Introduction to real analysis

Course notes written by Samuel Coskey;

Based on material from "Understanding Analysis",

written by Stephen Abbott

Contents

PART I

The theory of the real number line

### §1. What are rational and real numbers?

We begin the course by briefly pondering a very deep question: what are numbers and what are they for?

The natural or counting numbers $1, 2, 3, \ldots$ are collectively denoted $\mathbb{N}$ and are central to many areas of mathematical study. *Combinatorics* is the study of counting finite sets. For example: how many ways are there to write the number 10 as a sum of smaller natural numbers? *Number theory* is the study of the arithmetic operations on the natural numbers. For example: how many prime numbers are there less than 10^{10}?

But in *calculus*, analysis, and geometry, numbers are used for measurement. Using the board as a plane, how can we measure the distance between two points? To answer this we would first need to choose a unit, such as a randomly chosen straight stick. With this in hand, we can try count how many copies of the unit stick fit along the straight line segment joining the two points. But what if the number of unit sticks doesn't come out whole?

In this case we can try to use *ratios* to measure the distance. For example, suppose we are trying to measure the distance d, and three unit sticks fit perfectly inside two copies of d. Then we say that $d = 3/2$ units. Sometimes this telescoping process takes a long time, for example if five units fit perfectly inside four copies of d. The set of all ratios of natural numbers (and their negatives) is called the *rational numbers* and given the symbol $\mathbb{Q}$ (for quotient).

This leads to the *big question* of the day: are rational numbers enough to measure all distances? The answer turns out to be a big NO, and this fact surprised many early mathematicians (and still blows minds of new mathematicians today). Ratios are good enough for most practical purposes because in the real world we can accept approximations. But there are ideal distances (numbers) which are not rational. The simplest classical example is $d =$ the diagonal of a square. Notice that d has the property that $d^2 = 2$.
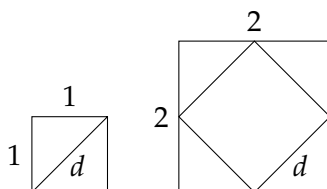
The argument is given in Plato's Meno. If we quadruplicate the square, we see that the square with sidelength d is exactly half the square with sidelength 2. Thus $d^2 = 2^2/2 = 2$.

1.1. THEOREM. *If $d^2 = 2$ then $d \notin \mathbb{Q}$.*

PROOF. Suppose d is rational and write $d = a/b$ where a, b are integers and the fraction is in lowest terms. Then $a^2 = 2b^2$. Notice that the right-hand side is even, and therefore so is the left-hand side. It follows that a is even. Hence the left-hand side is divisible by 4, and so is the right-hand side. It follows that b is even. Thus both a and b are even, contradicting that the fraction a/b is in lowest terms. $\square$

We conclude that the rational number system has *gaps*, such as the one at $\sqrt{2}$. All we can say at the moment is that some rational numbers are less than $\sqrt{2}$ and the rest are larger than $\sqrt{2}$. More accurately, since in our development the number $\sqrt{2}$ doesn't exist yet, we can only say that some numbers have square less than 2 and others have square greater than 2. This is an example of a property which *cuts* the rational number system into two halves.

1.2. DEFINITION. A subset $C \subset \mathbb{Q}$ is called a *cut* if it satisfies the following three properties.

   (a) (nontriviality) $C \neq \varnothing$ and $C \neq \mathbb{Q}$
   (b) (downward closure) if $q < r$ and $r \in C$ then $q \in C$
   (c) (no greatest element) if $q \in C$ then there is $r \in C$ such that $q < r$

For example, it is easy to check that the set $C = \{q \in \mathbb{Q} \mid q < 2\}$ satisfies all three properties of a cut. In fact, any set of the form $\{q \in \mathbb{Q} \mid q < r\}$ for some $r \in \mathbb{Q}$ is a cut. These cuts are somewhat obvious to define, and they are called *rational cuts*.

On the other hand, the cut at $\sqrt{2}$ considered previously would formally be defined as $C = \{q \in \mathbb{Q} \mid q < 0 \text{ or } q^2 < 2\}$. This is an example of an irrational cut.

1.3. DEFINITION. A cut C is said to be an *irrational cut* or a *gap* if the complement of the cut, $\mathbb{Q} \smallsetminus C$, has no least element.

In the next section, we will prove rigorously that the cut at $\sqrt{2}$ is really an example of an irrational cut. We will then describe a correspondence between cuts and real numbers, with rational cuts corresponding to rational numbers and irrational cuts corresponding to irrational numbers.

Activity for §1.

1. (see Abbott, exercise 1.4.1) Complete and prove:
    (a) The sum of two rational numbers is...

    (b) The product of two rational numbers is...

    (c) The sum of a rational number and an irrational number is...

    (d) The product of a rational number and an irrational number is... unless...

    (e) The product of two irrational numbers is...? (discuss)

2. (see Abbott, ex 8.6.1) For each set, either argue that it is a cut, or show that one of the conditions in the definition of cut is not satisfied.
    (a) $\{ q \in \mathbb{Q} \mid q < 2 \}$

    (b) $\{ q \in \mathbb{Q} \mid q \leq 2 \}$

    (c) $\{ q \in \mathbb{Q} \mid q^2 = 2 \}$

    (d) $\{ q \in \mathbb{Q} \mid q^2 < 2 \}$

    (e) $\{ q \in \mathbb{Q} \mid q < 0 \text{ or } q^2 < 2 \}$

READING FOR NEXT TIME. Abbott §8.6, start §1.3

§2. Cuts and least upper bounds

READING. Abbott §8.6, start §1.3

In the previous section we defined a cut in the rational number system, and stated that some cuts are "rational" while others are "irrational" (or "gaps"). However we have yet to actually prove that any cut is really irrational. According to our motivation, the cut at $\sqrt{2}$ should be a gap. The proof turns out to be more technical than one might have expected. It will provide us with a not-so-soft example of the kind of arguments we will see later in the class.

