstraction says in a concrete, explicit case.

Since $DF = \left[\begin{smallmatrix} \frac{\partial F}{\partial x} & \frac{\partial F}{\partial y} & \frac{\partial F}{\partial z} \end{smallmatrix}\right]$ is nonzero at every point, at least one of $\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y}, \frac{\partial F}{\partial z}$ is nonzero at every point. Let us consider only the case where $\frac{\partial F}{\partial x}(x_0, y_0, z_0) \neq 0$ since the other cases are similar. Since $\frac{\partial F}{\partial x}(x_0, y_0, z_0)$ is invertible (this is a 1×1 matrix), the implicit function theorem says that $x = g(y, z)$ for a C^1 function g defined on an open set $W \subseteq \mathbb{R}^2$ containing (y_0, z_0). For $(y, z) \in W$—thus near (y_0, z_0)—we then have $F(g(y, z), y, z) = 0$, so $(g(y, z), y, z)$ gives points on our surface. The fact that we can describe points satisfying $F(x, y, z) = 0$ parametrically by C^1 functions in terms of two independent parameters is then what we take to be the definition of "surface"; we get parametric equations

$$x = g(y, z),\ y = y,\ z = z \text{ for } (y, z) \in W$$

for points on $F(x, y, z) = 0$ near where $\frac{\partial F}{\partial x} \neq 0$, and near points where $\frac{\partial F}{\partial y}$ or $\frac{\partial F}{\partial z}$ are nonzero instead we get parametric equations of the form

$$x = x,\ y = g(x, z),\ z = z \quad \text{or} \quad x = x,\ y = y,\ z = g(x, y)$$

respectively. The point is that we can find such equations—which might have to vary as we move from region to region—valid near all points on $F(x, y, z) = 0$.

In the case of a sphere, we use $F(x, y, z) = x^2 + y^2 + z^2 - 1$. We have $DF = \left[\begin{smallmatrix} 2x & 2y & 2z \end{smallmatrix}\right]$, and near points where $\frac{\partial F}{\partial x} = 2x \neq 0$, the setup above expresses x as a function of y and z. In this case we can actually do so explicitly:

$$x^2 + y^2 + z^2 - 1 = 0 \rightsquigarrow x = \pm\sqrt{1 - y^2 - z^2} =: g(y, z).$$

The $\pm$ is uniquely determined by whether x is positive or negative, which is why this only works when x (equivanlently $\frac{\partial F}{\partial x} = 2x$) is nonzero. If $x = 0$, then we have to shift to expressing y or z as functions of the other two variables in a similar way depending on which is guaranteed to be nonzero. (We cannot have all of x, y, z be zero since $(0, 0, 0)$ is not on the sphere.)

Now we use the part of the implicit function theorem that tells us the derivatives of the implicitly-defined function $x = g(y, z)$, still in the $\frac{\partial F}{\partial x} \neq 0$ case. We have

$$Dg(y, z) = -DF_x(g(y, z), y, z)^{-1} DF_{(y,z)}(g(y, z), y, z),$$

which looks like (let us suppress the point at which we are evaluating from the notation for now)

$$\begin{bmatrix} \frac{\partial g}{\partial y} & \frac{\partial g}{\partial z} \end{bmatrix} = -\left[\frac{\partial F}{\partial x}\right]^{-1} \begin{bmatrix} \frac{\partial F}{\partial y} & \frac{\partial F}{\partial z} \end{bmatrix}, \text{ so } \frac{\partial g}{\partial y} = -\frac{F_y}{F_x} \text{ and } \frac{\partial g}{\partial z} = -\frac{F_z}{F_x}.$$

(At the end we use subscript notation for partial derivatives for ease of reading.) In the sphere example, since we have $g(y, z) = \pm\sqrt{1 - y^2 - z^2}$ explicitly, we can compute these partials directly:

$$\frac{\partial g}{\partial y} = \frac{-2y}{\pm 2\sqrt{1 - y^2 - z^2}} = -\frac{y}{x} \quad \text{and} \quad \frac{\partial g}{\partial z} = \frac{-2z}{\pm 2\sqrt{1 - y^2 - z^2}} = -\frac{z}{x}$$

where we use $x = \pm\sqrt{1 - y^2 - z^2}$. This matches the general computation since in this case $F(x, y, z) = x^2 + y^2 + z^2 = -1$, so that $F_x = 2x, F_y = 2x$, and $F_z = 2z$. Note also that these can

be found the usual implicit differentiation techniques you would have seen in a previous calculus course: if x is a function of y, differentiating $x^2 + y^2 + z^2 = 1$ with respect to y gives

$$2x\frac{\partial x}{\partial y} + 2y + 0 = 0, \text{ so } \frac{\partial x}{\partial y} = -\frac{y}{x},$$

and similarly when differentiating with respect to z. (The matrix version $Dg = -(DF_{\mathbf{y}})^{-1}(DF_{\mathbf{x}})$ of the implicit derivative is just the full formal version of old-school "implicit differentiation".)

