Math 320-2: Real Analysis
Northwestern University, Lecture Notes

Written by Santiago Cañez

These are notes which provide a basic summary of each lecture for Math 320-2, the second quarter of "Real Analysis", taught by the author at Northwestern University. The book used as a reference is the 4th edition of *An Introduction to Analysis* by Wade. Watch out for typos! Comments and suggestions are welcome.

Contents

Lecture 1: Convergence of Series

Today we started with a brief overview of the class, which will focus mainly on generalizing concepts from the previous quarter to other settings, such as the setting of *metric spaces*. We then moved on to talk about series of real numbers, in preparation for studying series of functions in a few weeks.

Definition. A *series* is an expression of the form $\sum_{n=1}^{\infty} a_n$, where (a_n) is a sequence of real numbers. Intuitively, we think of a series as an infinite sum. Given a series, its *sequence of partial sums* is the sequence (s_n) defined by

$$s_n = \sum_{k=1}^{n} a_k = a_1 + \cdots + a_n.$$

We say that the series $\sum a_n$ *converges* to $s \in \mathbb{R}$ if the sequence of partial sums (s_n) converges to s.

Important. A series is essentially thus a special type of sequence (namely, a sequence of partial sums), and questions about convergence of series are really questions about convergence of this sequence.

Geometric series. This is a standard example, which you would have seen in a previous calculus course. For a fixed $r \in \mathbb{R}$, the series $\sum_{n=0}^{\infty} r^n$ is a called a *geometric series*. The basic fact is that this series converges if $|r| < 1$ and diverges if $|r| \geq 1$. Indeed, the partial sums for $r \neq 1$ are concretely given by

$$s_n = 1 + r + r^2 + \cdots + r^n = \frac{1 - r^{n+1}}{1 - r},$$

and this sequence converges if and only if the r^{n+1} term in the numerator converges, which happens if and only if $|r| < 1$. In this case, $r^{n+1} \to 0$ so the sequence of partial sums converges to $\frac{1}{1-r}$ and so we write:

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}.$$

Cauchy criterion. As opposed to the case of a geometric series, in most instances it is not possible to compute the partial sums of a series directly, and thus we need another way to check for convergence. The point is that since convergence of a series boils down to convergence about a sequence, we can use what we already know about sequences from last quarter. In particular, we can say that a series converges if and only if its sequence of partial sums is Cauchy. Spelling this out in detail gives:

> A series $\sum a_k$ converges if and only if for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that for $m \geq n \geq N$, we have $|\sum_{k=n}^{m} a_k| = |a_n + \cdots + a_m| < \epsilon$.

This condition comes from applying the definition of a Cauchy sequence to $s_n = a_1 + \cdots + a_n$, in which case the $a_n + \cdots + a_m$ expression is the difference $s_m - s_{n-1}$ showing up in the Cauchy definition.

More examples. The Cauchy criterion can be used to show that all kinds of series converge, such as

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{n}, \quad \sum_{n=1}^{\infty} \frac{1}{n^2}, \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{\sin n}{n^2}.$$

Actually, we essentially already did these last quarter, only we didn't necessarily phrase them in terms of series but rather in terms of sequences of partial sums. In particular, we showed in class last quarter that the sequence

$$x_n = -1 + \frac{1}{2} - \frac{1}{3} + \cdots + \frac{(-1)^n}{n}$$

was Cauchy, and showing that

$$y_n = \frac{\sin 1}{1^2} + \frac{\sin 2}{2^2} + \cdots + \frac{\sin n}{n^2}$$

converges was on the final exam. These two facts say that the first and third series above converge, even though we may not know what they converge to. You can check the Lecture Notes for Math 320-1 to see the first example worked out, and the solutions to last quarter's final for the third.

Showing that

$$s_n = 1 + \frac{1}{2^2} + \cdots + \frac{1}{n^2}$$

converges was a homework problem last quarter, but for completeness (and to refresh your memory) we'll give the proof here, only now we'll rephrase it as using the Cauchy criterion above to show that the series $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges. Let $\epsilon > 0$ and choose $N \in \mathbb{N}$ such that $\frac{1}{N-1} < \epsilon$. If $m \geq n \geq N$, we have (with $a_n = \frac{1}{n^2}$):

$$\begin{aligned}
|a_n + \cdots + a_m| &= \left| \frac{1}{n^2} + \frac{1}{(n+1)^2} + \cdots + \frac{1}{(m-1)^2} + \frac{1}{m^2} \right| \\
&= \frac{1}{n^2} + \frac{1}{(n+1)^2} + \cdots + \frac{1}{(m-1)^2} + \frac{1}{m^2} \\
&\leq \frac{1}{n(n-1)} + \frac{1}{n(n+1)} + \cdots + \frac{1}{(m-2)(m-1)} + \frac{1}{(m-1)m} \\
&= \left(\frac{1}{n-1} - \frac{1}{n} \right) + \left(\frac{1}{n} - \frac{1}{n+1} \right) + \cdots + \left(\frac{1}{m-2} - \frac{1}{m-1} \right) + \left(\frac{1}{m-1} - \frac{1}{m} \right) \\
&= \frac{1}{n-1} - \frac{1}{m} \\
&\leq \frac{1}{n-1} \leq \frac{1}{N-1} < \epsilon.
\end{aligned}$$

Thus $\sum_{n=1}^{\infty} \frac{1}{n^2}$ satisfies the Cauchy criterion for convergence as claimed.

(Note that the inequality $\frac{1}{n^2} \leq \frac{1}{(n(n-1))} = \frac{1}{n-1} - \frac{1}{n}$ which is used here is one which I gave as a hint in the homework problem this example came from last quarter, and a similar hint was given in the final exam problem dealing with the third examples listed above. These are not inequalities which are obvious nor should they just "jump out" at you; indeed, these problems would have been very difficult without these hints.)

Observation and Harmonic series. Note that in all examples of convergent series $\sum a_n$ so far, it is true that the sequence a_n itself (not the sequence of partial sums) converges to 0. This is no accident, and is true for any convergent series: if $\sum a_n$ converges, then $a_n \to 0$. This makes sense intuitively: if the "infinite sum" $\sum a_n$ exists, it had better be true that the terms we are adding on at each step get smaller and smaller.

However, note that converse is not true: $a_n \to 0$ does NOT necessarily mean that $\sum a_n$ converges. The basic example of this is the so-called *harmonic series* $\sum \frac{1}{n}$. Here, $\frac{1}{n} \to 0$, but $\sum \frac{1}{n}$

does not converge. The book has one proof of this using integral comparisons, but here is another. Note that:

$$\begin{aligned}\frac{1}{3}+\frac{1}{4}\geq\frac{1}{4}+\frac{1}{4}&=\frac{1}{2}\\ \frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}\geq\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}&=\frac{1}{2}\\ \frac{1}{9}+\frac{1}{10}+\cdots+\frac{1}{15}+\frac{1}{16}\geq\underbrace{\frac{1}{16}+\cdots+\frac{1}{16}}_{8\text{ times}}&=\frac{1}{2},\end{aligned}$$

and so on. The point is that we can always group together terms in the sum $\sum\frac{1}{n}$ to get parts which are larger than $\frac{1}{2}$, and this implies that the sequence of partial sums is unbounded; in particular, the 2^n-th partial sum is larger than $\frac{n+2}{2}$. Since the sequence of partial sums is unbounded, it does not converge and hence the harmonic series diverges.

Other convergence tests. Check the book for other convergence tests, which may be familiar from a previous calculus course. In particular, we have the integral test, p-series test, comparison test, and limit comparison test. The alternating series test may also be familiar—we won't cover it explicitly but is in section 6.4 of the book.

From now on, feel free to use any of these tests when applicable. We'll use some of these tests from time to time, but not as much as you would have used them in a previous calculus course. Again, for us the main goal is to build up to studying series of functions, and so we'll really only use these tests as means towards that end.

Lecture 2: Lim Sup and Root Test

Today we spoke about the notion of the *lim sup* of a sequence and about the root test. The root test and the comparison test are probably the only series convergence tests we'll really care about later on, along with the Cauchy criterion.

Warm-Up. Suppose that (a_n) is a decreasing sequence that the series $\sum a_n$ converges. We claim that then the sequence (na_{2n}) converges to 0. (In class I first claimed instead that (na_n) converged to 0, but my argument didn't work. I'll point out below what the problem is. It is still true that (na_n) converges to 0, but you have to be more clever about how to show this.)

First note that since $\sum a_n$ converges, $a_n \to 0$ so since (a_n) is decreasing, all the a_n's must be nonnegative. Also note that since (a_n) is decreasing, we have:

$$na_{2n}=\underbrace{a_{2n}+a_{2n}+\cdots+a_{2n}}_{n\text{ times}}\leq a_n+a_{n+1}+\cdots+a_{2n}.$$

(This is the inequality which doesn't work when trying to instead show $na_n \to 0$: I tried to use $a_n+\cdots+a_n \leq a_n+\cdots+a_{2n}$, which assumes that a_n is smaller than everything coming after it, which requires that (a_n) be increasing instead of decreasing.) Since all the a_n's are nonnegative, the same inequality holds after taking absolute values of both sides.

Let $\epsilon > 0$. Since $\sum a_n$ converges, by the Cauchy criterion for convergence there exists N such that

$$\left|\sum_{k=n}^{m}a_k\right| < \epsilon \text{ for } m \geq n \geq N.$$

In particular, for $n \geq N$ we have $\left|\sum_{k=n}^{2n} a_k\right| < \epsilon$. Thus if $n \geq N$,

$$|na_{2n} - 0| = |na_{2n}| \leq \left|\sum_{k=n}^{2n} a_k\right| < \epsilon,$$

so $na_{2n} \to 0$ as claimed. (As mentioned above, it is also true that $na_n \to 0$. To show this you can first show that $na_{2n+1} \to 0$ by a similar method as above, and then use this to show that $(2n+1)a_{2n+1} \to 0$. This together with the fact that $2na_{2n} \to 0$ will imply the convergence of na_n. We won't give all details here—the problem as given in the Warm-Up is enough to get across some key ideas.)

Lim sup. Given a sequence (a_n), we define its *lim sup* (more formally called its *limit superior*) by:

$$\limsup_{n\to\infty} a_n = \lim_{n\to\infty} \left(\sup_{k\geq n} a_k\right),$$

that is, the lim sup of (a_n) is the ordinary limit of the sequence given by $\sup_{k\geq n} a_k$. (This is a sequence of supremums, hence the name "lim sup".) This sequence is defined by first taking the supremum of all terms in the original sequence, then the supremum of all terms except the first, then the supremum of all terms except the first two, and so on.

The point is that this lim sup *always* exists, even when the original sequence does not converge. Indeed, denoting the supremums we are using by b_n:

$$b_n = \sup_{k\geq n} a_k,$$

note that $b_1 \geq b_2 \geq b_3 \geq \ldots$ since at each step we are taking the supremum of fewer things, so the supremum either stays the same or gets smaller. Hence (b_n) is a decreasing sequence. If (a_n) is bounded, all b_n's exist as real numbers so in this case (b_n) is decreasing and bounded and thus converges; if (a_n) is unbounded, all b_n's are ∞, so in this case we say that $\lim b_n = \lim supa_n = \infty$. Thus, as claimed, $\limsup a_n$ always exists either as a real number or as ∞.

Examples. Consider the sequence $a_n = 3 + (-1)^{n+1} + \frac{1}{n}$, whose terms look like:

$$4 + \frac{1}{1}, 2 + \frac{1}{2}, 4 + \frac{1}{3}, 2 + \frac{1}{4}, 4 + \frac{1}{5}, \ldots.$$

Using the same notation as above, we have:

$$\begin{aligned} b_1 &= \sup_{k\geq 1} a_k = 4 + \frac{1}{1} \\ b_2 &= \sup_{k\geq 2} a_k = 4 + \frac{1}{3} \\ b_3 &= \sup_{k\geq 3} a_k = 4 + \frac{1}{3} \\ b_4 &= \sup_{k\geq 4} a_k = 4 + \frac{1}{5} \end{aligned}$$

and so on. Hence $\lim b_n = 4$, so $\limsup a_n = 4$.

For the sequence $x_n = \frac{(-1)^n}{n}$ which looks like $-1, \frac{1}{2}, -\frac{1}{3}, \frac{1}{4}, -\frac{1}{5}, \ldots$, we have:

$$b_1 = \sup_{k \geq 1} a_k = \frac{1}{2}$$

$$b_2 = \sup_{k \geq 2} a_k = \frac{1}{2}$$

$$b_3 = \sup_{k \geq 3} a_k = \frac{1}{4}$$

$$b_4 = \sup_{k \geq 4} a_k = \frac{1}{4}$$

$$b_5 = \sup_{k \geq 5} a_k = \frac{1}{6}$$

and so on, so $\limsup x_n = \lim b_n = 0$.

Key properties. Note that in the second examples above, (x_n) itself converges to 0, so actually it is no coincidence that $\limsup x_n = 0$ as well: if $\lim a_n$ exists, then $\limsup a_n = \lim a_n$. In other words, the lim sup of a convergent sequence is the ordinary limit. The point again is that lim sup exists for *all* sequences, even ones which don't converge.

Here are two other key properties of lim sups: if $\limsup a_n < x$, then $a_n < x$ for n large enough; and if $\limsup a_n > x$, then $a_n > x$ for infinitely many n. Both of these come from properties of convergent sequences (applied to the sequence of b_n's) and properties of supremums. Check the book for full details, and for the proof that $\limsup = \lim$ for a convergent sequence.

Root Test. Now with the notion of the lim sup of a sequence we can state a powerful series convergence test, which you no doubt saw in a calculus course, although probably not phrased in terms of lim sups.

Suppose that $\sum a_n$ is a series and set $r := \limsup |a_n|^{1/n}$, which as stated before always exists. Then: if $r < 1$, the series $\sum a_n$ converges absolutely; and if $r > 1$, the series $\sum a_n$ diverges. (We consider $r = \infty$ to be larger than 1.) If $r = 1$, then the root test gives us no information.

(Recall that to say $\sum a_n$ converge *absolutely* means that the series $\sum |a_n|$ converges; we'll come back to the notion of absolute convergence and its importance next time.)

The proof uses properties of lim sup given above as well as properties of geometric series. It's in the book, but let's go ahead and reproduce it here for completeness, and to point out one subtlety I missed in class. First suppose that $r < 1$ and pick $r < x < 1$. Since $\limsup |a_n|^{1/n} < x$, $|a_n|^{1/n} < x$ for large enough n, and hence $|a_n| < x^n$ for large enough n. Since $\sum x^n$ converges (because $x < 1$), the comparison test implies that $\sum |a_n|$ converges as well, so $\sum a_n$ converges absolutely as claimed. If instead $r > 1$ and we pick $r > x > 1$, then $\limsup |a_n|^{1/n} > x$ implies $|a_n|^{1/n} > x$ for infinitely many n, and thus $|a_n| > x^n$ for infinitely many n. Since $x > 1$, x^n is unbounded and hence this means that $|a_n|$ cannot converge to 0. Therefore neither does a_n, so $\sum a_n$ diverges.

(In class after getting $|a_n| > x^n$, I went on to say that since $\sum x^n$ diverges, so does $\sum |a_n|$ by comparison and stopped there. However, the divergence of $\sum |a_n|$ does NOT imply the divergence of $\sum a_n$ since it is very well possible for $\sum |a_n|$ to diverge but for $\sum a_n$ to converge, as we'll see next time. So, my argument in class was not complete, but the above proof works fine.

Ratio Test. I didn't state the ratio test in class, but I'll go ahead and state it here for reference. Note, however, that although similar in spirit to the root test, the ratio test is actually weaker, in particular since it only applies when $\lim \frac{|a_{n+1}|}{|a_n|}$ exists. For this reason, the root test will be more useful to use later on.

Suppose that $\sum a_n$ is a series of nonzero numbers and that $r := \lim \frac{|a_{n+1}|}{|a_n|}$ exists. (Note that r could be ∞.) Then: if $r < 1$, $\sum a_n$ converges absolutely; while if $r > 1$, $\sum a_n$ diverges. As with the root test, $r = 1$ gives no information about the convergence or divergence of $\sum a_n$.

Lecture 3: Absolute Convergence

Today we spoke about absolutely convergent series, where the key point is that these are the series for which rearranging terms does not affect the value of the sum. This will be important later on when we talk about power series and other series of functions.

Warm-Up. We determine the values of $p \in \mathbb{R}$ for which $\sum_{n=1}^{\infty} \frac{n^p}{p^n}$ converges absolutely using the root test. (I guess p should be nonzero so that the denominator actually makes sense. We have for $a_n = \frac{n^p}{p^n}$:

$$|a_n|^{1/n} = \left(\frac{n^p}{|p|^n}\right)^{1/n} = \frac{(n^{1/n})^p}{|p|}.$$

Now, $n^{1/n} = e^{\frac{1}{n}\log n}$, and since the exponent here converges to 0 (say by L'Hopital's rule) and $x \mapsto e^x$ is continuous, we have that $n^{1/n} \to e^0 = 1$. Thus

$$\limsup |a_n|^{1/n} = \lim |a_n|^{1/n} = \frac{1}{p}.$$

Hence according to the root test, if $|p| < 1$ the given series diverges since $\frac{1}{|p|} > 1$, while if $|p| > 1$ the given series converges absolutely since $\frac{1}{|p|} < 1$.

For $p = 1$, the given series is $\sum n$, which diverges, while if $p = -1$ the given series is $\sum \frac{(-1)^n}{n}$, which converges conditionally but not absolutely. Indeed, that $\sum \frac{(-1)^n}{n}$ converges can be seen using the Cauchy criterion or more simply by using the alternating series test; it does not converge absolutely since $\sum \frac{1}{n}$ diverges.

Manipulating series. Some standard arithmetic operations for series make sense, while others do not. For instance, if $\sum a_n$ and $\sum b_n$ are each convergent, then $\sum(a_n + b_n)$ converges and

$$\sum(a_n + b_n) = \sum a_n + \sum b_n,$$

and if $\sum a_n$ converges and $c \in \mathbb{R}$, then $\sum(ca_n)$ converges and

$$\sum(ca_n) = c\sum a_n.$$

In particular, this second fact should be viewed as an infinite sum version of the usual distributive property: $c(a_1 + a_2 + \cdots) = ca_1 + ca_2 + \cdots$ and so on.

However, note what happens if we try to "multiply" two series. To be precise, say we want to multiply two *power* series:

$$\sum_{n=0}^{\infty} a_n x^n \text{ and } \sum_{n=0}^{\infty} b_n x^n.$$

Writing this out, we get something like:

$$(a_0 + a_1x + a_2x^2 + \cdots)(b_0 + b_1x + b_2x^2) = a_0b_0 + a_0b_1x + a_0b_2x^2 + \cdots + a_1b_0x + a_1b_1x^2 + \cdots$$

Going by our intuition with finite sums, we might expect that when multiplying these infinite sums together we should be able to group like-terms, so that we get:

$$a_0b_0 + (a_0b_1 + a_1b_0)x + (a_0b_2 + a_1b_1 + a_2b_0)x^2 + \cdots .$$

However, to do this requires that we can *rearrange* terms in our original sum, since in order to "group" say the a_0b_1x and a_1b_0x terms, we need to move the a_1b_0x term to the left a bunch of times until we have $a_0b_1x + a_1b_0x$; similarly for other terms we would want to regroup.

This doesn't seem like a big deal given our experience with finite sums, but it turns out that this is a big deal when working with infinite sums: in fact, rearranging the terms of a series does not affect its convergence *if and only if* the series is absolutely convergent! Rearranging the terms of a conditionally convergent series (definition to come) can indeed affect its convergence, as we'll see. In the case of power series, it will thus be important to know that when a power series does converge it actually does so absolutely.

When rearranging doesn't work. To show what can go wrong when rearranging the terms of a non-absolutely convergent series, consider the series

$$-1 + \frac{1}{2} - \frac{1}{3} + \frac{1}{4} - \frac{1}{5} + \frac{1}{6} \cdots = \sum_{n=1}^{\infty} \frac{(-1)^n}{n},$$

which actually converges to $S = -\ln 2$. (This can be justified using Taylor series, which we'll look at later.)

Now, multiplying through by 2 gives a series which converges to $2S$:

$$2S = -2 + 1 - \frac{2}{3} + \frac{2}{4} - \frac{2}{5} + \frac{2}{6} + \cdots .$$

In this new sum, rearrange and regroup terms which have the same denominator: the -2 and 1 combine to give -1, the $\frac{2}{4}$ gives $\frac{1}{2}$, the $-\frac{2}{3}$ and $\frac{2}{6}$ combine to give $-\frac{1}{3}$, and so on. (In general, a term of the form $\frac{1}{2(2k)}$ in the original sum gives $\frac{2}{2(2k)} = \frac{1}{2k}$ in the new sum, and terms of the form $-\frac{1}{2k+1}$ and $\frac{1}{2(2k+1}$ in the original sum combine to give $-\frac{2}{2k+1} + \frac{2}{2(2k=1)} = -\frac{1}{2k+1}$ in the new one.) The key point is that after these regroupings the resulting sum

$$-1 + \frac{1}{2} - \frac{1}{3} + \frac{1}{4} - \frac{1}{5} + \frac{1}{6} \cdots$$

is the *same* as the original sum, which converged to S. Thus we would seem to have $2S = S$, implying $S = 0$, which is nonsense since we said earlier that $S = -\ln 2$.

The problem is that once we started rearranging and regrouping terms, there was no reason to expect that the resulting series would still converge to the same value as the non-rearranged sum, so while the rearranged sum does indeed converge to S it no longer must have the value $2S$ as well. The series in this case is conditionally convergent, which as we'll see explains why rearranging terms does not necessarily preserve the value of the sum.

Definition. The series $\sum a_n$ is said to *converge absolutely* if $\sum |a_n|$ converges. If $\sum a_n$ converges but not absolutely, we say it is *conditionally convergent.*

Absolute convergence implies convergence. If a series $\sum a_n$ is absolutely convergent, then it converges in the ordinary sense: since $|a_n + \cdots + a_m| \leq |a_n| + \cdots + |a_m|$ for $m \geq n$, the

Cauchy criterion for the series $\sum |a_n|$ implies the Cauchy criterion for the series $\sum a_n$. So, absolute convergence is *stronger* than ordinary convergence.

Theorem. And now the key point, which is why we care about absolutely convergent series: if $\sum a_n$ converges absolutely and $\sum b_n$ is any rearrangement of $\sum a_n$, then $\sum b_n$ is also convergent and $\sum a_n = \sum b_n$. So, rearranging the terms of an absolutely convergent series still gives a convergent series, which converges to the same value as the non-rearranged series.

To be clear, saying that $\sum b_n$ is a rearrangement of $\sum a_n$ means that the b_n's are simply the a_n's, only occurring in a possibly different order than they do in the original $a_1 + a_2 + a_3 + \cdots$ sum. The book has a proof of this, but here is a hopefully slightly simpler to follow proof; the book's proof also uses a subtle point, which I'll elaborate on after my proof.

Proof of Theorem. Let $A_n = a_1 + \cdots + a_n$ denote the partial sums of $\sum a_n$ and $B_n = b_1 + \cdots + b_n$ the partial sums of $\sum b_n$. Furthermore, let $A = \sum a_n$ and $B = \sum b_n$. Let $\epsilon > 0$ and pick $N \in \mathbb{N}$ such that

$$\sum_{k=n}^{m} |a_k| < \frac{\epsilon}{2} \text{ for } m \geq n \geq N \text{ and } |A_N - A| < \frac{\epsilon}{2},$$

which we can do in the first case since $\sum |a_n|$ converges and in the second since $A_n \to A$. (Last quarter we would have said there exists an index N_1 guaranteeing the first condition and an index N_2 guaranteeing the second, so we take $N = \max\{N_1, N_2\}$. We should hopefully be used to such things by now that we can avoid explicitly phrasing this in terms of a maximum of two indices.)

Now, pick $M \in \mathbb{N}$ large enough so that

$$\{a_1, \dots, a_N\} \subseteq \{b_1, \dots, b_M\}.$$

In other words, the specific terms $a_1, \dots, a_N$ occur somewhere among the b_n's, so we are going out as far among the b_n's as need to to make sure we are past all of these a_n's, which is possible since there are only finitely many a_n's among $a_1, \dots, a_N$. For $\ell < M$ then, the partial sum $B_\ell = b_1 + \cdots + b_\ell$ thus includes each of $a_1, \dots, a_N$ among its terms, so the difference

$$B_\ell - A_N$$

only includes a_n's within the range in the original series $\sum a_n$ where $n > N$. Hence

$$\begin{aligned} |B_\ell - A_N| &= \left|\sum (a_n\text{'s not among } a_1, \dots, a_N)\right| \\ &\leq \sum |a_n\text{'s not among } a_1, \dots, a_N| \\ &\leq \sum_{k=N}^{\ell} |a_k| \end{aligned}$$

where in the second step we use the triangle inequality and in the third the fact that adding on more nonnegative terms to a nonnegative sum can only make it larger. This last sum involves only terms to which the Cauchy criterion for $\sum a_n$ applies, so this last sum is less than $\frac{\epsilon}{2}$.

Thus for $\ell > M$, we have:

$$|B_\ell - A| = |B_\ell - A_N + A_N - A| \leq |B_\ell - A_N| + |A_N - A| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

so $B_\ell \to A$. Hence $\sum b_n$ converges to $A = \sum a_n$ as claimed. □

Remark. This is pretty much the same as the book's proof, only that the book doesn't phrase this in terms of the Cauchy criterion for $\sum a_n$ and rather uses an inequality of the form

$$\sum_{k=n}^{\infty} |a_k| < \frac{\epsilon}{2}.$$

Then, at some point the book uses something like

$$\left|\sum_{k=n}^{\infty} a_k\right| \leq \sum_{k=n}^{\infty} |a_k|,$$

which is essentially the triangle inequality only now applied to an *infinite* sum, whereas so far we've only spoken about the triangle inequality as applied to finite sums. This is fine, but requires justification which the book glosses over. Here it is: since the partial sums $a_n + a_{n+1} \cdots + a_m$ converge, the absolute values $|a_n + a_{n+1} + \cdots + a_m|$ converge by continuity of the absolute value function, and since $|a_n + a_{n+1} + \cdots + a_m| \leq |a_n| + |a_{n+1}| + \cdots + |a_m|$, we get $|\sum_{k=n}^{\infty} a_k| \leq \sum_{k=n}^{\infty} |a_k|$ as claimed. (Note how the continuity of the absolute value function was important here!)

Theorem. (We actually covered this next fact in the following lecture in class, but I'll put it here in the notes where it fits better.) And now, we come to the fact that when rearranging the terms of a conditionally convergent series, all kinds of crazy things can happen: not only can some rearrangements give different values for the sum (in fact, *any* real number can be obtained as such a sum), other rearrangements might not even converge.

Here is the statement. Suppose that $\sum a_n$ is conditionally convergent. Then for any $x \in \mathbb{R}$, there exists a rearrangement of $\sum a_n$ which converges to x. Moreover, there exists a rearrangement which diverges to ∞ and there exists a rearrangement which diverges to $-\infty$.

Proof. We will use the following facts, whose proofs are in the book: if $\sum a_n$ is conditionally convergent, there are infinitely many a_n which are positive and infinitely many which are negative, and moreover the series of all positive terms diverges to ∞ while the series of all negative terms diverges to $-\infty$. We'll denote the positive terms in $\sum a_n$ by b_n's and the negative terms by c_n's.

Begin by adding together positive terms $b_1, b_2, \ldots, b_{n_1}$ until we first get something larger than x; so,

$$x < b_1 + \cdots + b_{n_1} \text{ but } b_1 + \cdots + b_{n_1 - 1} \leq x.$$

(Note that if $x \leq 0$, then we only need the first b_1 to get something larger than x, and that if by adding together b_n's we actually hit x on the nose, we keep adding to get a sum which is *strictly* larger than x. Also, this is possible to do since the sum of all positive terms diverges to ∞.) Set $R_1 = b_1 + \cdots + b_{n_1}$. Since

$$b_1 + \cdots + b_{n_1 - 1} \leq x < b_1 + \cdots + b_n,$$

we have

$$|R_1 - x| < |(b_1 + \cdots + b_n) - (b_1 + \cdots + b_{n_1 - 1}| = b_{n_1}.$$

Next, starting adding on *negative* terms $c_1, \ldots, c_{n_2}$ until we first get something smaller than x; so,

$$(b_1 + \cdots + b_{n_1}) + (c_1 + \cdots + c_{n_2}) < x \text{ but } x \leq (b_1 + \cdots + b_{n_1}) + (c_1 + \cdots + c_{n_2 - 1}).$$

Set $R_2 = (b_1 + \cdots + b_{n_1}) + (c_1 + \cdots + c_{n_2}) = R_1 + (c_1 + \cdots + c_{n_2})$. Then

$$|R_2 - x| < |c_{n_2}|.$$

Continue on in this manner, first adding on positive terms until we first get

$$x < R_1 + R_2 + (b_{n_1+1} + \cdots + b_{n_2}) \quad \text{(call this sum on the right } R_1 + R_2 + R_3\text{)},$$

then adding on negative terms until we first get

$$R_1 + R_2 + R_3 + (c_{n_1+1} + \cdots + c_{n_2}) < x,$$

and so on, denoting the new pieces of sums we're constructing at each step by R_n's. (So for n odd, R_n is obtained from R_{n-1} by adding on positive terms while for n odd it is obtained by adding on negative terms.) Note that

$$|R_3 - x| < b_{n_2}, \; |R_4 - x| < |c_{n_2}|,$$

and in general $|R_n - x|$ is bounded by the final term we add on, using an absolute value to deal with the cases where this final term is one of the negative c_n's.

The idea is that we are constructing sums that alternate "bouncing" to the right and to the left of x, and the key point is that as we go on the terms we are adding on are getting smaller and smaller since the a_n's converge to 0 given that $\sum a_n$ converges—the b_n's and c_n's after all are made up of a_n's. Given $\epsilon > 0$, pick an index large enough so that all b_{n_i}'s and $|c_{n_i}|$'s are smaller than ϵ (again possible since $a_n \to 0$), and then past this index we have

$$|R_k - x| < \text{ (either } b_{n_k} \text{ or } |c_{n_k}|\text{) } < \epsilon,$$

so $R_k \to x$. The R_k's are the partial sums of a series obtained by rearranging the terms of $\sum a_n$, and thus this rearrangement converges to x.

To get a rearrangement which diverges to ∞, start by adding together enough positive terms to get a sum larger than 100, then add one a single negative term, then add on positive terms to get a sum larger than 1000, then add another negative term, then positive ones to get a sum larger than 10000, and so on. Since at each step the sum of the positive terms we include are getting arbitrary large while we add on only a single negative term, this sum will diverge to ∞. The same process only flipping the roles of the positive and negative terms gives a series which diverges to $-\infty$. □

Important. Rearranging the terms of a series does not affect that the fact that it converges nor what the value of its sum if and only if the series is absolutely convergent. Thus, the adding together of infinitely many numbers is guaranteed to be commutative only for absolutely convergent sums.

Lecture 4: Sequences of Functions

Today we started talking about sequences of functions, and what it might mean for such a sequence to converge. This is one of the main topics of the course, and will provide the first sense in which can generalize concepts from last quarter to other settings.

Warm-Up. Touching on something we looked at last time, say we want to define what it means to take the product of two convergent series $\sum a_n$ and $\sum b_n$. We would hope that this product can itself be considered as a series:

$$\left(\sum_{n=0}^{\infty} a_n\right)\left(\sum_{n=0}^{\infty} b_n\right) = \sum_{n=0}^{\infty} c_n$$

for some numbers c_n. Trying to multiply out the expression

$$(a_0 + a_1 + a_2 + a_3 + \cdots)(b_0 + b_1 + b_2 + b_3 + \cdots)$$

as you normally would using the distributive property, we see that one nice way of rewriting this product is by grouping individual products terms which have indices adding up to the same value; for instance, we might try

$$(a_0 + a_1 + a_2 + \cdots)(b_0 + b_1 + b_2 + \cdots) = a_0b_0 + (a_0b_1 + a_1b_0) + (a_0b_2 + a_1b_1 + a_2b_0) + \cdots,$$

which we actually saw last time when trying to multiply together power series. Thus we might try to define

$$\left(\sum_{n=0}^{\infty} a_n\right)\left(\sum_{n=0}^{\infty} b_n\right) = \sum_{n=0}^{\infty} c_n, \text{ where } c_n = \sum_{k=0}^{n} a_k b_{n-k}.$$

However, since coming up with this expression requires rearranging terms, we have to be careful about whether or not such rearrangements are possible. Indeed, we claim that for conditionally convergent series this definition does not necessarily work.

In particular, take $\sum a_n$ and $\sum b_n$ to be the same series where $a_n = b_n = \frac{(-1)^n}{\sqrt{n+1}}$. We claim that the series $\sum c_n$ obtained by defining c_n as above does not converge,so that this product $(\sum a_n)(\sum b_n)$ is not actually defined, even though $\sum a_n = \sum b_n$ actually converge. Using the expression for c_n above we have:

$$c_n = a_0b_n + a_1b_{n-1} + \cdots + a_{n-1}b_1 + a_nb_0 = \sum_{k=0}^{n} \frac{(-1)^k}{\sqrt{k+1}} \frac{(-1)^{n-k}}{\sqrt{n-k+1}} = \sum_{k=0}^{n} \frac{(-1)^n}{\sqrt{(k+1)(n-k+1)}}.$$

Since $0 \leq k \leq n$, we have $(k+1)(n-k+1) \leq (n+1)(n+1)$, so that each denominator in the above expression is less than or equal to $n+1$. Thus

$$c_n = \sum_{k=0}^{n} \frac{(-1)^n}{\sqrt{(k+1)(n-k+1)}} \geq \sum_{k=0}^{n} \frac{(-1)^n}{n+1} = (-1)^n,$$

since at the end we are left with a sum not involving the indexing variable k, meaning that we are adding together the fixed expression $\frac{(-1)^n}{n+1}$ to itself $n+1$ times. For n even, this gives that $c_n \geq 1$, so $c_n \not\to 0$ and thus $\sum c_n$ does not converge as claimed.

Remark. Again, the issue above is that we tried to rearrange the terms of an infinite sum, and for the conditionally convergent series $\sum \frac{(-1)^n}{\sqrt{n+1}}$ this rearrangement affects the convergence. However, if we start with absolutely convergent series $\sum a_n$ and $\sum b_n$, it turns out that the series $\sum c_n$ defined as above does converge, so in this case

$$\left(\sum_{n=0}^{\infty} a_n\right)\left(\sum_{n=0}^{\infty} b_n\right) = \sum_{n=0}^{\infty} c_n$$

makes sense. The series $\sum c_n$ in this case is called the *Cauchy product* of $\sum a_n$ and $\sum b_n$, and in fact it turns out that this works even if only one of $\sum a_n$ or $\sum b_n$ is absolutely convergent and the other only conditionally convergent.