2.1. THEOREM. *The cut $C = \{\, q \in \mathbb{Q} \mid q < 0 \text{ or } q^2 < 2 \,\}$ is an irrational cut.*

PROOF. We begin by verifying that C has the three properties of a cut.
(a) (nontriviality) Note that $0 \in C$ and $5 \notin C$.
(b) (downward closure) Suppose that $r \geq 0$ and that $r \in C$, so that we have $r^2 < 2$. Let $q < r$ be given. Clearly if $q < 0$ we have $q \in C$. On the other hand if $q \geq 0$ then we have $q^2 < r^2 < 2$ so again $q \in C$.
(c) (no greatest element) Suppose that $q \geq 0$ and $q \in C$, so that $q^2 < 2$. We seek to find a small rational number $\epsilon > 0$ such that $(q + \epsilon)^2 < 2$ also.
To see what ϵ has to be, we work *backwards*.

$$(q + \epsilon)^2 < 2 \iff q^2 + 2q\epsilon + \epsilon^2 < 2$$
$$\impliedby q^2 + 2q\epsilon + \epsilon < 2$$
$$\iff \epsilon < (2 - q^2)/(2q + 1)$$

There is a clever trick going on in the second implication (the $\impliedby$ one) which we will comment on in a moment. But overall, this calculation is telling us that we will be safe if we choose $\epsilon < (2 - q^2)/(2q + 1)$.

Now, in the second implication we replaced ϵ^2 with ϵ to simplify things and make it easier to solve for ϵ. To do this, we actually need to make the extra assumption that $\epsilon < 1$, since then $\epsilon^2 < \epsilon$ and the $\impliedby$ holds true. This is acceptable because we haven't yet chosen ϵ, so we just need to additionally promise to choose $\epsilon < 1$.

We now know that our proof will succeed if we choose ϵ to be any positive rational number such that both $\epsilon < 1$ and also $\epsilon < (2 - q^2)/(2q + 1)$. To verify that this works, we use our preparatory work above *forwards* as follows.

$$(q + \epsilon)^2 = q^2 + 2q\epsilon + \epsilon^2$$
$$< q^2 + 2q\epsilon + \epsilon$$
$$= q^2 + (2q + 1)\epsilon$$
$$= q^2 + (2q + 1)(2 - q^2)/(2q + 1)$$
$$= 2$$

Thus we conclude that $r = q + \epsilon$ is a rational number such that $q < r$ and $r^2 < 2$, which completes the proof of (c).

Finally, to see that C is a gap, we note that the same argument as in part (c) above can be used to show that the set $\{\, q \in \mathbb{Q} \mid q > 0 \text{ and } q^2 > 2 \,\}$ has no least element. And since Theorem 1.1 implies that $\sqrt{2}$ is itself not rational, we can conclude that $\mathbb{Q} \smallsetminus C$ has no least element. $\square$

Let us look at the number systems we know so far, from smallest to largest. Each one has progressively more power than the last.

- $\mathbb{N}$: you can add but not subtract.
- $\mathbb{Z}$: you can add and subtract, but not divide.
- $\mathbb{Q}$: you can add, subtract, and divide, but there are "gaps".

The real number system $\mathbb{R}$ will be defined in such a way that it fills the gaps of $\mathbb{Q}$. In order to fill these gaps, we need the terminology of least upper bounds.

2.2. DEFINITION. If A is any set of numbers, then an *upper bound* for A is a number b such that for all $a \in A$, $a \le b$. If b is an upper bound for A then it is called the *least upper bound* or *supremum* of A if it is least among all possible upper bounds of A.

2.3. EXAMPLE.   ◦ The set $(-5, 5]$ has a maximum element 5. Therefore it has least upper bound 5.
- The set $(-5, 5)$ has no maximum element. But the set has upper bounds, for example 7, 23.6, and 100. The least possible upper bound is 5.
- The set $\{\, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \dots \,\}$ has no maximum element. But it again has upper bounds, and the least possible upper bound is 1.

Our definitions of rational and irrational cuts can now be rephrased as follows. A rational cut is one which has a supremum in $\mathbb{Q}$, and an irrational cut is one which does not have a supremum in $\mathbb{Q}$. We can therefore "fill" the gaps in $\mathbb{Q}$ by inserting a newly created supremum for every irrational cut.

2.4. DEFINITION. The *real number system* consists of the rational numbers $\mathbb{Q}$ together with a supremum for every cut of $\mathbb{Q}$ that doesn't already have a supremum (that is, for every irrational cut).

In summary, to form the real number system we first built the rational number system, observed that it has gaps, and then manually *filled in* all these gaps.

It is important to emphasize that with our definition of real numbers, an irrational number is literally *defined* by its cut. That is, we don't define $\sqrt{2}$ by saying it is a positive number whose square is 2. That would tell you what property the number has, but not where to actually find it. Instead we define $\sqrt{2}$ to be the supremum of the cut $C = \{\, q \in \mathbb{Q} \mid q < 0 \text{ or } q^2 < 2 \,\}$.

We close this section with a characterization of the supremum of a set that will be useful in the future.

2.5. LEMMA. *If A is a set, then α is the supremum of A if and only if the following two conditions are satisfied:*

- *(α is an upper bound) for all $a \in A$ we have $a \leq \alpha$, and*
- *(α is the least such) for all $\epsilon > 0$ there exists $a \in A$ such that $\alpha - \epsilon \leq a$.*

You can find the proof in Lemma 1.3.8 of the text.