The implicit function $x = g(y, z)$ is C^1 and hence differentiable at (y_0, z_0), so

$$\lim_{(y,z)\to(y_0,z_0)} \frac{g(y,z) - g(y_0,z_0) - Dg(y_0,z_0)\left[\begin{smallmatrix} y-y_0 \\ z-z_0 \end{smallmatrix}\right]}{\sqrt{(y-y_0)^2 + (z-z_0)^2}} = 0.$$

The graph of the linear approxmation

$$x = g(y_0, z_0) + Dg(y_0, z_0)\left[\begin{smallmatrix} y-y_0 \\ z-z_0 \end{smallmatrix}\right]$$

appearing in the numerator is a plane, and indeed this is what we *define* to be the "tangent plane" to the graph of g, and hence to the surface $F(x, y, z) = 0$, at $(x_0 = g(y_0, z_0), y_0, z_0)$. (The definition of differentiable for a function $\mathbb{R}^2 \to \mathbb{R}^2$ is just the statement a valid tangent plane to the graph exists.) If we multiply out $Dg(y_0, z_0)\left[\begin{smallmatrix} y-y_0 \\ z-z_0 \end{smallmatrix}\right]$, we get

$$x = x_0 + \frac{\partial g}{\partial y}(y_0, z_0)(y - y_0) + \frac{\partial g}{\partial z}(y_0, z_0)(z - z_0)$$

as the tangent plane to the surface at (x_0, y_0, z_0).

In the sphere case, this becomes

$$x = x_0 - \frac{y_0}{x_0}(y - y_0) - \frac{z_0}{x_0}(y - y_0)$$

when $x_0 \neq 0$. Multiplying through by x_0 and rearranging gives

$$x_0(x - x_0) + y_0(y - y_0) + z_0(z - z_0) = 0,$$

which is indeed the typical equation for tangent planes to sphere. In the general case, with the implicit derivatives computed before, our tangent plane becomes

$$x = x_0 - \frac{F_y(x_0, y_0, z_0)}{F_x(x_0, y_0, z_0)}(y - y_0) - \frac{F_z(x_0, y_0, z_0)}{F_x(x_0, y_0, z_0)}(y - y_0).$$

(Near points where we instead express y in terms of x, z or z in terms of x, y, we get similar expressions.) After multiplying through by $F_x(x_0, y_0, z_0)$ and rearranging, we get

$$F_x(x_0, y_0, z_0)(x - x_0) + F_y(x_0, y_0, z_0)(y - y_0) + F_z(x_0, y_0, z_0)(z - z_0) = 0,$$

which is then the usual tangent plane equation to the surface $F(x, y, z) = 0$ at (x_0, y_0, z_0) given in a multivariable calculus course. The geometric interpretation given in such a course is that this equation says that the gradient vector $\nabla F = (F_x, F_y, F_z)$ is orthogonal to the tangent plane, and hence the surface, at a given point, so we have now derived this interpretation from first principles using the implicit function theorem.

Proof of implicit function theorem. Recall the setup of the implicit function theorem: F is a C^1 function $\mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$, $(\mathbf{x}, \mathbf{y}) = (\mathbf{a}, \mathbf{b})$ satisfies $F(\mathbf{a}, \mathbf{b}) = 0$, and $DF_{\mathbf{y}}(\mathbf{a}, \mathbf{b})$ is invertible. Define the function $\widetilde{F} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m \times \mathbb{R}^n$ by

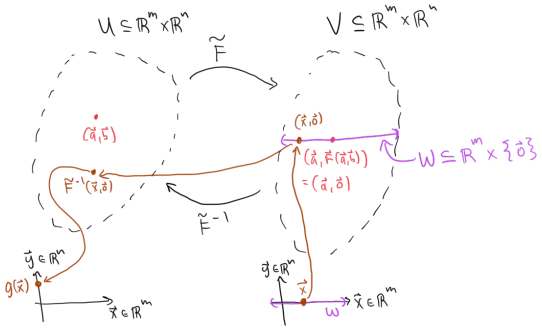
$$\widetilde{F}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, F(\mathbf{x}, \mathbf{y})).$$