Pointwise convergence. A sequence (f_n) of functions is an infinite list

$$f_1,\ f_2,\ f_3,\ f_4,\ \ldots$$

of functions $f_n : E \to \mathbb{R}$, where E is some domain in $\mathbb{R}$. (So, all functions in the sequence are considered to have the same domain.) What should such a sequence converge to? Since sequences of numbers converge to numbers, we should expect that sequences of functions converge to functions.

Here is our first attempt at defining what it means for a sequence of functions to converge: a sequence (f_n) of functions is said to *converge pointwise* to the function f if

$$\lim_{n\to\infty} f_n(x) = f(x) \text{ for any } x \in E,$$

where E is the common domain of all functions in question. Here is what this definition says: taking a fixed $x \in E$ and plugging it into all of our functions gives a sequence $(f_n(x))$ of *numbers*, and the definition says that for any x this sequence of numbers should converge to the number given by $f(x)$, which is the value of the limit function at x. We call f the *pointwise limit* of the sequence (f_n).

Remark having nothing to do with mathematics. Fun fact: my computer was autocorrecting "pointwise" to "pointless", so the first draft of these notes had numerous references to "pointless limits". I wonder if this reflects some general opinion of Apple Inc. towards real analysis.

Example 1. Consider the sequence (f_n) of functions $f_n : \mathbb{R} \to \mathbb{R}$ where the n-th function is defined by

$$f_n(x) = \frac{1}{n}\sin x.$$

So, this is the sequence of functions given by

$$\sin x,\ \frac{1}{2}\sin x,\ \frac{1}{3}\sin x,\ \frac{1}{4}\sin x,\ \ldots.$$

For a fixed $x \in \mathbb{R}$, $\sin x$ is some fixed number and hence

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \frac{1}{n}\sin x = 0 \text{ for any } x \in \mathbb{R}.$$

Thus the sequence (f_n) converges pointwise to the constant function which has the value 0 everywhere, since this pointwise limit must satisfy $f(x) = \lim_{n\to\infty} f_n(x) = 0$ for all $x \in \mathbb{R}$.

Example 2. Define (g_n) where $g_n : \mathbb{R} \to \mathbb{R}$ by

$$g_n(x) = \frac{nx+1}{n} + \cos\left(\frac{x}{n}\right) - \sqrt{x^2 + -\frac{1}{n}}.$$

As before, to determine the pointwise limit (if it exists) we keep x fixed and take $n \to \infty$. The three terms making up $g_n(x)$ each converge for a fixed $x \in \mathbb{R}$ as follows:

$$\frac{nx+1}{n} \to x,\ \cos\left(\frac{x}{n}\right) \to \cos 0 = 1,\ \sqrt{x^2 + \frac{1}{n}} \to \sqrt{x^2} = |x|,$$

where for the second we use the fact that $y \mapsto \cos y$ is continuous. Thus

$$g_n(x) = \frac{nx+1}{n} + \cos\left(\frac{x}{n}\right) - \sqrt{x^2 + -\frac{1}{n}} \to x + 1 + |x| \text{ for any } x \in \mathbb{R},$$

so the sequence (g_n) converges pointwise to the function $g : \mathbb{R} \to \mathbb{R}$ defined by

$$g(x) = x + 1 - |x|.$$

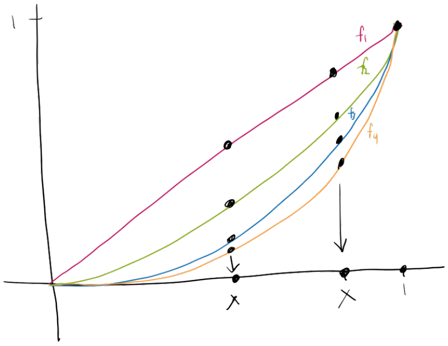
Example 3. (This is a standard example we'll see pop-up again and again.) Define (h_n) where $h_n : [0,1] \to \mathbb{R}$ by $h_n(x) = x^n$, so we have the sequence of functions

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

For $0 \leq x < 1$ we have $\lim x^n = 0$ as we saw last quarter, while for $x = 1$ we have $\lim x^n = \lim 1 = 1$. Thus the sequence (h_n) converges pointwise to the function $h : [0,1] \to \mathbb{R}$ defined by

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

Here is the picture to have in mind. Take $0 \leq x < 1$ on the x-axis and look at all the points you get on the graphs of $h_1, h_2, h_3, \ldots$ corresponding to this. As n gets larger, the y-values of these points (i.e. the values $f_n(x)$) are getting closer and closer to the x-axis where $y = 0$, reflecting the fact that $h_n(x) = x^n \to 0$ for such x. However, at $x = 1$, we get the point $(1,1)$ on all of the graphs, and hence this y-value stays fixed at 1, reflecting the the fact that $h_n(x) \to 1$ when $x = 1$.



Thus visually the graph of the pointwise limit is describing what happens "vertically" to points on the graphs of the h_n's at fixed values of x as n gets larger and larger.

Now, say that we considered the same functions only defined on all of $[0, \infty)$ instead of just $[0, 1]$. In this case, the sequence (h_n) would *not* converge pointwise since $\lim f_n(x) = \lim x^n$ does not exist for $x > 1$. The upshot is that this notion of pointwise convergence depends on the domain we are considering for our functions: given functions might converge pointwise on one domain but not on another.

Important. To determine the pointwise limit of a sequence of functions (f_n), compute $\lim_{n\to\infty} f_n(x)$ for fixed x in the domain of the f_n; the value obtained at a fixed x defines the value of the pointwise limit at that x. If for some x in the given domain the sequence of numbers $(f_n(x))$ does not converge, then (f_n) does not converge pointwise on that domain.

Pointwise convergence isn't so nice. Note that in Example 3, all of the functions $h_n(x) = x^n$ being considered are continuous, and yet their pointwise limit is not! In this case, the pointwise limit fails to be continuous only at $x = 1$, but it is in fact possible to come up with examples where the pointwise limit of continuous functions is *nowhere* continuous. Thus, it is not true that the pointwise limit of continuous functions is necessarily itself continuous. This is not so good, since it would be awesome if the limit of functions with a certain property still had that same property. (We'll see why this would be awesome as we go on.)

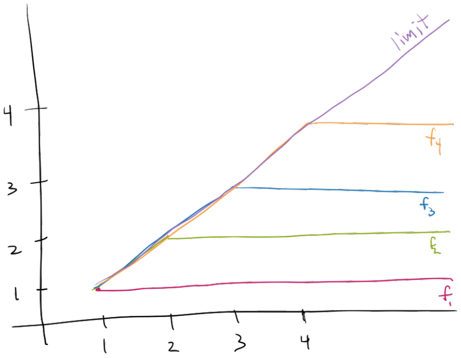
Similarly, it turns out that the pointwise limit of integrable functions is not necessarily integrable, that the pointwise limit of differentiable functions is not necessarily differentiable, and that the pointwise limit of bounded functions is not necessarily bounded. The book has examples for integrability and differentiability, and we'll give an example for boundedness below. Thus, continuity, integrability, differentiability, and boundedness are all properties which are not necessarily preserved under pointwise convergence, which illustrates that pointwise convergence alone isn't going to be that good a property to have. Next time we'll see a better notion of convergence for sequences of functions, where much nicer things happen.

But to give one last comment: indeed, there is no reason to expect that pointwise convergence should have any nice such properties. After all, pointwise convergence, as the name suggests, is a "pointwise" condition, meaning that it depends on what's happening at each point one at a time, but that what happens at one point has no bearing on what happens at other points; i.e. determining the value of $\lim f_n(x)$ at a fixed x does not depend on the values of f_n at other points. However, all of the properties of functions mentioned above (continuity, integrability, differentiability, boundedness) *do* depend not only on individual points but on the behavior of multiple points all at once; for instance, determine if a function f is continuous at some x depends on the behavior of f not only at x but at points nearby as well, while determining if a function is integrable depends on the behavior of f on an entire interval. Hence, such properties in general should not be expected to behave nicely with respect to "pointwise" definitions.

Example 4. Consider the sequence (f_n) of functions $f_n : [1, \infty) \to \mathbb{R}$ defined by

$$f_n(x) = \min\{n, x\} \text{ for } x \geq 1.$$

We claim that these functions are all bounded, but they converge pointwise to the function $f(x) = x$, which is not. To get a feel for these functions we determine the first few explicitly. The function f_1 is defined by $f_1(x) = \min\{1, x\}$, but since we are only considering $x \geq 1$ this minimum is always 1 so f_1 is the constant function 1. Now, f_2 has the value x for $1 \leq x \leq 2$, after which point it remains constant at 2, and f_3 has the value x for $1 \leq x \leq 3$ after which it remains constant at 3. This pattern continuous: in general, f_n starts off the same as $f(x) = x$ until we hit $x = n$, at which point f_n remains constant at n:



Thus we see that all of these functions are indeed bounded. At a fixed x, the values $f_n(x)$ go from one integer to another, until eventually they remain constant at x: for instance, for $x = 10.5$, we have $f_1(x) = 1,\ f_2(x) = 2, \ldots f_{10}(x) = 10$, and $f_n(x) = x$ for $n > 10$. Thus for any $x \in [1, \infty)$, the sequence of numbers $(f_n(x))$ is eventually constant at x, so $f_n(x) \to x$ for any x. Hence the pointwise limit of this sequence is the function $f(x) = x$, which is not bounded on $[1, \infty)$.

Important. Common properties functions may have are not necessarily preserved under pointwise convergence, meaning that if a sequence (f_n) converges pointwise to a function f and each f_n has some given property (i.e. continuity, integrability, differentiability, or boundedness), it is not always true that f also has that same property.

Lecture 5: Uniform Convergence

Today we started talking about uniform convergence, which is a stronger notion of convergence for sequences of functions than pointwise converge. Uniform convergence is much better behaved than pointwise convergence, in that many nice properties of functions end up being preserved.

Warm-Up. We determine the pointwise limit of the sequence of functions defined by:

$$f_n(x) = x\cos\left(\frac{1}{n}\right) + \frac{1+\frac{1}{n}}{x} \text{ for } x \in (0,\infty).$$

For a fixed $x \in (0,\infty)$, as $n \to \infty$ we have

$$x\cos\left(\frac{1}{n}\right) \to x\cos 0 = x \text{ and } \frac{1+\frac{1}{n}}{x} \to \frac{1}{x}.$$

Thus (f_n) converges pointwise to the function f on $(0,\infty)$ defined by $f(x) = x + \frac{1}{x}$.

Uniform convergence. Let us write out what it means for (f_n) to converge pointwise to f on some domain E:

for any $x \in E$ and any $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \epsilon$ for $n \geq N$.

Of course, the portion beginning with "for any $\epsilon > 0$" is just the definition of what it means for the sequence of numbers $(f_n(x))$ to converge to the number $f(x)$, as required in pointwise convergence.

Now, with one small change we get our new definition: a sequence of functions (f_n) is said to *converge uniformly* to a function f on a domain E if:

for any $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \epsilon$ for $n \geq N$ and all $x \in E$.

We call f the *uniform limit* of the sequence (f_n).

Note the difference: in the definition of pointwise convergence the index N might depend on $x \in E$ in that different points require different indices, while in uniform convergence there is a single N which works *for all $x \in E$ simultaneously*, which is the key point! If you treat the index N showing up in the definition of convergence for a sequence of numbers as some sort of "measure" of how rapidly the sequence is converging (larger indices indicate a slower convergence), then here we are saying that in uniform convergence the sequences $(f_n(x))$, in a sense, converge "at the same rate" as x ranges throughout E—i.e. in a "uniform" way. So, what happens at one point of E *is* related to what happens at other points—an idea which is missing in the notion of pointwise convergence.

Checking for uniform convergence. Even though uniform convergence is what we really care about, we still had to go through the process of first speaking about pointwise convergence anyway, the reason being that uniform convergence implies pointwise convergence. Indeed, if there is one index N which satisfies the requirement of uniform convergence for all $x \in E$ at once, this same N

applied to a fixed $x \in E$ will satisfy the requirement of pointwise convergence. In other words, if $f_n \to f$ uniformly, then $f_n \to f$ pointwise as well.

Thus, in order to check for uniform convergence, we must first check for pointwise convergence since the only possible candidate for the uniform limit is the pointwise limit: if our sequence does not converge pointwise, then it does not converge uniformly either, while if it does converge pointwise we are left checking further to see if the convergence to the pointwise limit is actually uniform.

Example 1. Define the sequence (f_n) on $\mathbb{R}$ by $f_n(x) = \frac{1}{n}\sin x$. We saw previously that this sequence converged pointwise to the constant zero function, and now we claim that this convergence is indeed uniform. Checking the definition, for any $\epsilon > 0$ pick $N \in \mathbb{N}$ such that $\frac{1}{N} < \epsilon$. Then for any $x \in \mathbb{R}$ we have

$$\left|\frac{1}{n}\sin x - 0\right| = \frac{1}{n}|\sin x| \leq \frac{1}{n} \leq \frac{1}{N} < \epsilon,$$

so we conclude that $\frac{1}{n}\sin x \to 0$ uniformly on $\mathbb{R}$. Note that, as required, the index N here only depends on ϵ.

Also, note that the reason why we were able to make this work is that we were able to find a bound on $|\frac{1}{n}\sin x - 0|$ (playing the role of $|f_n(x) - f(x)|$) which was independent of x, namely $\frac{1}{n}$ in this case. This is a common idea when working with explicit examples.

Important. To check whether a sequence of functions (f_n) converges uniformly on some domain E, first determine the pointwise limit f (if it exists), and then see whether (f_n) actually converges uniformly to that pointwise limit on E. Practically, this will usually require that we find a bound on $|f_n(x) - f(x)|$ which does not depend on x.

Example 2. Consider now the sequence from Warm-Up:

$$f_n(x) = x\cos\left(\frac{1}{n}\right) + \frac{1+\frac{1}{n}}{x} \text{ for } x \in (0,\infty).$$

Previously we determined that this converged pointwise to the function $f : (0,\infty) \to \mathbb{R}$ given by $f(x) = x + \frac{1}{x}$. Now we see whether or not this convergence is uniform.

We would like to find a bound on $|f_n(x) - f(x)|$ which does not depend on x. Playing around a bit, we have:

$$\begin{aligned} |f_n(x) - f(x)| &= \left|\left(x\cos\frac{1}{n} + \frac{1+\frac{1}{n}}{x}\right) - \left(x + \frac{1}{x}\right)\right| \\ &= \left|\left(x\cos\frac{1}{n} - x\right) + \left(\frac{1+\frac{1}{n}}{x} - \frac{1}{x}\right)\right| \\ &\leq \left|x\cos\frac{1}{n} - x\right| + \left|\frac{1+\frac{1}{n}}{x} - \frac{1}{x}\right| \\ &= x\left|\cos\frac{1}{n} - 1\right| + \frac{1}{xn} \end{aligned}$$

where $|x| = x$ since $x > 0$. But now we see a problem: from this along we will not be able to find a bound on $|f_n(x) - f(x)|$ which is independent of x since on $(0,\infty)$ x can get arbitrarily large or arbitrarily small! This means that we cannot find a bound for the x in the first term nor a bound for the $\frac{1}{x}$ in the second term. Of course, this is not enough to show that this sequence does not converge uniformly, but it suggests that this might be the case. (This sequence in fact does not converge uniformly, but it takes more work to show this precisely.)

Instead, let us now consider the same sequence only with the functions defined on some interval (a, b) where $0 < a < b$. We claim that on this interval the convergence is uniform. (Thus, just as the notion of pointwise convergence depends on the domain in question, so too does the notion of uniform convergence.) Using the same inequalities as above, the point is that we now consider only $x \in (a, b)$, so $x < b$ and $\frac{1}{x} < \frac{1}{a}$. Thus for such x we get:

$$|f_n(x) - f(x)| \leq b \left|\cos \frac{1}{n} - 1\right| + \frac{1}{an},$$

which as we wanted is a bound independent of $x \in (a, b)$. We can now choose appropriate indices to make each piece smaller than $\frac{\epsilon}{2}$ and we will get uniform convergence. Here is a formal proof:

Let $\epsilon > 0$. Since $\frac{1}{n} \to 0$ and $x \mapsto \cos x$ is continuous, $\cos \frac{1}{n} \to \cos 0 = 1$ so there exists N_1 such that

$$\left|\cos \frac{1}{n} - 1\right| < \frac{\epsilon}{2b} \text{ for } n \geq N_1.$$

Pick $N > N_1$ large enough to also guarantee that $\frac{1}{n} < a\epsilon$ for $n \geq N$. Then for $n \geq N$ and any $x \in (a, b)$, we have (using the inequalities derived above):

$$|f_n(x) - f(x)| \leq b \left|\cos \frac{1}{n} - 1\right| + \frac{1}{an} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

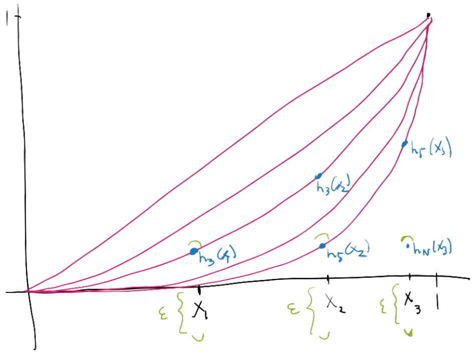
Thus $f_n \to f$ uniformly on (a, b) as claimed.

Non-example. Consider the sequence $h_n(x) = x^n$ on $[0, 1]$ from last time, when we determined that this converged pointwise to the function $h : [0, 1] \to \mathbb{R}$ defined by

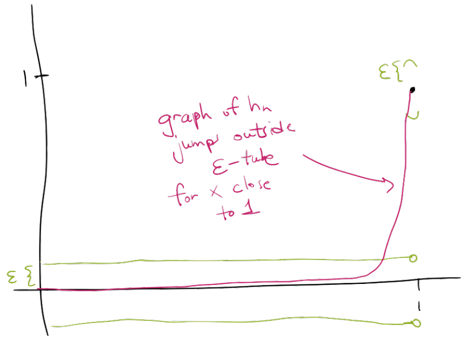
$$h(x) = \begin{cases} 0 & 0 \leq x < 1 \\ 1 & x = 1. \end{cases}$$

Now, this convergence is not uniform: we will see next time that a uniform limit of continuous functions must itself be continuous, which is not true here. However, let us try to see "visually" why this convergence should not be uniform.

Fix $\epsilon > 0$. At a fixed $0 \leq x_1 < 1$, draw a vertical interval around $h(x_1) = 0$ of half-length ϵ. Then using pointwise convergence we can find N large enough so that $h_N(x_1)$ lies within ϵ away from $0 = h(x_1)$, i.e. so that $h_N(x_1)$ lies within this vertical interval. Now take x_2 which is closer to 1 than x_1 and note that the same N no longer works, so we need a larger N to guarantee that $h_N(X_2)$ is still within ϵ away from $0 = h(x_2)$. The point is that as $x \to 1$ the index N needed to guarantee pointwise convergence is getting larger and larger, so there is not a single N which works for all $0 \leq x < 1$ at once:



Visually, these vertical intervals sweep out an "ϵ-tube" around the graph of h—meaning a "tube" which at each point x moves a vertical distance of ϵ above and below the corresponding point on the graph of h—and the point is that no matter what ϵ is eventually the graphs of h_n must jump "outside" of this tube:

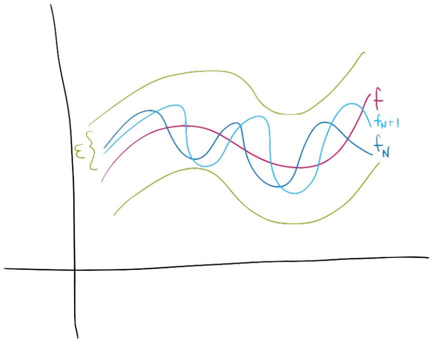


This is what prevents there from being a single index N satisfying the required inequalities in the definition of uniform convergence for all x at once.

Visualizing uniform convergence. The pictures above give us a nice way to visualize the idea behind uniform convergence in general. Suppose that $f_n \to f$ uniformly and take any "ϵ-tube" around the graph of f as above. The condition

$$|f_n(x) - f(x)| < \epsilon \text{ for } n \geq N \text{ and all } x$$

in the definition of uniform convergence says precisely that for large enough n, the entire graph of f_n lies fully within this tube!

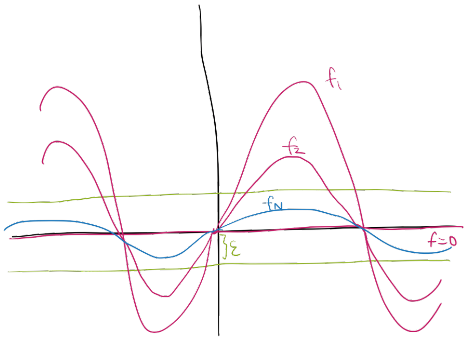


Indeed, this inequality can be interpreted as

$$f_n(x) \in (f(x) - \epsilon, f(x) + \epsilon),$$

and we should visualize the interval $(f(x) - \epsilon, f(x) + \epsilon)$ as making up the vertical pieces of the tube, where the entire tube is swept out by these vertical intervals as x varies throughout our domain.

Back to Example 1. Finally we come back to the sequence $f_n(x) = \frac{1}{n}\sin x$ on $\mathbb{R}$ of Example 1, which we showed converged uniformly to the constant zero function. Indeed, we can now see why this makes sense visually. The graphs of the f_n look like sine curves only shrunk vertically as n gets larger. Thus, given any ϵ-tube around the x-axis (which is the graph of the constant zero function), eventually the graph of f_n is fully contained within this tube:



Important. Visually, to say that $f_n \to f$ uniformly means that given any tube around the graph of f, the graph of f_n lies fully within that tube for n past some index. Thus, the graph of f_n is "close" to the graph of f for large enough n, and only gets "closer" as n increases. View this as analogous to the picture for convergence of a sequence of numbers in terms intervals around the limit: given any such interval, eventually all terms in your sequence are inside of it.

Lecture 6: More on Uniform Convergence

Today we continued talking about uniform convergence, and looked at properties of functions which are "preserved" under uniform convergence. The point in the coming week is that these ideas will give us new ways to show that functions have certain properties, when verifying such properties may not be possible to do directly. Later when we talk about so-called metric spaces we'll revisit these facts from another point of view.

Warm-Up 1. We show that the sequence (f_n) on $\mathbb{R}$ defined by $f_n(x) = \sqrt{x^2 + \frac{1}{n}}$ converges uniformly. The candidate for the uniform limit is the pointwise limit, so first we determine that. For a fixed $x \in \mathbb{R}$, we have

$$\sqrt{x^2 + \frac{1}{n}} \to \sqrt{x^2} = |x|,$$

where we use the fact that the square root function is continuous in order to say that

$$\lim_{n\to\infty} \sqrt{x^2 + \frac{1}{n}} = \sqrt{\lim_{n\to\infty} (x^2 + \frac{1}{n})}.$$

Thus, the sequence (f_n) converges pointwise to the absolute value function $f(x) = |x| = \sqrt{x^2}$.

To show that this convergence is actually uniform, we must find a bound on $|f_n(x) - f(x)|$ which does not depend on x. We use the inequality

$$|\sqrt{a} - \sqrt{b}| = \sqrt{|a - b|},$$

which we derived last quarter. (In particular, we used this previously to show that the square root function was uniformly continuous.) Thus we have:

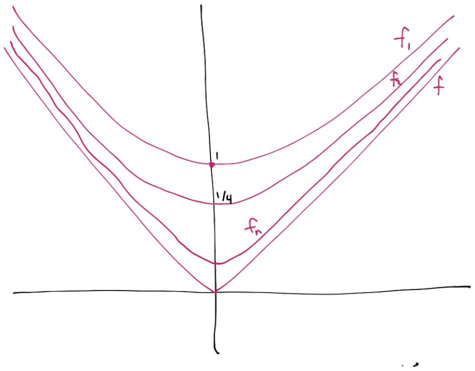
$$|f_n(x) - f(x)| = \left|\sqrt{x^2 + \frac{1}{n}} - \sqrt{x^2}\right| \leq \sqrt{\left|(x^2 + \frac{1}{n}) - x^2\right|} = \sqrt{\frac{1}{n}}.$$

Hence for $\epsilon > 0$, picking N such that $\frac{1}{N} < \epsilon^2$ gives

$$|f_n(x) - f(x)| \leq \sqrt{\frac{1}{n}} \leq \sqrt{\frac{1}{N}} < \sqrt{\epsilon^2} = \epsilon$$

for $n \geq N$ and all $x \in R$. Thus $f_n \to f$ uniformly as claimed.

Let us visualize this uniform convergence as well. The graphs of the functions f_n look kind of like parabolas which are approaching the graph of the absolute value function as n gets larger:



Thus given any ϵ-tube around the graph of $f(x) = |x|$, it makes visual sense that the graph of $f_n(x) = \sqrt{x^2 + \frac{1}{n}}$ is fully within this tube once n is large enough.

Warm-Up 2. (This problem was on an old final of mine.) Suppose that for all n, $f_n : [1,3] \to \mathbb{R}$ is a decreasing function and that the sequence (f_n) converges pointwise to 0. We claim that this convergence is actually uniform. (So, a rare instance in which pointwise convergence does imply uniform convergence, at least under the additional assumption that our functions are all decreasing.)

Given $\epsilon > 0$, we need to come up with the inequality

$$|f_n(x)| < \epsilon \text{ for large enough } n \text{ and all } x \in [1,3].$$

The point here is that since each f_n is decreasing, we know that

$$f_n(3) \leq f_n(x) \leq f_n(1),$$

so we are able to focus solely on bounding $f_n(3)$ and $f_n(1)$. Applying the pointwise convergence condition to these alone will give us what we want. Here's the proof.

Let $\epsilon > 0$. Since $f_n \to 0$ pointwise, $f_n(1) \to 0$ and $f_n(3) \to 0$. Thus there exists $N \in \mathbb{N}$ such that

$$|f_n(3)| < \epsilon \text{ and } |f_n(1)| < \epsilon \text{ for } n \geq N.$$

(Technically the given assumptions give us possibly different indices guaranteeing these two inequalities, but all we need to do is take their maximum.) Then for $n \geq N$ and $x \in [1,3]$, we have:

$$-\epsilon < f_n(3) \leq f_n(x) \leq f_n(1) < \epsilon, \text{ so } |f_n(x) - 0| = |f_n(x)| < \epsilon$$

as required. Hence $f_n \to 0$ uniformly.

Continuity preserved. And now we come to see why we care about uniform convergence, with this first fact being perhaps the most crucial:

> If (f_n) is a sequence of continuous functions converging uniformly to a function f on E, then f is continuous on E as well. (More generally, at any point at which the f_n are continuous, f is also continuous.).

Thus, we say that continuity is *preserved* under uniform continuity. (We saw from the example $h_n(x) = x^n$ on $[0,1]$) that this is not true for pointwise convergence alone.) Practically, this will tells us that in order to show a certain function is continuous, we need only show that we can "approximate" it to whatever degree of accuracy we want using continuous functions, which it turns out is often simpler to carry out for complicated functions than trying to show continuity directly.

The proof is in the book, but let's outline the basic idea here. To check continuity of f at a point x_0, we want to come up with the inequality

$$|f(x) - f(x_0)| < \epsilon$$

for all x within some δ away from x_0. We have two things to work with: uniform convergence of f_n to f, which will give us inequalities of the form

$$|f_n(y) - f(y)| < \text{ whatever we want for all } y,$$

and continuity of f_n, which gives us inequalities of the form

$$|f_n(x) - f_n(x_0)| < \text{ whatever we want for } x \text{ within some } \delta \text{ away from } x_0.$$

The point is that we can bound $|f(x) - f(x_0)|$ in terms of these types of absolute values using:

$$|f(x) - f(x_0)| = |(f(x) - f_n(x)) + (f_n(x) - f_n(x_0)) + (f_n(x_0) - f(x_0)),$$

where we "work" from $f(x)$ to $f(x_0)$ using the types of terms we have some control over. Applying an "$\frac{\epsilon}{3}$-trick" to this gives us what we want, and the δ we need comes from the continuity of an appropriate f_n. Again, check the book for full details.

Note that reason why this works is because uniform convergence tells us something about what's happening at all points at once (which we need in order to bound $|f(x) - f_n(x)|$ and $|f_n(x_0) - f(x_0)|$ above simultaneously), as opposed to the point-by-point behavior of pointwise convergence. Visually, you can't have the graph of a function be arbitrarily close to the graph of a continuous function and still have a "jump", indicating a discontinuity. (Of course, not all discontinuities are jump discontinuities, but this picture is the intuitive one to have in mind.)

Integrability preserved. Next we look at integration, where the basic fact is:

> If (f_n) is a sequence of integrable functions on $[a, b]$ converging uniformly to a function f, then f is integrable on $[a, b]$ as well. Moreover, the sequence of numbers obtained by integrating the f_n converges to the number obtained by integrating f:
>
> $$\int_a^b f_n(x)\,dx \to \int_a^b f(x)\,dx.$$

Again, the analogous statement is not true for pointwise convergence alone. The proof of this is in the book, and involves working with good ol' upper and lower sums.

Example. We use the above fact to compute $\lim_{n\to\infty} \int_0^2 e^{x^2/n}\,dx$. The point is that $\int_0^2 e^{x^2/n}\,dx$ is not something we can compute directly since e^{x^2} does not have an elementary antiderivative, so we need a more clever approach. The key is that we would like to be able to say that:

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

which makes the computation simple. However, this interchanging of the limit as $n \to \infty$ and the integration *depends* on knowing that the sequence $e^{x^2/n}$ converges uniformly! (This is the result above which says that under uniform convergence, the integrals of the f_n converge to the integral of f.) So, we first show that $e^{x^2/n}$ indeed converges uniformly.

For a fixed $x \in [0, 2]$, we have

$$e^{x^2/n} \to e^0 = 1,$$

so the pointwise limit is the function function 1. To see that this is uniform limit as well, let $\epsilon > 0$ and use continuity of the exponential function to pick $N \in \mathbb{N}$ such that

$$\left|e^{4/n} - 1\right| < \epsilon \text{ for } n \geq N.$$

(To be clear, we use the fact that the exponential function is continuous and $\frac{4}{n} \to 0$, $e^{4/n} \to e^0 = 1$.) Then for $n \geq N$ and $x \in [0, 2]$, we have:

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

where we use the fact that the exponential function is increasing. Thus $e^{x^2/n} \to 1$ uniformly as claimed, and we have:

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

which is our required value. (Note again how impossible this would likely be to compute without using the fact that uniform convergence preserves integrals.)

Differentiability (with additional assumption) preserved. Now, we move on how differentiability behaves with respect to uniform convergence, where things aren't as straightforward as they were for continuity and integrability. Indeed, the sequence in the first Warm-Up shows that in fact differentiability is NOT preserved under uniform convergence in general: the functions $f_n(x) = \sqrt{x^2 + \frac{1}{n}}$ are all differentiable at 0, but their uniform limit $f(x) = |x|$ is not.

The issue is that uniform convergence has to do with functions being "close" to one another, but two functions which are "close" can still change (which is what the derivative measures) in vastly different ways; for instance, a constant function does not change and has derivative zero, but a function whose graph rapidly oscillates up and down and yet remains close to this constant will experience rapid rates of increase and decrease, so that its derivative will behave very differently from the constant zero function.

However, all is not lost, as with one additional assumption we get a nice relation between derivatives and uniform convergence:

> If (f_n) is a sequence of differentiable functions converging uniformly to a function f on some domain, AND the sequence of derivatives (f_n') converges uniformly to a function g, then f is itself differentiable and $f' = g$.

The condition that (f_n') converges uniformly says that there is some control over how wildly the derivatives f_n' can behave, and with this control the original uniform limit is in fact differentiable. Saying that $f' = g$ where g is the uniform limit of (f_n') is simply saying that under these assumptions we do have that

$$f_n' \to \text{ the derivative of } f,$$

so that the limit of the derivatives is the derivative of the limit, analogously to what we had for integration. The proof of this fact is in the book, but as opposed to the proof for the analogous property of integrals (which is not hard to follow), this proof is indeed hard to follow. So, don't worry about fully understanding the proof, but the statement is definitely one you should be familiar with.

Example. Going back to the the sequence $f_n(x) = \sqrt{x^2 + \frac{1}{n}}$ which converged uniformly to $f(x) = |x|$, we can now see what the issue is: the derivatives of the f_n are given by

$$f_n'(x) = \frac{x}{\sqrt{x^2 + \frac{1}{n}}} \text{ for all } x \in \mathbb{R},$$

but this sequence of derivatives does not converge uniformly. Indeed, the pointwise limit of the f_n' is the function g defined by $g(0) = 0$ and for $x \neq 0$, $g(x) = \frac{x}{\sqrt{x^2}} = \pm 1$, depending on whether $x > 0$ or $x < 0$, and since this function is not continuous but each of the f_n' are, the convergence $f_n' \to g$ is not uniform. Hence the additional assumption in the above theorem that (f_n') converges uniformly fails in this example.

Important. Continuity and integrability are preserved under uniform convergence, meaning that if all functions f_n in our sequence have these properties, so does their uniform limit. Moreover, the integral of the limit is the limit of the integrals. In the case where the f_n are differentiable, if in addition the derivatives f_n' converge uniformly, then the uniform limit of the f_n is differentiable and the derivative of the limit is the limit of the derivatives.

Lecture 7: Series of Functions

Today we spoke about series of functions, generalizing what we saw previously for series of numbers. This will serve as the foundation of what we'll do with power series soon and Fourier series next quarter.

Warm-Up. We compute

$$\lim_{n\to\infty}\int_1^3 \frac{nx^{99}+5}{x^3+nx^{66}}\,dx.$$

As in a similar example from last time, the actual computation isn't so difficult—the point is recognizing that this requires we know the sequence of functions being integrated converges uniformly. In this case, the sequence we're interested in is

$$f_n(x)=\frac{nx^{99}+5}{x^3+nx^{66}}.$$

The pointwise limit of this sequence on $[1,3]$ is the function f defined by $f(x)=x^{33}$. We have:

$$|f_n(x)-f(x)|=\left|\frac{nx^{99}+5}{x^3+nx^{66}}-x^{33}\right|=\left|\frac{5-x^{36}}{x^3+nx^{66}}\right|\le\frac{3^{36}-5}{n}\text{ for }x\in[1,3].$$

Thus given $\epsilon>0$, picking N such that $\frac{3^{36}-5}{N}<\epsilon$ will give us uniform convergence $f_n\to f$ on $[1,3]$.