Activity for §2.

1. (Abbott, ex 1.3.8) Find the supremum and infemum of each of the following sets. You do not need to give a detailed proof, but please explain your reasoning.

   (a) $\{\, n \in \mathbb{N} \mid n^2 < 10 \,\}$

   (b) $\{\, n/(m+n) \mid m, n \in \mathbb{N} \,\}$

   (c) $\{\, n/(2n+1) \mid n \in \mathbb{N} \,\}$

   (d) $\{\, n/m \mid m, n \in \mathbb{N} \text{ and } m + n \le 10 \,\}$

2. Suppose that A and B are subsets of $\mathbb{R}$ and that $A \subset B$. Which of the following is correct, and why? "any upper bound for A is also an upper bound for B", or; "any upper bound of B is also an upper bound for A."

   Now additionally suppose that $\sup(A)$ and $\sup(B)$ exist. Which inequality is right, and why? $\sup(A) \le \sup(B)$, or; $\sup(B) \le \sup(A)$.

Reading for next time. Abbott §1.3, 1.4

§3. The Axiom of Completeness

READING. Abbott, §1.3, 1.4

In the previous section we showed how one can construct the real number system; in this section we take a high-level approach and show how to define the real number system by its properties alone.

Although we introduced the real numbers as a way to measure distances, we of course know that the real numbers should carry arithmetic operations, too. In fact, it should be an *ordered field*, which means we should be able to add, subtract, multiply, divide, and compare numbers, and these operations should satisfy the usual axioms of associativity, commutativity, and so forth. The rational number system has all these properties, and the real number system does too. The only thing we haven't done is actually *define* the concepts of plus, times, and comparison for real numbers. Then in principle we have to verify that the usual axioms are actually true.

For example, if A and B are cuts then we define

$$A + B = \{\, a + b \mid a \in A \text{ and } b \in B \,\}$$

One should then verify that $A + B$ is a cut, that $A + B = B + A$, that $(A + B) + C = A + (B + C)$, and so forth. Similarly we define

$$A < B \iff A \subset B \text{ and } A \neq B$$

One should then verify that $A < B < C \implies A < C$, and $A < B \implies A + C < B + C$, and so forth. We will skip the remaining details of these tedious definitions and verifications.

Since we constructed $\mathbb{R}$ by filling in gaps in the rational number system $\mathbb{Q}$, it is natural to ask whether $\mathbb{R}$ itself has any gaps. In other words, could it be that by filling in gaps we introduced new ones? Thankfully the answer to this question is no, and $\mathbb{R}$ does not have any gaps.

3.1. DEFINITION. An ordered field is said to satisfy the *Axiom of Completeness* if every bounded subset has a least upper bound, that is, a supremum.

For example, $\mathbb{Q}$ is an ordered field that is not complete, since we checked that the cut corresponding to $\sqrt{2}$ has no rational supremum. On the other hand the real field $\mathbb{R}$ does turn out to be complete. The results of this discussion are summarized in the following big theorem.

3.2. THEOREM. *The real number system $\mathbb{R}$ is an ordered field satisfying the Axiom of Completeness. Moreover, $\mathbb{R}$ is the unique such field up to isomorphism.*

As soon as one accepts this theorem, it is actually never necessary to speak of cuts again. In the future we can describe a real number in many ways: algebraically, as a limit, as a summation, or of course by its decimal expansion.

Although we are omitting the greater part of the proof of the above theorem, let us verify the Axiom of Completeness holds in the real number system.

PROOF OF COMPLETENESS. Suppose that S is a bounded set of real numbers (cuts). We claim that the union $U = \bigcup S$ is a cut too. Indeed, since S is bounded, there is a cut B such that $U \subset B$. It follows that U is nontrivial. Next it is easy to check that the union of downward closed sets is again downward closed. Thirdly if U had a greatest element x, then x must lie in some cut $A \in S$. But then x would be the greatest element of A, a contradiction.

Now, since $A \subset U$ for every $A \in S$, we see that U is an upper bound for S. Moreover U is the least upper bound for S. Indeed, any other upper bound U' satisfies $A \subset U'$ for every $A \in S$, and therefore $U = \bigcup S \subset U'$. In other words, $U \leq U'$, as desired. $\square$

From this point on, we will never need cuts again, and the real number system $\mathbb{R}$ is simply the unique ordered field extending the rational number system and satisfying the Axiom of Completeness. From this abstract definition we can prove all the facts we will need about the real number system. We begin with the following.

3.3. THEOREM (Archimedian property). *For any real number x, there exists a natural number n such that $x < n$. That is, while there are real numbers between integers, there aren't any real numbers beyond the integers to the right.*

PROOF. Suppose to the contrary that x is a real number such that for all $n \in \mathbb{N}$ we have $n \leq x$. This says x is an upper bound for $\mathbb{N}$, in particular it would mean $\mathbb{N}$ is bounded. By the Axiom of Completeness, $\mathbb{N}$ should have a least upper bound, call it α.

Now, since α is the *least* upper bound, $\alpha - 1$ is not an upper bound for $\mathbb{N}$. Hence there exists $n \in \mathbb{N}$ such that $\alpha - 1 < n$. It follows that $\alpha < n + 1$, contradicting that α is an upper bound for all natural numbers. $\square$
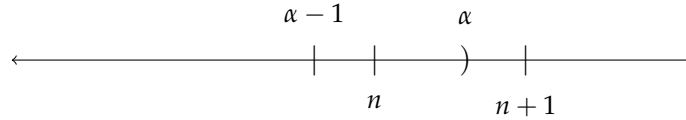
As a consequence of this simple fact, we can now formally establish one fact we have previously stated without proof. While the rational number system is not sufficient to measure any distance, it is sufficient to *approximate* any distance. In the following statement, we let $x < y$ and imagine that they are very close to one another.