Since $\widetilde{F}$ now maps between spaces of the same dimension, we can hope to apply the inverse function theorem, with the intuition being that the inverse will send $(\mathbf{x}, F(\mathbf{x}, \mathbf{y}))$ to $(\mathbf{x}, \mathbf{y})$, so that if we know $F(\mathbf{x}, \mathbf{y})$ and $\mathbf{x}$, we should be able to recover $\mathbf{y}$ in $F(\mathbf{x}, \mathbf{y})$ from knowing $\mathbf{x}$.

We have

$$D\widetilde{F}(\mathbf{a}, \mathbf{b}) = \begin{bmatrix} I & 0 \\ DF_{\mathbf{x}}(\mathbf{a}, \mathbf{b}) & DF_{\mathbf{y}}(\mathbf{a}, \mathbf{b}) \end{bmatrix},$$

where the I comes from differentiating the first components of $(\mathbf{x}, F(\mathbf{x}, \mathbf{y}))$ with respect to $\mathbf{x}$, 0 from differentiating these first $\mathbf{x}$ components with respect to $\mathbf{y}$, and the $DF_{\mathbf{x}}$ and $DF_{\mathbf{y}}$ from differentiating the remaining $F(\mathbf{x}, \mathbf{y})$-components of $(\mathbf{x}, F(\mathbf{x}, \mathbf{y}))$. Since $DF_{\mathbf{y}}(\mathbf{a}, \mathbf{b})$ is invertible, the $(m+n) \times (m+n)$ matrix above is invertible too, so the inverse function theorem applies. We get open sets $U, V \in \mathbb{R}^m \times \mathbb{R}^n$, with $(\mathbf{a}, \mathbf{b}) \in U$ and $(\mathbf{a}, F(\mathbf{a}, \mathbf{b})) = (\mathbf{a}, \mathbf{0})$ in V, on which $\widetilde{F}$ is invertible with C^1 inverse:



The open set W claimed to exist in the implicit function theorem, which will be the domain of the implicitly-defined function, is defined to be

$$W := \{\mathbf{x} \in \mathbb{R}^m \mid (\mathbf{x}, \mathbf{0}) \in V\}.$$

This contains the $\mathbf{x}$-coordinates of points in the intersection of $V \subseteq \mathbb{R}^m \times \mathbb{R}^n$ above with $\mathbb{R}^m \times \{\mathbf{0}\}$ (taking $\{\mathbf{0}\}$ in the second factor is what will enforce the $F(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ equation), and is open in $\mathbb{R}^m$ since V was open in $\mathbb{R}^m \times \mathbb{R}^n$. (The distance between two points in W is the same as the distance between the corresponding points in V since the $\mathbf{y}$-coordinate of each is $\mathbf{0}$, so a radius giving an open ball contained in V also gives an open ball contained in W.) To obtain the implicit function $\mathbf{y} = g(\mathbf{x})$, we must extract the $\mathbf{y}$-coordinate of a point $(\mathbf{x}, \mathbf{y})$ for which $F(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ where $\mathbf{x} \in W$, but by the construction of W this comes from applying the inverse of $\widetilde{F}$ to $(\mathbf{x}, \mathbf{0}) \in V$ and taking the $\mathbf{y}$-coordinate of the result. Thus, we define $g : W \to \mathbb{R}^n$ by

$$g(\mathbf{x}) = pr_{\mathbf{y}}(\widetilde{F}^{-1}(\mathbf{x}, \mathbf{0})) \text{ for } \mathbf{x} \in W,$$

where $pr_{\mathbf{y}}$ denotes the projection map $(\mathbf{x}, \mathbf{y}) \to \mathbf{y}$ onto the $\mathbf{y}$-coordinate. This g is C^1 since it is the composition of the C^1 maps $\mathbf{x} \mapsto (\mathbf{x}, \mathbf{0})$, $\widetilde{F}^{-1}$, and $pr_{\mathbf{y}}$, and satisfies

$$F(\mathbf{x}, g(\mathbf{x})) = \mathbf{0} \text{ for all } \mathbf{x} \in W$$

by the definition of $\widetilde{F}$ and definition of g using $\widetilde{F}^{-1}$.