Since each f_n is integrable on $[1,3]$, so is the uniform limit f and:

$$\lim_{n\to\infty}\int_1^3\frac{nx^{99}+5}{x^3+nx^{66}}\,dx=\int_1^3\left(\lim_{n\to\infty}\frac{nx^{99}+5}{x^3+nx^{66}}\right)dx=\int_1^3 x^{33}\,dx=\frac{3^{34}-1}{34}$$

is the desired value.

Uniformly Cauchy. Analogously to what we know about sequences of numbers, we can phrase uniform convergence of a sequence of functions in terms of a "Cauchy" condition, which gives the notion of a sequence being *uniformly Cauchy*:

> A sequence of functions (f_n) on some domain E is uniformly Cauchy if for any $\epsilon>0$ there exists $N\in\mathbb{N}$ such that for $m,n\ge N$, $|f_n(x)-f_m(x)|<\epsilon$ for all $x\in E$.

The fact that a single N works for all x at once is what makes the sequence "uniformly" Cauchy as opposed to merely "pointwise" Cauchy.

The fact is that a sequence (f_n) is uniformly Cauchy if and only if it is uniformly convergent, just as we had for sequences of numbers. The proof is in the book, but here is the basic idea for the forward direction. Given a uniformly Cauchy sequence (f_n), we need a candidate for the what the uniform limit should be, which we know should be the pointwise limit. The key is that the uniformly Cauchy condition implies pointwise Cauchy, meaning that for each $x\in E$ the sequence of numbers $(f_n(x))$ is Cauchy and hence converges; what this converges to defines the value of the pointwise limit f at x, and then the goal is to show that $f_n \to f$ uniformly and not just pointwise. Again, check the book for full details.

Series of functions. A *series of functions* is an infinite sum $\sum f_n$ of functions f_n, all defined on some common domain. As with series of numbers, we define convergence of a series of functions in terms of convergence of its sequence of partial sums:

$$s_n = f_1 + \cdots + f_n.$$

However, now that these partial sums are themselves functions, we have to be careful about what type of convergence we ask for: the series $\sum f_n$ converges

- *pointwise* to f on E if the sequence of partial sums converges pointwise to f on E;
- *uniformly* to f on E if the sequence of partial sums converges uniformly to f on E;
- *absolutely* and pointwise/uniformly to f on E if $\sum |f_n|$ converges pointwise/uniformly on E.

Important. A series of functions $\sum f_n$ converges uniformly (or pointless) if its sequence of partial sums—which is a sequence of functions—converges uniformly (or pointwise).

Main examples. A *power series* is a series of functions of the form

$$\sum_{n=0}^{\infty} a_n (x - x_0)^n,$$

where the functions we are adding up are polynomials. We will learn next time everything there is to learn about convergence (pointwise and uniform) of power series.

A *Fourier series* is a series of functions of the form

$$\sum_{n=0}^{\infty} \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right),$$

where the functions we are adding up are trig functions. We will learn everything there is to know (at least regarding convergence) about Fourier series next quarter.

Checking convergence. As with sequences of functions, we will mainly be interested in knowing about the uniform convergence of series of functions. If we're lucky, the sequence of partial sums of a given series are possible to compute explicitly—this is rarely the case apart from Example 1 below. Otherwise, we can develop a Cauchy criterion for series of functions analogous to the one for series of numbers, by writing out what it means for the sequence of partial sums to be uniformly Cauchy:

> A series $\sum f_n$ converges uniformly on E if and only if for any $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that for $m \geq n \geq N$, $|\sum_{k=n}^{m} f_k(x)| < \epsilon$ for all $x \in E$.

Again, the fact that "for all $x \in E$" is at the end is what makes this "uniform". We'll see an example of this in action below, which is still a bit tedious.

Example 1. We determine the convergence of $\sum_{n=0}^{\infty} x^n$, which essentially just repeats what we know about geometric series. In this case, the partial sums are given by:

$$1 + x + \cdots + x^n = \frac{1 - x^{n+1}}{1 - x} \text{ for } x \neq 1,$$

and this sequence converges pointwise on $(-1,1)$ to the function $f(x) = \frac{1}{1-x}$. Thus we say that the series $\sum_{n=0}^{\infty} x^n$ converges pointwise on $(-1,1)$ to $\frac{1}{1-x}$, and write

$$\sum_{n=0}^{\infty} x_n = \frac{1}{1-x} \text{ pointwise for } |x| < 1.$$

We will discuss the uniform convergence of this series next time when talking about power series.

Example 2. We claim that the series $\sum_{n=1}^{\infty} \frac{\sin nx}{n^2}$ converges uniformly on all of $\mathbb{R}$. Indeed, note that for $m \geq n$, we have:

$$\left| \frac{\sin nx}{n^2} + \cdots + \frac{\sin mx}{m^2} \right| \leq \left| \frac{\sin nx}{n^2} \right| + \cdots + \left| \frac{\sin mx}{m^2} \right| \leq \frac{1}{n^2} + \cdots + \frac{1}{m^2} \text{ for all } x \in \mathbb{R}.$$

For $\epsilon > 0$, since $\sum_{k=1}^{\infty} \frac{1}{k^2}$ converges there exists N such that

$$\frac{1}{n^2} + \cdots + \frac{1}{m^2} < \epsilon \text{ for } m \geq n \geq N,$$

and thus for this same index we also have $\left| \sum_{k=n}^{m} \frac{\sin kx}{k^2} \right| < \epsilon$ for $m \geq n \geq N$ and all $x \in \mathbb{R}$. Hence $\sum \frac{\sin nx}{n^2}$ converges uniformly on $\mathbb{R}$ by the Cauchy criterion as claimed.

Weierstrass M-Test. Having to check the Cauchy condition as above every single time is a pain, but the proof above suggests a way around it: in the end all we needed has the fact that $\sum \frac{1}{k^2}$ was convergent and that the terms of this series of numbers bounded the corresponding terms in the given series of functions. This idea leads to what's known as the *Weierstrass M-Test*:

> Suppose that $\sum f_n$ is a series of functions and that M_n are bounds on the f_n on some domain E: $|f_n(x)| \leq M_n$ for all $x \in E$. Then if the series of numbers $\sum M_n$ converges, the series of functions $\sum f_n$ converges uniformly and absolutely on E.

The point is that we can show uniform convergence by using an appropriate series of *numbers*, which are usually simpler to work with. The proof of this is in the book, but the basic idea is given in Example 2: we bound

$$|f_n(x)| + \cdots + |f_m(x)| \leq M_n + \cdots + M_m,$$

and then use the Cauchy criteria on the series of numbers $\sum M_n$. (We get absolute convergence because in this inequality we are actually bounding the partial sums of the series $\sum |f_n|$ and not just of $\sum f_n$.)

Back to Example 2. Now Example 2 becomes simpler: we have $\left| \frac{\sin nx}{n^2} \right| \leq \frac{1}{n^2}$ for all $x \in \mathbb{R}$, so since $\sum \frac{1}{n^2}$ converges the series $\sum \frac{\sin nx}{n^2}$ converges uniformly (and absolutely) on $\mathbb{R}$ by the M-test.

Important. In practice, the Weierstrass M-test is the most useful way of showing that series of functions converge uniformly. All we need to do is bound the functions in our series by numbers which themselves form a convergent series.

Example 3. Finally, we show that $\sum \frac{1}{n} \sin\left(\frac{x}{n}\right)$ converges uniformly on any bounded interval $[-R, R]$. The key inequality we use here

$$|\sin y| \leq |y| \text{ for all } y \in \mathbb{R},$$

which can be established using, say, the Mean Value Theorem. Applying this in our case gives:

$$\left|\frac{1}{n}\sin\frac{x}{n}\right| \leq \frac{1}{n}\frac{|x|}{n} = \frac{R}{n^2} \text{ for all } x \in [-R, R].$$

Thus since $\sum \frac{R}{n^2}$ converges (note that R here is a constant), the Weierstrass M-test implies that $\sum \frac{1}{n}\sin\frac{x}{n}$ converges uniformly on $[-R, R]$.

Note that the bound $\left|\frac{1}{n}\sin\frac{x}{n}\right| \leq \frac{1}{n}$, using the fact that sine is bounded by 1, would have gotten us nowhere since $\sum \frac{1}{n}$ does not converge. Also note that we would not be able to show uniform convergence on all of $\mathbb{R}$ at once since in the inequality above we would not be able to bound $|x|$ by a single fixed quantity.

Lecture 8: Power Series

Today we started talking about power series, which give us an important type of series of functions. We'll see that determining the uniform convergence of these is fairly straightforward.

Warm-Up. We show that $\sum_{k=0}^{\infty} e^{-kx}$ converges uniformly on any $[a, \infty)$ where $a > 0$. Since the exponential function is increasing, we have:

$$\left|e^{-kx}\right| = \frac{1}{e^{kx}} \leq \frac{1}{e^{ka}} = \left(\frac{1}{e^a}\right)^k \quad \text{for all } x \in [a, \infty).$$

Thus since $\sum \left(\frac{1}{e^a}\right)^k$ converges (geometric series $\sum r^n$ with $|r| < 1$), the Weierstrass M-test implies that $\sum e^{-kx}$ converges uniformly on $[a, \infty)$. Note that we would not be able to get uniform convergence on $[0, \infty)$: here the bound we would get is $|e^{-kx}| \leq 1$, but $\sum 1$ does not converge so the M-test does not apply.

Continuity/integrability/differentiability of series. Using what we know about the relation between uniform convergence of sequences of functions and continuity, integrability, and differentiability, we can state the following basic facts about series. Suppose that $\sum f_n$ converges uniformly to f on some domain. Then

- if the f_n are continuous on E, $f = \sum f_n$ is continuous on E,
- if the f_n are integrable on $[a, b]$, $f = \sum f_n$ is integrable on $[a, b]$ and
$$\int_a^b f(x)\,dx = \int_a^b \left(\sum f_n(x)\right)\,dx = \sum \left(\int_a^b f_n(x)\,dx\right),$$
- if the f_n are differentiable on (a, b) *and* $\sum f_n'$ converges uniformly on (a, b), then $f = \sum f_n$ is differentiable on (a, b) and
$$f' = \left(\sum f_n\right)' = \sum f_n'.$$

These come from applying the analogous statements about sequences of functions to the sequence of partial sums; the extra assumption in the third property comes from the analogous extra assumption for sequences of functions. Note that the second and third properties say that "infinite sums" are interchangeable with integrals and derivatives, at least under the right assumptions.

Example 1. Consider the series of functions $\sum_{n=0}^{\infty} \frac{x^n}{n!}$, which is of course an example of a power series. It turns out that this series converges pointwise on all of $\mathbb{R}$ and it converges uniformly on any closed interval $[a, b] \subset \mathbb{R}$. (We'll see why when we talk about power series in general in a bit.) Thus, the function to which this series converges is automatically continuous and integrable on any $[a, b]$. (This series actually converges to e^x, which is actually continuous on all of $\mathbb{R}$, but this can deduced from what we said here by allowing the closed intervals $[a, b]$ to get larger and larger.)

Now, the series obtained by taking term-by-term derivatives is:

$$\sum_{n=1}^{\infty} \frac{nx^{n-1}}{n!} = \sum_{n=1}^{\infty} \frac{x^{n-1}}{(n-1)!}.$$

But this is the same as the original series, only here indexed to start at $n = 1$ instead of $n = 0$. Thus this series of term-by-term derivatives also converges uniformly on any $[a, b]$, so the function defined by the original series is differentiable and thus equals its own derivative! Of course, we already know that e^x is continuous, integrable, differentiable, and equals its own derivative, but the point here is that we can derive these facts solely from the power series definition of e^x.

Convergence of power series. Recall that a *power series* is a series of functions of the form

$$\sum a_n (x - x_0)^n,$$

where we say that this series is *centered* at x_0. The number

$$R := \frac{1}{\limsup |a_n|^{1/n}} \geq 0$$

is called the *radius of convergence* of the power series, where we interpret this as $R = \infty$ when the denominator is 0 and as $R = 0$ when the denominator is ∞. The main fact is the following, which justifies the term "radius of convergence":

> With R defined as above, the series $\sum a_n (x - x_0)^n$ converges pointwise and absolutely on $(x_0 - R, x_0 + R)$, and possibly at one or both of the endpoints $x_0 - R$ and $x_0 + R$. We interpret this interval as $(-\infty, \infty)$ when $R = \infty$ and as $\{x_0\}$ when $R = 0$.

This fact follows from the root test: for a fixed x, the convergence of the power series is determined by whether

$$\limsup |a_n (x - x_0)^n|^{1/n} = |x - x_0| \limsup |a_n|^{1/n}$$

is smaller or larger than 1, so whether $|x - x_0| < R$ or $|x - x_0| > R$ where $R = \frac{1}{\limsup |a_n|^{1/n}}$. (Note that the fact we have an explicit expression for the radius of convergence is one reason why the notion of lim sup is useful.) When $|x - x_0| = R$ the root test gives no information, so the series may or may not converge at one or both of $x_0 - R$ and $x_0 + R$. Also, so far these are only pointwise conditions since we applied the root test with a fixed x.

Example 2. Consider the power series $\sum_{n=0}^{\infty} x^n$. We saw last time that this converged pointwise on $(-1, 1)$ to $\frac{1}{1-x}$. We can now also derive this convergence from the fact that

$$\limsup |a_n|^{1/n} = \limsup 1^{1/n} = 1,$$

so the radius of convergence is indeed 1. However, now we point out that this convergence cannot be uniform on all of $(-1, 1)$: for each n, the nth partials sum is bounded on $(-1, 1)$ by $n + 1$ since

$$|1 + x + \cdots + x^n| \leq 1 + |x| + \cdots + |x|^n \leq \underbrace{1 + \cdots + 1}_{n \text{ times}} = n + 1 \text{ for } |x| < 1,$$

so if the convergence were uniform on $(-1,1)$ the limit $\frac{1}{1-x}$ would also be bounded on $(-1,1)$, which it is not. Thus, it is not true that power series in general converge uniformly on their entire interval of convergence (x_0-R, x_0+R), even though they do so pointwise. (We normally won't care much about what's happening at the endpoints.)

Now, the function $\frac{1}{1-x}$ is bounded on any *smaller* closed interval $[-R,R] \subseteq (-1,1)$ within the interval of convergence, so the above issue is no longer a problem. Indeed, we have $|x^n| \le R^n$ for $x \in [-R,R]$ and since $R<1$, $\sum R^n$ converges so the Weierstrass M-test implies that $\sum x^n$ does converge uniformly on $[-R,R]$. Thus, even though we do not have uniform convergence on the entire interval of convergence, we do have it on *any* smaller closed interval contained within the interval of convergence.

Theorem. The previous example illustrates what happens in general. Let $\sum a_n(x-x_0)^n$ be a power series with radius of convergence R. Then $\sum a_n(x-x_0)^n$ converges uniformly on any $[a,b]$ within the interval of convergence (x_0-R, x_0+R).

Proof. The proof is in the book, but we give one here anyway, at least in the case where the closed interval we're taking looks like $[x_0-r, x_0+r] \subset (x_0-R, x_0+R)$ for $0<r<R$. (This will just make some of the notation simpler.)

For $x \in [x_0-r, x_0+r]$, we have

$$|a_n(x-x_0)^n| \le |a_n|r^n = |a_n([x_0+r]-x_0)^n|.$$

Since the number x_0+r is within the interval of convergence of the given power series, the series of numbers $\sum a_n([x_0+r]-x_0)^n$ converges absolutely, so by the M-test the power series $\sum a_n(x-x_0)^n$ converges uniformly on $[x_0-r, x_r+r]$ as claimed. (The point is that we are evaluating the given power series at a point x_0+r within the interval of convergence to get the convergent series of numbers we want in order to apply the M-test.) □

Important. A power series converges pointwise on its entire interval of convergence, and uniformly on any closed interval contained within the interval of convergence.

Continuity and integrability. Since the terms $a_n(x-x_0)^n$ making up a power series are always continuous and integrable on closed intervals $[a,b]$ within the interval of convergence, and since we have uniform convergence on such closed intervals, the function to which a power series converges is always continuous and integrable on such intervals and we can compute integrals by integrating term-by-term. In fact, the function $f(x)=\sum a_n(x-x_0)^n$ is also continuous on the entire interval of convergence: given any $c \in (x_0-R, x_0+R)$, take a closed interval $[a,b] \subset (x_0-R, x_0+R)$ with $a<c<b$ and then note that continuity on $[a,b]$ in particular implies continuity at c, so f is continuous on all of (x_0-R, x_0+R).

This is why the restoration that power series are only guaranteed to be uniformly convergent on closed intervals contained within their interval of convergence is not a big deal: since such closed intervals can be made to get closer and closer to the endpoints of the entire interval of convergence, we still get nice properties of the power series over its entire interval of convergence as well.

Important. The function to which a power series converges is always continuous on the entire interval of convergence and integrable on any closed interval $[a,b]$ within the interval of convergence, and

$$\int_a^b \left(\sum_{n=0}^\infty a_n(x-x_0)^n\right) dx = \sum_{n=0}^\infty \left(\int_a^b a_n(x-x_0)^n\, dx\right) = \sum_{n=0}^\infty \frac{a_n(x-x_0)^{n+1}}{n+1}\Bigg|_a^b.$$

Lecture 9: Analytic Functions

Today we started talking about analytic functions, which essentially are the functions defined by convergent power series. Such functions have especially nice properties, leading to them being the "holy grail" of functions: if you're working in some area and come across an analytic function, you shout out in joy that you have such an awesome function to work with.

Warm-Up. We determine the radius of convergence of the power series

$$\sum_{k=0}^{\infty} \frac{x^{3k}}{k+1}$$

and the explicit function to which this power series converges on its interval of convergence. The thing to be careful of here is that this series as written is technically *not* in the form of a power series due to the exponent being $3k$ instead of simply k. The point is that in order to determine the radius of convergence using a lim sup, we need the coefficients a satisfying

$$\sum_{k=0}^{\infty} \frac{x^{3k}}{k+1} = \sum_{n=0}^{\infty} a_n x^n,$$

and it is not true that $a_k = \frac{1}{k+1}$ as one might guess based on the original expression.

In fact, we have $a_{3k} = \frac{1}{k+1}$ and $a_n = 0$ for $n \neq 3k$, so that the sequence $|a_n|^{1/n}$ actually looks like:

$$1,\ 0,\ 0,\ \left(\frac{1}{2}\right)^{1/3},\ 0,\ 0,\ \left(\frac{1}{3}\right)^{1/6},\ \ldots \text{ and so on.}$$

This is the sequence we need to take the lim sup of. However, this lim sup will indeed be fully determined by the nonzero terms since $\sup_{k\geq n} |a_k|^{1/k}$ will always be one of these nonzero terms. Since

$$\lim_{k\to\infty} \left(\frac{1}{k+1}\right)^{1/3k} = 1,$$

which can be seen using

$$\left(\frac{1}{k+1}\right)^{1/3k} = e^{\frac{1}{3k}\log\frac{1}{k+1}}$$

and L'Hopital's rule, we have $\limsup |a_n|^{1/n} = 1$ so that the given power series has radius of convergence $\frac{1}{1} = 1$.

Now, to determine explicitly the function to which this series converges on $(-1, 1)$, we start with the fact that

$$\sum_{n=0}^{\infty} y^n = \frac{1}{1-y} \text{ for } |y| < 1.$$

The key is that we can manipulate the left side of this expression to get the series we want. Indeed, integrating term-by-term (more technically, we integrate both sides from 0 to a fixed $y \in (-1, 1)$, which we can do since we are within the interval of convergence) gives

$$\sum_{n=0}^{\infty} \frac{y^{n+1}}{n+1} = -\log|1-y| \text{ for } |y| < 1.$$

Now we substitute $y = x^3$ to get

$$\sum_{n=0}^{\infty} \frac{x^{3k+3}}{k+1} = -\log|1-x^3| \text{ for } |x^3| < 1, \text{ or equivalently } |x| < 1.$$

(Note that going from $|x^3| < 1$ to $|x| < 1$ in this step gives us another way to derive the radius of convergence of our series: it comes from making the substitution $y = x^3$ in the geometric series $\sum y^n$, whose radius of convergence we already know.) Finally, factoring out $x^3 \neq 0$ from the left side and dividing gives the required value for our original series:

$$\sum_{n=0}^{\infty} \frac{x^{3k}}{k+1} = \begin{cases} -\frac{\log|1-x^3|}{x^3} & -1 < x < 1, x \neq 0 \\ 0 & x = 0, \end{cases}$$

where the value for $x = 0$ comes simply from evaluating the original series at $x = 0$. Thus the given series converges pointwise to the function defined by the right side above on $(-1, 1)$ and uniformly to it on any closed interval $[a, b] \subset (-1, 1)$.

Derivatives of power series. Now we look at the differentiability of power series. As usual, here we have to be careful since we need an additional assumption in order to make everything work; that is, if we want to conclude that $\sum a_n(x-x_0)^n$ is differentiable, we have to know in advance that the term-by-term derivative $\sum na_n(x-x_0)^{n-1}$ is uniformly convergent. However, the basic fact is that this is never something we have to check:

> Given a power series $\sum_{n=0}^{\infty} a_n(x-x_0)^n$ with radius of convergence R, the term-by-term derivative $\sum_{n=1}^{\infty} na_n(x-x_0)^{n-1}$ also has radius of convergence R. Thus, the function to which the original series converges is differentiable on its interval of convergence and its derivative is given by this term-by-term derivative.

The claim about the radius of convergence follows from the fact that

$$\limsup |na_n|^{1/n} = \limsup |a_n|^{1/n}$$

is always true: the book has a proof of this in general, but in the simpler case where $\lim |a_n|^{1/n}$ exists this follows from the fact that $\lim n^{1/n} = 1$.

Thus we get that our original power series is differentiable on any $[a, b] \subset (x_0 - R, x_0 + R)$, since these are types of intervals where we have uniform convergence. But now, as in the case of continuity, since these intervals can be made to fill up the entire interval of convergence, we get differentiability on all of $(x_0 - R, x_0 + R)$: to be concrete, fix $c \in (x_0 - R, x_0 + R)$ and take $[a, b] \in (x_0 - R, x_0 + R)$ with $a < c < b$; since we have uniform convergence on $[a, b]$ we have differentiability on $[a, b]$ and hence in particular at c.

Note that this all justifies the derivatives of $\sum \frac{x^n}{n!}$ we computed in an example last time.

Smooth functions. Thus a power series is always differentiable on its interval of convergence. Since its derivative is then also a power series with the same interval of convergence, we get that our original series is twice-differentiable on this same interval. The second derivative is again a power series with the same radius of convergence, so our original series is three-times differentiable, and so on: the function to which a power series converges on its interval of convergence is in fact *infinitely-differentiable.*

In addition to the term "infinitely-differentiable", we also use the term *smooth* to refer to such functions, or we use the notation "f is C^∞", where the ∞ indicates the number of times this function is differentiable. The notation "$f \in C^\infty[a, b]$" means that f is a smooth function on $[a, b]$.

Analytic functions. Analytic functions are essentially those which can be expressed as convergent power series. Here is a precise definition:

> A function f is *analytic* on a domain E if for any $x_0 \in E$, there exists a power series $\sum a_n(x-x_0)^n$ centered at x_0 which converges to f near x_0, where "near x_0" means that there exists an interval $(x_0-\delta, x_0+\delta)$ on which the series converges to f.

A few remarks are in order. First, since power series are always infinitely-differentiable, it follows that in order for f to be analytic it must first be smooth; in other words, analytic implies smooth. (BUT, it is not true that smooth implies analytic, as we'll see.) Second, note that the power series used in the definition differs from point to point since the center changes from point to point—we'll see that we can get around this somewhat, but it will not be true that there will always be a *single* power series which equals f throughout the entire domain. (So, we would say that an analytic function is one which is *locally expressible* by a power series, where the "locally" is used to emphasize the series needed might differ from region to region.) Finally, the definition says nothing about what the power series must look like, although we'll see that we really have no choice: there can only be one power series which can satisfy this definition at a given point.

Examples. Most well known functions you see in a calculus course are actually analytic, including $e^x, \sin x$, and $\cos x$; let's look at e^x. Recall that

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \text{ for all } x \in \mathbb{R},$$

which we actually proved at one point last quarter when discussing Taylor's Theorem, but which we'll justify again later. Now, here we are only looking at a single power series centered at 0, whereas the definition of analytic requires a convergent power series *at each point* in our domain, which is $\mathbb{R}$ in this case. So, in order to conclude that e^x is analytic it seems that we would need to look at power series centered at any possible point, which is tedious.

However—and here is one simplification to the definition of analytic—in fact if we have a power series converging to our function on some given interval, then at any point in that interval we will in fact be able to find a power series centered at it which will converge to our function. In other words, if $f(x) = \sum a_n(x-x_0)^n$ in some interval around x_0, then f *is* analytic on that entire interval without having to actually check for power series centered at points other than x_0 in that interval. In the example above, since the given equality holds on all of $\mathbb{R}$, this alone implies that e^x is analytic on all of $\mathbb{R}$ even though so far we only have a power series centered at 0.

Important. To check whether a function f is analytic on some domain, it is enough to know that we can cover that domain by intervals, on each of which we can express f as the sum of a convergent power series. More precisely, for a given power series $\sum a_n(x-x_0)^n$ which converges on an interval I, the center can always be "shifted" to give a convergent power series centered at any other point in I. Thus, we do not have to literally check the definition of analytic at every single point in our domain, just points which are not yet included in intervals of convergence of power series centered at *other* points.

Remark. The fact about shifting centers is Theorem 7.46 in the book, but here is the basic idea. Say that we know

$$f(x) = \sum a_n(x-x_0)^n$$

for x in some interval I. Take any other $y_0 \in I$ and note that

$$x - x_0 = (x - y_0) + (x_0 - y_0).$$

Thus, using a binomial expansion:

$$(x - x_0)^n = [(x - y_0) + (x_0 - y_0)]^n = \text{ a sum involving powers of } x - y_0$$

where pieces involving powers of $(x_0 - y_0)$ contribute to the coefficients. Thus making these substitutions gives

$$\sum a_n (x - x_0)^n = \text{ a sum involving powers of } x - y_0,$$

which will give a power series centered at y_0 instead. Check the book for full details, and note that the fact that power series converge absolutely within their intervals of convergence is important since making this center-shifting idea work involves having to rearrange the terms of a series.

More examples. The function to which the example in the Warm-Up converges is analytic on $(-1, 1)$. Indeed, we showed that this function was equal to the value of the power series

$$\sum_{k=0}^{\infty} \frac{x^{3k}}{k+1}$$

on $(-1, 1)$, and the "center shifting" result outlined above implies that this function is thus analytic on all of $(-1, 1)$.

Similarly, the function $\frac{1}{1-x}$ is analytic on $(-1, 1)$ since

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n \text{ on } (-1, 1).$$

However, it turns out that this function is analytic elsewhere as well, and in fact is analytic on all of $\mathbb{R} \backslash \{1\}$. (It certainly cannot be analytic on an interval which contains 1 since it is not defined at 1, and there is no way to give it a value at to make it differentiable at 1.) The point is that if we want to express $\frac{1}{1-x}$ as a convergent power series at some point outside of $(-1, 1]$, we need to use a different series. This illustrates a point mentioned previously, that it is not always the case that a single power series will express a given analytic function over an entire domain, but in general that we need different series over different regions of that domain. We'll come back to this example next time and explicitly work out power series representations elsewhere.

Taylor series. Finally we come to the fact that if a given function is expressible as a power series, there is only one series which can do the job. Suppose that

$$f(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots .$$

Plugging in x_0 on the right side gives only a_0 since all other terms will contain a power of $x_0 - x_0 = 0$. Thus $f(x_0) = a_0$ is the constant term. Next, taking derivatives gives

$$f'(x) = a_1 + 2a_2(x - x_0) + \cdots ,$$

so $f'(x_0) = a_1$. In general, taking n-th derivatives gives

$$f^{(n)}(x) = n! a_n + (\text{terms with } (x - x_0)), \text{ so } f^{(n)}(x_0) = n! a_n.$$

Thus the coefficients of the given power series *must* be given by $a_n = f^{(n)}(x_0)/n!$, meaning that the power series in question must be the *Taylor series* of f centered at x_0:

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n.$$

Hence, in the definition of analytic, we can omit the reference to the existence of "a" power series and replace this specifically by the Taylor series centered at the given point. All of the series describing analytic functions we've seen so far are examples of Taylor series centered at 0.

Important. We can rephrase the definition of analytic using the notion of a Taylor series: a smooth function f is analytic on E if for any $x_0 \in E$, the Taylor series of f centered at x_0 has positive radius of convergence and converges to f on its interval of converges. Or more simply by shifting centers: a smooth function f is analytic on E if we can cover E by intervals such that on each interval f equals its own Taylor series centered at some point in that interval.

Lecture 10: More on Analytic Functions

Today we continued talking about analytic functions, looking at more examples and methods for showing a function is analytic.

Warm-Up. We show that $\frac{1}{1-x}$ analytic on $\mathbb{R}\backslash\{1\}$, as we claimed last time. Recall that previously we showed this function is analytic on $(-1,1)$ since

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n \text{ for } x \in (-1,1).$$

Now, let $x_0 \neq 1$, and rewrite $\frac{1}{1-x}$ as follows:

$$\frac{1}{1-x} = \frac{1}{(1-x_0)-(x-x_0)} = \frac{1}{1-x_0}\left(\frac{1}{1-\frac{x-x_0}{1-x_0}}\right).$$

Using $\frac{1}{1-y} = \sum_{n=0}^{\infty} y^n$ for $|y|<1$ with $y = \frac{x-x_0}{1-x_0}$, we have:

$$\frac{1}{1-x} = \frac{1}{1-x_0}\sum_{n=0}^{\infty}\left(\frac{x-x_0}{1-x_0}\right)^n = \sum_{n=0}^{\infty}\frac{1}{(1-x_0)^{n+1}}(x-x_0)^n \text{ when } \left|\frac{x-x_0}{1-x_0}\right| < 1.$$

For $x_0 > 1$, this expresses $\frac{1}{1-x}$ as a convergent power series on $(1, 2x_0-1)$, while for $x_0 < 1$ this expresses $\frac{1}{1-x}$ as a convergent power series on $(2x_0-1, 1)$. Thus $\frac{1}{1-x}$ can be expressed as a convergent power series on intervals covering $\mathbb{R}\backslash\{1\}$, so it is analytic on this domain.

Smooth does not imply analytic. Certainly, the Taylor series of a function around a given point is defined as long as that function is infinitely differentiable at that point. However, being able to write down this Taylor series is not enough to show that function is analytic, since it *may not* converge to the function in question! The key point in the definition of analytic is that the given function *itself* can be written as the sum of a convergent power series.

Consider the function f on $\mathbb{R}$ defined by

$$f(x) = \begin{cases} e^{-1/x} & x > 0 \\ 0 & x \leq 0. \end{cases}.$$

Using the chain rule and other properties of derivatives, this function is certainly infinitely-differentiable at any $x \neq 0$. In fact it is also infinitely-differentiable at 0; for instance,

$$\lim_{x \to 0} \frac{f(x) - f(0)}{x - 0} = 0$$

since for $x \leq 0$, the fraction we are taking the limit of is identically zero, while for $x > 0$ we are taking the limit of

$$\lim_{x \to 0^+} \frac{e^{-1/x}}{x},$$

which is also 0. (Any limit of the form $\lim_{x \to 0} e^{-1/x}/p(x)$, where $p(x)$ is a polynomial, is 0; this takes some effort to prove, but the intuition is that $e^{-1/x} \to 0$ as $x \to 0$ much faster than any polynomial.) Thus f is differentiable at 0 with $f'(0) = 0$ and:

$$f'(x) = \begin{cases} \frac{e^{-1/x}}{x^2} & x > 0 \\ 0 & x \leq 0. \end{cases}$$

For the second derivative we have:

$$\lim_{x \to 0} \frac{f'(x) - f'(0)}{x - 0} = \lim_{x \to 0^+} \frac{e^{-1/x}}{x^3} = 0$$

for the same reason as before. Thus f is twice-differentiable and $f''(0) = 0$. Continuing in this way, we can in fact determine that f is infinitely-differentiable on $\mathbb{R}$ and $f^{(n)}(0) = 0$ for all n.

Hence the Taylor series of f centered at 0 is

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n = \sum_{n=0}^{\infty} 0,$$

which converges on all of $\mathbb{R}$ to the constant zero function. But, f is nonzero for $x > 0$, so this Taylor series—although convergent—does not converge to f. This shows that f is *not* analytic on any interval containing 0: to be analytic requires that f be expressible as a power series, but the only possible series which can satisfy this requirement—the Taylor series—does not equal f. The book has another example of a non-analytic smooth function, and Homework 3 has another.

Example. Now we show that $f(x) = e^x$ is analytic on $\mathbb{R}$. This uses *Taylor's Theorem* from last quarter, and will lead to a general method for trying to show that functions are analytic. (Take note though that this general method is not easy to apply in general, and really only works well for functions whose derivatives are easily computable.)

We claim that

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \text{ for all } x \in \mathbb{R},$$

where the right side is the Taylor series of e^x centered at 0. (Of course, this series is sometimes taken to be the *definition* of e^x, but here we are assuming that e^x is defined in some other manner, in which case this equality is far from obvious.) For a fixed $x \in \mathbb{R}$, Taylor's Theorem says that there exists c between 0 and x such that

$$|f(x) - (n\text{-th order Taylor polynomial})| = \left| e^x - \left(1 + x + \cdots + \frac{x^n}{n!} \right) \right| = \left| \frac{f^{(n+1)}(c)}{(n+1)!} x^{n+1} \right|.$$

Since f equals its own derivative, we have:

$$\left| \frac{f^{(n+1)}(c)}{(n+1)!} x^{n+1} \right| = e^c \frac{|x|^{n+1}}{(n+1)!},$$

which converges to 0 as $n \to \infty$. (This was a fact we saw last quarter: for any $a \in \mathbb{R}$, $a^n/n! \to 0$.) The Taylor polynomials here are the partial sums of the Taylor series, so this shows that the given Taylor series converges pointwise to f on $\mathbb{R}$, and this implies that f is analytic on $\mathbb{R}$ as claimed. (In fact, using the general properties of power series we've seen, this implies that $\sum x^n/n!$ converges uniformly to e^x on any bounded interval.)

Theorem. The use of Taylor's Theorem in the example above can be generalized to other functions. The result is:

> If f is smooth on E and there exists $M > 0$ such that $|f^{(n)}(x)| \leq M^n$ for $n \geq 1$ and all $x \in E$, then f is analytic on E.

The point is that the given inequality gives a bound on the remainder term in Taylor's Theorem, so that $|f(x) - (n\text{-th order Taylor polynomial})|$ is bounded by something of the form

$$\frac{|Mx|^{n+1}}{(n+1)!},$$

which always converges to 0 as $n \to \infty$, implying that the sequence of Taylor polynomials converges pointwise to f. Check the book for full details.