3.4. THEOREM (Rational density). *For any real numbers x, y, if $x < y$ then there exists $q \in \mathbb{Q}$ such that $x < q < y$.*

PROOF. Using the Archimedean principle, we can find $n \in \mathbb{N}$ such that $n > 1/(y-x)$. Intuitively, this guarantees that increments of size $1/n$ cannot step over the interval (x,y).

Now let $m \in \mathbb{N}$ be the smallest natural number such that $x < m/n$ (again use the Archimedean principle). Note that by this minimality we have $m/n - 1/n \leq x$. Putting

these equations together, we have

$$x < m/n$$
$$\leq x + 1/n$$
$$< x + (y - x) = y$$

Thus the rational number $q = m/n$ lies in the interval (x, y).   □

To see that this implies we can approximate real numbers by rational numbers as closely as we like, let x be a given real number and apply the rational density theorem to intervals of the form $(x - \epsilon, x + \epsilon)$ where ϵ is small.

Activity for §3.

1. (Abbott, exercise 1.4.5) Prove the statement known as *irrational density*: for every two real numbers a and b with $a < b$, there exists an *irrational* number r satisfying $a < r < b$.

2. Prove or give a counterexample: If A and B are sets of real numbers such that $a < b$ for all $a \in A$ and $b \in B$, then $\sup(A) < \inf(B)$.

READING FOR NEXT TIME. Abbott, §1.5

§4. Completeness and the size of the reals

READING. Abbott, §1.5

In this section we will continue with consequences of the completeness axiom. In particular we will establish one of the most surprising and historically disruptive consequences: the fact that there are more irrational numbers than there are rational numbers! In order to make this statement precise, we need to introduce the notion of countable and uncountable sets.

4.1. DEFINITION. A set A is *countable* if it can be enumerated in a sequence a_n, whose items are numbered by natural numbers n. Intuitively, A is countable if all its elements can be ordered up in an infinite shopping list:

> 1. _____
> 2. _____
> 3. _____
> $\vdots$

4.2. EXAMPLE. The set of natural numbers $\mathbb{N}$ is countable, as is any subset of $\mathbb{N}$.

4.3. EXAMPLE. The set of integers $\mathbb{Z}$ is countable. To see this, one can simply alternate between positive and negative integers: $0, +1, -1, +2, -2, +3, -3, \ldots$.

4.4. EXAMPLE. The set of rational numbers $\mathbb{Q}$ is countable. To see this, let us first show that the set of positive rational numbers $\mathbb{Q}_+$ is countable. Although the numerator and denominator make it seem like there are two "dimensions" of information, we can get around this by fixing the sum of the two and then letting it increase: begin with $\frac{1}{1}$, then do $\frac{1}{2}, \frac{2}{1}$, then $\frac{1}{3}, \frac{2}{2}, \frac{3}{1}$, then $\frac{1}{4}, \frac{2}{3}, \frac{3}{2}, \frac{4}{1}$, and so forth. The list does have duplicates such as $\frac{1}{1}$ and $\frac{2}{2}$, but this is allowed, and in any case we could run through and delete them if we want. Finally we can include the negatives as well by applying the same method as we did with $\mathbb{Z}$.

The following result can be handy for showing sets are countable.

4.5. LEMMA. ○ *If A, B are countable, then the union $A \cup B$ is countable.*
   ○ *If $A_1, A_2, \ldots$ is a sequence of countable sets, then $\bigcup A_n$ is countable.*

PROOF. For the first statement, since A is countable we can enumerate its elements $a_1, a_2, a_3, \ldots$ and since B is countable we can enumerate its elements $b_1, b_2, b_3, \ldots$. It follows that we can enumerate the elements of $A \cup B$ in the sequence $a_1, b_1, a_2, b_2, a_3, b_3 \ldots$. (Formally we are describing c_n where $c_{2n} = b_n$ and $c_{2n-1} = a_n$.) Thus $A \cup B$ is countable.

We leave the proof of the second statement as an exercise. □

To show how this lemma can come in handy, we can use it recast the proof that $\mathbb{Z}$ is countable. It is obvious that the set $\mathbb{Z}_+$ of nonnegative integers is uncountable, and ditto

for the negative integers $\mathbb{Z}_-$. Since $\mathbb{Z}$ is the union of the two sets, it follows that $\mathbb{Z}$ is countable.

We now come to the "surprising" fact hinted at above.

4.6. THEOREM. *The set of real numbers $\mathbb{R}$ is **not** countable.*

In order to prove this theorem, we need the following consequence of the Axiom of Completeness for the real numbers called the *nested interval property*. This is a fact which we'll also use a few more times throughout the course.

4.7. THEOREM (Nested interval property). *Suppose that I_n is a sequence of closed and bounded intervals, that is, intervals of the form $[a, b]$. Moreover suppose that the intervals I_n are nested, that is, $I_{n+1} \subset I_n$. Then we have $\bigcap I_n \neq \emptyset$.*

PROOF. For naming purposes, let a_n be the left endpoint and b_n be the right endpoint of the interval I_n. Notice that the set of left endpoints $\{\, a_n \mid n \in \mathbb{N} \,\}$ is bounded above, with any of the b_n being such an upper bound. By the Axiom of Completeness, the set $\{\, a_n \mid n \in \mathbb{N} \,\}$ has a least upper bound, call it α.

Then given any n, we have $a_n \leq \alpha$, since α is an upper bound for the set $\{\, a_n \mid n \in \mathbb{N} \,\}$. On the other hand, we also have $\alpha \leq b_n$, since α is the *least* upper bound for the set $\{\, a_n \mid n \in \mathbb{N} \,\}$. It follows that α lies in the interval $I_n = [a_n, b_n]$. Since n was arbitrary, we conclude that α lies in $\bigcap I_n$. This completes the proof. □

We are now ready to prove that the set of real numbers is uncountable.