The fact that $\mathbf{y} = g(\mathbf{x})$ is the unique element which satisfies $F(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ comes from injectivity of $\widetilde{F}$: if $\mathbf{y}'$ is another such element, then

$$\widetilde{F}(\mathbf{x}, \mathbf{y}') = (\mathbf{x}, F(\mathbf{x}, \mathbf{y}')) = (\mathbf{x}, \mathbf{0}) = (\mathbf{x}, F(\mathbf{x}, \mathbf{y})) = \widetilde{F}(\mathbf{x}, \mathbf{y}),$$

so $\mathbf{y}' = \mathbf{y} = g(\mathbf{x})$. In particular, $g(\mathbf{a}) = \mathbf{b}$ since $\mathbf{b}$ satisfies $F(\mathbf{a}, \mathbf{b}) = 0$, thus completing the proof.

Implicit implies inverse. In this course we have derived the implicit function theorem as a consequence of the inverse function theorem, which is what Rudin does and is a common thing to do. But probably just as common is to do it the other way around: prove the implicit function theorem via a contraction fixed-point argument, and derive the inverse function theorem as a consequence. The inverse and implicit function theorems are thus equivalent to one another, which is not surprising since they both amount to solving the same types of equations, with the only differences between the number of variables vs the number of equations.

So, assuming we have proved the implicit function theorem in some *other* way, let us derive the inverse function theorem as a consequence. Given f as in the setup of the inverse function theorem, set $F(\mathbf{x}, \mathbf{y}) := f(\mathbf{x}) - \mathbf{y}$, which the motivation being that $F(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ is then the inverse function setup $f(\mathbf{x}) = \mathbf{y}$ where we want to express $\mathbf{x}$ in terms of $\mathbf{y}$. With $\mathbf{a}$ as in the inverse function setup, we have $F(\mathbf{a}, f(\mathbf{a})) = f(\mathbf{a}) - f(\mathbf{a}) = \mathbf{0}$. Since

$$DF_{\mathbf{x}}(\mathbf{a}, f(\mathbf{a})) = Df(\mathbf{a})$$

(note that we differentiate with respect to $\mathbf{x}$ here instead of $\mathbf{y}$ since it is $\mathbf{x}$ that we are trying to solve for in terms of $\mathbf{y}$ as opposed to the other way around; so, the roles of $\mathbf{x}$ and $\mathbf{y}$ here are reversed compared to the original way we stated the implicit function theorem) is invertible, the implicit function theorem says that there exists an open $W \ni f(\mathbf{a})$ and a C^1 function g on W such that

$$F(g(\mathbf{y}), \mathbf{y}) = \mathbf{0} \text{ for all } \mathbf{y} \in W.$$

Using the definition of F, this becomes $f(g(\mathbf{y})) - \mathbf{y} = \mathbf{0}$, or

$$f(g(\mathbf{y})) = \mathbf{y} \text{ for all } \mathbf{y} \in W.$$

Thus g satisfies one of the requirements needed to be the inverse of f, with the remaining requirement being that $g(f(\mathbf{x}))$ should equal $\mathbf{x}$ for all $\mathbf{x}$.

Since $f(g(\mathbf{y})) = \mathbf{y}$ for all $\mathbf{y} \in W$, $g(\mathbf{y})$ is always in the preimage $f^{-1}(W)$ of W under f. This preimage is open since f is continuous, so g is a function $W \to f^{-1}(W)$ and f then goes $f^{-1}(W) \to W$. We have $\mathbf{a} \in f^{-1}(W)$ since $f(\mathbf{a}) \in W$, so $f^{-1}(W)$ and W are the sets $U \ni \mathbf{a}$ and $V \ni f(\mathbf{a})$ claimed to exist in the inverse function theorem. For any $\mathbf{x} \in f^{-1}(W)$, so $f(\mathbf{x}) \in W$ is in the domain of f, we have

$$F(\mathbf{x}, f(\mathbf{x})) = f(\mathbf{x}) - f(\mathbf{x}) = \mathbf{0}.$$

But the uniqueness property of the implicit function g, this means that $\mathbf{x}$ must be $g(f(\mathbf{x}))$, so

$$g(f(\mathbf{x})) = \mathbf{x} \text{ for all } \mathbf{x} \in f^{-1}(W).$$

Thus $g : W \to f^{-1}(W)$ is indeed the inverse of $f : f^{-1}(W) \to W$, so the proof is complete.