Examples. Since the derivatives of $\sin x$ and $\cos x$ are all bounded by $M = 1$, this result implies that these functions are analytic on all of $\mathbb{R}$:

$$\sin x = \sum_{n=0}^{\infty} (-1)^{n+1} \frac{x^{2n+1}}{(2n+1)!} \text{ and } \cos x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} \text{ for all } x \in \mathbb{R}.$$

The right sides are the well-known Taylor series of these functions centered at 0, and the above theorem gives the equalities.

However, as noted earlier, the reason why the theorem works nicely in the case of $e^x, \sin x$, and $\cos x$ is that these all have nice derivatives, making the bounds M^n required in the theorem straightforward to obtain. For functions with not-so-simple-to-compute derivatives, this theorem may not be so helpful.

Other facts. Using the fact that sums of convergent series are convergent, it follows that the sum of analytic functions is analytic. Also, since the derivative of a power series is again a power series, the derivatives of analytic functions are also analytic. In addition, it turns out that products and quotients of analytic functions are also analytic, at least as long the as the denominator is nonzero in the case of quotients. We'll look at the case of quotients next time, but here is idea behind the proof for products.

Suppose that f and g are both analytic on some domain. Then for each x_0 in that domain, we can express each as convergent power series near x_0:

$$f(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n \text{ and } g(x) = \sum_{n=0}^{\infty} b_n (x - x_0)^n.$$

(These might have different radii of convergence, but they for sure both convergent for the smaller of these radii, so we may as well assume they both have the same radius of convergence.) Since power series converge absolutely, we can multiply these series together and rearrange terms to obtain:

$$\begin{aligned} f(x)g(x) &= [a_0 + a_1(x-x_0) + a_2(x-x_0)^2 + \cdots][b_0 + b_1(x-x_0) + b_2(x-x_0)^2 + \cdots] \\ &= a_0b_0 + (a_0b_1 + a_1b_0)(x-x_0) + (a_0b_2 + a_1b_1 + a_2b_0)(x-x_0)^2 + \cdots \\ &= \sum_{n=0}^{\infty} c_n(x-x_0)^n \text{ where } c_n = \sum_{k=0}^{n} a_k b_{n-k}. \end{aligned}$$

This series turns out to have the same radius of convergence as the series we multiplied together, which shows that $f(x)g(x)$ can also be expressed as the sum of a convergent power series, and this shows that the product of analytic functions is analytic. Check the book for full details, and note again the role which absolute convergence plays in rearranging terms.

Important. To show that a function is analytic on some domain, we can:

- show that the function is a sum, product, or quotient of analytic functions,
- start with some well-known analytic function and manipulate it and its Taylor series to get our function (such as in the case of the Warm-Up today, or the Warm-Up from last time),
- show that for some $M > 0$, the n-th derivative of our function is bounded by M^n.

Of these, the first two methods are the simplest to carry out; the third is really only useful for functions with easy-to-compute-derivatives, meaning derivatives for which we can find explicit formulas.

Lecture 11: Yet More on Analytic Functions

Today we finished talking about analytic functions, looking at a few more properties and giving a sense as to why their important. The culmination of these ideas belongs to the realm of *complex analysis*, which we gave a way-too-brief introduction to.

Warm-Up. We show that the function f defined by

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

is analytic on $\mathbb{R}$. (Note that the value at 0 is simply $\lim_{x\to 0} \frac{e^x-1}{x}$, which makes f continuous on all of $\mathbb{R}$.) Now, this function is indeed smooth, but the derivatives get tedious to compute and no nice general pattern emerges; this makes trying to bound the derivatives using powers of a number M difficult, so the theorem derived last time based on Taylor's Theorem will not be very useful.

Instead we can easily work out how to express this function as the sum of a convergent power series. Indeed, starting with the Taylor series for e^x centered at 0, which converges for all x, we have:

$$e^x - 1 = \left(1 + x + \frac{x^2}{2!} + \cdots\right) - 1 = x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots.$$

Thus dividing by $x \neq 0$ gives:

$$\frac{e^x - 1}{x} = 1 + \frac{x}{2!} + \frac{x^2}{3!} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{(n+1)!} \text{ for nonzero } x \in \mathbb{R}.$$

Evaluating this series at $x = 0$ gives the value of 1, so this series indeed converges to f on all of $\mathbb{R}$, and thus f is analytic on $\mathbb{R}$.

Reciprocals. Last time we stated that the quotient of analytic functions is analytic. As a special case, we give the idea behind showing that the reciprocals of analytic functions are analytic. This implies the analogous result for quotients since if f and g are analytic (with g nonzero), f/g is the product of the analytic functions f and $1/g$, so is analytic itself.

Suppose that f is analytic and nonzero, so that we can write

$$f(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n$$

for x in some domain. We need to find coefficients such that the power series $\sum_{n=0}^{\infty} b_n (x - x_0)^n$ satisfies

$$[a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots][b_0 + b_1(x - x_0) + b_2(x - x_0)^2 + \cdots] = 1,$$

which is what is required in order to say that $\sum_{n=0}^{\infty} b_n (x - x_0)^n$ is the reciprocal of f. Multiplying out the left side above, collecting like-terms, and comparing coefficients with the (finite) series on the right, we get the equations:

$$a_0 b_0 = 1,\ a_0 b_1 + a_1 b_0 = 0,\ a_0 b_2 + a_1 b_1 + a_2 b_0 = 0,$$

and in general $a_0 b_n + \cdots + a_n b_0 = 0$. Since f is nonzero, $a_0 \neq 0$ since otherwise $f(x_0)$ would be zero. Thus we get

$$b_0 = \frac{1}{a_0}, \text{ and then } b_1 = -\frac{a_1 b_0}{a_0},$$

and continuing we can use the equations above to solve for each b_n in terms of the previously determined b_k's. Thus the series $\sum_{n=0}^{\infty} b_n (x - x_0)^n$ asked for above does exist, and if it converges it in fact converges to $1/f$.

The difficult part about this is showing that the series derived for $1/f$ indeed converges. This is true, and further this series has the same radius of convergence as the one for f, but we'll skip the proof here, which is quite involved.

Identity Theorem. Here is an important property of analytic functions, which says that the behavior of an analytic function over a certain region completely determines how it behaves over every region. In class we organized this proof with a lemma first, but here I'll avoid this since I think this is simpler to follow. (The Identity Theorem is precisely Theorem 7.56 in the book, only here we give a somewhat different proof.)

Suppose that f and g are both analytic on some domain E and that $f(x) = g(x)$ for all x in some interval $I \subseteq E$. The claim is that $f = g$ on all of E. By way of contradiction, suppose that it is not true that $f(x) = g(x)$ for all $x \in E$ and pick $x_0 \in I$. Since f and g are analytic, so is $f - g$ and hence we can express $f - g$ as a convergent power series near x_0:

$$f(x) - g(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n \text{ for } x \in (x_0 - \delta, x_0 + \delta) \subset I.$$

If all coefficients a_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 on all of E; but this is not true since we are assuming that $f - g$ is *not* always zero on all of E. Thus this series must have some nonzero coefficients. Since $x_0 \in I$, we have $f(x_0) - g(x_0) = 0$, and since this gives the constant term in the series above we know that the first nonzero coefficient must be some a_k where $k \geq 1$.

With this k, we can then write our series as

$$f(x) - g(x) = a_k(x - x_0)^k + a_{k+1}(x - x_0)^{k+1} + \cdots,$$

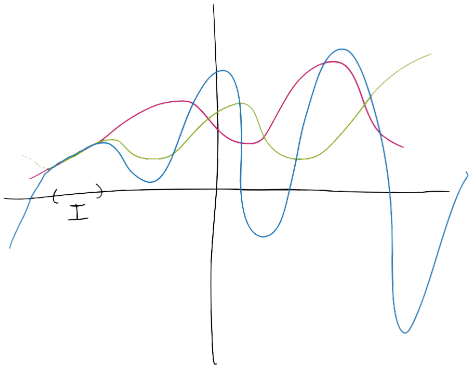
where we've ignored all the zero terms occurring before we hit $a_k(x - x_0)^k$. Factoring out $(x - x_0)^k$ we get

$$f(x) - g(x) = (x - x_0)^k \underbrace{[a_k + a_{k+1}(x - x_0) + a_{k+2}(x - x_0)^2 + \cdots]}_{h(x)},$$

where we denote the function to which the power series in brackets converges to by h. Since $a_k \neq 0$, $h(x_0) \neq 0$, and since h is continuous (as are the sums of any convergent power series), we have that h is nonzero on some interval around x_0. Hence on this interval the only way in which $f - g$ can be zero is for the first factor $(x - x_0)^k$ to be zero, which only occurs at x_0. Thus, the only point on this interval at which $f - g$ is zero is x_0, contradicting the fact that $f(x) = g(x)$ for all $x \in I$ since I contains x_0. Thus $f = g$ on all of E as claimed.

Important. If f and g are analytic and agree on some interval, no matter how incredibly small (imagine an interval the size of an electron!) that interval may be, then f and g must agree everywhere. So, the behavior of an analytic function over one interval fully determines what is must look like on some interval incredibly far away.

This is in sharp contrast to non-analytic functions, where it is possible for such functions to be the same over some interval but not elsewhere. For instance, we can have smooth functions whose graphs look like:



These functions agree over the interval I but not everywhere; analytic functions cannot possibly have such behavior.

Complex Analysis. Let C denote the set of continuous functions, and C^k the set of functions which are k-times continuously differentiable. We then have the containments:

$$C \supset C^1 \supset C^2 \supset C^3 \supset \cdots C^\infty \supset \{\text{analytic functions}\},$$

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 beginning in the subject known as *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:

$$f'(z) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}.$$

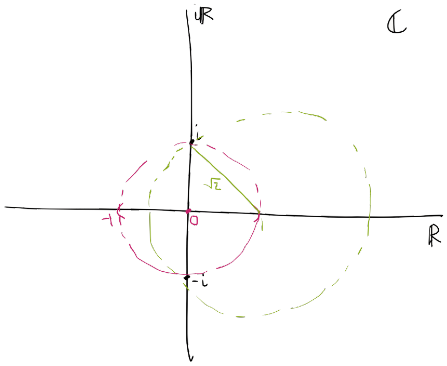
Here, $f(z)$ denotes a complex function $f : \mathbb{C} \to \mathbb{C}$, z is a complex variable, and z_0 a complex number. This limit makes perfect sense in this settings as well, and all the same derivative rules we had for real functions work in this setting too: the derivative of z^n is nz^{n-1}, the derivative of $\sin z$ is $\cos z$ (once we've defined what it means to take sin or cos of a complex number), and so on.

Here is the punchline. In complex analysis, the distinctions given in the containments above for real functions disappears: 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.

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 $\pm i$. 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 x_0 has radius of convergence equal to the radius of the largest possible disk which can be drawn centered at x_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 $\sqrt{2}$.

Important. Real analytic functions are the only types of real functions which can be extended to complex differentiable functions. Complex analysis is awesome, and here we've only given a brief glimpse; take Math 325 to learn more.

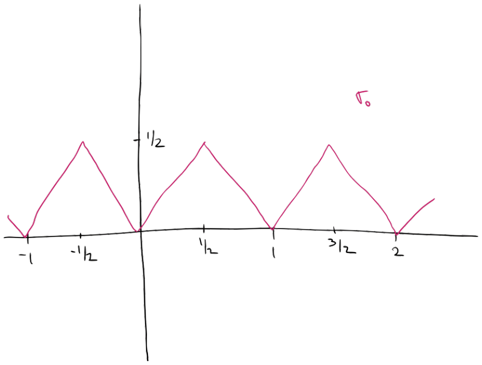
Lecture 12: Metric Spaces

Today we started talking about the notion of a metric space, which is the topic we'll focus on the rest of the quarter. Metric spaces provide a natural setting in which to talk about generalizations of many concepts we've seen, and in particular we'll see that this provides another point of view on uniform convergence.

Continuous but nowhere-differentiable function. Before moving on to metric spaces, we give one final example using material we've seen so far of a function with some bizarre-looking properties. Define the function $\sigma_0 : \mathbb{R} \to \mathbb{R}$ by first setting

$$\sigma_0(x) = \begin{cases} x & 0 \leq x \leq \frac{1}{2} \\ 1-x & \frac{1}{2} < x \leq 1 \end{cases}$$

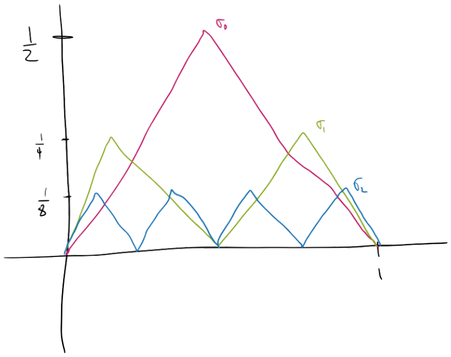
and then extending σ_0 to the rest of $\mathbb{R}$ so that it has period 1; so, the graph of σ_0 looks like:



where we have "peaks" at half-integer points and "valleys" at the integers. Note that the peaks occur at a height of $\frac{1}{2}$. Then for each $k \geq 1$ set

$$\sigma_k(x) := \frac{1}{2^k}\sigma(2^k x),$$

which visually has the effect of shrinking the peaks by a factor of $\frac{1}{2^k}$ and horizontally-compressing the peaks so that there are 2^k peaks within each region where previously there was one peak in the graph of σ_0:



Define the function $f : \mathbb{R} \to \mathbb{R}$ by

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

First, this is well-defined: we have $|\sigma_k(x)| \leq \frac{1}{2^{k+1}}$ for all $x \in \mathbb{R}$, so this series converges uniformly on $\mathbb{R}$ by the Weierstrass M-test. Furthermore, since each σ_k is continuous on $\mathbb{R}$, f is continuous on all of $\mathbb{R}$.

Now here is the surprising part: f is *nowhere* differentiable! While we certainly know examples of continuous functions which are continuous everywhere but not everywhere differentiable—each of the σ_k's for instance—up until now all such functions we know are still differentiable at most points and only fail to be differentiable at some scattered points. In fact, any function you start to draw by hand will be differentiable at some points since by drawing it you will undoubtedly draw some "smooth" portion and the drawn function will be differentiable at such "smooth" points. It is impossible to draw by hand, or really even by computer, a function which is continuous and yet nowhere differentiable.

The proof that f is nowhere differentiable is in the book in the final section of Chapter 7, which we didn't cover, but here is the intuition. Note that σ_0 is not differentiable at any integer since its graph has a sharp "peak" or "valley" at such points. Then, σ_1 has even more peaks and valleys, and so fails to be differentiable at even more points, and so on: the points at which the σ_k's fall to be differentiable get "bunched up" as k gets larger, and this effect then prevents their sum from being differentiable anywhere.

Fun fact. So, we have an example of a continuous but nowhere differentiable function on $\mathbb{R}$. While such an example might seem to be a rarity (based on our experience with continuous functions thus far), it is actually the rule rather than the exception: it is a fact that "most" continuous functions from $\mathbb{R}$ to $\mathbb{R}$ will be nowhere differentiable! Thus, "most" continuous functions behave like the strange example above, again counter to our intuition based on working with functions like e^x, trig functions, and polynomials, meaning that if you pick a continuous function at random, it is more likely than not to be nowhere differentiable.

We will not give a proof of this, but we will get to a point where we can say precisely what "most" means in this context. It turns out that giving a precise definition of "most" here depends on the types of *topological* concepts we'll soon get into with metric spaces.

Motivation for metric spaces. A metric space is essentially a space where we have a notion of distance between points defined; no more, no less. The point is that if you look back through many of the concepts we saw last quarter—sequence convergence, continuity, limits—you'll notice that all were phrased in terms of expressions like $|x - y|$, which gives the distance between two points on the real number line. If you replace these distances with some other type of "distance", you automatically get notions of convergence and continuity in more general contexts, and many of the same theorems we saw earlier will hold more generally in other types of "spaces". Thus, these more general types of spaces can be studied using the same types of ideas we had for $\mathbb{R}$.

Remark on references. The book covers metric spaces in Chapter 10, although much of what it does in Chapters 8 and 9 for $\mathbb{R}^n$ specifically is subsumed by this later material. So, we won't explicitly cover anything from Chapters 8 or 9, but will see it all as part of our study of general metric spaces.

Also, the book doesn't go into as much detail in Chapter 10 as I would like, so in addition to the book take note of the "*Notes on Metric Spaces*" posted on canvas. These are notes I wrote up a few years ago for a similar course taught elsewhere, and are closer in line with how I think a lot of this should be taught. Be sure to use these as a reference as well, in addition to the lecture notes.

Definition. Let X be a set. A *metric* on X is a function $d : X \times X \to \mathbb{R}$ such that

- $d(p, q) \geq 0$ for all $p, q \in X$ and $d(p, q) = 0$ if and only if $p = q$,
- $d(p, q) = d(q, p)$ for all $p, q \in X$,
- $d(p, q) \leq d(p, r) + d(r, q)$ for all $p, q, r \in X$.

A *metric space* is X together with a chosen metric d; we often use the notation (X, d) to denote a metric space, or simply X if the metric is clear from context. (But don't forget that the metric is part of required data.)

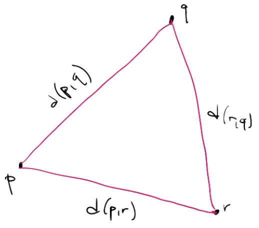
The intuition is simple: d should be thought of as a "distance function" which gives the distance $d(p, q)$ between two points $p, q \in X$. The first condition in the definition says that these "distances" are always nonnegative and equal 0 only when we are computing the distance from a point to itself and the second condition says that the distance from p to q should be the same as the distance from q to p, both of which are clearly properties which "distance" should satisfy.

The third property is called the *triangle inequality* and is the most important one: it says that the "distance" from p to q should be the give the shortest way of going from p to q in the sense that going through some "intermediate" point r can only increase the overall distance. Again, it makes intuitive sense that a notion of "distance" should satisfy this. We'll see what this looks like for $\mathbb{R}^2$ below, which will explain where the name "triangle inequality" comes from.

Example 1. The *Euclidean* metric on $\mathbb{R}^n$ is defined by the usual notion of distance in these spaces:

$$d((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sqrt{|x_1 - y_1|^2 + \cdots + |x_n - y_n|^2}.$$

In particular, on $\mathbb{R}^1$ this gives the distance $d(x, y) = |x - y|$ we've been using all along. Verifying that this is a metric is fairly straightforward, although the triangle inequality takes some work to establish. Instead of doing this thoroughly in general, let us point out what the triangle inequality is true in $\mathbb{R}^2$. In this case, we have a picture like:



in which case the triangle inequality says that the length of one side of this triangle is smaller than or equal to the sum of the lengths of the other two sides, which is clear from what we know about triangles. As alluded to earlier, this is where the name "triangle inequality" comes from.

The case of $\mathbb{R}^2$ with the Euclidean metric is probably the most important example of a metric space to keep in mind as far as intuition goes; in particular, whenever we draw a picture meant to illustrate some general property of metric spaces, it will be a picture in $\mathbb{R}^2$.

Example 2. To emphasize the role which the metric plays in all this, note that in addition to the Euclidean metric we have other possible metrics we can put on $\mathbb{R}^2$. (These definitions generalize to $\mathbb{R}^n$ but we'll state them only for $\mathbb{R}^2$ to keep things simpler.)

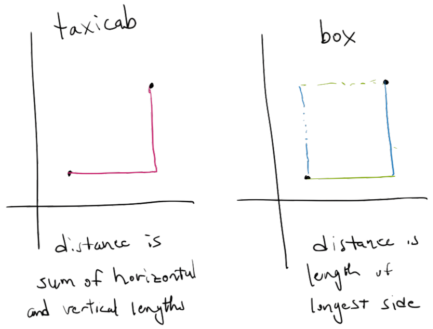
The *taxicab* metric on $\mathbb{R}^2$ is defined by adding together the distance between the x-coordinates of two points to the distance between their y-coordinates:

$$d((x_1, y_1), (x_2, y_2)) = |x_1 - x_2| + |y_1 - y_2|,$$

and the *box* metric on $\mathbb{R}^2$ is defined by taking the maximum of the distance between the x-coordinates of two points and the distance between their y-coordinates:

$$d((x_1, y_1), (x_2, y_2)) = \max\{|x_1 - x_2|, |y_1 - y_2|\}.$$

You will verify on the homework that both of these are indeed metrics. The names of these two metrics come from the following picture:


taxicab
distance is sum of horizontal and vertical lengths
box
distance is length of longest side

The distance between two points with respect to the taxicab metric is the distance you have to travel to get from point to the other if you can only move vertically and horizontally but not "diagonally" (as if you were driving a taxicab on grid-like streets), and the distance between two points with respect to the box metric is the length of the largest side of the rectangle (i.e. "box") with one corner at the first point and opposite corner at the other.

Example 3. For any set X, the *discrete* metric on X is defined by setting the distance between distinct points to always be 1:

$$d(p,q) = \begin{cases} 1 & p \neq q \\ 0 & p = q. \end{cases}$$

The first two requirements in the definition of a metric are straightforward to check, and the triangle inequality comes from looking at the possible values the terms in the expression

$$d(p,q) \leq d(p,r) + d(r,q)$$

can have: if the left side is 0 then the inequality holds no matter what the right side is, while if the left side is 1, meaning $p \neq q$, then at least one of the term on the right is also 1 (since either $p \neq r$ or $q \neq r$ or both), so again the inequality holds.

This will be a useful example to keep in mind, as it can give some insight as to what the various definitions we will be seeing mean. The name comes from the fact that distinct points are always separated by a minimum fixed positive distance, which is not true for "continuous" spaces like $\mathbb{R}$ with the Euclidean metric where distances can get arbitrarily small.

Example 4. We will use the notation $C_b(E)$ to denote the space of bounded real-valued functions on some domain $E \subseteq \mathbb{R}$:

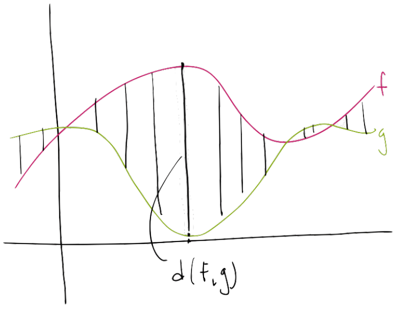
$$C_b(E) := \{f : E \to \mathbb{R} \mid f \text{ is bounded}\}.$$

The *sup metric* on $C_b(E)$ is defined as the supremum of the distances between the values of f and g at common points in E:

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

Note that since f and g are bounded, $f - g$ is bounded so that this supremum actually exists. We will verify next time that this is a metric. Apart from the Euclidean metrics, this is going to be our most important example.

Here is the interpretation of the sup metric. Look at the graphs of f and g and take all the vertical distances between various points on these graphs:



The distance between these functions with respect to the sup metric is the "largest" such vertical distance, where largest is in quotation marks since such a *maximum* distance may not actually be attained, which is why we're taking the supremum in general. The smaller the "sup distance" between two functions is the "closer" their graphs are to one another.

Important. A metric space is a set with a defined notion of distance between elements. The most important examples are $\mathbb{R}$ and $\mathbb{R}^2$ with the standard Euclidean metrics, and the space of bounded functions $C_b(E)$ on some domain with the sup metric. The best intuition for metric spaces in general comes from drawing pictures in $\mathbb{R}^2$, although (as we'll see) this specific metric space has properties that others in general do not.

Lecture 13: Sequences in Metric Spaces

Today we continued talking about metric spaces, introducing a few more basic definitions. Chief among these are the notion of a "ball" in a metric space of what it means for a sequence to converge.

Warm-Up. We verify that the sup metric is indeed a metric. Recall that this is the metric on the set $C_b(E)$ of bounded functions on some domain $E \subseteq \mathbb{R}$ defined by

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

First, since $|f(x) - g(x)| \geq 0$ for all $x \in E$, the supremum of these values $d(f,g)$ is also nonnegative. If $d(f,g) = 0$ then we must have

$$|f(x) - g(x)| = 0 \text{ so } f(x) = g(x) \text{ for all } x \in E.$$

This verifies the first property in the definition of a metric. Since $|f(x) - g(x)| = |g(x) - f(x)|$ for all $x \in E$,

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

which is the second property.

Finally we verify the triangle inequality. Let $f, g, h \in C_b(E)$. For any $x \in E$ we have

$$|f(x) - h(x) \leq d(f,h) \text{ and } |h(x) - g(x)| \leq d(h,g)$$

since the values on the et of each inequality are among the values the right side is the supremum of. Thus for all $x \in E$,

$$|f(x) - g(x)| = |f(x) - h(x) + h(x) - g(x)| \leq |f(x) - h(x)| + |h(x) - g(x)| \leq d(f,h) + d(h,g).$$

Hence $d(f,h) + d(h,g)$ is an upper bound for the set of all values $|f(x) - g(x)|$ for $x \in E$, and so is bigger than or equal to the least upper bound of such values, which is $d(f,g)$. Thus

$$d(f,g) \leq d(f,h) + d(h,g)$$

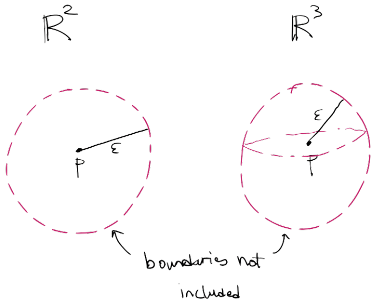
as required, so we conclude that d is a metric on $C_b(E)$.

Definitions. Let (X,d) be a metric space and $Y \subseteq X$. Then restricting the metric on X to only allow ourselves to plug in points of Y gives a metric on Y, and in this case we call Y a *subspace* of X. So, a subspace of a metric space is nothing but a subset, but with the *same* metric as on the larger space. (For instance, $\mathbb{Q}$ with the usual Euclidean distance is a subspace of $\mathbb{R}$ with the usual Euclidean distance, but $\mathbb{Q}$ with the discrete metric is not.)

For $p \in X$ and $\epsilon > 0$, the *ball of radius* ϵ (or the ϵ*-ball*) in X centered at p is the set $B_\epsilon(p)$ of points of X whose distance to p is less than r:

$$B_\epsilon(p) = \{q \in X \mid d(p,q) < \epsilon\}.$$

The name comes from the picture of what these sets look like in $\mathbb{R}^2$ or $\mathbb{R}^3$ with the standard Euclidean metrics: in $\mathbb{R}^2$ the ball of radius ϵ around a point is the (open) disk of radius ϵ centered at that point, and in $\mathbb{R}^3$ these ϵ-balls look like honest solid spheres (without boundary):



Example 1. Consider $\mathbb{R}^2$ with the taxicab metric. The ball of radius 1 centered at the origin is defined by

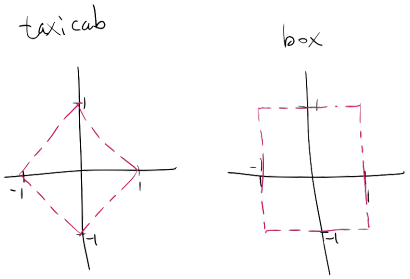
$$B_1((0,0)) = \{(x,y) \in \mathbb{R}^2 \mid d((0,0),(x,y)) < 1\} = \{(x,y) \in \mathbb{R}^2 \mid |x| + |y| < 1\}.$$

The inequality $|x| + |y| < 1$ in $\mathbb{R}^2$ describes a diamond-shaped region, which is thus the "ball" of radius 1 around the origin origin with respect to the taxicab metric. Each of the points on the diamond itself (the boundary) are at a distance 1 from $(0,0)$ and so are not in this ball.

With respect to the box metric, the ball of radius 1 centered at the origin is

$$B_1((0,0)) = \{(x,y) \in \mathbb{R}^2 \mid \max\{|x|,|y|\} < 1.$$

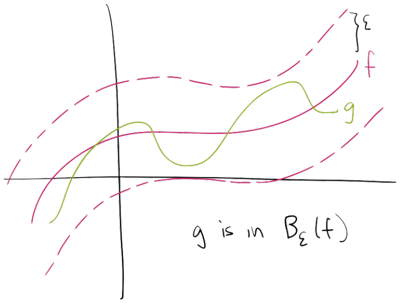
The condition $\max\{|x|,|y|\} < 1$ describes a square (i.e. box) centered at the origin of width and length equal to 2. The points on the boundary square are not included in the ball of radius 1 since they are at a distance 1 from the origin.



Example 2. Let (X, d) be a discrete metric space. For any $0 < r \leq 1$ and a fixed $p \in X$, the only possible distance $d(p, q)$ which satisfies $d(q, p) < r$ is $d(p, q) = 0$ since the only possible values for $d(p, q)$ are 0 or 1 in this case. Since $d(p, q) = 0$ if and only if $p = q$, the r-ball around p for $0 < r \leq 1$ consists only of p. For any $r > 1$, *any* $q \in X$ satisfies $d(q, p) < r$, so any ball of radius larger than 1 in this case consists of the entire space X. To summarize:

$$B_r(p) = \begin{cases} \{p\} & 0 < r \leq 1 \\ X & r > 1. \end{cases}$$

Example 3. Consider $C_b(E)$ with the sup metric. This space itself is not possible to visualize, but we can still draw pictures of the ϵ-balls represent. For $f \in C_b(E)$ and $\epsilon > 0$, draw the ϵ-tube around the graph of f, a concept we previously discussed when looking at uniform convergence. Then $B_\epsilon(f)$ almost consists of those functions g whose graphs lie fully within this tube:



The point being that the requirement $d(f, g) < \epsilon$ says that all vertical distances $|f(x) - g(x)|$ as x ranges throughout E should be less than ϵ. However, note that a function g whose graph gets arbitrarily close to the boundary of the ϵ-tube will in fact have $d(f, g) = \epsilon$, and so such a function is not in the ϵ-ball centered at f. Thus $B_\epsilon(f)$ more precisely consists of those functions g whose graphs are contained within the ϵ-tube around the graph of f *and* do not come arbitrarily close the boundary of this tube.

Important. The ball of radius $r > 0$ around $p \in X$ is the set of all points of X whose distance to p is less than r. We'll come to see this concept helps to formulate many of the definitions and ideas we'll see when dealing with metric spaces.

Sequences. And now we come to one of the motivations we had for introducing metric spaces: the idea that we can take previous definitions and generalize them simply by replacing the distance $|x - y|$ between points of $\mathbb{R}$. In particular, we have the following.

Let X be a metric space. A *sequence* in X is an infinite list (p_n) of elements of X:

$$p_1,\ p_2,\ p_3,\ p_4,\ \ldots.$$

We say that the sequence (p_n) *converges* to $p \in X$ if for any $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that

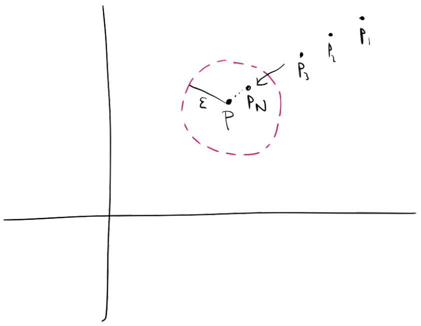
$$d(p_n, p) < \epsilon \text{ for } n \geq N.$$

The intuition is as follows: the condition $d(p_n, p) < \epsilon$ says that $p_n \in B_\epsilon(p)$, so the definition says that given any ball around p, no matter how small its radius, eventually all terms in the sequence (p_n) are in that ball. This captures the idea that the terms p_n are getting closer and closer to p.

If a sequence converges in X we say it is *convergent*, and if not we say it is *divergent*, or that it *diverges* in X.

Remark. Note that the specific metric space in question matters. For instance, take a sequence (r_n) of rationals converging to $\sqrt{2}$ with respect to the Euclidean metric. This sequence is convergent in $\mathbb{R}$ but it is considered to be *divergent* in $\mathbb{Q}$ since the thing to which it converges does not exist in the metric space $\mathbb{Q}$. (We'll soon see that this is an example of a *Cauchy sequence* in $\mathbb{Q}$ that does not converge in $\mathbb{Q}$.)

Examples. In $\mathbb{R}^2$ with the Euclidean metric, a convergent sequence looks like:



Indeed, as we described earlier when giving the intuition behind the definition of convergence, any ball we draw around p, no matter how small, has the property that all terms p_n past some index are in it. (This picture also illustrates something we mentioned earlier, that pictures drawn in $\mathbb{R}^2$ will help to clarify many metric concepts we'll come across.)

Now consider a discrete space (X, d). If $p_n \to p$ in X, we must have

$$d(p_n, p) < \frac{1}{2}$$

for all n past some index. However, the only way a distance can be smaller than $1/2$ in a discrete space is for it to zero, so the above condition gives

$$d(p_n, p) = 0 \text{ for all } n \text{ past some index,}$$

which in turn says that $p_n = p$ for all n past some index. Thus, a sequence in a discrete space is convergent if and only if it is *eventually constant*, meaning that all terms past some index are the same.

Back to uniform convergence. Finally we look at an example which starts to illustrate why we are looking at metric spaces now at this particular point in this year-long course. Consider $C_b(E)$ with the sup metric, and suppose that $f_n \to f$. Thus for any $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that

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

Since for any $x \in E$, $|f_n(x) - f(x)|$ is smaller than or equal to this supremum, we get

$$|f_n(x) - f(x)| < \epsilon \text{ for } n \geq N \text{ and all } x \in E,$$

which is precisely the definition of what it means for (f_n) to converge to f uniformly on E!

Conversely, if $f_n \to f$ uniformly on E, then for any $\epsilon_1 > 0$, there exists $N \in \mathbb{N}$ such that

$$|f_n(x) - f(x)| < \epsilon_1 \text{ for } n \geq N \text{ and all } x \in E.$$

This implies that $d(f_n, f) \leq \epsilon_1$ for $n \geq N$. For a fixed $\epsilon > 0$, applying this condition to some $\epsilon_1 < \epsilon$ gives

$$d(f_n, f) < \epsilon \text{ for all } n \geq N,$$

which means that $f_n \to f$ with respect to the sup metric. Thus we come to the conclusion that...(drumroll)...convergence with respect to the sup metric means *precisely* the same thing as uniform convergence! Thus, everything we did previously with uniform convergence can now be viewed as studying properties of the metric space $C_b(E)$, and as we develop further metric space material we'll be able to go back and apply it to uniform convergence.

Important. Convergence with respect to the sup metric means the same thing as uniform convergence, which is one of the main reasons why studying metric spaces in general will be useful.

Lecture 14: Completeness

Today we spoke about Cauchy sequences in the metric space setting and what it means for a space to be "complete". Complete spaces are the ones where sequences which "look" like they should converge actually do converge.