PROOF OF THEOREM 4.6. Suppose towards a contradiction that $\mathbb{R}$ is countable. Then its elements can be enumerated in a sequence x_n. Now inductively construct a nested sequence of closed, bounded intervals as follows. Let I_1 be any closed, bounded interval not containing x_1. Let I_2 be any closed subinterval of I_1 not containing x_2. Inductively, let $I_n \subset I_{n-1}$ be any closed interval not containing x_n. This completes the construction.

Now by the NIP, there exists an element $\alpha \in \mathbb{R}$ such that for all n, we have $\alpha \in I_n$. Since I_n does not contain x_n and $\alpha \in I_n$, we have that $\alpha \neq x_n$. Since this is true for all n, the existence of α contradicts the claim that every element of $\mathbb{R}$ was enumerated in the sequence x_n! □

4.8. COROLLARY. *The set of irrational numbers is uncountable.*

PROOF. Suppose towards a contradiction that the set of irrational numbers is countable. Then $\mathbb{R} = \mathbb{Q} \cup (\mathbb{R} \smallsetminus \mathbb{Q})$ is countable, being the union of two countable sets. This contradicts Theorem 4.6. □

Activity for §4.

1. Let A be the set of integer lattice points in the plane, that is, $A = \{ (x,y) : x, y \in \mathbb{Z} \}$. Draw a picture of the set A. Show that A is countable by explaining how to enumerate all its elements in a list.

2. Find a nested family of nonempty bounded intervals with empty intersection.

3. Find a nested family of nonempty closed intervals with empty intersection.

READING FOR NEXT TIME. Abbott, §1.6

§5. More on countability and uncountability

READING. Abbott, §1.6

In the previous section we saw that among the infinite sets, some of them are not too big (countable sets) while others are much larger (uncountable sets). Showing that a set is countable requires a direct argument—write down how to enumerate all its elements in a sequence. Showing that a set is uncountable often requires an indirect argument—assume the set is countable and find a contradiction.

We were able to show that $\mathbb{R}$ is uncountable using the Nested Interval Property. But that proof was very particular to the set of real numbers. How would we show that some other given set is uncountable? It turns out there is another and more systematic way to prove that $\mathbb{R}$ is uncountable, due the mathematician Cantor. The method is called *diagonalization*, and uses the decimal digit system.

5.1. THEOREM. *The unit interval $(0,1)$ is uncountable.*

PROOF. Suppose towards a contradiction that $(0,1)$ is countable. Then the elements of $(0,1)$ can all be enumerated in a sequence a_n. We need to analyze the digits of each entry of this sequence, so let us fix some notation for them:

$$a_1 = 0.a_{11}\, a_{12}\, a_{13} \cdots$$
$$a_2 = 0.a_{21}\, a_{22}\, a_{33} \cdots$$
$$a_3 = 0.a_{31}\, a_{32}\, a_{33} \cdots$$
$$\vdots$$

We now define a "diagonal" real number $d \in (0,1)$ with a decimal expansion d_n with the property that for all n

$$d_n \neq a_{nn}$$

We can even give an explicit definition of the digits d_n that achieves this property. For instance we may let $d_n = 2$ when $a_{nn} \neq 2$ and let $d_n = 3$ when $a_{nn} = 2$. It then follows that for all n, we have $d \neq a_n$. Thus d is an element of $(0,1)$ which does not appear in the sequence a_n, contradicting our assumption that every element of $(0,1)$ appears in the sequence a_n. $\qquad \square$

In the exercises you will be asked to use diagonalization on other kinds of sets.

The dichotomy between countable and uncountable sets suggests that one should further investigate the possibilities for the size of an infinite set. Recall that for finite sets we write $|A|$ for the number of elements of A. However for infinite sets this definition of $|A|$ isn't a good one anymore. What is needed is the following.

5.2. DEFINITION. Suppose that A, B are sets. Then we introduce the following notation:

○ We write $|A| \leq |B|$ if there is a one-to-one function $A \to B$

○ We write $|A| = |B|$ if there exists a one-to-one, onto function $A \to B$
○ We write $|A| < |B|$ if we have $|A| \leq |B|$ and $|A| \neq |B|$.

This definiton is consistent with the "number of elements" definition of $|A|$ when A is finite. However, while the terms "number of elements", "more", and "fewer" all lose their meaning for infinite sets, the notions of one-to-one and onto functions remain meaningful.

We call $|A|$ the *cardinality* of A. It is important to understand that we never defined what $|A|$ actually is, only how it compares with other set cardinalities $|B|$. Using set theory we can also define what $|A|$ actually is, but it isn't necessary for this class.

Cantor's theorem can be generalized to show that for any set there is a set of even larger cardinality. Therefore there is no bound on the number of different sizes of infinity! Recall that if A is any set, the *power set* $\mathcal{P}(A)$ is the set of all subsets of A. Of course if A is finite, then we have $|\mathcal{P}(A)| = 2^{|A|}$ so in particular $|A| < |\mathcal{P}(A)|$. It turns out that this works for infinite sets too.

5.3. THEOREM. *If A is any set, then $|A| < |\mathcal{P}(A)|$.*

PROOF. To show that $|A| \leq |\mathcal{P}(A)|$ we must exhibit a one-to-one function $A \to \mathcal{P}(A)$. For this, we can use the function $i(a) = \{a\}$.