Lagrange multipliers. Finally, we give one of the other main applications (in addition to making geometric notions like "surface" or similar higher-dimensional objects precise) of the implicit function theorem, or rather we give the simplest case of another main application. Given a function

which we to optimize among points satisfying some constraint, the *method of Lagrange multipliers* characterizes the points at which local extreme values occur. In the simplest setup, we have a C^1 function $f : \mathbb{R}^2 \to \mathbb{R}$ and a C^1 "constraint" function $g : \mathbb{R}^2 \to \mathbb{R}$. The claim is that if among points satisfying the constraint $g(x, y) = 0$ the function f has a local extremum at (x_0, y_0), and if $Dg(x_0, y_0)$ is nonzero, then $Df(x_0, y_0)$ is a scalar multiple of $Dg(x_0, y_0)$:

$$Df(x_0, y_0) = \lambda Dg(x_0, y_0) \text{ for some } \lambda \in \mathbb{R}.$$

The method of Lagrange multipliers works by then finding the points satisfying such an equation, and the optimal points sought will be among these points. (You will prove the simplest "two constraint" version of this on the last homework, and, if so inclined, the most general version with any number of constraints as a "bonus" problem.)

In the case at hand, what we want is some λ that satisfies

$$\frac{\partial f}{\partial x}(x_0, y_0) = \lambda \frac{\partial g}{\partial x}(x_0, y_0) \quad \text{and} \quad \frac{\partial f}{\partial y}(x_0, y_0) = \lambda \frac{\partial g}{\partial y}(x_0, y_0).$$

If $Dg(x_0, y_0) = \begin{bmatrix} \frac{\partial g}{\partial x}(x_0,y_0) & \frac{\partial g}{\partial y}(x_0,y_0) \end{bmatrix}$ is nonzero, then at least one of $\frac{\partial g}{\partial x}(x_0, y_0)$ or $\frac{\partial g}{\partial y}(x_0, y_0)$ is nonzero; let us consider only the case were $\frac{\partial g}{\partial x}(x_0, y_0) \neq 0$ as the other case is very similar. Then we have no choice as to what λ must be in the first equation as we can solve for it explicitly:

$$\lambda = \frac{f_x(x_0, y_0)}{g_x(x_0, y_0)}$$

where we use subscript notation for partial derivatives. The claim, then, is that this *same* λ also satisfies the desired second equation. To show this we will find a way to relate the derivatives of f with to those of g using the derivative of an implicitly-defined function.

Since the "partial Jacobian matrix" $Dg_x(x_0, y_0) = \frac{\partial g}{\partial x}(x_0, y_0)$ is invertible (as a 1×1 matrix), the implicit function theorem says that in the constraint equation

$$g(x, y) = 0$$

we have $x = h(y)$ for some C^1 function h defined on a neighborhood of y_0, which also satisfies $x_0 = h(y_0)$. Moreover, the derivative of h with respect to y at y_0 is

$$h'(y_0) = -g_x(h(y_0), y_0)^{-1} g_y(h(y_0), y_0) = -\frac{g_y(x_0, y_0)}{g_x(x_0, y_0)}.$$

On the other hand, since $(h(y), y)$ satisfies the constraint equation for all y, the function $f(h(y), y)$ of y has derivative zero because it a local extremum at y_0 by assumption. By the chain rule, this thus gives

$$\begin{bmatrix} f_x(h(y_0), y_0) & f_y(h(y_0), y_0) \end{bmatrix} \begin{bmatrix} h'(y_0) \\ 1 \end{bmatrix} = 0, \text{ or } f_x(h(y_0), y_0)h'(y_0) + f_y(h(y_0), y_0) = 0.$$

Hence $f_y(x_0, y_0) = -f_x(x_0, y_0)h'(y_0)$, so substituting the other expression for $h'(y_0)$ we derived before gives

$$f_y(x_0, y_0) = -\left(-\frac{g_y(x_0, y_0)}{g_x(x_0, y_0)}\right) f_x(x_0, y_0) = \left(\frac{f_x(x_0, y_0)}{g_x(x_{,}y_0)}\right) g_y(x_0, y_0),$$

which is precisely the desired equation $f_y(x_0, y_0) = \lambda g_y(x_0, y_0)$! Thus, for this $\lambda = f_x(x_0, y_0)/g_x(x_0, y_0)$ we have $Df(x_0, y_0) = \lambda Dg(x_0, y_0)$ as claimed. (The implicit function theorem in this case is thus a tool used to be able to compare certain derivatives to others. The general case of Lagrange multipliers has "essentially" the same proof.)

Thanks for reading!