Warm-Up 1. We show that a convergent sequence in an arbitrary metric space (X, d) is bounded. Now, the first thing to note is that we have to clarify what it means to say that a subset of an arbitrary metric is bounded; the definition in $\mathbb{R}$ was that there exists $M > 0$ such that $|x| \leq M$ for all x in that subset, but this won't work in general since we do not have the notion of an "absolute value" in an arbitrary metric space. The key is that we can rephrase the definition we had in $\mathbb{R}$ by saying that the subset in question is fully contained within an interval of finite length, and this we can generalize to an arbitrary metric space by replacing "interval of finite length" with "ball of finite radius". Thus we say:

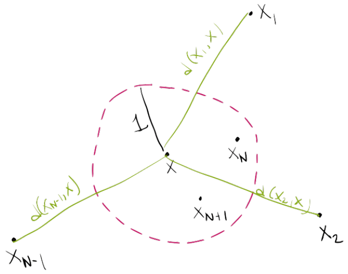
A subset S of X is *bounded* if there exists $p \in S$ and $r > 0$ such that $S \subseteq B_r(p)$.

You'll see on the homework that the point p the ball is centered at here really doesn't matter, in that once we know S is bounded we can find such a ball centered at *any* point of S; the important thing is that there exists a ball of finite radius—regardless of where it's centered—which fully contains S.

Suppose that $x_n \to x$ in X. Then there exists $N \in \mathbb{N}$ such that $d(x_n, x) < 1$ for $n \geq N$, which can be rephrased as saying

$$x_n \in B_1(x) \text{ for } n \geq N.$$

Thus we have a ball of finite radius containing at least all terms in our sequence starting with the N-th one. The idea is now to make this radius large enough so that the corresponding ball includes *all* terms of (x_n). The picture (drawn in $\mathbb{R}^2$ to get some intuition) to have in mind is the following:



We can get a ball which includes x_1 by increasing our current radius of 1 to $1+d(x_1,x)$, then we can make the ball include x_2 by increasing our radius if need be to make it at least as large as $1+d(x_2,x)$, and so on. Thus if we define

$$r := \max\{d(x_1,x), d(x_2,x), \dots, d(x_{N-1},x)\} + 1 > 0,$$

we claim that $B_r(x)$ will contain all terms of (x_n). Indeed, for $1 \le k \le N-1$ we have

$$d(x_k,x) < d(x_k,x) + 1 \le r$$

so $x_k \in B_r(x)$ for $1 \le k \le N-1$, and for $n \ge N$ we have

$$d(x_n,x) < 1 \le r$$

so $x_n \in B_r(x)$ for $n \ge N$. Thus $x_n \in B_r(x)$ for all n, so (x_n) is bounded as claimed.

Remark. Note the strategy above: we used a picture drawn in $\mathbb{R}^2$ to get a sense for what's going on in general, but the actual proof in the end uses only properties of the arbitrary metric d in question and nothing special about $\mathbb{R}^2$. This is a common theme we'll see when working through questions dealing with arbitrary metric spaces.

Warm-Up 2. Let (x_n, y_n) be a sequence of points in $\mathbb{R}^2$. We claim that $(x_n, y_n) \to (x, y) \in \mathbb{R}^2$ with respect to the standard metric on $\mathbb{R}^2$ if and only if $x_n \to x$ and $y_n \to y$ in $\mathbb{R}$ with respect to the standard metric on $\mathbb{R}$. Thus, questions about sequence convergence in $\mathbb{R}^2$ (with respect to the standard metric) can be phrased completely in terms of convergence of the sequences of numbers obtained by looking at each coordinate one at a time. A similar result is true for $\mathbb{R}^n$ in general.

Suppose that $(x_n, y_n) \to (x, y)$ and let $\epsilon > 0$. Then there exists N such that

$$\sqrt{|x_n - x|^2 + |y_n - y|^2} < \epsilon \text{ for } n \ge N.$$

But $|x_n - x|$ and $|y_n - y|$ are each smaller than or equal to this square root, so for $n \ge N$ we have

$$|x_n - x| = \sqrt{|x_n - x|^2} \le \sqrt{|x_n - x|^2 + |y_n - y|^2} < \epsilon$$

and

$$|y_n - y| = \sqrt{|y_n - y|^2} \le \sqrt{|x_n - x|^2 + |y_n - y|^2} < \epsilon.$$

Hence $x_n \to x$ and $y_n \to y$ in $\mathbb{R}$.

Conversely suppose that $x_n \to x$ and $y_n \to y$ in $\mathbb{R}$ and let $\epsilon > 0$. Then there exist N_1, N_2 such that

$$|x_n - x| < \frac{\epsilon}{\sqrt{2}} \text{ for } n \geq N_1 \text{ and } |y_n - y| < \frac{\epsilon}{\sqrt{2}} \text{ for } n \geq N_2.$$

Thus for $n \geq \max\{N_1, N_2\}$ we get

$$\sqrt{|x_n - x|^2 + |y_n - y|^2} < \sqrt{\frac{\epsilon^2}{2} + \frac{\epsilon^2}{2}} = \epsilon,$$

so $(x_n, y_n) \to (x, y)$ in $\mathbb{R}^2$ as desired.

Cauchy sequences. Now we can say what it means for a sequence in a metric space to be Cauchy, simply by taking the definition we had for $\mathbb{R}$ previously and replacing absolute values by distances:

> A sequence (p_n) in a metric space (X, d) is *Cauchy* if for any $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that $d(p_n, p_m) < \epsilon$ for $m, n \geq N$.

The intuition is the same one we had for Cauchy sequences in $\mathbb{R}$: a sequence being Cauchy means that its terms are getting "bunched" up closer and closer together, which suggests that the sequence looks like it *should* converge, although it might not as we'll soon see.

Examples. This definition of Cauchy gives the usual one in the case of $\mathbb{R}$ with the standard metric, and it gives the notion of *uniformly Cauchy* we saw for sequences of functions in the case of $C_b(E)$ with the sup metric.

Consider now a discrete metric space. If (x_n) is Cauchy in this case, then we get

$$d(x_n, x_m) < \frac{1}{2} \text{ for large enough } n, m,$$

which means that $d(x_n, x_m) = 0$ for large enough n, m since this is the only possible distance smaller than $\frac{1}{2}$ in a discrete space. Thus we see that a sequence is Cauchy in a discrete space if and only if it is eventually constant.

In particular, note that the sequence $\left(\frac{1}{n}\right)$ in $\mathbb{Q}$ is Cauchy with respect to the standard metric, but is NOT Cauchy with respect to the discrete metric. The point is that the definition of Cauchy (as with the definition of convergence) depends on the metric being used and not just on the sequence in question.

Cauchy but not convergent. Consider now the sequence in $\mathbb{Q}$ defined as follows. Take the decimal expansion of $\sqrt{2}$:

$$\sqrt{2} = 1.4142\ldots$$

and set $r_1 = 1, r_2 = 1.4, r_3 = 1.41$, and so on at each step we take one more digit in the decimal expansion. We claim that the resulting sequence (r_n) is Cauchy in $\mathbb{Q}$ with respect to the standard metric but is NOT convergent in $\mathbb{Q}$. Indeed, viewed as sequence in $\mathbb{R}$ we have $r_n \to \sqrt{2}$, so (r_n) is convergent in $\mathbb{R}$ and hence Cauchy in $\mathbb{R}$. Since we are considering $\mathbb{Q}$ with the same Euclidean metric as on $\mathbb{R}$ here, (r_n) is also Cauchy in $\mathbb{Q}$. (The point being that the definition of Cauchy only depends on the metric and the terms of the sequence in question, and not on any kind of "limit".) However, (r_n) is not convergent in $\mathbb{Q}$ since the thing to which it should converge does not exist in $\mathbb{Q}$ since $\sqrt{2}$ is not rational.

More generally, any sequence of rationals which converges to an irrational will be an example of Cauchy sequence in $\mathbb{Q}$ which does not converge in $\mathbb{Q}$. This emphasizes that fact that in the definition of "convergent", the limit should actually exist in the space we are looking at and not solely in some "larger" space.

Important. A Cauchy sequence in an arbitrary metric space is one whose terms are getting closer and closer to one another, and looks as if it should converge. However, it is not true that Cauchy sequences in arbitrary metric spaces must converge as this depends on the metric and space in question.

Completeness. A metric space (X, d) is said to be *complete* if every Cauchy sequence in X converges in X. The "converges in X" part is key in that it requires that the limit of the sequence itself belong to the same space as the terms of the sequence. Thus, complete spaces are the ones where sequences which "look" like they should converge actually do converge, giving us a way to show that a sequence converges in a complete space without knowing what the limit will be.

Examples. The previous example shows that $\mathbb{Q}$ with the standard metric is not complete. However, note that $\mathbb{Q}$ with the discrete metric *is* complete, as is any discrete metric space X. Indeed, if (x_n) is Cauchy with respect to a discrete metric, then (x_n) is eventually constant so that there exists $x \in X$ such that $x_n = x$ for large enough n, which implies that $x_n \to x$. The point is that whether or not a space is complete depends on the metric being used: the same set can be complete with respect to one metric but not with respect to another.

More examples. The fact from last quarter that Cauchy sequences in $\mathbb{R}$ always converge shows that $\mathbb{R}$ is a complete metric space with respect to the standard metric. If you go back through the proof of this fact from last quarter you'll see that the underlying reason why this works is that $\mathbb{R}$ has what we previously called *completeness property*, which is the fact that nonempty subsets of $\mathbb{R}$ which are bounded above always have least upper bounds. Indeed, you can view our new notion of "complete" in the metric space sense to be a generalization of this previous completeness property, which as stated does not apply to metric spaces in which there is no notion of one element being larger than another. It can be shown that if you take the fact that $\mathbb{R}$ is complete in the metric space as a given, you can derive the previous completeness property of $\mathbb{R}$, so that the two notions of "complete" in the case of $\mathbb{R}$ are equivalent.

We can also see that $\mathbb{R}^n$ is complete in general with respect to the Euclidean metric. For instance, if (x_n, y_n) is Cauchy in $\mathbb{R}^2$, it can be shown using an idea similar to that in the second Warm-Up that (x_n) and (y_n) are Cauchy in $\mathbb{R}$. But since $\mathbb{R}$ is complete, (x_n) and (y_n) both converge in $\mathbb{R}$, say to x and y respectively. Thus again using the second Warm-Up, since $x_n \to x$ and $y_n \to y$, we get $(x_n, y_n) \to (x, y)$ in $\mathbb{R}^2$, showing that a Cauchy sequence in $\mathbb{R}^2$ always converges in $\mathbb{R}^2$. The same idea works for any $\mathbb{R}^n$, after generalizing the second Warm-Up to $n \geq 3$.

Yet more examples. The space $C_b(E)$ of bounded functions on some domain with respect to the sup metric is complete. Indeed, if (f_n) is Cauchy with respect to the sup metric, then (f_n) is uniformly Cauchy. We saw earlier this quarter that uniformly Cauchy sequences of functions are always uniformly convergent, which means that (f_n) converges with respect to the sup metric as well. Hence any Cauchy sequence in $C_b(E)$ is convergent, so $C_b(E)$ is complete.

Now consider the space of continuous functions on an interval $[a, b]$, which we denote by $C([a, b])$:

$$C([a, b]) = \{f : [a, b] \to \mathbb{R} \mid f \text{ is bounded.}\}$$

Since continuous functions on a closed interval are always bounded (Extreme Value Theorem), we can consider $C([a, b])$ to be a subspace of $C_b([a, b])$ so that the sup metric is defined on it as well. If (f_n) is a Cauchy sequence in $C([a,b])$, then (f_n) is also a Cauchy sequence in $C_b([a,b])$, so it converges in $C_b([a,b])$. Since the uniform limit of continuous functions is always continuous, the limit of this sequence is actually in $C([a,b])$, so the space of continuous functions on $[a,b]$ is complete as well. (We'll see next time that this is a special case of the general fact that a *closed* subspace of a complete metric space is itself complete.)

Important. A complete metric space is one where Cauchy sequences do always converge. The most important examples of complete spaces will be $\mathbb{R}^n$ and $C_b(E)$. The basic example of a noncompete space is $\mathbb{Q}$ with the standard metric, or say an open bounded interval (a,b) also with the standard metric.

Completion. The examples of $\mathbb{Q}$ in $\mathbb{R}$ and of $C([a,b])$ in $C_b([a,b])$ suggest that the only thing which determines whether or not a Cauchy sequence is convergent is whether its limit in the "larger" space exists in the "smaller" space. For this to make sense in general we would have to know that any metric space whatsoever always sits "inside" of another one which *is* complete. This is true in these examples, where any subspace of $\mathbb{R}$ sits inside of the complete space $\mathbb{R}$ and any subspace of $C_b(E)$ sits inside of the complete space $C_b(E)$.

However, once we get to an arbitrary metric space things are not so clear since there is no obvious candidate for what the "larger" space should be. In the case of $\mathbb{Q}$ we know ahead of time what the set of real numbers is, but how could you construct something analogous to this when all you have is a set X with a metric and no further information about what X looks like? It turns out that such a "larger" space always exists, and that there is a "smallest" such space:

> For any metric space (X,d) there exists a complete metric space $\overline{X}$ containing X as a subspace which is the smallest such space in the sense that if M is any complete metric space containing X as a subspace, $\overline{X} \subseteq M$. This $\overline{X}$ is called the *completion* of X with respect to d.

For instance, the competition of $\mathbb{Q}$ with respect to its standard metric is $\mathbb{R}$, and if X is a metric which is already complete, its completion is itself. Note that in the case of $\mathbb{Q}$, $\mathbb{C}$ (the set of complex numbers with the usual Euclidean distance) is also a complete space containing $\mathbb{Q}$ as a subspace, but is not the smallest such space since $\mathbb{R} \subseteq \mathbb{C}$.

We will not prove that the completion always exists since this will not be a notion we'll use again in the rest of the course. For those who are interested, there will be some additional notes on Canvas which go through the construction of the completion and prove that it has all the properties we claim it does. The hard part is in coming up with what type of set the completion should even be, and the answer is that $\overline{X}$ contains as elements Cauchy sequences in X, so we are working with a space whose "points" are themselves Cauchy sequences, and this takes a while to wrap your head around. Again, check the additional notes for details if you are interested.

Lecture 15: Open and Closed Sets

Today we started talking about the notions of *open* and *closed* subsets of a metric space. These so-called "topological" notions provide a concise way to phrase many properties of metric spaces and continuous functions we will be looking at.

Warm-Up. Suppose that (p_n) is a Cauchy sequence in a metric space X and that (p_{n_k}) is a convergent subsequence. We show that (p_n) converges as well. This is something we saw last quarter in the case of the metric space $\mathbb{R}$ and formed the basis of the proof that Cauchy sequences in $\mathbb{R}$ always converge; the general proof is obtained simply by taking the proof in that special case and replacing absolute values by metrics.

Say that $p_{n_k} \to p \in X$ and let $\epsilon > 0$. Since (p_n) is Cauchy there exists N such that

$$d(p_n, p_m) < \frac{\epsilon}{2} \text{ for } n \geq N.$$

Since $p_{n_k} \to p$ there exists K such that

$$d(p_{n_K}, p) < \frac{\epsilon}{2}.$$

By making K larger if necessary we may also assume that $n_k \geq N$, so that p_{n_K} is within the range of terms where the first inequality above holds. Then for $n \geq N$ we have

$$d(p_n, p) \leq d(p_n, p_{n_K}) + d(p_{n_K}, p) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

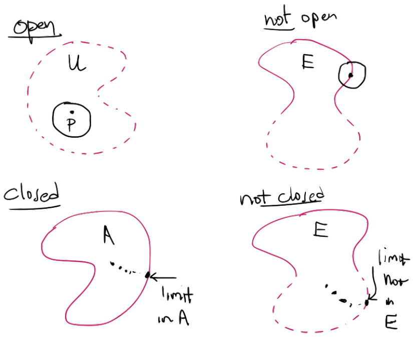
so $p_n \to p$ as claimed. (We will soon see that this is the proof of the fact that so-called *compact* spaces are always complete.)

Open and closed. Let X be a metric space. We say that a subset $U \subseteq X$ is *open* in X if for any $p \in U$ there exists $r > 0$ such that $B_r(p) \subseteq U$. We say that a subset $A \subseteq X$ is *closed* in X if whenever $a_n \to x \in X$ and all $a_n \in A$, then $x \in A$ as well.

The intuition is as follows. If U is open and $p \in U$, the definition says that we can surround p by an entire ball which remains fully contained in U; thus, this says that points which are "close enough" to an element of an open set are themselves also in that open set, so that an open set in a sense "surrounds" all of its points. If A is closed, the definition says if $x \in X$ has the property that there is a sequence of points of A converging to it, then x is also in A; thus, we cannot "escape" a closed set by taking limits of elements inside of it, so that points which are "arbitrarily close" to an element of a closed set are themselves also in that closed set.

Example 1. An open interval (a, b) is an open subset of $\mathbb{R}$ and a closed interval $[a, b]$ is a closed subset of $\mathbb{R}$. Indeed, given $x \in (a, b)$ we can imagine visually that there is an open interval we can draw around x which is fully contained in (a, b), and the claim about closed intervals is just the fact that if $x_n \to x$ and $a \leq x_n \leq b$, then $a \leq x \leq b$. Both of these facts are just special cases of the fact that *open* balls in an arbitrary metric space are always open and *closed* balls are always closed, which we'll prove shortly.

More generally, it is easy to picture what open and closed subsets of $\mathbb{R}^2$ look like, which give a lot of good intuition for these concepts in general. Consider the following subsets of $\mathbb{R}^2$, where dotted curves indicate that those points are not in the subset in question while solid curves indicate that they are:



In the first picture on the top left, given a point $p \in U$ we have drawn a ball around it which is fully contained inside of U, showing that U is open in $\mathbb{R}^2$. In the second picture on top, for a point on the "boundary" of E we can see that any ball we draw around it will contain something not in E, so no such ball will be fully contained in E and hence E is not open in $\mathbb{R}^2$. For the first picture in the second row, we can imagine that if there is a sequence of points of A converging to some $x \in \mathbb{R}^2$, then x itself is in A so A is closed in $\mathbb{R}^2$. (The point is that we cannot have in this case a sequence of points in A which converges to something outside of A.) In the final picture, we have drawn a sequence of points in E converging to something not in E (since the point it converges to is on the "dotted" piece of the boundary), so E is not closed in $\mathbb{R}^2$. In general, a subset of $\mathbb{R}^2$ is open if it contains none of its "boundary" and a subset of $\mathbb{R}^2$ is closed if it contains all of its "boundary". (We'll make the notion of "boundary" precise later on.)

Example 2. The following examples emphasize the idea that the notions of open and closed sets are *relative* ones, in that they depend on what "larger" metric space our sets are sitting inside of. Say that the metric space we are considering is the interval $[-1, 3)$ with the Euclidean metric. We claim that $[-1, 2)$ is actually open (!) in $[-1, 3)$ and that $[2, 3)$ is closed (!) in $[-1, 3)$. Now, of course neither of these smaller intervals are open or closed in $\mathbb{R}$, but the point is that asking whether they are open or closed in the metric space $\mathbb{R}$ is different than asking whether they are open or closed in the metric space $[-1, 3)$.

Surely, for $x \in (-1, 2)$ we can draw an open interval around it which is fully contained in $[-1, 2)$. So, the only point we have to worry about when asking if $[-1, 2)$ is open in $[-1, 3)$ is $x = -1$. We claim that the ball in $[-1, 3)$ of radius 1 centered at -1 *is* contained in $[-1, 2)$. The key point is that when we take a ball around a point, this ball by definition only contains points from our "larger" metric space, which is $[-1, 3)$ in this scenario. The ball in the metric space $[-1, 3)$ of radius 1 centered at -1 is

$$B_1(-1) = \{x \in [-1, 3) \mid |x + 1| < 1\} = [-1, 0),$$

since the elements of $[-1, 0)$ are the only numbers satisfying this inequality *among the points of* $[-1, 3)$. (The ball in $\mathbb{R}$ of radius 1 centered at -1 is $(-2, 0)$, but the points in $(-2, -1)$ do not exist in the "larger" metric space $[-1, 3)$ we are considering here.) Thus the ball in $[-1, 3)$ of radius 1 around $-1$is indeed contained in $[-1, 2)$, so $[-1, 2)$ is open in $[-1, 3)$.

Now, to check that $[2, 3)$ is closed in $[-1, 3)$, we want to know that any sequence in $[2, 3)$ which converges in the "larger" space $[-1, 3)$ has its limit inside of $[2, 3)$. It might seem that this is not true for sequences which converge to 3, but the point is that since 3 is not an element of our larger space, such sequences are *not* to be considered in the definition of closed, which only asks that sequences of the subset which are *assumed* to converge in the larger space actually have their limits in the subset. Thus, sequences in $\mathbb{R}$ which might converge to 3 are irrelevant when asking about a set being closed in $[-1, 3)$. It is definitely true that a sequence of points satisfying $2 \leq x_n < 3$ which converges to something satisfying $-1 \leq x < 3$ must in fact satisfy $2 \leq x < 3$, which is what it means to say that $[2, 3)$ is closed in $[-1, 3)$.

Example 3. We claim that the set E of rational numbers between $-\sqrt{2}$ and $\sqrt{2}$:

$$E := \{r \in \mathbb{Q} \mid -\sqrt{2} < r < \sqrt{2}\}$$

is both open *and* closed in $\mathbb{Q}$, even though it is neither open nor closed in $\mathbb{R}$. If $x \in E$, the open interval U in $\mathbb{R}$ of radius $r = \min\{\sqrt{2} - x, x - (-\sqrt{2})\}$ is fully contained in $(-\sqrt{2}, \sqrt{2})$. Then the ball in $\mathbb{Q}$ of radius r around x consists of only the *rational* numbers in U, and this ball is fully contained in E. (As in the previous example, the irrational numbers in U do not exist in the "larger" space $\mathbb{Q}$ we are considering, so they do not appear in the ball of radius r centered at x in the metric space $\mathbb{Q}$.) This shows that E is open in $\mathbb{Q}$.

To show that E is closed in $\mathbb{Q}$, suppose that (r_n) is a sequence in E which converges to some $r \in \mathbb{Q}$. Since $r_n \in E$, we have $-\sqrt{2} < r_n < \sqrt{2}$, so the limit r satisfies $-\sqrt{2} \leq r \leq \sqrt{2}$. But since $r \in \mathbb{Q}$, $r \neq \pm\sqrt{2}$ so that we actually have $-\sqrt{2} < r\sqrt{2}$, showing that $r \in E$. (Again, sequences in E which converge to either $-\sqrt{2}$ or $\sqrt{2}$ in $\mathbb{R}$ are irrelevant when asking if E is closed in $\mathbb{Q}$, since in this case we only care about sequences which are assumed to converge in $\mathbb{Q}$ to begin with.) Thus E is closed in $\mathbb{Q}$ as claimed.

So we have a nontrivial example of a subset of $\mathbb{Q}$ which is both open and closed in $\mathbb{Q}$. Subsets of a metric space which are both open and closed are called *clopen* subsets (no doubt a very imaginative name), and whether or not nontrivial such sets exist leads to an important concept: any metric space X has the empty set and all of X itself as clopen subsets, and if these are the only ones we say that X is *connected*. We'll come back to looking at connected spaces later on, but this example shows that $\mathbb{Q}$ is not connected.

Important. Given a metric X, $U \subseteq X$ is open in X if we can surround any point of U by a ball which remains within U, and $A \subseteq X$ is closed in X if the limit of any sequence in A which converges in X is itself in A. Both of these definitions depend on the larger space X, meaning that a given set can be open or closed (or both) in one space but not in another.

Proposition. Now we come to the generalization of the fact that, in $\mathbb{R}$, open intervals are open and closed intervals are closed. Let (X, d) be a metric space. Then for any $p \in X$ and any $r > 0$, the open ball $B_r(p)$ of radius r centered at p is open in X and the *closed* ball $M_r(p)$ of radius r centered at p:

$$M_r(p) = \{q \in X \mid d(q,p) \leq r\}$$

is closed in X. (As opposed to the notation $B_r(p)$ for open balls, there is no standard notation for closed balls. Some people use $\overline{B_r(p)}$, which is bad since it may cause confusion with the notion of the *closure* of a set which we'll soon look at, and I've also seen $B_r[p]$, which might be okay except that the difference in notation between $[p]$ and (p) is can be hard to detect. So, I'm using M's to denote closed balls, but don't except to see this notation elsewhere.)

As usual, a picture drawn in $\mathbb{R}^2$ gives the necessary intuition. Draw an open ball centered at p and take some other q in this ball. To show that this open ball is open, we have to come up with a ball centered at q which remains within the ball centered at p. It is easy to draw this in picture, where say the radius obtained by taking r minus $d(p,q)$ will give a ball around q which indeed is contained inside the original ball. This suggests that in general the ball of radius $r - d(p,q)$ centered at q will give what we need, and justifying this precisely relies on the triangle inequality.

Proof. Let $p \in X$ and $r > 0$, and let $q \in B_r(p)$. We claim that the ball of radius $s = r - d(p,q)$ satisfies $B_s(q) \subseteq B_r(p)$, which will show that $B_r(p)$ is open in X. First, note that since $q \in B_r(p)$, $d(p,q) < r$ so that $r - d(p,q) > 0$ and hence s is indeed positive.

To show that $B_s(q) \subseteq B_r(p)$, let $x \in B_s(q)$. Then $d(x,q) < s$. Thus by the triangle inequality:

$$d(x,p) \leq d(x,q) + d(q,p) < s + d(q,p) = (r - d(q,p)) - d(q,p)) = r,$$

so $x \in B_r(p)$. Hence any element of $B_s(q)$ is also in $B_r(p)$, so $B_s(q) \subseteq B_r(p)$ as claimed. We conclude that $B_r(p)$ is open in X.

To show that the closed ball $M_r(p)$ is closed in X, suppose that (x_n) is a sequence of points in $M_r(p)$ converging to some $x \in X$. We must show that $x \in M_r(p)$, i.e. that $d(x,p) \leq r$. For any n we have

$$d(x,p) \leq d(x,x_n) + d(x_n,p) \leq d(x,x_n) + r$$

since $d(x_n,p) \leq r$ because the x_n come from $M_r(p)$. Since $x_n \to x$, $d(x,x_n) \to 0$ so taking the limit as $n \to \infty$ in this inequality gives

$$d(x,p) \leq 0 + r = r,$$

showing that $d(x,p) \leq r$ as desired. Hence $x \in M_r(p)$ so $M_r(p)$ is closed in X. □

Important. Open balls are always open and closed balls are always closed, which is good since otherwise our use of the terms "open" and "closed" to describe these would prove to be quite confusing.

Lecture 16: More on Open and Closed Sets

Today we continued talking about open and closed subsets of metric spaces. In particular, we looked at ways of constructing open and closed sets from other open and closed sets, where the main point was a characterization of "open" in terms of "closed" and vice-versa.

Warm-Up. We show that a closed subset of a complete space X is itself a complete subspace. Indeed, suppose that $A \subseteq$ is closed and that (a_n) is a Cauchy sequence in A. Since we are viewing A as a subspace of X here, (a_n) is then also Cauchy in X since the metric we are using on A is the same as that on X, and the notion of "Cauchy" only depends on this metric. Since X is complete, (a_n) converges to some $x \in X$, and since A is closed in X we then have $x \in A$, so that (a_n) converges in A. Hence a Cauchy sequence in A always converges in A, so A is complete.

Note that we've already seen instances of this previously. In particular, since closed intervals are closed in $\mathbb{R}$, we get that closed intervals are always complete, and since $C([a,b])$ is closed in $C_b([a,b])$, we get that $C([a,b])$ is complete. (The fact that $C([a,b])$ is closed in $C_b([a,b])$ is the statement that the uniform limit of continuous functions is continuous. Similarly, since the uniform limit of integrable functions is integrable, the space of integrable functions on $[a,b]$ is closed in $C_b([a,b])$. In all of these statements we are using the sup metric.)

Proposition. It turns out that the notions of "open" and "closed" are in a sense "opposite" to one another in that a set has one of these properties if and only if its complement has the other. To be precise, we prove: a subset U of a metric space X is open in X if and only if its complement U^c is closed in X, and $A \subseteq X$ is closed in X if and only if A^c is open. The key point is that if you negate the definition of open you (essentially) get the definition of closed, and vice-versa.

Proof. By taking contrapositives, to show that U is open if and only if U^c is closed it is equivalent to show that U is not open if and only if U^c is not closed. Thus, suppose that U is not open in X. Then there exists $p \in U$ such that no ball around it is fully contained in U, meaning that any ball around it contains an element of U^c. In particular, for any $n \in \mathbb{N}$ there exists $x_n \in U^c$ such that $x_n \in B_{\frac{1}{n}}(p)$, so $d(x_n,p) < \frac{1}{n}$. This gives a sequence (x_n) of points in U^c which converge to $p \in U$ (since $d(x_n,p) < \frac{1}{n}$ implies $d(x_n,p) \to 0$), showing that U^c is not closed.

Conversely, suppose that U^c is not closed. Then there exists a sequence (x_n) in U^c which converges to some p not in U^c, so to some $p \in U$. Since $x_n \to p$, any ball around p will contain all x_n past some index, so any such ball will thus contain an element of U^c and will not be fully contained in U. This shows that U is not open.

The claim that A is closed if and only if A^c is open now follows by applying the above result to $U = A^c$: $U = A^c$ is open if and only if $U^c = (A^c)^c = A$ is closed. □

Example. Since open intervals in $\mathbb{R}$ are always open, the union

$$\bigcup_{n\in\mathbb{Z}} (n, n+1)$$

is open in $\mathbb{R}$ as a consequence of the following theorem. But this is the complement of $\mathbb{Z}$ in $\mathbb{R}$, so $\mathbb{Z}$ is closed in $\mathbb{R}$. (We can also see this without using open sets by showing that the only sequences in $\mathbb{Z}$ which converge in $\mathbb{R}$ are those which are eventually constant, and so when a sequence in $\mathbb{Z}$ does converge it must converge to an integer.)

Theorem. Now we look at how open sets behave when taking unions and intersections. The fact is that the union of arbitrarily many open sets is always open, and that the intersection of *finitely* many open sets is always open. (Careful: the intersection of infinitely many open sets might not be open, as we'll see in an example to follow.)

Proof. Suppose that $\{U_\alpha\}$ is a collection of arbitrarily many open subsets of a metric space X, indexed by α in some indexing set I, and let $p \in \bigcup_\alpha U_\alpha$. Then there exists β such that $p \in U_\beta$. Since U_β is open in X, there exists $B_r(p) \subseteq U_\beta$, and since $U_\beta \subseteq \bigcup_\alpha U_\alpha$ we then have

$$B_r(p) \subseteq \bigcup_\alpha U_\alpha,$$

showing that $\bigcup_\alpha U_\alpha$ is open in X as claimed.

Now suppose that $U_1, \ldots, U_n$ are finitely many open subsets of X and that $p \in U_1 \cup \cdots \cup U_n$. Then $p \in U_i$ for each $i = 1, \ldots, n$. Since U_i is open, for each $i = 1, \ldots, n$ there exists $r_i > 0$ such that $B_{r_i}(p) \subseteq U_i$. Set $r = \min\{r_1, \ldots, r_n\}$, which is positive since it is the minimum of positive numbers. Then for each i, $r \leq r_i$ so

$$B_r(p) \subseteq B_{r_i}(p) \subseteq U_i.$$

Thus $B_r(p) \subseteq U_1 \cup \cdots \cup U_n$, showing that $U_1 \cup \cdots \cup U_n$ is open in X. □

Remark. Note well what goes wrong in the above proof if we try to intersect infinitely many open sets. In this case we end up with infinitely many positive radii r_i which we try to take the minimum of. First, such a minimum may not exist, but this is easy to get around by taking the infimum of the r_i instead. However, the problem is that the infimum of infinitely many positive numbers could be zero, in which case the above proof doesn't work since the final radius r must be positive.

Example. For each $n \in \mathbb{N}$, the interval $(-\frac{1}{n}, \frac{1}{n})$ is open in $\mathbb{R}$. However, the intersection of all such intervals:

$$\bigcap_{n\in\mathbb{N}} \left(-\frac{1}{n}, \frac{1}{n}\right) = \{0\}$$

only consists of zero since zero is the only number satisfying $-\frac{1}{n} < x < \frac{1}{n}$ for all n, as a consequence of the Archimedean property of $\mathbb{R}$. Thus this gives an example showing that the intersection of infinitely many open sets might not be open. (As the remark above suggests, the issue is that the radii $\frac{1}{n}$ used here have infimum equal to 0.)

Theorem. Similarly, we can look at how closed sets behave when taking unions and intersections. A similar phenomena as for open sets occurs, only in this case the "arbitrarily many" and "finitely many" get switch around: the intersection of arbitrarily many closed subsets of a metric space X is always closed and the union of finitely many closed subsets of X is always closed. Again be careful: the union of infinitely many closed sets might not be closed.

Proof 1. Since a set is closed if and only if its complement is open, we this result immediately from the analogous result for open sets by taking complements. Recall *DeMorgan's Laws* in set theory which say that

$$\left(\bigcap_\alpha A_\alpha\right)^c = \bigcup_\alpha A_\alpha^c \text{ and } \left(\bigcup_\alpha A_\alpha\right)^c = \bigcap_\alpha A_\alpha^c.$$

Given an arbitrary collection $\{A_\alpha\}$ of closed sets, we have that each A_α^c is open, so by the previous Theorem $\bigcup A_\alpha^c$ is open, but since this union is the complement of $\bigcap A_\alpha$, we get that $\bigcap A_\alpha$ is closed. Similarly, if $A_1, \ldots, A_n$ are finitely many closed sets, then each A_i^c is open so $A_1^c \cap \cdots \cap A_n^c$ is open, and since this equals the complement of $A_1 \cup \cdots \cup A_n$, we get that $A_1 \cup \cdots \cup A_n$ is closed. □

Proof 2. For good measure, here is a proof using only the definition of closed which doesn't rely on analogous properties of open sets. Suppose that $\{A_\alpha\}$ is a collection of arbitrarily many closed subsets of X and let (x_n) be a sequence in $\bigcap A_\alpha$ which converges to some $x \in X$. Then for each n, $x_n \in A_\alpha$ for all α, so (x_n) can be viewed as a sequence in A_α for each α. Since for each α, A_α is closed, we get that the limit of $x_n \in A_\alpha$ is in A_α, so $x \in A_\alpha$ for each α. Thus $x \in \bigcap A_\alpha$, showing that $\bigcap A_\alpha$ is closed.