To show that $|A| \neq |\mathcal{P}(A)|$, we have to show that there is no one-to-one, onto function $f \colon A \to \mathcal{P}(A)$. For this, suppose that $f \colon A \to \mathcal{P}(A)$ is any function. Then we consider the "diagonal" set D:

$$D = \{\, a \in A \mid a \notin f(a) \,\}$$

We claim that D is not in the range of f, and therefore that f is not onto. To see this, suppose towards a contradiction that there exists $a_0 \in A$ such that $D = f(a_0)$. Then by the definition of D, we have that $a_0 \in D$ iff $a_0 \notin f(a_0)$. And since $D = f(a_0)$ we can write this as $a_0 \in D$ iff $a_0 \notin D$, which is a clear contradiction. □

It is natural to ask to what extent cardinalities behave like ordinary finite counting numbers. For instance, is it possible to have two sets which are *incomparable* in size, that is $|A| \nleq |B|$ and $|B| \nleq |A|$? If we assume all of the standard axioms about sets (including the axiom of choice) then this situation is impossible.

5.4. THEOREM. *If A, B are any sets, then exactly one of the following is true: $|A| < |B|$, $|A| = |B|$, or $|B| < |A|$.*

Another natural question is whether the cardinality of $\mathbb{R}$ is the *smallest* uncountable cardinality. Many very narrow or thin subsets of $\mathbb{R}$ still have the full cardinality of $\mathbb{R}$. Just consider $(0, 1)$ or $(0, \epsilon)$ or the set of irrationals. So the question is whether for any set A of real numbers, if A is uncountable then is it necessarily true that $|A| = |\mathbb{R}|$?

This question is known as the *continuum hypothesis* and has great historical significance. After remaining open for many decades, it was finally settled in a surprising way: the answer is *undecidable* using the accepted axioms of mathematics! In other words, there

is a universe of mathematics in which the continuum hypothesis is true, and also a universe of mathematics in which it is false.

Since this stunning resolution to the continuum hypothesis was found in 1965, the occurrence and nature of undecidable statements has remained of key interest to mathematicians.

Activity for §5. In the following, let S_∞ be the set of all *infinite* words made with the letters a, b. So S_∞ contains elements like $aaaa \cdots$, $ababa \cdots$, $aabaabaab \cdots$, also non-repeating strings like $abaabaaabaaaab \cdots$, and even strings with no discernible pattern whatsoever.

1. Write any ten elements of S_∞. Include at least ten letters in each element (with the understanding that the words continue indefinitely, either according to a pattern or arbitrarily).

$$s_1 = \underline{\hspace{7cm}} \cdots$$

$$s_2 = \underline{\hspace{7cm}} \cdots$$

$$s_3 = \underline{\hspace{7cm}} \cdots$$

$$s_4 = \underline{\hspace{7cm}} \cdots$$

$$s_5 = \underline{\hspace{7cm}} \cdots$$

$$s_6 = \underline{\hspace{7cm}} \cdots$$

$$s_7 = \underline{\hspace{7cm}} \cdots$$

$$s_8 = \underline{\hspace{7cm}} \cdots$$

$$s_9 = \underline{\hspace{7cm}} \cdots$$

$$s_{10} = \underline{\hspace{7cm}} \cdots$$

2. Write the first ten letters of the *diagonal* word d with the property that for all n, d disagrees with s_n on the nth letter.

3. Write the definition of d in the abstract (not just in this example).

$$d = \begin{cases} \underline{\hspace{1.5cm}} & \text{if } s_{nn} = \underline{\hspace{1.5cm}} \\ \underline{\hspace{1.5cm}} & \text{if } s_{nn} = \underline{\hspace{1.5cm}} \end{cases}$$

4. Write the full proof that S_∞ is uncountable.

READING FOR NEXT TIME. Abbott, §2.1

Sequences and series

§6. Infinity and arithmetic

READING. Abbott, §2.1

A sequence of real numbers consists of one real number a_n for each natural number n. Formally we can think of a sequence as a *function* from $\mathbb{N}$ to $\mathbb{R}$. However sequences are so special that we rarely use the functional notation $a(n)$ and instead prefer the special notation a_n. We have already used sequences, for instance in the definition of a countable set. We will see later in the course that sequences play a huge role in analysis. In this section we give further motivation for studying sequences and the sibling concept of series.

To begin consider the simple addition problem $\pi + e$, which we write out in decimal notation as

$$3.14159265\cdots$$
$$+2.71828182\cdots$$

It is unclear how to begin this problem. Our usual method of carrying would have us start at the right-most digit, but there isn't one! It is possible to handle this issue by regarding π and e as cuts. But it is even more elegant to regard each of π and e as the limit of a sequence, and perform the addition term-by-term.

Another problem with addition is the prospect of adding together infinitely many numbers at once. Sometimes this clearly isn't possible; for example the sum

$$1 + 1 + 1 + 1 + \cdots$$

is infinite and has no value in the real numbers. On the other hand the familiar Zeno paradox asks whether the sum

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots$$

has any value in the real numbers, and we usually agree that has a value of 2. It seems that sometimes it makes sense to add together infinitely many numbers, and other times it doesn't.

But this conclusion deserves a great deal of scrutiny. To see why, consider the follow-
ing simple proof that the Zeno series has a value of 2. We first notice that

$$\left(1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots\right) - 1 = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots$$
$$= \frac{1}{2}\left(1 + \frac{1}{2} + \frac{1}{4} + \cdots\right)$$

Thus we see that if $x = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots$, then x satisfies the equation $x - 1 = \frac{1}{2}x$ and
therefore $x = 2$.

Now let us apply this proof to the series $1 + 2 + 4 + 8 + \cdots$. Again write

$$(1 + 2 + 4 + 8 + \cdots) - 1 = 2 + 4 + 8 + \cdots$$
$$= 2(1 + 2 + 4 + \cdots)$$

This time we conclude that if $x = 1 + 2 + 4 + 8 + \cdots$ then x satisfies the equation $x - 1 =
2x$ and therefore that $x = -1$, which is nonsense!