Now suppose that $A_1, \ldots, A_n$ are finitely many closed subsets of X and that (x_n) is a sequence in $A_1 \cup \cdots \cup A_n$ which converges to some $x \in X$. Then for each n, $x_n \in A_1 \cup \cdots \cup A_n$. Since there are infinitely many terms in (x_n) but only finitely many sets A_i, at least one of these sets must contain an infinite number of the x_n's; call this set A_i. (Otherwise, if each of $A_1, \ldots, A_n$ contained only finitely many of the x_n's, then there would only be finitely many terms in (x_n), contradicting our definition of a sequence as an infinite list.) The terms from (x_n) which belong to A_i then form a subsequence (x_{n_k}) of (x_n), and since $x_n \to x$ we have that this subsequence also converges to x. But since A_i is closed, the limit of this subsequence must be in A_i, so $x \in A_i$. Thus $x \in A_1 \cup \cdots \cup A_n$, so $A_1 \cup \cdots \cup A_n$ is closed as claimed. □

Remark. Note well what goes wrong in the above proof if we try to take the union of infinitely many closed sets. In this case it is certainly possible that each of the A_i's we are taking the union of contains only finitely many terms from our sequence—for instance each A_i might only contain one term from our sequence—in which case we do not get an entire subsequence (x_{n_k}) inside one of the A_i's as we needed in the proof above. □

Examples. For each $n \in \mathbb{N}$, the interval $[-1 + \frac{1}{n}, 1 - \frac{1}{n}]$ is closed in $\mathbb{R}$, but their union:

$$\bigcup_{n\in\mathbb{N}} \left[-1 + \frac{1}{n}, 1 - \frac{1}{n}\right] = (-1, 1)$$

is not. (That this is the correct union follows from the fact that $-1 + \frac{1}{n} \to -1$ from the right and $1 - \frac{1}{n} \to 1$ from the left.) As in the remark above, the issue is that for the sequence $x_n = \frac{1}{n}$ in this union, each of the intervals we are taking the union of only contains finitely many of the terms in the sequence (x_n).

Important. A set is open if and only if its complement is closed, and a set is closed if and only if its complement is open. The union of any number of open sets is open, and the intersection of any number of closed sets is closed. The intersection of finitely many open sets is open, and the union finitely many closed sets is closed, but the intersection of infinitely many open sets is not necessarily open and the union of infinitely many closed sets is not necessarily closed.

Uses of open and closed sets. We didn't talk about this in class, but let us say a bit about why the notions of open and closed sets are useful ones. Next quarter we will want to talk about what it means for a function between higher-dimensional spaces to be differentiable, which will depend on the notion of a higher-dimensional limit. The point is that in order for this to be a good and worthwhile idea, it will be best if we can approach a given point $p \in \mathbb{R}^m$ from *all* possible directions. (You probably remember from a multivariable calculus course that for a function $f : \mathbb{R}^2 \to \mathbb{R}$ the idea of approaching a point from different directions was essential when determining whether or not a multivariable limit exists.) But to be able to approach $p \in \mathbb{R}^m$ from all possible directions, we have to know that "all possible directions" are included within the domain of our function, which amounts to saying that we can surround p by a ball which is fully contained in the domain, which is precisely saying that the domain of our function is *open* in $\mathbb{R}^m$! Thus, functions defined on open domains inside $\mathbb{R}^m$ are the types of functions for which the notion of "limits" and "differentiability" will make the most sense.

In addition, often we'll be carrying out some kind of process, and we'll want to know that we don't "leave" our domain after taking limits. For instance, maybe we'll be looking at a collection of points satisfying some equation like

$$4x^2 + 3y^2 - 2z^2 = 1,$$

which you might recognize from multivariable calculus as describing a so-called "hyperboloid of one-sheet", and we will need to know that taking limits on this hyperboloid keeps us on the hyperboloid. But this is precisely saying that we need to know the hyperboloid is *closed* in $\mathbb{R}^3$, which it is. Thus, it is only on closed subsets of $\mathbb{R}^m$ that we can take limits of various equations or non-strict inequalities and know that the resulting points will still satisfy those same equations or non-strict inequalities.

Lecture 17: Interior, Closure, and Boundary

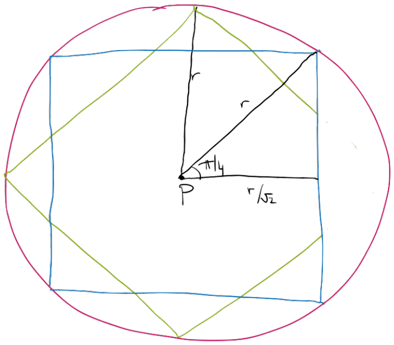
Today we continued looking at more "topological" concepts, and in particular saw ways of rephrasing the definitions of open and closed.

Warm-Up 1. We show that a subset U of $\mathbb{R}^2$ is open with respect to the Euclidean metric if and only if it is open with respect to the taxicab metric and if and only if it is open with respect to the box metric. The key point is that within any ball with respect to any of these metrics we can always find a ball with respect to one of the other two.

To be clear, recall that balls with respect to the taxicab metric look like diamonds and balls with respect to the box metric look like boxes, and suppose that $B_r^E(p)$ is a ball with respect to the Euclidean metric. (We'll use the notation B^E, B^{taxi}, and B^{box} to indicate what metric we are taking a ball with respect to.) Then we claim that

$$B_r^{taxi}(p) \subseteq B_r^E(p) \text{ and } B_{\frac{r}{\sqrt{2}}}^{box}(p) \subseteq B_r^E(p).$$

This is clear from the following picture:



(The circle, diamond, and box should all be drawn with dotted curves, but that it tedious to do on my iPad, so I didn't.) If U is open with respect to the Euclidean metric, then for $p \in U$ there exists $B_r^E(p) \subseteq U$, so we have

$$B_r^{taxi}(p) \subseteq B_r^E(p) \subseteq U \text{ and } B^{box}_{\frac{r}{\sqrt{2}}}(p) \subseteq B_r^E(p) \subseteq U,$$

which shows that U is open with respect to the taxicab metric and the box metric as well.

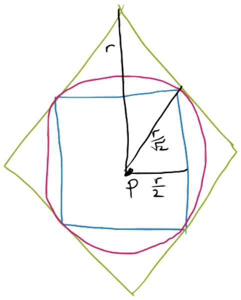
Similarly, if U is open with respect to the taxicab metric, then the facts that

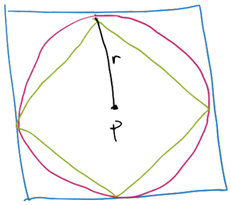
$$B^E_{\frac{r}{\sqrt{2}}}(p) \subseteq B_r^{taxi}(p) \text{ and } B^{box}_{\frac{r}{2}}(p) \subseteq B_r^{taxi}(p)$$

imply that U is also open with respect to the Euclidean and box metrics, and if U is open with respect to the box metric, then

$$B_r^E(p) \subseteq B_r^{box}(p) \text{ and } B_r^{taxi}(p) \subseteq B_r^{box}(p).$$

imply that U is open with respect to the Euclidean and taxicab metrics. All of these containments follow from the pictures:





Note that we could have also derived this fact from the result that a set is open if and only if its complement is closed. On the previous homework you showed that $(x_n, y_n) \to (x, y)$ with respect to any of these three metrics if and only if $(x_n, y_n) \to (x, y)$ with respect to the other two, which implies that a set is closed with respect to any of these metrics if and only if it is closed with respect to the other two. Then taking complements gives the result of this Warm-Up.

Warm-Up 2. Now we give another proof that closed balls are closed in any metric space, by showing that their complements are open. (We previously showed this using the sequence definition of closed.) Recall that the closed ball of radius $r > 0$ around p in a metric space X is

$$M_r(p) := \{q \in X \mid d(q,p) \leq r\}.$$

Thus we must show that

$$M_r(p)^c := \{q \in X \mid d(q,p) > r\}$$

is open in X. For $q \in M_r(p)^c$, a picture drawn in $\mathbb{R}^2$ suggests that the radius $s := d(q,p) - r$ looks as if it should satisfy $B_s(q) \subseteq M_r(p)^c$, which we now verify.

First note that since $q \in M_r(p)^c$, $d(q,p) > r$ so that $s = d(q,p) - r$ is indeed positive. Now, let $x \in B_s(q)$. From the triangle inequality we have

$$d(q,p) \leq d(q,x) + d(x,p),$$

which after rearranging gives the "reverse triangle inequality"

$$d(x,p) \geq d(q,p) - d(q,x).$$

Since $x \in B_s(q)$, $d(q,x) < s$ so $d(q,p) - d(q,x) > d(q,p) - s$. Thus

$$d(x,p) \geq d(q,p) - d(q,x) > d(q,p) - s = d(q,p) - [d(q,p) - r] = r,$$

which shows that $x \in M_r(p)^c$. Hence $B_s(q) \subseteq M_r(p)^c$, so $M_r(p)^c$ is open in X and hence $M_r(p)$ is closed as claimed.

Definitions. Let X be a metric space and E a subset. We say that $p \in X$ is

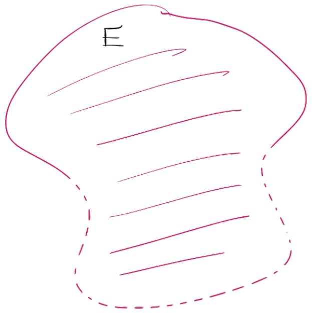
- an *interior point* of E if there exists $B_r(p) \subseteq E$,
- a *limit point* of E if there exists a sequence (x_n) in E converging to p,
- a *boundary point* of E if any $B_r(p)$ contains an element of E and an element of E^c.

The *interior* of E is the set E^o of all interior points of E, the *closure* of E is the set $\overline{E}$ of all limits point of E, and the *boundary* of E is the set ∂E of all boundary points of E.

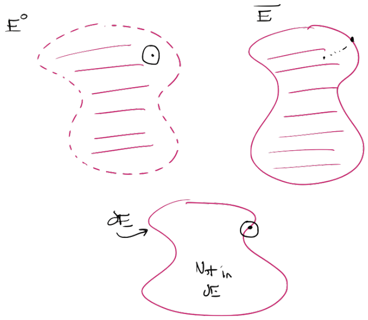
Some observations. First, since $p \in B_r(p)$, an interior point of E must actually belong to E so that we always have $E^o \subseteq E$. The reverse containment $E \subseteq E^o$ is precisely what it means to say that E is open, so we can rephrase the definition of open as saying that E equals its own interior. Second, since for $p \in E$, the constant sequence (p) is a sequence in E converging to p, p is a limit point of E so we always have $E \subseteq \overline{E}$. The reverse containment $\overline{E} \subseteq E$ is precisely what it means to say that E is closed, so we can rephrase the definition of closed as saying that E equals its own closure.

We'll see what these notions look like visually in the case of $\mathbb{R}^2$ below, but the intuitions are as follows: an interior point of E is one for which points "close enough" to it are also in E (so an interior point is fully "surrounded" by points of E), a limit point of E is one which is arbitrarily close to elements of E, and a boundary point of E is one which is arbitrarily close to elements of E and of its complement E^c.

Examples. Consider the following subset E of $\mathbb{R}^2$, where as usual a solid curve means that those points are included in E and a dotted curve means that those points are not included in E.



The interior of E is (as the name suggests) E without the solid curve, since these are the only points of E around which a ball can be drawn which is fully contained in E; the closure of E is E together with the points on the dotted curve since these are all the points in $\mathbb{R}^2$ for which a sequence in E can be found which converges to it (the name closure suggests we are "closing up" E); and the boundary of E consists of the points along the solid curve *and* the dotted curve since these are the points around which any ball will contain elements of E and elements not in E, so in particular the boundary of E is geometrically what you would normally think of as the boundary.



Note that in this case, E^o is itself open in $\mathbb{R}^2$ and $\overline{E}$ is itself closed in $\mathbb{R}^2$, which are both general facts we'll come back to later. Also note that E^o contains no part of ∂E and $\overline{E}$ contains all of ∂E, which will soon give us a characterization of open and closed sets in terms of their boundaries.

Example. We determine the interior, closure, and boundary of $\mathbb{Q}$ in $\mathbb{R}$ with the standard metric. First, the interior of $\mathbb{Q}$ is empty since no interval around a rational can ever be fully contained in $\mathbb{Q}$ because any such interval will always contain an irrational. Second, any $x \in \mathbb{R}$ is a limit point of $\mathbb{Q}$ since a sequence of rationals converging to x can always be found, so the closure of $\mathbb{Q}$ in $\mathbb{R}$ is $\mathbb{R}$. Finally, for any $x \in \mathbb{R}$, any interval around x will contain both rationals and irrationals, so any $x \in \mathbb{R}$ is a boundary point of $\mathbb{Q}$ and thus the boundary of $\mathbb{Q}$ in $\mathbb{R}$ is $\mathbb{R}$.

To summarize these reasons more succinctly: $\mathbb{Q}^o = \emptyset$ since the irrationals are dense in $\mathbb{R}$, $\overline{\mathbb{Q}} = \mathbb{R}$ since the rationals are dense in $\mathbb{R}$, and $\partial\mathbb{Q} = \mathbb{R}$ since both the rationals and irrationals are dense in $\mathbb{R}$. (Here we are using "dense" in the sense that for any $x < y$ there exists an element from the set in question between them. We'll see next time a more general notion of dense.)

Proposition. As mentioned previously, we immediately have that E is open in X if and only if $E = E^o$ and E is closed in X if and only if $E = \overline{E}$. The picture of what boundary points look like in $\mathbb{R}^2$ suggests the following, which we'll now prove: E is open in X if and only if $E \cap \partial E = \emptyset$ and E is closed in X if and only if $\partial E \subseteq E$.

Proof. First we show that E is open in X if and only if $E \cap \partial E = \emptyset$. To prove the forward direction by contrapositive, suppose that $E \cap \partial E \neq \emptyset$. Then there exists $p \in E \cap \partial E$. Since $p \in \partial E$, any ball around p contains an element not in E, so no such ball is fully contained in E. Hence since there exists $p \in E$ such that no ball around it is contained in E, E is not open in X.

Conversely suppose that $E \cap \partial E = \emptyset$ and let $p \in E$. Then $p \notin \partial E$ since $E \cap \partial E$ contains no elements. If $p \notin \partial E$, then there exists a ball $B_r(p)$ with either does not contain any element of E or does not contain any element of E^c. But such a ball always contains at least $p \in E$, so we must have that there is a ball $B_r(p)$ which contains no element of E^c, which means that $B_r(p)$ contains only elements of E and hence $B_r(p) \subseteq E$. Thus E is open.

Now we prove that E is closed in X if and only if $\partial E \subseteq E$. Suppose that E is closed in X and let $q \in \partial E$. Then any ball around q contains an element of E, so in particular for each $n \in \mathbb{N}$ there exists $x_n \in E$ such that $x_n \in B_{\frac{1}{n}}(q)$. This gives a sequence (x_n) in E such that $d(x_n, q) < \frac{1}{n}$, which implies that $x_n \to q$. Since E is closed, we thus have $q \in E$ so that $\partial E \subseteq E$ as claimed.

We prove the converse by contrapositive; to be clear, we show that if E is not closed in X, then $\partial E \nsubseteq E$. If E is not closed there exists $q \in E^c$ such that there is a sequence (x_n) in E converging to it. Then, since $x_n \to q$, any $B_r(q)$ will contain x_n past some index, so that $B_r(q)$ always contains an element of E (namely some term from the sequence (x_n)) and an element of E^c (namely q), so q is a boundary point of E. Thus q is an element of ∂E with is not in E, so $\partial E \nsubseteq E$ as required. □

Important. A subset E of a metric space X is open in X if and only if it equals its own interior, and if and only if it contains no piece of its boundary. The subset E is closed in X if and only if it equals its own closure, and if and only if it contains all of its boundary.

Proposition. Finally, we state two more basic set equalities, which help to further understand intuitively what a boundary is. First, as the pictures previously drawn in $\mathbb{R}^2$ suggest, we claim that for any $E \subseteq X$, $\partial E = \overline{E} \backslash E^o$. Thus, the boundary is what you get when you remove the interior from the closure. This is in the book as Theorem 10.39, so you can read the proof there; we'll look at a Warm-Up next time which explains the property numbered in the book's proof as (2).

The second fact, which we will prove, is the equality $\partial E = \overline{E} \cap \overline{E^c}$, so that the boundary consists precisely of the points common to the closure of E and the closure of its complement. Again, this makes sense intuitively for the pictures drawn before.

Proof. Let $x \in \partial E$. Then any ball around x contains an element of E and an element of E^c. In particular, for any $n \in \mathbb{N}$ there exist $p_n \in E$ and $q_n \in E^c$ such that

$$d(p_n, x) < \frac{1}{n} \text{ and } d(q_n, x) < \frac{1}{n}.$$

This gives a sequence (p_n) in E which converges to x and a sequence (q_n) in E^c converging to x, so $x \in \overline{E}$ and $x \in \overline{E^c}$. Thus $\partial E \subseteq \overline{E} \cap \overline{E^c}$.

Conversely, let $x \in \overline{E} \cap \overline{E^c}$. Then $x \in \overline{E}$ and $x \in \overline{E^c}$, so there exist sequences (p_n) in E and (q_n) in E^c converging to x. Since $p_n \to x$, any ball around x contains a term $p_N \in E$ and since $q_n \to x$, any ball around x contains a term $q_M \in E^c$. Thus any ball around x contains an element of E and of E^c, so $x \in \partial E$. Thus $\overline{E} \cap \overline{E^c} \subseteq \partial E$, so $\partial E = \overline{E} \cap \overline{E^c}$ as claimed. □

Remark. All of these proofs, although somewhat tedious, are good practice in that they force you to really understand the various definitions involved and how they relate to one another, and are good examples of using geometric intuition in $\mathbb{R}^2$ to understand metric space concepts in general.

Lecture 18: Denseness

Today we continued talking about interiors and closures, and defined what it means for a space to be dense in another space.

Warm-Up. For a subset E of a metric space X, we show that $p \in X$ is a limit point of E if and only if every open ball around p contains an element of E. First suppose that p is a limit point of E, so that there exists a sequence (x_n) in E converging to p. Then any open ball around p will contain all terms from this sequence beyond some index, so any open ball around p contains an element of E.

Conversely suppose that every open ball around p contains an element of E. Then for each $n \in \mathbb{N}$ there exists $x_n \in E$ such that $x_n \in B_{1/n}(p)$. This gives a sequence (x_n) in E such that $d(x_n, p) < \frac{1}{n}$ for all n, which implies that $x_n \to p$. Hence p is a limit point of E as required.

Remark. Note that the way we're using "limit point" in this course differs from the way in which others often use it, where they say that $p \in X$ is a limit point of $E \subseteq X$ if every open ball around p contains an element of E *distinct* from p. In other words, the requirement is that when $p \in E$, p itself is not allowed to be the element of E contained in $B_r(p)$. According to this definition, $\mathbb{Z}$ has no limit points in $\mathbb{R}$ since a ball of radius $\frac{1}{2}$ around $n \in \mathbb{Z}$ contains no integer apart from n, while according to our definition the limit points of E are the integers themselves.

This is the approach taken in one of the most famous analysis books of all time: *Principles of Mathematical Analysis* by Walter Rudin. This is something to keep in mind if you try to compare our arguments to those you might find elsewhere. Note, however, that even though various authors might disagree on the precise meaning of the term "limit point", *everyone* agrees on what "closure" means.

Proposition. Let E be a subset of a metric space X. Here are some additional things of which to be aware. First, the interior of E is always open in X and the closure of E is always closed in X. In addition, E^o is the "largest" open subset of X which is contained in E, and $\overline{E}$ is the "smallest" closed subset of X containing E. To be precise, "largest" in the case of E^o means that any open subset U of X contained in E is itself contained in E^o, and "smallest" in the case of $\overline{E}$ means that any closed subset A of X which contains E itself contains $\overline{E}$.

Proof. We'll only prove that E^o is always open in X; the fact that $\overline{E}$ is always closed was a homework problem. Suppose that $p \in E^o$. Then p is an interior point of E so there exists $B_r(p) \subseteq E$. We claim that any element of $B_r(p)$ is an interior point of E, so that $B_r(p)$ is actually contained in E^o.

Let $q \in B_r(p)$. Since $B_r(p)$ is open in X, there exists $B_s(q) \subseteq B_r(p)$. But $B_r(p) \subseteq E$, so also $B_s(q) \subseteq E$, showing that q is indeed an interior point of E. Thus $B_r(p) \subseteq E^o$, so E^o is open in X as claimed. □

Denseness. We say that a subset A of a metric space X is *dense* in X if $\overline{A} = X$, or in other words if every element of X is a limit point of A. Concretely, this means that for any $p \in X$ there exists a sequence (x_n) in a converging to p. Thus, to say that A is dense in X means that we can approximate points of X to whatever accuracy we want using elements of A, so that elements of A are in a sense "everywhere".

Examples. Last quarter we used the word "dense" when saying that $\mathbb{Q}$ was dense in $\mathbb{R}$, which we took to mean that between any two distinct real numbers there exists a rational number. We also saw this implies that given any real number there exists a sequence of rationals converging to it, which says that $\mathbb{Q}$ is dense in $\mathbb{R}$ in our new sense as well.

Indeed, this new notion of dense is a direct generalization of the previous notion in $\mathbb{R}$ in the following way. Saying that $\mathbb{Q}$ is dense in $\mathbb{R}$ in the previous sense can be rephrased as saying that any nonempty open interval (a, b) contains a rational number. The analogous claim about a subset A of a metric space X is that any open ball in X contains an element of A, and you will show on the homework that this is indeed equivalent to A being dense in X. Note that this gives another sense in which a dense set is "everywhere", in that no matter how small an open ball and no matter which center you take, you will always find an element of A.

Similarly, the fact that between any two distinct real numbers there is an irrational number gives the fact that the set of irrationals is also dense in $\mathbb{R}$ in our new sense of word "dense".

Important. A subset A of X is dense in X if for any $p \in X$ there exists a sequence in A converging to p. Equivalently, any open ball in X contains an element of A, which intuitively means that no matter how close you "zoom" in on any portion of X, you will find elements of A dispersed throughout.

Remark. You might recall that last quarter we had some examples (or homework problems) showing that the behavior of a continuous function on $\mathbb{R}$ was completely determined by its behavior on $\mathbb{Q}$ (or $\mathbb{R}\backslash\mathbb{Q}$), and we'll see that the same is true for continuous functions between metric spaces in general. This is why denseness will be a useful concept for us.

Weierstrass Approximation Theorem. The examples of $\mathbb{Q}$ and $\mathbb{R}\backslash\mathbb{Q}$ in $\mathbb{R}$ (or analogues of these in $\mathbb{R}^n$) will likely be the most important examples of dense sets. But, here is another important and perhaps surprising example: the space of polynomial functions is dense in the space $C[a, b]$ of real-valued continuous functions on $[a, b]$ with respect to the sup metric.

Concretely, this says that for any continuous function f, there exists a sequence of polynomials p_n which converges to f uniformly. Even more concretely, the claim is that for any $\epsilon > 0$ there exists a polynomial p such that

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

which says that we can approximate any continuous function to any degree of accuracy using polynomials. We previously saw similar results involving Taylor polynomials, but these only made sense for functions which were infinitely-differentiable, whereas now we're saying we can do something similar for *any* continuous function whatsoever. Being able to approximate arbitrary continuous functions using polynomials is good since polynomials in general are fairly straightforward to work with.

This result is also sometimes called the "Stone-Weierstrass Theorem", although the name "Stone-Weierstrass Theorem" usually refers to a much more general fact characterizing dense subsets of $C[a, b]$ of which the Weierstrass Approximation Theorem is a special case. The proofs of any of these theorems are quite difficult and require some amazing ingenuity. In the case of the space polynomials being dense, the idea is to construct polynomials p_n which agree with a given continuous function f at more and more points in a way which makes $|p_n(x) - f(x)|$ a quantity we can control. Again, this is far from easy, so we'll skip the details and take the result for granted. If interested, the book contains proofs of all of these.

Another example. It is also true that the subset of so-called *trigonometric polynomials* is dense in $C[a,b]$, where a trigonometric polynomial is a function of the form

$$\sum_{k=0}^{n}(a_n \cos kx + b_n \sin kx),$$

or maybe with some additional constants thrown in somewhere. This is also hard to prove in general, unless you use the general Stone-Weierstrass Theorem given in the book. We'll see next quarter that this is somewhat related to the notion of a Fourier series.

Lecture 19: Compact Sets

Today we started talking about the notion of compactness, which you could say is the ultimate reason why we're talking about metric spaces at all. Indeed, the special properties which continuous functions on compact sets have are *the* reason why calculus works, which we saw a glimpse of last quarter when looking at continuous functions on $[a,b]$.

Warm-Up. We show that $\mathbb{Q}^2$ dense in $\mathbb{R}^2$, where $\mathbb{Q}^2$ denotes the set of points in $\mathbb{R}^2$ where both coordinates are rational. Suppose that $(x,y) \in \mathbb{R}^2$. Since $\mathbb{Q}$ is dense in $\mathbb{R}$, there exist sequences (r_n) and (s_n) in $\mathbb{Q}$ such that $r_n \to x$ and $s_n \to y$. Then $(r_n, s_n) \to (x,y)$, so (x,y) is a limit point of $\mathbb{Q}^2$ and hence $\mathbb{Q}^2$ is dense in $\mathbb{R}^2$ with respect to the Euclidean, taxicab, and box metrics.

We can also show this using the characterization of dense sets in terms of open balls. Suppose that $B_r(x,y)$ is an open ball in $\mathbb{R}^2$, say with respect to the box metric for simplicity. (With respect to the other metrics you can apply the following argument to a ball with respect to the box metric contained in $B_r(x,y)$, which we have shown before always exists.) Since $\mathbb{Q}$ is dense in $\mathbb{R}$, the interval $(x-r, x+r)$ contains a rational p and the interval $(y-r, y+r)$ contains a rational s. Then (p,s) is an element of $\mathbb{Q}^2$ contained in $(x-r,x+r) \times (y-r,y+r)$, which is equal to $B_r(x,y)$ with respect to the box metric. Hence any open ball in $\mathbb{R}^2$ contains an element of $\mathbb{Q}^2$, so $\mathbb{Q}^2$ is dense in $\mathbb{R}^2$ as claimed.

Sequential definition of compactness. We say that a subset K of a metric space X is *compact* if any sequence in K has a convergent subsequence in K; to be clear, this requires that the limit of the convergent subsequence referred to be in K itself. (We call this the "sequential" definition of compactness, to distinguish it from an equivalent characterization we'll see next time.)

We'll talk about the intuition behind compact sets more later, but for now the idea is as follows: given any random collection of points in K, we can always find some points among these which approach *something* in K, so that the original points maybe weren't so random after all. (This is incredibly vague, and we'll give a better sense as to what idea compactness is mean to capture next time.)

Non-examples. $\mathbb{R}$ is not compact, since the sequence $x_n = n$ is a sequence in $\mathbb{R}$ with no convergent subsequence due to the fact that any subsequence is unbounded.

Also, no nonempty open interval (a,b) is compact since $x_n = a + \frac{b-a}{n+1}$ is a sequence in (a,b) with no convergent subsequence. The issue is that x_n itself converges to a, and hence will any subsequence, so no subsequence will converge *in* (a,b) since $a \notin (a,b)$.

Closed intervals are compact. Closed intervals $[a,b]$ are always compact, which is essentially a rephrasing of the Bolzano-Weierstrass Theorem. Indeed, if (x_n) is a sequence in $[a,b]$, then (x_n) is bounded so it has a convergent subsequence by the Bolzano-Weierstrass Theorem, and the limit of this convergent subsequence is in $[a,b]$ since $[a,b]$ is closed in $\mathbb{R}$. This is why we already saw a glimpse of the importance of compact sets last quarter, where any result we saw which depended on using closed intervals were really results about compactness.

Compact implies closed and bounded. Generalizing properties of closed intervals in $\mathbb{R}$, it is true in general that a compact subset K of a metric space X is always closed and bounded in X. Note carefully however, as we'll see in a bit, that closed and bounded does *not* necessarily imply compactness in general, although it does in $\mathbb{R}^n$.

Proof. Suppose that $K \subseteq X$ is compact. We first show that K is closed in X. Take a sequence (p_n) in K which converges to some $p \in X$. Since K is compact, (p_n) has a subsequence (p_{n_k}) which converges to an element of K, but since $p_n \to p$ this subsequence can only converge to p, so the element of K which this subsequence converges to must be p. Hence $p \in X$, showing that K is closed in X.

To show that K is bounded, suppose for a contradiction that it was unbounded. Then no ball of any radius around a fixed point $p \in K$ can contain all of K. Thus we can find a point x_1 of K not in $B_1(p)$, a point x_2 of K not in $B_2(p)$, and in general a point x_n of K not in $B_n(p)$. This gives a sequence (x_n) of K satisfying $d(x_n, p) > n$, which implies that any subsequence of (x_n) is unbounded and hence not convergent. This contradicts compactness of K, so K must be bounded. □

Heine-Borel Theorem. In turns out that in $\mathbb{R}^n$, the properties of compact sets given above completely characterize those sets which are compact: a subset K of $\mathbb{R}^n$ is compact if and only if K is closed and bounded. Thus, compact subset of $\mathbb{R}^n$ are fairly easy to describe, and easy to visualize in the case of $\mathbb{R}, \mathbb{R}^2$, and $\mathbb{R}^3$. We'll give the proof for $\mathbb{R}^2$, which illustrates the basic idea.

Proof. The forward direction, that compact implies closed and bounded, was the statement of the previous proposition and is true in any metric space. Thus we need only justify the backwards direction. To this end, suppose that $K \subseteq \mathbb{R}^2$ is closed and bounded, and let (x_n, y_n) be an arbitrary sequence in K. The goal is to produce a subsequence (x_{n_k}, y_{n_k}) which converges in K itself.

Since K is bounded, the sequence (x_n, y_n) is bounded and hence the sequence (x_n) of x-coordinates (x_n) alone is also bounded in $\mathbb{R}$. By the Bolzano-Weierstrass Theorem, this then has a convergent subsequence (x_{n_k}), converging to some $x \in \mathbb{R}$. (At this point in class, I completely glossed over the following subtlety: if we similarly take a convergent subsequence (y_{n_ℓ}) of (y_n), there is no guarantee that (x_{n_k}, y_{n_ℓ}) will be a subsequence of the original (x_n, y_n) since the indices n_k and n_ℓ don't necessarily match up. If for instance (x_{n_k}) consists of the odd-indexed terms of (x_n) and (y_{n_ℓ}) the even-indexed terms of (y_n), then we get things like (x_1, y_2) in (x_{n_k}, y_{n_ℓ}), which don't belong to the original sequence (x_n, y_n). Thus we have to be more careful about how we construct the subsequence we need, in that we need the indices between the x_n's and y_n's to agree. Again, note that I didn't explicitly go through this in class when first talking about the Heine-Borel Theorem, but instead we cleared it up the following lecture during the Warm-Up.)

Consider now the terms (y_{n_k}) of the sequence (y_n) of y-coordinates corresponding to the subsequence (x_{n_k}) we got above. Since (x_n, y_n) is bounded, (y_n) and hence (y_{n_k}) is also bounded, so again by the Bolzano-Weierstrass Theorem there exists a convergent subsequence $(y_{n_{k_\ell}})$ converging to some $y \in \mathbb{R}$. (This is a "subsequence of a subsequence of (y_n)", quite a mouthful!) Since $x_{n_k} \to x$, the subsequence $(x_{n_{k_\ell}})$ of terms corresponding to the $y_{n_{k_\ell}}$ also converges to x, so the sequence $(x_{n_{k_\ell}}, y_{n_{k_\ell}})$ converges to (x, y) in $\mathbb{R}^2$. This is now a subsequence of (x_n, y_n), and hence consists of elements of K, so since K is closed in $\mathbb{R}^2$ the limit (x, y) is also in K, showing that (x_n, y_n) has a convergent subsequence in K. Thus K is compact. □

Remark. The proof for $\mathbb{R}^n$ in general is similar, only we have to continue taking subsequent of subsequences of subsequences over and over again, and we end up with crazy notation like $x_{n_{k_{\ell_s}}}$ and so on. So, it gets tedious to write all out, so we won't.

Converse does not hold in general. But now a warning, even being closed and bounded implies compactness in $\mathbb{R}^n$, this is not true in general. For a first example, consider the set A of rational numbers between $-\sqrt{2}$ and $\sqrt{2}$. We have seen previously that this is closed in $\mathbb{Q}$, and it is certainly bounded, but we claim that it is not compact. Indeed, take a sequence (r_n) of rationals in this set converging to $\sqrt{2}$ in $\mathbb{R}$. Then this is a sequence in A with no convergent subsequence in A, since any subsequence will converge to $\sqrt{2}$, which is not in A. Thus A is not compact.

For a second example, consider the closed ball $M_1(0)$ of radius 1 around the constant zero function in $C[0,1]$ equipped with the sup metric:

$$M_1(0) := \{f \in C[0,1] \mid |f(x)| \leq 1 \text{ for all } x \in [0,1]\}.$$

The sequence of functions $f_n(x) = x^n$ is in this closed ball, but has no convergent subsequence. This is because in this setting a convergent subsequence would be a uniformly convergent subsequence, and we have seen that no sequence of such functions can converge uniformly. To be precise, the pointwise limit of any subsequence of (f_n) would be the function which is 0 on $[0,1)$ and 1 at 1, which is not continuous and hence this pointwise convergence cannot be uniform. Thus $M_1(0)$ is not compact.

Important. Compact subsets of a metric space are always closed and bounded, but not necessarily the other way around in general. However, in $\mathbb{R}^n$, closed and bounded *does* imply compact.

Proposition. One final general fact: if X itself is compact and K is closed in X, then K is compact as well. Indeed, suppose that (p_n) is a sequence in K. Since X is compact, this has a subsequence (p_{n_k}) converging to some $p \in X$. But K is closed in X, so since p_{n_k} is also a sequence in K, the limit p of this sequence is in K. Thus the sequence (p_n) in K has a convergent subsequence in K, so K is compact.

Lecture 20: More on Compactness

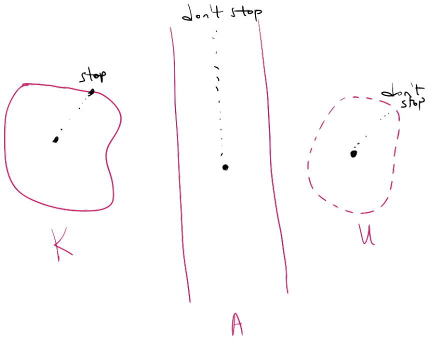
Today we continued talking about compact sets, giving an equivalent definition in terms of so-called *open covers*. While this definition is harder to grasp at first, I think it is better suited for understanding the intuition behind compactness.