Finally, even if we do accept some infinite sums as valid, we still should decide whether
the familiar laws of associativity and commutativity of addition are applicable to the
terms of infinite sums. To see the potential difficulty, consider two ways of grouping
the terms of the same infinite summation:

$$S_1 = (1 - 1) + (1 - 1) + (1 - 1) + \cdots$$
$$S_2 = 1 + (-1 + 1) + (-1 + 1) + \cdots$$

Using the first grouping, we obtain a value of 0. Using the second grouping, we obtain a
value of 1.

Commutativity has similar issues. For example, consider two ways of summing the
same series:

$$S_1 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + - \cdots$$
$$S_2 = 1 + \frac{1}{3} - \frac{1}{2} + \frac{1}{5} + \frac{1}{7} - \frac{1}{4} + + - \cdots$$

In the second summation we first add two of the positive terms before then adding one of
the negative terms. A quick approximation with a calculator reveals that $S_1 \approx 0.69$ while
$S_2 \approx 1.03$. This time we conclude that even if infinite summations are possible, one must
be very careful when applying the usual laws of arithmetic to them!

Activity for §6.

1. Consider the following infinite summation:

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

- ○ Try to reorder the terms to make the sum as close as possible to 0.5. Use at least 15 terms to be sure you are actually close.
- ○ Try to reorder the terms to make the sum as close as possible to 1.
- ○ Try to reorder the terms to make the sum as close as possible to 1.5.

Were you (approximately) successful in each case? If so, write down your ordering of the terms. What method did you use to get it? If not, what barrier did you encounter?

2. This time consider the following infinite summation:

$$1 - \frac{1}{2} + \frac{1}{4} - \frac{1}{8} + \frac{1}{16} - + \cdots$$

- ○ Try to reorder the terms to make it come as close as possible to 0.5. Use at least 12 terms to be sure you are actually close.
- ○ Try to reorder the terms to make it come as close as possible to 1.
- ○ Try to reorder the terms to make it come as close as possible to 1.5.

Were you (approximately) successful in each case? If so, write down your ordering of the terms. What method did you use to get it? If not, what barrier did you encounter?

READING FOR NEXT TIME. Abbott, §2.2

§7. Convergence

READING. Abbott, §2.2

Recall that a *sequence* is a list of one real number a_n for every natural number n. The most important characteristic of a sequence is whether it converges or diverges. Intuitively a sequence a_n converges to a limit point L if the sequence eventually comes as close to L as one could ever wish.

7.1. DEFINITION. The sequence a_n *converges* to L if the following holds:

$$(\forall \epsilon > 0)\ (\exists N \in \mathbb{N})\ (\forall n \geq N)\ |a_n - L| < \epsilon$$

In this case we write both $\lim a_n = L$ and also $a_n \to L$, and we say that L is the *limit* of the sequence.

Since the definition is both of central importance and also somewhat subtle, we will analyze it in great detail.

- The $(\forall \epsilon > 0)$ clause. Intuitively we think of ϵ as a very small *tolerance*. Thus the definition of convergence begins by saying that for all possible tolerances, something will happen.
- The $(\exists N \in \mathbb{N})\ (\forall n \geq N)$ clause. Intuitively this means that *eventually* something will happen. This is because it says that there is some index such that for all later indices, that thing will be true.
- The $|a_n - L| < \epsilon$ clause. Intuitively we think of this as saying that a_n lies within a small *neighborhood* of L. Formally, if x is any real number and $\epsilon > 0$, then the *ϵ-neighborhood* of x, denoted $N_\epsilon(x)$, is the open interval centered at x of radius ϵ. In symbols:

$$N_\epsilon(x) = (x - \epsilon, x + \epsilon) = \{\, y \in \mathbb{R} \mid |y - x| < \epsilon \,\}$$

  Thus the conclusion of the definition of convergence says precisely that a_n lies within the ϵ-neighborhood of L.

To summarize, the statement that a_n converges to L says

"for all tolerances ϵ, the sequence eventually lies within the ϵ-neighborhood of L."

So how does one show that a given sequence converges to a specified limit? One way to think of verifying the three-quantifier definition holds true is to play a three-step game. Your enemy takes the first turn and gives you some (small) value of ϵ. You respond by calculating a (large enough) natural number N. Your enemy then responds by choosing a number $n \geq N$. Finally, you win if you can show that $L - \epsilon < a_n < L + \epsilon$.

7.2. EXAMPLE. Let us confirm from the definition that $\lim 1/\sqrt{n} = 0$. Suppose that ϵ is given to us. We must find a value N (depending on this ϵ) with the property that

whenever $n \geq N$ we have $-\epsilon < 1/\sqrt{n} < \epsilon$. Since $1/\sqrt{n}$ is always positive, we only have to worry about the second inequality.

To get an idea of how this is going to work, if $\epsilon = .1$ then we would need to take $N = 101$. Similarly if $\epsilon = .01$ then we would need to take $N = 10,001$. In general we can figure out the needed relationship by working backwards. We want

$$1/\sqrt{n} < \epsilon \iff \sqrt{n} > 1/\epsilon$$
$$\iff n > 1/\epsilon^2$$

This tells us we should choose $N > 1/\epsilon^2$. Technically N has to be a natural number, but we can just take N to be the first natural number $N > 1/\epsilon^2$.