Warm-Up. Suppose that X and Y are compact subsets of $\mathbb{R}$. We show that $X \times Y \subseteq \mathbb{R}^2$ is compact with respect to any of our three favorite metrics on $\mathbb{R}^2$. For a first approach, since X and Y are closed in $\mathbb{R}$, $X \times Y$ is closed in $\mathbb{R}^2$. Indeed, if $(x_n, y_n) \to (x, y) \in \mathbb{R}^2$ where each $(x_n, y_n) \in X \times Y$, then $x_n \to x$ and $y_n \to y$ in $\mathbb{R}$, so since each $x_n \in X$ and X is closed in $\mathbb{R}$, $x \in X$, and since each $y_n \in Y$ and Y is closed in $\mathbb{R}$, $y \in \mathbb{R}$. Thus $(x, y) \in X \times Y$, so $X \times Y$ is closed as claimed. Also, since X and Y are bounded, so is $X \times Y$; to be precise, if $X \subseteq (p - r, p + r)$ for some $p \in X$ and $r > 0$ and $Y \subseteq (q - s, q + s)$ for some $q \in Y$ and $s > 0$, then $X \times Y \subseteq B_{\max\{r,s\}}(p, q)$ with respect to the box metric, so $X \times Y$ is bounded. Hence by the Heine-Borel Theorem, $X \times Y$ is compact since it is closed and bounded in $\mathbb{R}^2$.

For a second approach. we can essentially repeat the proof of the Heine-Borel Theorem in $\mathbb{R}^2$. Take a sequence (x_n, y_n) in $X \times Y$. Since X is compact, the sequence (x_n) in X has a convergence subsequence (x_{n_k}) in X, and then since Y is compact, the sequence (y_{n_k}) in Y has a convergent subsequence $(y_{n_{k_\ell}})$ in Y. Then $(x_{n_{k_\ell}}, y_{n_{k_\ell}})$ is a convergent subsequence of (x_n, y_n) in $X \times Y$, so $X \times Y$ is compact.

Compact implies complete. One more quick fact: compact spaces are always complete. Indeed, if K is compact and (p_n) is a Cauchy sequence in K, then (p_n) has a convergent subsequence (p_{n_k}) in K. But then the Warm-Up from February 11th shows that (p_n) itself converges to the same limit as (p_{n_k}), so Cauchy sequences always converge and hence K is complete.

Intuition. Before giving an alternate characterization of compact sets, let us say a bit about the idea which compact sets are meant to capture. I claim that compact sets are those which are, in a sense, not "too large". We'll see using the open cover approach to compactness what this really means, but here's a rough idea to have in mind. As usual, we focus on $\mathbb{R}^2$, in which case a compact is simply one which is closed and bounded:



In the case of the compact set K, the point is that if you are in K are start walking in any direction while remaining in K, you will eventually reach a stopping point, namely once you reach the boundary. Thus, K is not "too large" in the sense that you eventually stop when walking through K. Now, in the case of the set A which is drawn, when walking vertically through A you will *never* reach a stopping point since A is unbounded in this direction. So, in a sense A is "too large physically" to be able to stop walking.

What about the set U? This is bounded but still not compact, and the way in which U is "too large" in this case is more subtle. Again, imagine starting in U and walking in a certain direction. The point is that if you are restricted to remain in U throughout your entire walk, you will never stop since you cannot reach the boundary as this boundary is not in U itself. The issue is not that U is too large physically, it is that U is "too large" in the sense of "time", in that you will never reach the boundary in a finite amount of time if you are supposed to remain in U throughout. So, U is "too large" in a different sense than A was, but is still not compact.

To get an idea for where these ideas will eventually lead us, think back to how continuous functions behave on closed intervals. The Extreme Value Theorem says that such functions always have a maximum and a minimum, but the same is not necessarily true of continuous functions on open intervals. Thus, from this point of view, open intervals are "too large" in that a continuous function on it might be unbounded (too large physically) or be bounded but without a maximum nor minimum (too large in the sense of time). The fact that continuous functions on closed intervals always had maximums and minimums was crucial in developing a good theory of integration, and indeed we will see next quarter that the same is true in higher dimensions: compact subsets of $\mathbb{R}^n$ are the kinds of spaces over which we can give a good definition of Riemann integration.

Covering definition of compactness. Let K be a subset of a metric space X. By an *open cover* of K we mean a collection $\{U_\alpha\}$ of open subsets of X which cover K in the sense that K is contained in their union:

$$K \subseteq \bigcup_\alpha U_\alpha.$$

We say that K is *compact* if every open cover of K has a finite subcover, where by *finite subcover* we mean finitely many of the U_α's which still cover K:

$$K \subseteq U_1 \cup \cdots \cup U_n.$$

Thus, compactness of K means that every open cover can be reduced to a finite sub cover. Intuitively, any possibly "infinite amount of data" $\{U_\alpha\}$ can be replaced by a "finite amount of data" $U_1, \ldots, U_n$, which more correctly captures the sense in which compact sets are not "too large".

One point of possible confusion: the definition does not say that K has *a* finite open cover, but rather that *any* open cover and be reduced to a finite one. Indeed, we can always view X itself as a one-element open cover of any of its subsets so that any possible subset has a finite cover; the key is that whether or not we can always find such finite covers no matter what arbitrary open cover we start with.

Theorem. Before looking at examples and uses of this new definition, rest assured that it truly is equivalent to the previous one: a space K satisfies the sequence definition of compactness if and only if it satisfies the covering definition of compactness. This is somewhat mind-blowing since the two definitions seem so different from another, but it is true.

We will prove that the covering definition implies the sequential definition, but the proof that the sequential definition implies the covering definition is truly a work of art and is very difficult. We'll leave it to other sources, if you're interested in looking it up. (Note that our book does not prove this, and indeed doesn't even seem to mention the sequential definition of compactness at all. I can suggest other books to look at if you really want to see how this works.)

Proof. (This proof is taken from the book *Real Mathematical Analysis* by Pugh.) Suppose that $K \subseteq X$ satisfies the covering definition of compactness. For a contradiction, suppose also that K does not satisfy the sequential definition of compactness. Then there exists a sequence (p_n) of distinct points in K with no convergent subsequence in K. For a fixed $x \in K$, it then follows that there is some open ball around x which contains only finitely many terms from (p_n), since if every open ball around x contained infinitely many such terms we could take balls of shrinking radii to get a subsequence of (p_n) converging to x. Denote the radius of the ball around x which contains only finitely many of the p_n's by $r_x > 0$.

Then the collection of such balls $\{B_{r_x}(x)\}$ as x varies in K forms an open cover of K. Since K satisfies the covering definition of compactness, this cover has a finite subcover, say:

$$B_{r_1}(x_1), \ldots, B_{r_\ell}(x_\ell).$$

Since these open balls together cover K and all the terms from (p_n) are in K, each of the terms from (p_n) must be in at least one of these open balls. But there are only finitely many open balls here, and infinitely many terms in (p_n), so at least one of these balls must contain an infinite number of the p_n's, which contradicts the choice of $r_1, \ldots, r_\ell$ as radii of balls which contained only finitely many of the terms in (p_n). Thus K does satisfy the sequential definition of compactness. □

Important. A space is compact according to the definition in terms of sequences if and only if it is compact according to the definition in terms of open covers.

Examples. The collection of open intervals $\{(-n, n)\}_{n \in \mathbb{N}}$ is an open cover of $\mathbb{R}$ with no finite subcover, so $\mathbb{R}$ is not compact. Indeed, any finite number of these open intervals

$$(-n_1, n_1), \ldots, (-n_\ell, n_\ell)$$

will have as their union the interval $(-N, N)$ where $N = \max\{n_1, \ldots, n_\ell\}$, so no finite number of these intervals can cover all of $\mathbb{R}$. Also, the interval $(0, 1)$ is not covering compact since the intervals $(\frac{1}{n}, 1)$ together form an open cover with no finite subcover.

However, we know that a closed interval $[a, b]$ is compact according to our previous definition, so it should according to our new definition as well. Note that the type of open cover used above in the case of $(0, 1)$ will no longer give a counterexample since the endpoints a and b are now required to be among the points covered by the open cover. Showing that $[a, b]$ is compact using the covering definition is tricky, but if interested you can find a proof in the "Notes on Compactness" available at `http://math.northwestern.edu/~scanez/archives/real-analysis/notes.php`. The sequential definition is much simpler to apply in this case.

Another example. Any finite subset $K = \{x_1, \ldots, x_n\}$ of a metric space is compact. Indeed, suppose that $\{U_\alpha\}$ is an open cover of K. Pick U_1 from this collection which containing x_1, U_2 which contains x_2. and so on. This then gives a finite number of open sets $U_1, \ldots, U_n$ from this collection which still cover all of K, so this is a finite subcover of $\{U_\alpha\}$ and hence K is compact.

This can also be proved using the sequential definition as follows. Take a sequence (p_n) in K. Since K only has finitely many points, one of these finitely many points must be repeated infinitely often among the p_n's, and the subsequence of (p_n) of these terms is then a convergent subsequence of (p_n), so K is sequentially compact as well. Note that, although the sequential definition is not difficult to verify here, verifying the covering definition is a little more straightforward.

Closed in compact is compact. Now we reprove some properties we saw last time, only now using the covering definition of compactness. We show if K is closed in X and X is compact, then K is compact as well. Take an arbitrary open over $\{U_\alpha\}$ of K. Since K is closed in X, K^c is open so then

$$\{U_\alpha\} \cup \{K^c\}$$

is an open cover of X. (The first collection covers all of K and K^c covers the rest of X.) Since X is compact, this has a finite subcover, say

$$U_1 \cup \ldots \cup U_n \cup K^c.$$

Then the sets $U_1, \ldots, U_n$ cover K since any element in K is also in X and hence is in one of the sets $U_1, \ldots, U_n, K^c$, but certainly does not belong to K^c. Thus $U_1, \ldots, U_n$ is a finite subcover of the open cover $\{U_\alpha\}$ of K, so K is compact.

Compact implies closed and bounded. Finally we show that a compact subset K of a metric space X is always closed and bounded. Note in the proof how compactness allows us to replace an infinite amount of data with a finite amount of data, which is really key point behind compactness.

First we show that K is bounded. Let $p \in K$ and consider the collection $\{B_r(p)\}_{r>0}$ of all opens balls centered at p. These together cover all of K since eventually any point of K will be in one of these balls once r is large enough, so since K is compact this cover has a finite subcover, say:

$$B_{r_1}(p), \ldots, B_{r_n}(p).$$

Setting $s = \max\{r_1, \dots, r_n\} > 0$, we then have:

$$K \subseteq B_{r_1}(p) \cup \cdots \cup B_{r_n}(p) = B_s(p),$$

so K is bounded as claimed. (Note: replacing an infinite number of radii with finitely many allows us to take their maximum.)

Now we show that K is closed by showing that K^c is open. Let $p \in K^c$. Our goal is to show there is a ball around p which is contained in K^c. For any $x \in K$, we can find open balls V_x and U_x around x and p respectively which do not intersect each other, say by taking their common radius to be $r = \frac{d(x,p)}{2}$. (Draw a picture!) As $x \in K$ varies through all possible points, we then get an open cover $\{U_x\}$ of K. Since K is compact, this has a finite subcover, say:

$$K \subseteq U_{x_1} \cup \cdots \cup U_{x_n}.$$

The corresponding V's then all contain p and we claim that their intersection:

$$V_{x_1} \cap \cdots \cap V_{x_n}$$

is fully contained in K^c. Indeed, V_{x_i} is contained in the complement of U_{x_i}, so

$$V_{x_1} \cap \cdots \cap V_{x_n} \subseteq (U_{x_1})^c \cap \cdots \cap (U_{x_n})^c = (U_{x_1} \cup \cdots \cup U_{x_n})^c \subseteq K^c$$

as desired. But $V_{x_1} \cap \cdots \cap V_{x_n}$ is the intersection of finitely many open sets, so is itself and open and hence since $p \in V_{x_1} \cap \cdots \cap V_{x_n}$, there exists a ball $B_s(p)$ contained in this intersection, and hence contained in K^c. Thus K^c is open in X, so K is closed.

This is a tricky proof to follow at first, but drawing a picture helps to see what the rationale behind the U's and V's are. Again, note the fact that compactness of K allowed us to replace an infinite amount of data (the U_x's in general) by a finite amount (the U_{x_i}'s), and hence we were able to take their intersection and guarantee that the resulting set is still open.

Important. The covering definition of compactness in practice allows us to replace an infinite amount of data with a finite amount. Thus, compact sets are not "too large" in the sense that questions about infinitely many things can be reduced to questions about finitely many things.

Lecture 21: Connected Sets

Today we spoke about the notion of a connected space, which intuitively means a space which consists of a single "piece". We'll see later that this leads to a generalization of the Intermediate Value Theorem, and is a concept which will pop up when looking at higher dimensional derivatives and integrals next quarter.

Warm-Up. Suppose that K is a compact subset of $\mathbb{R}^2$ with the property that for all $p \in K$ there exists an open ball $B_r(p)$ around p which intersects K only at p, so an open ball with the property that the only element of K in that ball is p itself. We show that K must be finite.

To make the dependence of the open balls in question on the point it is centered at clear, we denote the radius of the ball by r_p so that $B_{r_p}(p)$ is the ball around $p \in K$ in which the only element of K is p itself. The collection of open balls $\{B_{r_p}(p)\}$ as p ranges through K form an open cover of K, so since K is compact this has a finite subcover:

$$B_{r_{p_1}}(p_1), \dots, B_{r_{p_n}}(p_n).$$

Now, being a cover of K, every element of K is contained in at least one of the open balls, but the only element of K in the first ball is p_1, the only element of K in the second ball is p_2, and so on, showing that the only elements of K are $p_1, \ldots, p_n$ and hence that K is finite as claimed. Note that the same reasoning works for a compact set K of an arbitrary metric space X, not simply a compact subset of $\mathbb{R}^2$, so the claim generalizes.

As an additional remark, it follows that for an *infinite* compact subset K of a metric space X, there must exist at least one point $p \in K$ such that *every* open ball around it contains an element of K different from p itself; as mentioned on the set of practice problems for the second midterm, such a point p is called an *accumulation point* of K, so what we have shown here can be rephrased as saying that any infinite compact subset of a metric space has an accumulation point.

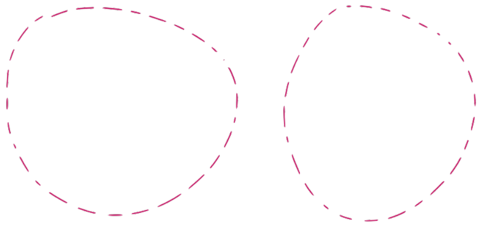
Connectedness. A metric space X is said to be *disconnected* if there exist disjoint, nonempty open subsets U and V of X such that

$$X = U \cup V.$$

(Recall that saying U and V are disjoint means that $U \cap V = \emptyset$.) Intuitively, this means that we can "split" X into two pieces, U and V. A space is *connected* if it is not disconnected (beware the double negative!), meaning that such a "splitting" is not possible.

We'll draw some pictures of what disconnected and connected spaces look like, which will match our intuitive notion as to what these terms might mean. But, here is another way to phrase the definition of connected in a way which is more practically useful: X is connected if whenever we have $X = U \cup V$ with U and V disjoint and open in X, one of U or V must be empty.

Examples. The space consisting of the union of the following open disks is disconnected:



Indeed, each open disk in question is open in this space, is nonempty, and has nothing in common with the other one. Visually we can see the way in which disconnected spaces can be "split" up into multiples "pieces". Each individual open disk, however, would be an example of a connected space.

The space $X = [0, 1] \cup [2, 3]$ is also disconnected, with respect to the standard absolute value metric. At first this might not seem to match the definition since the definition says we need to break our space up into *open* sets, but the point is that these sets are only required to be open in X itself. In particular, the intervals $[0, 1]$ and $[2, 3]$ are indeed open in X so we conclude that X is disconnected.

Finally, $\mathbb{Z}$ is disconnected, which again should be visually clear. To be precise, *every* subset of $\mathbb{Z}$ is open in $\mathbb{Z}$ so

$$\mathbb{Z} = \{\text{negative integers}\} \cup \{\text{nonnegative integers}\}$$

exhibits $\mathbb{Z}$ as the union of two nonempty, open and disjoint subsets, so $\mathbb{Z}$ is disconnected. In fact, the only connected subsets of $\mathbb{Z}$ are those which consist of a single point or are empty.

Clopen subsets. In the decomposition $X = U \cup V$ into disjoint open subsets, note that we can view each subset as the complement of the other. Since complements of open sets are closed, we see that U and V are both open *and* closed in X, so they are *clopen* subsets of X. Thus, saying that X is disconnected implies that it has a nonempty proper clopen subset, while conversely if X has a nonempty property clopen subset A,

$$X = A \cup A^c$$

exhibits X as the union of two nonempty, disjoint, open subsets, so X is disconnected.

Thus we have that X is disconnected if and only if it has a nonempty property clopen subset, or equivalently X is connected if and only if the only clopen subsets of X are $\emptyset$ and X itself. This gives a more succinct way of saying what disconnected/connected mean, although the definition we first gave is visually clearer. In the case of $X = [0,1] \cup [2,3]$, both $[0,1]$ and $[2,3]$ are clopen subsets of X, while in the case of the integers *every* subset is clopen.

Important. X is disconnected if we can write it as $X = U \cup V$ for nonempty and disjoint open subsets U, V of X. X is connected if whenever $X = U \cup V$ with U, V open and disjoint, one of U or V must be empty. Equivalently, X is disconnected if it has a proper, nonempty clopen subset, and X is connected if the only clopen subsets of X are $\emptyset$ and X itself.

Why we care. To motivate why we care about the notion of connected sets, consider the following question: if $f : U \to \mathbb{R}$ is a differentiable function on an open subset U of $\mathbb{R}$ with $f'(x) = 0$ for all $x \in U$, is it true that f must be constant? Your experience in calculus might lead you to believe that this is true, but in fact it is only true if U is connected! Indeed, take the function $f : U \to \mathbb{R}$ on $U = (-2,-1) \cup (1,2)$ defined by

$$f(x) = \begin{cases} 1 & x \in (-2,-1) \\ -1 & x \in (1,2). \end{cases}$$

This is differentiable and has derivative equal to zero throughout U, but is clearly not constant, the issue being that U here is disconnected. In general, having derivative zero everywhere throughout a region only implies that your function is constant on each "connected piece" of that region, but the constant over different pieces can differ.

A similar thing will be true when we consider higher-dimensional derivatives, so the distinction between connected and disconnected spaces will pop-up next quarter as well, although only in the setting of $\mathbb{R}^n$ where things are easier to visualize.

Theorem. Any interval I in $\mathbb{R}$ is connected, which should make intuitive sense visually. This is proved in the book using the Intermediate Value Theorem, but here we give a proof which avoids this. The one fact we need, which we haven't mentioned explicitly before, is that a compact subset of $\mathbb{R}$ always has a maximum element and a minimum element: indeed, since a compact subset is bounded, it has a supremum and an infimum, and since a compact set is closed, it will contain its supremum and its infimum. Note that by "interval" we mean any type: open, closed, half-open, half-closed, bounded or unbounded, so that in particular $\mathbb{R} = (-\infty, \infty)$ itself is connected.

Proof. For a contradiction, suppose that $I = U \cup V$ where U and V are nonempty, disjoint, and open in I. Pick $x \in U$ and $y \in V$, and assume without loss of generality that $x < y$ and consider the smaller interval $[x,y] \subseteq I$. Then we have

$$[x,y] = ([x,y] \cap U) \cup ([x,y] \cap V).$$

Since $[x,y] \cap U$ is open in $[x,y]$, its complement $[x,y] \cap V$ is closed in $[x,y]$ and hence compact since a closed subset of a compact space is always compact. Thus $[x,y] \cap V$ has a minimum element, call it $b \in [x,y] \cap V$.

Consider now the interval $[x,b] \subseteq [x,y]$. We have

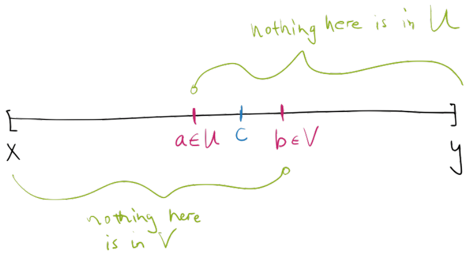
$$[x,b] = ([x,b] \cap U) \cup ([x,b] \cap V),$$

so the same argument as above shows that $[x,b] \cap U$ is compact and hence has a maximum element, call it $a \in [x,b] \cap U$. We have $a \leq b$, and thus $a < b$ since $a \neq b$ given that $a \in U$, $b \in V$, and U and V are disjoint. Now, take any $a < c < b$. Since

$$x \leq a < c < b \leq y,$$

$c \in [x,y]$, and thus either $c \in [x,y] \cap U$ or $c \in [x,y] \cap V$. However, the first is not possible since c is greater than the largest element a of the first set, and the second is not possible since c is smaller than the smallest element b of the second set. Thus we have a contradiction, so I must have been connected to begin with.

Here is a picture to illustrate where all the different elements considered above come from:



The point is that b is the smallest element of V in $[x,y]$ and a the largest element of U in $[x,y]$, so nothing between them is in $[x,y]$, which contradicts a basic property of intervals. □

Remark. The above proves that intervals are always connected, but in fact the converse is true as well: any connected subset of $\mathbb{R}$ must be an interval. You can check the book for a proof of this fact, which is quite short.

Important. A subset of $\mathbb{R}$ is connected if and only if it is an interval.

Lecture 22: Continuous Functions

Today we starting talking about continuous functions between metric spaces, which will be our last topic. We'll see that many properties we saw last quarter of continuous functions in the setting of $\mathbb{R}$ generalize to arbitrary spaces.

Warm-Up. Suppose that A and B connected subsets of $\mathbb{R}^2$ which are not disjoint. We show that $A \cup B$ connected. This generalizes to other metric spaces as well, and gives a quick way of verifying that various sets are indeed connected.

Suppose that $A \cup B = U \cup V$ where U and V are open in $A \cup B$ and disjoint. We must show that one of U or V is empty. Since A and B are not disjoint, there exists $p \in A \cap B$, and hence this element is in $U \cup V$ so either $p \in U$ or $p \in V$. Without loss of generality suppose that $p \in U$, in which case we must show that V is empty.

Now, we can write A as

$$A = (A \cap U) \cup (A \cap V)$$

since any element of A must be in U or V. Since U and V are open in the larger space $A \cup B$, these intersections are each open in A. (In general, it is true that if $A \subseteq X$ and U is open in X, then $A \cap U$ is open in A, and in fact all open subsets of A arise in this way.) Since A is connected, one of these two open sets must be empty, and since $p \in A \cap U$ we must have $A \cap V = \emptyset$. Similarly, writing B as

$$B = (B \cap U) \cup (B \cap V)$$

and using the fact that $p \in B \cap U$, the fact that B is connected implies that $B \cap V = \emptyset$. But now we can conclude that $V = \emptyset$: if not, a point of V would be in A or B since $V \subseteq U \cup V = A \cup B$, in which case this point would be either in $A \cap V$ or $B \cap V$, neither of which are possible since both of these intersections are empty. Thus we conclude that $A \cup B$ is connected as claimed.

Continuity. Suppose that (X, d_X) and (Y, d_Y) are metric spaces. We say that a function $f : X \to Y$ is *continuous at* $p \in X$ if either of the following equivalent conditions hold:

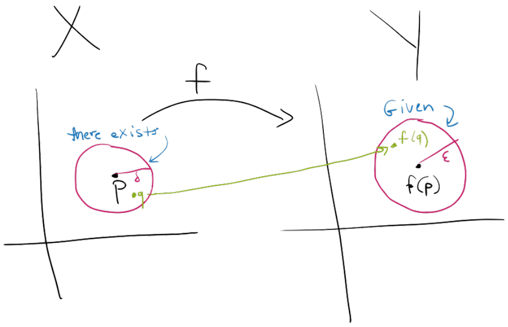
- whenever $p_n \to p$ in X, $f(p_n) \to f(p)$ in Y, or
- for any $\epsilon > 0$ there exists $\delta > 0$ such that if $d_X(q, p) < \delta$, then $d_Y(f(q), f(p)) < \epsilon$.

We say that f is *continuous* on X if it is continuous at each point of X.

You should recognize these two conditions as mimicking the sequential and ϵ-δ characterizations of what continuity meant for functions $f : \mathbb{R} \to \mathbb{R}$ or with domain some subset of $\mathbb{R}$. Indeed, all we did was to take those characterizations and replace the absolute value distance used in $\mathbb{R}$ with more general metrics. Note that, as opposed to the case of functions $\mathbb{R} \to \mathbb{R}$, there are *two* metrics at play here: the one on X and the one on Y. In particular, in the sequential characterization the metric on X is used when defining what $p_n \to p$ means and the metric on Y is used when defining what $f(p_n) \to f(p)$ means. Changing one or both of these metrics could very well change whether a given function is continuous or not.

The proof that the two conditions given above are equivalent is the same as the proof of the analogous fact for functions $f : \mathbb{R} \to \mathbb{R}$, only we replace $|p - q|$ or $|f(p) - f(q)|$ throughout by $d_X(p, q)$ or $d_Y(f(p), f(q))$ respectively. Check the book for full details.

Visualizing continuity. The picture to have mind is most easily obtained using the ϵ-δ characterization, which we can rephrase as: given $\epsilon > 0$, there exists $\delta > 0$ such that if $q \in B_\delta(p)$, then $f(q) \in B_\epsilon(f(p))$. (Note that these are open balls in possibly *different* metric spaces.) In other words, given any open ball around $f(p)$ in Y, there exists an open ball around p which is sent entirely into the given ball around $f(p)$ under f:



In other words, given any measure as to how close (as determined by the metric) we want to end up near $f(p)$, we can always come in close enough to p in the domain to guarantee that we end up where we want. We'll see next time that this point of view will lead to a characterization of continuity phrased solely in terms of open sets.

Important. Saying that $f : X \to Y$ is continuous at $p \in X$ means that given any ball around $f(p)$ in Y we can find a ball around p in X which is fully sent into the original given ball around $f(p)$.

Example 1. The function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x, y) = x + y - xy$ is continuous with respect to the Euclidean metrics on $\mathbb{R}^2$ and $\mathbb{R}$. Indeed, fix $(a, b) \in \mathbb{R}^2$ and suppose that $(a_n, b_n) \to (a, b)$ in $\mathbb{R}^2$. Then $a_n \to a$ and $b_n \to b$ in $\mathbb{R}$. Using what we know about real sequences from last quarter, we then have $a_n b_n \to ab$ so

$$f(a_n, b_n) = a_n + b_n - a_n b_n \to a + b - ab = f(a, b).$$

Thus f is continuous at (a, b), and since $(a, b) \in \mathbb{R}^2$ was arbitrary, f is continuous on $\mathbb{R}^2$.

Using what we know about convergence in $\mathbb{R}^2$ with respect to the Euclidean, taxicab, or box metrics, the same reasoning shows that f is still continuous if we change the metric on $\mathbb{R}^2$ to be either the taxicab or box metric instead. In general, a functions $\mathbb{R}^n \to \mathbb{R}^m$ is continuous with respect to some choices of these metrics if and only if it is continuous with respect to the other choices. (For instance, we can use the taxicab metric on $\mathbb{R}^n$ and the Euclidean metric on $\mathbb{R}^m$, or the box metric on $\mathbb{R}^n$ and the taxicab metric on $\mathbb{R}^m$, and so on.)

Also, it is more generally true that the sums and products of continuous functions $f : X \to \mathbb{R}$ are also continuous, as are quotients as long as the denominator is nonzero. Note that a statement like this does not necessarily make sense for a function $f : X \to Y$ where Y is an arbitrary metric space since "addition" and "multiplication" do not necessarily make sense in a random space Y. But, in the cases where it does make to sense to speak of addition or multiplication in Y, sums and products of continuous functions are indeed continuous.

Example 2. We show that the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x, y) = \begin{cases} \frac{x^2y^2}{x^2+y^2} & (x, y) \neq (0, 0) \\ 0 & (x, y) = (0, 0) \end{cases}$$

is continuous at the origin. (It is continuous elsewhere since it is a quotient of continuous functions with nonzero denominator.) Let $\epsilon > 0$. We want to find $\delta > 0$ such that

$$d((x, y), (0, 0)) < \delta \implies |f(x, y) - f(0, 0)| < \epsilon$$

where d denotes the metric on $\mathbb{R}^2$, say the Euclidean metric, and where we use the standard metric on $\mathbb{R}$. Cleaning this up, we want $\delta > 0$ such that

$$\sqrt{x^2 + y^2} < \delta \implies \left| \frac{x^2y^2}{x^2 + y^2} \right| < \epsilon$$

for $(x, y) \neq (0, 0)$. (At the origin the expression $|f(x, y) - f(0, 0)|$ is zero, and so is certainly smaller than whatever ϵ is.)

But if $(x, y) \neq (0, 0)$, at least one of x or y is nonzero, so assuming that is x is the nonzero one we have:

$$\left| \frac{x^2y^2}{x^2 + y^2} \right| \leq \frac{x^2y^2}{x^2} = y^2$$

where the first inequality is true since we made the denominator smaller or kept it the same. Since

$$|y| = \sqrt{y^2} \leq \sqrt{x^2 + y^2},$$

choosing $\delta = \sqrt{\epsilon}$ gives

$$|y| = \sqrt{y^2} \leq \sqrt{x^2 + y^2} < \sqrt{\epsilon}$$

so

$$\left|\frac{x^2y^2}{x^2 + y^2}\right| \leq \frac{x^2y^2}{x^2} = y^2 < \epsilon.$$

A similar reasoning works in the case when $x = 0$ and instead y is nonzero, so we conclude that $\delta = \sqrt{\epsilon}$ satisfies

$$\sqrt{x^2 + y^2} < \delta \implies \left|\frac{x^2y^2}{x^2 + y^2}\right| < \epsilon$$

and hence that f is continuous at the origin as claimed. Again, you can also make this work if you use the taxicab or box metrics on $\mathbb{R}^2$ instead of the Euclidean metric.

Example 3. Suppose that X is a discrete metric space. Then *any* function $f : X \to Y$ is continuous no matter what Y is. Indeed, using the sequence characterization we start with $p_n \to p$ in X. But the only convergent sequences in a discrete space are those which are eventually constant, so we have that $p_n = p$ for large enough n. Thus $f(p_n) = f(p)$ for large enough n, so $f(p_n) \to f(p)$ in Y since $(f(p_n))$ is eventually constant. Thus $p_n \to p$ implies $f(p_n) \to f(p)$, so f is continuous at any $p \in X$ and hence on all of X.

Using the ϵ-δ characterization, for any $\epsilon > 0$ take $\delta = \frac{1}{2}$ and suppose that $d_X(q, p) < \frac{1}{2}$. Since d_X is the discrete metric, we can only have $d_X(q, p)$ smaller than 1 if it is zero, in which case $q = p$. But if $q = p$ we certainly have $d_Y(f(q), f(p)) = 0 < \epsilon$, so $d_X(q, p) < \frac{1}{2}$ implies $d_Y(f(p), f(q)) < \epsilon$ and hence f is continuous at any $p \in X$.

As a contrast, if Y is discrete the only continuous functions $f : \mathbb{R}^n \to Y$ are the constant ones, where we are using the Euclidean metric on $\mathbb{R}^n$. Indeed, in order to have $d_Y(f(p), f(q)) < \frac{1}{2}$ we must have $f(p) = f(q)$, so the ϵ-δ characterization of continuity implies that if $f : \mathbb{R}^n toY$ is continuous, f must be constant.

Important. Continuity depends on the metrics we use, in that a function $f : X \to Y$ may be continuous with respect to one choice of metrics on X and Y but not with respect to other choices. In the case of $\mathbb{R}^n$, however, a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous with respect to some choices of the Euclidean, box, or taxicab metrics on the domain and codomain if and only if it is continuous with respect to the other choices.

Example 4. Take $C[0, 1]$ with the sup metric d and fix $a \in [0, 1]$. We claim that the function $E : C[0, 1] \to \mathbb{R}$ defined by $E(f) = f(a)$ is continuous. (This map E is called *evaluation at a* since it takes as input a function and outputs the value of this function at the fixed a.) Indeed, let $\epsilon > 0$ and set $\delta = \epsilon$. If $d(f, g) < \delta$, then

$$|f(a) - g(a)| \leq \sup_{x \in [0,1]} |f(x) - g(x)| = d(f, g) < \delta = \epsilon.$$

But $|f(a) - g(a)|$ is $|E(f) - E(g)|$, so $d(f, g) < \delta$ implies $|E(f) - E(g)| < \epsilon$, so E is continuous as claimed. Of course, since we cannot really visualize $C[0, 1]$, it might be hard at first to picture this fact geometrically, but the point is that if f and g are "close" in the sense the sup metric (think in term of their graphs being close to one another), then $f(a)$ and $g(a)$ should be close in $\mathbb{R}$.

(In fact, since δ only depends on ϵ and not on any specific $f \in C[0,1]$ are checking continuity at, E is actually *uniformly continuous*, where the definition of uniformly continuous is the same as the one we had last quarter in the setting of $\mathbb{R}$ only that we replace absolute value distances by arbitrary metrics.)

Lecture 23: More on Continuity

Today we continued talking about continuity, and in particular saw that we can rephrase the definition of continuity solely in terms of open sets. We also saw that continuous functions send connected spaces to connected spaces, leading to a generalization of the Intermediate Value Theorem from last quarter.

Warm-Up. We can express any function $f : \mathbb{R}^2 \to \mathbb{R}^2$ as

$$f(x,y) = (f_1(x,y), f_2(x,y))$$

for some functions $f_1, f_2 : \mathbb{R}^2 \to \mathbb{R}$, simply by taking f_1 to be the function defining the first coordinate of $f(x,y)$ and f_2 to be the function defining the second coordinate. We claim that f is continuous (with respect to any of the Euclidean, taxicab, or box metrics) if and only if f_1 and f_2 are continuous.

In one direction, suppose that f is continuous. Let $\mathrm{pr}_1 : \mathbb{R}^2 \to \mathbb{R}$ and $\mathrm{pr}_2 : \mathbb{R}^2 \to \mathbb{R}$ denote the *projections* of $\mathbb{R}^2$ onto the first and second coordinates respectively:

$$\mathrm{pr}_1(x,y) = x \text{ and } \mathrm{pr}_2(x,y) = y.$$

It is straightforward to check, say using sequences, that these projections are continuous. Then we can express f_1 and f_2 as the compositions

$$f_1 = \mathrm{pr}_1 \circ f \text{ and } f_2 = \mathrm{pr}_2 \circ f,$$

so f_1 and f_2 are continuous since compositions of continuous functions are always continuous.

Conversely suppose that $f_1, f_2 : \mathbb{R}^2 \to \mathbb{R}$ are continuous and fix $(a,b) \in \mathbb{R}^2$. Let $\epsilon > 0$. Since f_1 is continuous at (a,b) there exists $\delta_1 > 0$ such that

$$d((x,y),(a,b)) < \delta_1 \implies |f_1(x,y) - f_1(a,b)| < \frac{\epsilon}{\sqrt{2}}$$

where d is, say, the Euclidean metric on $\mathbb{R}^2$, and since f_2 is continuous at (a,b) there exists $\delta_2 > 0$ such that

$$d((x,y),(a,b)) < \delta_2 \implies |f_2(x,y) - f_2(a,b)| < \frac{\epsilon}{\sqrt{2}}.$$

Suppose that $d((x,y),(a,b)) < \min\{\delta_1, \delta_2\}$, so that both implies above hold. Then:

$$\begin{aligned} d(f(x,y), f(a,b)) &= d([f_1(x,y), f_2(x,y)], [f_1(a,b), f_2(a,b)]) \\ &= \sqrt{|f_1(x,y) - f_1(a,b)|^2 + |f_2(x,y) - f_2(a,b)|^2} \\ &< \sqrt{\frac{\epsilon^2}{2} + \frac{\epsilon^2}{2}} \\ &= \epsilon, \end{aligned}$$

which shows that f is continuous at (a,b) and hence on all of $\mathbb{R}^2$.

This generalizes to functions $f : \mathbb{R}^n \to \mathbb{R}^m$ and the point is that such a function is continuous if and only if each "component" function $\mathbb{R}^n \to \mathbb{R}$ is continuous, making continuity of functions which map into higher-dimensional spaces a little simpler to deal with.

Theorem. A function $f : X \to Y$ is continuous if and only if $f^{-1}(U)$ is open in X whenever U is open in Y. Here, $f^{-1}(U)$ denotes the *preimage* of U under f, which is the set of all elements in the domain which get sent to something in U:

$$f^{-1}(U) := \{x \in X \mid f(x) \in U\}.$$

This gives a characterization of continuity phrased solely in terms of open sets, and indeed leads to a notion of continuity in more general contexts where "open set" makes sense without necessarily having a metric available; this is the type of thing you would learn about in a full blown *topology* course such as Math 344.

This might seem like a strange phrasing of continuity, and indeed the proof might be tricky to follow at first, but I claim that this is essentially the same as the ϵ-δ definition when phrased in terms of open balls. In particular, saying

$$q \in B_\delta(p) \implies f(q) \in B_\epsilon(f(p))$$

means that q is in the preimage of $B_\epsilon(f(p))$, so $B_\delta(p)$ is an open ball around p which is fully contained in this preimage. The proof is tricky only because it requires jumping back and forth between various definitions, but the key point is that mentioned above.

Proof. Suppose that f is continuous and that $U \subseteq Y$ is open. Let $p \in f^{-1}(U)$. Then $f(p) \in U$, so since U is open in Y there exists $\epsilon > 0$ such that $B_\epsilon(f(p)) \subseteq U$. Now, since f is continuous at p, there exists $\delta > 0$ such that

$$q \in B_\delta(p) \implies f(q) \in B_\epsilon(f(p)).$$

But since $B_\epsilon(f(p)) \subseteq U$, this says that anything in $B_\delta(p)$ is sent to something in U, so that all of $B_\delta(p)$ is contained in the preimage of U. Thus there exists an open ball $B_\delta(p) \subseteq f^{-1}(U)$, so $f^{-1}(U)$ is open in X.

Conversely suppose that the preimage of any open subset of Y is open in X. Let $p \in X$ and let $\epsilon > 0$. Since $B_\epsilon(f(p))$ is open in Y, its preimage $f^{-1}(B_\epsilon(f(p)))$ is open in X. Thus since $p \in f^{-1}(B_\epsilon(f(p)))$, there exists $\delta > 0$ such that $B_\delta(p) \subseteq f^{-1}(B_\epsilon(f(p)))$. This means that any $q \in B_\delta(p)$ is in the preimage of $B_\epsilon(f(p))$, so for any such q we have $f(q) \in B_\epsilon(f(p))$. Thus

$$q \in B_\delta(p) \implies f(q) \in B_\epsilon(f(p)),$$

showing that f is continuous at p. □

Important. A function $f : X \to Y$ is continuous if and only the preimage of any open subset of Y is open in X. This is nothing but a rephrasing of the ϵ-δ definition of continuity.

Examples. I claim that although the above result seems new, we have essentially already used it numerous times last quarter and this quarter too. In particular, we have seen before that if $f : \mathbb{R} \to \mathbb{R}$ is continuous and $f(a) \neq 0$, then there exists an entire interval around a on which f is still nonzero. This is nothing but a consequence of the fact that the preimage of the set of nonzero real numbers is open in $\mathbb{R}$. Indeed, if f is continuous, then since $\mathbb{R}\backslash\{0\} = (-\infty, 0) \cup (0, \infty)$ is open in $\mathbb{R}$, $f^{-1}(\mathbb{R}\backslash\{0\})$ is open in $\mathbb{R}$. If $f(a) \neq 0$, a is in this preimage, so since this preimage is open there exists

$$(a-\delta, a+\delta) \subseteq f^{-1}(\mathbb{R}\backslash\{0\}).$$

But then anything in this interval is sent to something in $\mathbb{R}\backslash\{0\}$, so anything in this interval is sent to something nonzero and thus f is nonzero on this entire interval.

Similarly, if $f(a) > 2$, say, there exists an entire interval around a on which f is still strictly larger than 2. This comes from the fact that $(2, \infty)$ is open in $\mathbb{R}$, so if f is continuous its preimage $f^{-1}(2, \infty)$ is also open in $\mathbb{R}$, and thus for a in this preimage there is an interval around a which remains in the preimage. Then anything in this interval is sent to something in $(2, \infty)$ and we have our result.

It is also true that $f : X \to Y$ is continuous if and only if the preimage of any *closed* subset of Y is closed in X, as you will show on a practice problem for the final. As a consequence, the subset of $\mathbb{R}^2$ consisting of all points satisfying

$$xye^{x^2y-1} + \cos(\sin e^{xy-1}) + y^3e^{\sin xy} = 1$$

is closed in $\mathbb{R}^2$ since this can be characterized as the preimage of the closed subset $\{1\} \subseteq \mathbb{R}$ under the continuous functions $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x, y) = xye^{x^2y-1} + \cos(\sin e^{xy-1}) + y^3e^{\sin xy}$.

Theorem. Suppose that $f : X \to Y$ is continuous and that $A \subseteq X$ is connected. Then $f(A) \subseteq Y$ is connected as well. Thus, a continuous function sends connected sets to connected sets.

Proof. Suppose that $f(A) = U \cup V$ with U and V open in $f(A)$ and disjoint. We must show that one of U or V is empty. Since f is continuous, $f^{-1}(U)$ and $f^{-1}(V)$ are open in A and we have

$$A = f^{-1}(U) \cup f^{-1}(V).$$

(Technically, here we are not really considering the original $f : X \to Y$ but rather its *restriction* to A, which is often denoted by $f|_A : A \to f(A)$. This is a minor point which we will not dwell on.) Since U and V are disjoint, their preimages are disjoint, so since A is connected we thus have that one of $f^{-1}(U)$ or $f^{-1}(V)$ is empty. Without loss of generality, say that $f^{-1}(U)$ is empty. Then U is empty as well, since if not there would exist $f(p) \in U \subseteq f(A)$ and this p would then be in $f^{-1}(U)$, which is empty. Thus we conclude that $f(A)$ is connected. □

Intermediate Value Theorem. Finally we have the following result, which is a generalization of the Intermediate Value Theorem we had last quarter. Suppose that $f : A \to \mathbb{R}$ is continuous and that A is connected. Suppose further than $f(a) < f(b)$ in $\mathbb{R}$. Then for any $c \in \mathbb{R}$ such that $f(a) < c < f(b)$, there exists $p \in A$ such that $f(p) = c$.

Hence, any "intermediate value" between $f(a)$ and $f(b)$ is attained as a value of f, and we say that f has the *intermediate value property*. The Intermediate Value Theorem from last quarter is just the case of this more general version when A is an interval in $\mathbb{R}$, which is always connected.

Proof. Since f is continuous and A is connected, $f(A)$ is a connected subset of $\mathbb{R}$, so $f(A)$ must be an interval. But then if $f(a) < f(b)$ in this interval, any c between $f(a)$ and $f(b)$ remains in the interval, so $c \in f(A)$ and hence there exists $p \in A$ which is sent to c under f. □

Important. Continuous functions send connected sets to connected sets. As a consequence, any continuous function from a connected domain into $\mathbb{R}$ has the intermediate value property.

Lecture 24: Continuity and Compactness

Today we spoke about the special properties possessed by continuous functions on compact domains. As usual, these are generalizations of some facts we saw last quarter regarding the behavior of continuous functions $[a, b] \to \mathbb{R}$.

Warm-Up. We show that at any instant, there exist two *antipodal* points on the surface of the Earth at which the temperatures are exactly the same. (Saying that two points are antipodal means that they directly opposite one another, so for instance the north and south poles are antipodal points.)

The key point is that the surface of the Earth is connected—say when the modeled by a sphere—and temperatures vary continuously. To be clear, define f : Earth $\to \mathbb{R}$ to be the function

$$f(p) = (\text{temperature at } p) - (\text{temperature at } p')$$

where we use p' to denote the antipodal point of p. Saying that temperature varies continuously means that if we move away from a given point by a small amount, the temperature will also change by a small amount. Also, changing p by a small amount changes the antipodal point p' by a small mount, so we also have that the temperature at p' changes by a small amount. Putting this all together implies that f is continuous.

Since f is real-valued and continuous on a connected domain, it has the intermediate value property. Take any point p on the surface of the Earth. If the temperature at p is the same as that at p', there is nothing to show. Otherwise, either the temperature at p is larger than the temperature at p' or smaller than it. Thus either

$$f(p) > 0 \text{ or } f(p) < 0.$$

But if the temperature at p is larger than the temperature at its antipodal point, the temperature at the antipodal point is smaller than the temperature at its antipodal point, which is p. Thus if $f(p) > 0$ we have $f(p') < 0$, and similarly if $f(p) < 0$ we have $f(p') > 0$. Thus either way, 0 is an intermediate value between $f(p)$ and $f(p')$, so there exists q on the surface of the Earth such that $f(q) = 0$, meaning that the temperature at q is the same as the temperature at its antipodal point q', and we are done.

A modification of this will also show that there exist antipodal points at which both the temperatures *and* air pressure are the same, and similarly you can come up with tons of other fun facts about the world we live in.

Theorem. Suppose that $f : X \to Y$ is continuous and that $K \subseteq X$ is compact. Then $f(K) \subseteq Y$ is compact as well. Thus, the image of a compact set under a continuous function is always compact. We give two proofs, using the two characterizations of compactness we have seen.

Proof 1. Suppose that (q_n) is a sequence in $f(K)$. Then each q_n can be written as $q_n = f(p_n)$ for some $p_n \in K$ since each q_n is in the image of K. Since K is compact, the sequence (p_n) in K has a convergent subsequence, say $p_{n_k} \to p \in K$. Since f is continuous, we then have $f(p_{n_k}) \to f(p) \in f(K)$, so $(f(p_{n_k}))$ is a convergent subsequence of $(q_n) = (f(p_n))$ in $f(K)$, so $f(K)$ is compact. □

Proof 2. Suppose that $\{U_\alpha\}$ is an open cover of $f(K)$, so

$$f(K) \subseteq \bigcup_\alpha U_\alpha.$$

Since f is continuous, each preimage $f^{-1}(U_\alpha)$ is open in X, so the collection $\{f^{-1}(U_\alpha)\}$ forms an open cover of K:

$$K \subseteq \bigcup_\alpha f^{-1}(U_\alpha).$$

To be clear, for any $p \in K$, $f(p) \in f(K)$ is in some U_α, so that p is then in some preimage $f^{-1}(U_\alpha)$. Now, since K is compact, this open cover has a finite subcover, say:

$$K \subseteq f^{-1}(U_1) \cup \cdots \cup f^{-1}(U_n),$$

which implies that

$$f(K) \subseteq U_1 \cup \cdots \cup U_n.$$

Again, to be clear, anything in $f(K)$ is of the form $f(p)$ or some $p \in K$, and if this p is in $f^{-1}(U_i)$, $f(p)$ is in U_i. Thus $\{U_1, \ldots, U_n\}$ is a finite subcover of the open cover $\{U_\alpha\}$ of $f(K)$, so $f(K)$ is compact. □

Example. The final quiz of this quarter asked to show that if A and B are compact subsets of $\mathbb{R}$, then $A + B := \{a + b \mid a \in A \text{ and } b \in b\}$ was compact as well. Here is a "highbrow" way of approaching this using the theorem above.

The function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x, y) = x + y$ is continuous and $A \times B \subseteq \mathbb{R}^2$ is compact by a previous Warm-Up we did. Thus the image of $A \times B$ under f, which is precisely $A + B$, is compact as claimed.

Remark. We note here that compact and connected sets behave similarly under continuous functions, in that the image of each is of the same type. However, this does not generalize to other properties a set may have; in particular, it is NOT true that the image of an open set under a continuous function is open nor that the image of a closed set is closed. Rather, with open and closed sets, it is their *preimages* which are of the same type as the original set.

The moral is that compact and connected sets behave well in the "forward" direction with respect to continuity, but open and closed sets are only guaranteed to behave well in the "backward" direction. In addition, it is not true that the preimage of a connected set is connected nor that the preimage of a compact set is compact.

Extreme Value Theorem. Any continuous function $f : K \to \mathbb{R}$ on a compact domain K attains a maximum value and a minimum value. Thus this is a generalization of the Extreme Value Theorem from last quarter, which is the special case of this more general fact when $K = [a, b]$ is a closed interval in $\mathbb{R}$.

Proof. If K is compact and $f : K \to \mathbb{R}$ is continuous, then $f(K) \subseteq \mathbb{R}$ is compact by the previous theorem. Thus $f(K)$ is bounded and closed in $\mathbb{R}$. The fact that it is bounded implies that $\sup f(K)$ and $\inf f(K)$ both exist as finite real numbers, and the fact that $f(K)$ is closed implies that $\sup f(K)$ and $\inf f(K)$ are both in $f(K)$. Thus there exists $p \in K$ such that $f(p) = \sup f(K)$, so p gives the maximum value of f, and there exists $q \in K$ such that $f(q) = \inf f(K)$, so q gives the minimum value of f. □

Important. A continuous function sends compact sets to compact sets. As a consequence, any continuous function $K \to \mathbb{R}$ with compact domain achieves maximum and minimum values.

Example. Given a subset A of $\mathbb{R}^2$, we define the *diameter* of A to the supremum of all possible distances between points of A:

$$\operatorname{diam} A := \sup\{d(p,q) \mid p,q \in A\},$$

where d is the Euclidean metric. Visually this measures exactly what it sounds like, so that for a disk, for instance, the diameter is just the usual notion of diameter. (Of course, this definition works just as well for subsets of arbitrary metric spaces.)

We claim that if A is compact, then the diameter of A is actually attained as the distance between specific points, so that there exist $a, b \in A$ such that $d(a,b) = \operatorname{diam} A$. This is not true for arbitrary subsets A of $\mathbb{R}^2$, for instance an open disk. To prove this, consider the function $f : A \times A \to \mathbb{R}$ defined by $f(p,q) = d(p,q)$. Since $A \times A$ is compact (in $\mathbb{R}^4$), it suffices to show that this function is continuous, since then the Extreme Value Theorem implies that f has a maximum, which then gives the desired diameter.

To show that f is continuous (actually uniformly continuous), let $\epsilon > 0$. We will denote the taxicab metric on $\mathbb{R}^4$ by d_4 and the taxicab metric on $\mathbb{R}^2$ by d_2. Keeping in mind that $f = d_2$, we must show that there exists $\delta > 0$ such that

$$d_4((p_1,q_1),(p_2,q_2)) < \delta \implies |d_2(p_1,q_1) - d_2(p_2,q_2)| < \epsilon.$$

Here, we are thinking of a point (x,y,z,w) in $\mathbb{R}^4$ as consisting of a pair of points $p = (x,y)$ and $q = (z,w)$ in $\mathbb{R}^2$. Note that with this notation, we can express the taxicab metric on $\mathbb{R}^4$ in terms of the taxicab metric on $\mathbb{R}^2$ as:

$$d_4((p_1,q_1),(p_2,q_2)) = d_2(p_1,p_2) + d_2(q_1,q_2).$$

By the "reverse triangle inequality" we have

$$|d_2(p_1,q_1) - d_2(p_2,q_1)| \leq d_2(p_1,p_2)$$

and

$$|d_2(p_2,q_1) - d_2(p_2,q_2)| \leq d_2(q_1,q_2).$$

Thus for $\delta = \epsilon$, if $d_4((p_1,q_1),(p_2,q_2)) < \delta$ we get:

$$\begin{aligned} |d_2(p_1,q_1) - d_2(p_2,q_2)| &= |d_2(p_1,q_1) - d_2(p_2,q_1) + d_2(p_2,q_1) - d_2(p_2,q_2)| \\ &\leq |d_2(p_1,q_1) - d_2(p_2,q_1)| + |d_2(p_2,q_1) - d_2(p_2,q_2)| \\ &\leq d_2(p_1,p_2) + d_2(q_1,q_2) \\ &= d_4((p_1,q_1),(p_2,q_2)) < \delta = \epsilon. \end{aligned}$$

Hence $f = d_2$ is (uniformly) continuous, and thus as mentioned previously, since $A \times A$ is compact, $f : A \times A \to \mathbb{R}$ has a maximum value as desired.

Theorem. Finally, we come to the fact that any continuous function on a compact domain is uniformly continuous, which generalizes what saw for continuous functions $[a,b] \to \mathbb{R}$ last quarter. To be clear about notation, suppose that $f : X \to Y$ is continuous and that X is compact. Recall that to say f is uniformly continuous means that for any $\epsilon > 0$ there exists $\delta > 0$ such that

$$d_K(p,q) < \delta \implies d_Y(f(p),f(q)) < \epsilon.$$

We give two proofs, one using sequences and one using open covers.

Proof 1. We take the same proof in Chapter 3 of the book of the analogous fact about continuous functions $[a, b] \to \mathbb{R}$, and replace absolute value distances by our metrics. For a contradiction, suppose that f was not uniformly continuous. Then there exists $\epsilon > 0$ such that for any $\delta > 0$ we can find $p, q \in X$ such that

$$d_X(p, q) < \delta \text{ but } d_Y(f(p), f(q)) \geq \epsilon.$$

In particular, for each $n \in \mathbb{N}$ we can find $p_n, q_n \in X$ such that

$$d_X(p_n, q_n) < \frac{1}{n} \text{ and } d_Y(f(p_n), f(q_n)) \geq \epsilon.$$

Since X is compact, the sequence (p_n) in X has a convergent subsequence $p_{n_k} \to p \in X$, and then the sequence (q_{n_k}) in X has a convergent subsequence $q_{n_{k_\ell}} \to q \in X$. Since f is continuous we must then have

$$f(p_{n_{k_\ell}}) \to f(p) \text{ and } f(q_{n_{k_\ell}}) \to f(q).$$

However, we also have

$$d_X(p, q) \leq d_X(p, p_{n_{k_\ell}}) + d_X(p_{n_{k_\ell}}, q_{n_{k_\ell}}) + d_X(q_{n_{k_\ell}}, q).$$

Since $d_X(p_{n_{k_\ell}}, q_{n_{k_\ell}}) < \frac{1}{n_{k_\ell}}$ by the way in which p_n and q_n were chosen, each term on the right goes to 0 as $n_{k_\ell} \to \infty$, so $0 \leq d_X(p, q) \leq 0$. Thus we must have $p = q$ and so $f(p) = f(q)$ as well. But this a contradiction since we cannot have $(f(p_{n_{k_\ell}}))$ and $(f(q_{n_{k_\ell}}))$ converging to the same thing while also having $f(p_{n_{k_\ell}})$ and $f(q_{n_{k_\ell}})$ at a distance in Y of at least $\epsilon > 0$ apart from each other at all times. Thus f must have been uniformly continuous to begin with. □

Proof 2. The proof using open covers is given as part of the proof of Theorem 10.52 in the book. So, rather than repeat it all here, let me just comment on the thought process behind the proof given there. I'll use the book's notation, so that the metric on X is denoted by ρ and the metric on Y by τ. Also, we'll just take E in the book's notation to be the entire domain X, which is assumed to be compact.

Fix $\epsilon > 0$. By ordinary continuity, we know that for any $a \in X$ there exists $\delta(a) > 0$ (delta might depend on a) such that

$$\rho(x, a) < \delta(a) \implies \tau(f(x), f(a)) < \epsilon.$$

Here, a is fixed and x varies. Doing this for all $a \in X$ results in a corresponding $\delta(a)$ for each a, and then we consider the collection of open balls $\{B_{\delta(a)}(a)\}_{a \in X}$ in X given by these radii. This is an open cover of X since each $a \in X$ is in particular in the open ball $B_{\delta(a)}(a)$ centered at that point. Since X is compact, this open cover has a finite subcover

$$B_{\delta_1}(a_1), \ldots, B_{\delta_N}(a_N)$$

where δ_j denotes $\delta(a_j)$. Each element of X is in at least one of these open balls.

Now, we want to come up with a single $\delta > 0$ such that

$$\rho(x, y) < \delta \implies \tau(f(x), f(y)) < \epsilon.$$

The point of the finite cover obtained above is that now we have finitely many radii to deal with, so we can try to their minimum as the δ we need. However, we only know something about quantities of the form $\tau(f(x), f(a_j))$ where a_j is one of the centers of the finitely many open balls derived above. We would like to use something like

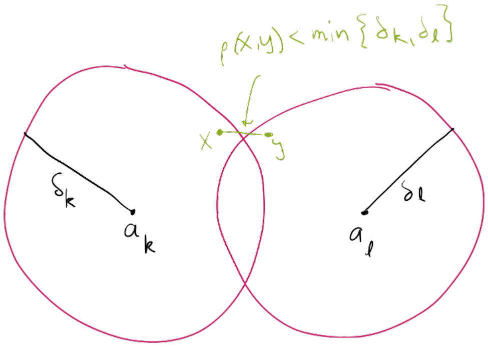
$$\tau(f(x), f(y)) \leq \tau(f(x), f(a_j)) + \tau(f(a_j), f(y)) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

so as a first fix we go back and replace the ϵ we used in coming up with the radii $\delta(a)$ by $\frac{\epsilon}{2}$, as the book does in its proof.

But now the problem is that, although we know that both x and y will each be in *some* open ball among the finitely many we obtained above, and so we will get *some* inequalities of the form

$$\tau(f(x), f(a_k)) < \frac{\epsilon}{2} \text{ and } \tau(f(a_\ell), f(y)) < \frac{\epsilon}{2},$$

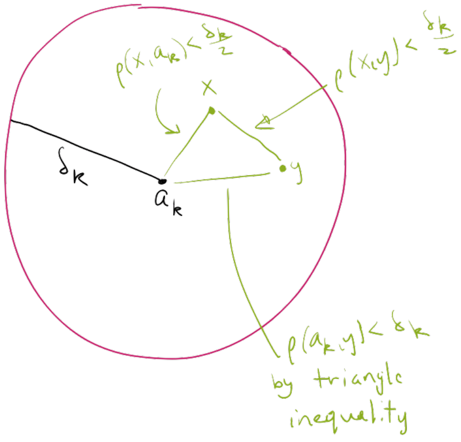
to get what we're doing to actually work we need the center points a_k and a_ℓ used here to be the *same*. Thus, we need to know that x and y are both in the *same* open ball among the finitely many obtained above. But when picking $\delta = \min\{\delta_1, \dots, \delta_N\}$, $\rho(x, y) < \delta$ does NOT guarantee that x and y will be in the same such ball, since we could have a picture like:



Thus this choice of δ is no good. But the fix is to go back and instead consider balls of radii $\frac{\delta(a)}{2}$, and after we get our finite subcover use $\delta = \min\{\frac{\delta_1}{2}, \dots, \frac{\delta_N}{2}\}$ instead as the book does. The point is that now having

$$\rho(x, y) \leq \min\left\{\frac{\delta_1}{2}, \dots, \frac{\delta_N}{2}\right\}$$

DOES guarantee that x and y will be in the same ball:



so that our approach will work out. The book's proof should be easier to digest now, and although it is still quite difficult, hopefully at least it's somewhat clearer now why the book uses $\frac{\epsilon}{2}$ and $\frac{\delta(a)}{2}$ instead of simply ϵ and $\delta(a)$ at the beginning. □

Important. A continuous function on a compact domain is automatically uniformly continuous.

Lecture 25: Contractions and Differential Equations

Our final day! Today we finished talking about metric spaces, closing out the quarter by talking about contractions and going through an application I mentioned back on the first day of class: using metric space material to show that differential equations have solutions. If nothing else, this should hopefully make it clear why considering abstract metric spaces can be useful.

Warm-Up. Given subsets A and B of $\mathbb{R}^2$, we define the *distance* between A and B as the infimum of all possible distances between a point of A and a point of B:

$$d(A,B) := \inf\{d(a,b) \mid a \in A \text{ and } b \in B\}.$$

(The same definition works for subsets of an arbitrary metric space.) Visually this gives the usual notion of distance you would expect between two subsets drawn in $\mathbb{R}^2$. We claim that if A and B are compact, this distance is realized by specific points, so that there exist $a \in A$ and $b \in B$ such that $d(a,b) = d(A,B)$, which again should make sense visually.

Consider the function $f : A \times B \to \mathbb{R}$ which sends a pair (a,b) to the distance $d(a,b)$ between a and b in $\mathbb{R}^2$. The same argument in the diameter example from last time shows that this function is continuous, so since $A \times B$ is compact, this function has a minimum by the Extreme Value Theorem; the pair (a,b) which gives this minimum then satisfies $d(a,b) = d(A,B)$ as desired.

Remark. Now that we have a sense of what it means to take the "distance" between compact subsets of $\mathbb{R}^2$, we can ask whether there is a sense in which this can be turned into a metric. Namely, if we let S denote the set of all compact subsets of $\mathbb{R}^2$:

$$S = \{K \subseteq \mathbb{R}^2 \mid K \text{ is compact}\},$$

is it possible to define some type of metric on this space? (Note that "points" of this space are actually themselves sets.) The "distance" between sets we defined above does not give a metric, since it is possible to have $d(A,B) = 0$ while $A \neq B$—indeed, any A and B which intersect will have $d(A,B) = 0$—and a metric requires that the distance between two "points" is zero if and only if those two "points" are the same.

But, it turns out that there are modifications we can make which will give an honest metric on the space of compact subsets of $\mathbb{R}^2$. (For this to work we actually do need to work with *compact* subsets as opposed to arbitrary subsets.) The resulting metric is called the *Hausdorff metric* on the space of compact subsets of $\mathbb{R}^2$. We won't give the definition here, but you can look it up on Wikipedia or elsewhere if interested. This turns out to be a very useful construction which measures the extent to which two compact sets are "close" to being the "same". In particular, you may have heard in recent years about the proof of something called the "Poincare conjecture", and although this is way beyond the scope of this course, the proof is essentially about studying properties of this Hausdorff metric, or rather a generalization known as the *Gromov-Hausdorff* distance. We live in exciting times.

Motivation from first day. Back on the first day of class this quarter I mentioned the following problem: how can we show that there exists a differentiable function f satisfying the differential equation

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

If had some simpler differential equation such as $f'(x) = f(x)$ we would know that any function of the form $f(x) = ce^x$ was a solution, and specifying an initial condition would single out what the value of the constant c should be. However, here we have no hope of writing down an explicit solution of this differential equation due to its complicated nature, so we need a more clever approach to show that a function satisfying the above conditions exists.

I claimed back then that after talking about metric spaces we would come back and show that there was in fact such a function, which we will now do. As a start, we note as we did then that we can rephrase this problem in the following way. After integrating both sides, we see that a function f satisfies

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

if and only if it satisfies

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

for some constant c. Indeed, if f is continuous, the second Fundamental Theorem of calculus implies that the integral expression is differentiable and that its derivative is the integrand evaluated at $t = x$; thus taking derivatives of both sides indeed gives 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

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

if and only if it satisfies the integral equation:

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

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 finally here is the amazing observation which makes everything work out: we can rephrase this integral equation as a fixed-point problem instead! Indeed, consider the metric space $C[1, k]$ of continuous functions $[1, k] \to \mathbb{R}$ for some to-be-determined constant $k > 1$ equipped with the sup metric, and define the map $T : C[1, k] \to C[1, k]$ by setting, for each $f \in C[1, k]$, Tf to be the function on $[1, k]$ whose value at $x \in [1, k]$ is:

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

Then, saying that f satisfies

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

is the same as saying that the function Tf equals f itself! Thus, showing that this integral equation has a solution for f is the same as showing that this map T has a fixed point!!!

And lo and behold, we had a problem on the final homework which dealt with fixed points, and is precisely what we need here in order to to finish of this problem. Thus, by rephrasing the problem of solving a differential equation as one about finding fixed points of a map from a metric space to itself, we can use all the metric space technology we have at our disposable to do something very concrete. Indeed, many modern approaches to solving differential equations are based on similar "fixed-point" ideas, and there are entire areas of mathematics devoted to this one topic.

Banach Contraction Principle. The metric space result we need to finish off our problem is the following, as given on the final homework: any contraction $f : X \to X$ from a complete metric space to itself has a unique fixed point, where by a *contraction* we mean a function f for which there exists $0 < C < 1$ such that

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

Intuitively, a contraction "shrinks" distances between points. Check the solutions to the final homework for a proof of this fact, which depends on properties of Cauchy sequences and continuous functions on general metric spaces.

This is actually a famous result, which usually goes by the name of *Banach Contraction Principle* or *Banach Fixed Point Theorem.* There is also a proof given in my *Notes on Metric Spaces.*

Example. Before returning to our differential equations problem, we give one more fun application of the Banach Contraction Principle, which answers the optional problem on the final homework. Take a map of Illinois and lay it out flat on a table. We show that there is a unique point on the map which lies directly over the physical location it is meant to represent.

Indeed, consider the function f from Illinois to itself which takes the entire state and shrinks and aligns it down onto the map. (Here we are considering Illinois as a subspace of the metric space given by the surface of the Earth with the usual distance between points on the surface of the Earth.) This map is clearly a contraction since after this shrinking process the resulting points on the map are way closer together than the two original points in Illinois were.

Since Illinois is a closed subset of the Earth and the Earth is complete (a sphere is a complete metric space), Illinois is complete as well. Thus the Banach Contraction Principle implies that this shrinking function has a unique fixed point, say $p \in$ Illinois. But $f(p) = p$ means that after shrinking Illinois down onto the map, the point p remains unchanged, meaning that p was already lying directly over the physical location it was meant to represent. Boo-yah!

Back to differential equations. Thus, since the space of continuous functions $C[1, k]$ is complete with respect to the sup metric d, we will be done with our differential equation problem if we can show that the map T is a contraction. And here is where the still to-be-determined constant k comes in: T will be a contraction on $C[1, k]$ for a well-chosen k, so that at the end we get a solution of our differential equation which will be defined on $[1, k]$.

For $f, g \in C[1, k]$ and $x \in [1, k]$, we have that $|(Tf)(x) - (Tg)(x)|$ equals

$$\left|\left(1 + \int_1^x [3|f(t)| - \log(e^{\sin(\cos t)} + 1)]\, dt\right) - \left(1 + \int_1^x [3|g(t)| - \log(e^{\sin(\cos t)} + 1)]\, dt\right)\right|,$$

which equals

$$\left|\int_1^x 3(|f(t)| - |g(t)|)\, dt\right|.$$

But using properties of integrals from last quarter and the reverse triangle inequality in the form $||a| - |b|| \leq |a - b|$, we have:

$$\begin{aligned}
|(Tf)(x) - (Tg)(x)| &= \left| \int_1^x 3(|f(t)| - |g(t)|)\,dt \right| \\
&\leq \int_1^x 3||f(t)| - |g(t)||\,dt \\
&\leq \int_1^x 3|f(t) - g(t)|\,dt.
\end{aligned}$$

But for each $t \in [1, k]$, $|f(t) - g(t)| \leq d(f, g)$ since $d(f, g)$ is the supremum of such expressions as t varies throughout $[1, k]$, and thus:

$$|(Tf)(x) - (Tg)(x)| \leq \int_1^x 3|f(t) - g(t)|\,dt \leq \int_1^x 3d(f, g)\,dt = 3(x - 1)d(f, g) \leq 3(k - 1)d(f, g),$$

where we use in the last step that the fact that $1 \leq x \leq k$.

Hence the number $3(k-1)d(f, g)$ is an upper bound for all expressions $|(Tf)(x) - (Tg)(x)|$ as x varies in $[1, k]$, so $3(k - 1)d(f, g)$ is larger than or equal to the supremum of such expresions, which is the sup distance $d(Tf, Tg)$:

$$d(Tf, Tg) \leq 3(k - 1)d(f, g).$$

Thus, as long as k satisfies $3(k - 1) < 1$, the map T as defined above is a contraction from $C[1, k]$ to itself. In particular, any $1 < k < \frac{4}{3}$ results in T being a contraction.

Working our way back to the beginning, for $1 < k < \frac{4}{3}$, the map $T : C[1, k] \to C[1, k]$ which sends $f \in C[1, k]$ to Tf defined by

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

is a contraction. (Note that the resulting function Tf is indeed continuous since it is defined as an integral of a continuous function, and such integrals are always continuous.) Since $C[1, k]$ is complete with respect to the sup metric, T has a unique fixed point, call it $f \in C[1, k]$. This f thus satisfies $Tf = f$, so

$$f(x) = 1 + \int_1^x [3|f(t)| - \log(e^{\sin(\cos t)} + 1)]\,dt \text{ for all } x \in [1, k].$$

This implies that $f(1) = 1$, and since the right side is differentiable by the second Fundamental Theorem of Calculus, f is as well and taking derivatives gives

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

so f satisfies the given differential equation with given initial condition. Tada!

Approximating the solution. As a final comment, the work above only shows that the given differential equation has a solution, but we can still ask whether it is possible to determine what the solution concretely is. Although we cannot determine the solution explicitly, it turns out we can in fact approximate the solution however well we want. The key is in the proof of the Banach Contraction Principle: the fixed point is obtained as the limit of the sequence defined by starting with any point and repeatedly applying contraction. Thus, in our case, if we start with any continuous function $g : [1, k] \to \mathbb{R}$ (say a constant one), the functions obtained by repeatedly applying T:

$$g,\ T(g),\ T(T(g)),\ \ldots,$$

will eventually provide better and better approximations of the fixed point of T, and hence of the solution of the differential equation we're interested in. This fact is at the core of modern approaches to solving differential equations numerically, say using a computer.

Moral. Metric space theory is awesome, and has some truly amazing applications.