Finally we can write the proof forwards as follows. Given any $\epsilon > 0$, we let N be any natural number such that $N > 1/\epsilon^2$. Then for all $n \geq N$ we have

$$n > 1/\epsilon^2 \implies \sqrt{n} > 1/\epsilon$$
$$\implies 1/\sqrt{n} < \epsilon$$

Since we always have $-\epsilon < 1/\sqrt{n}$, the proof is complete. $\qquad\square$

7.3. EXAMPLE. Let us confirm from the definition that $n/(n+1) \to 1$. So let ϵ be given; we wish to conclude that eventually $1 - \epsilon < n/(n+1) < 1 + \epsilon$. Since $n/(n+1)$ is always less than 1, we only have to worry about the first inequality. To see the needed value of N we work backwards:

$$1 - \epsilon < n/(n+1) \iff 1 - n/(n+1) < \epsilon$$
$$\iff 1/(n+1) < \epsilon$$
$$\iff n+1 > 1/\epsilon$$
$$\iff n > 1/\epsilon - 1$$

Thus the needed value of N is $1/\epsilon - 1$. To simplify the calculation, it will be good enough to take $N \geq 1/\epsilon$.

Now to write the proof forwards, let ϵ be given and let N be a natural number such that $N \geq 1/\epsilon$. Then for all $n > N$ we have

$$1 - n/(n+1) = 1/(n+1)$$
$$< 1/n$$
$$< \epsilon$$

Thus $1 - \epsilon < n/(n+1)$, as desired. $\qquad\square$

We should be careful to point out that we are not just giving a careful definition of convergence so that we can verify routine instances of convergence. The real reason is the converse: we verify the routine instances of convergence so that we can be sure our definition agrees with our intuition.

So far we have discussed only convergence of sequences. What happens when a sequence a_n does *not* converge to L? Formally negating the definition of convergence, this says

$$(\exists \epsilon > 0) \ (\forall N \in \mathbb{N}) \ (\exists n \geq N) \ |a_n - L| \geq \epsilon$$

Intuitively this says that there is some tolerance ϵ such that the sequence frequently lies outside the ϵ-neighborhood of L.

7.4. EXAMPLE. Consider the following sequence with the terms, and discuss whether it converges to 0 or not:

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

If ϵ has a given value of .5 it is possible to find a suitable response N which satisfies the definition of convergence. The same is true if $\epsilon = .25$. But if $\epsilon = .2$ then it is *not* possible, so we conclude that the sequence does not converge to 0

In general, we say that a sequence *diverges* if for every L, the sequence does not converge to L.

Activity for §7.

1. In each case, prove the given sequence converges to the proposed limit directly from the definition of convergence.

    (a) $\lim \dfrac{1}{6n^2 + 1} = 0$

    (b) $\lim \dfrac{3n - 1}{2n + 5} = \dfrac{3}{2}$

    (c) $\lim \dfrac{2}{\sqrt{n + 3}} = 0$

    (d) $\lim \dfrac{\sin(n)}{n} = 0$

    (e) $\lim 2^{1/n} = 1$

2. Prove the given sequence *does not* converge to the proposed limit directly from the *negation* of the definition of convergence.

    Sequence: $0.1, 0.01, 0.1, 0.01, 0.1, 0.01, \dots$. Proposed limit: 0.

READING FOR NEXT TIME. Abbott, §2.3

§8. Limit calculus

READING. Abbott, §2.3

In the previous section we gave a rigorous definition of the limit of a sequence, and checked that it corresponds with our intuition for several special examples. In this section we begin our study of *calculus*, that is, the rules for calculating with operators such as limits. Of course the core of calculus consists of the formulas and laws for derivatives and integrals. But limits obey laws as well, and understanding these limit laws is a prerequesite to understanding derivatives and integrals.

Before we begin recall the *triangle inequality* states that for any x, y we have $||x| - |y|| \leq |x - y| \leq |x| + |y|$. The triangle inequality is actually best visualized in two dimensions, where it says that if x is $|x|$ away from the origin and y is $|y|$ away from the origin, then the distance between x and y cannot be less than $||x| - |y||$ and not more than $|x| + |y|$. Since the sign of y doesn't matter, we can write:

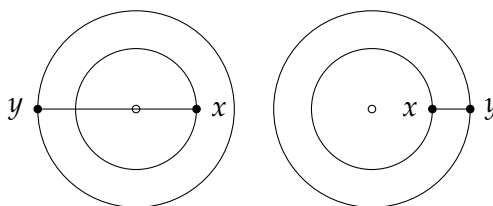$$||x| - |y|| \leq |x \pm y| \leq |x| + |y|$$

FIGURE 8.F1. Intuitive justification for the triangle inequalities. On the left, x and y are as far apart as possible at $|x| + |y|$ units. On the right, x and y are as close together as possible at $||x| - |y||$ units.

We can now state the laws for the calculus of limits. The proof is long, but the method of proof is very important and worth exploring at least until it becomes familiar (and somewhat tedious).

8.1. THEOREM (Limit calculus). *Suppose that* $\lim a_n = a$ *and* $\lim b_n = b$. *Then we have*

(a) $\lim a_n + b_n = a + b$
(b) $\lim k a_n = ka$ *for any* $k \in \mathbb{R}$
(c) $\lim a_n b_n = ab$
(d) $\lim a_n / b_n = a/b$ *provided* $b_n, b \neq 0$
(e) $\lim \sqrt{a_n} = \sqrt{\lim a_n}$ *provided* $a_n \geq 0$

PROOF.   (a) This is our first example of a so-called "$\epsilon/2$-argument." If $\epsilon > 0$ is given, we cleverly apply the convergence of a_n and b_n to the tolerance value $\epsilon/2$. Thus there exists an N_1 such that for all $n \geq N_1$ we have $|a_n - a| < \epsilon/2$, and there exsits an N_2 such that for all $n \geq N_2$ we have $|b_n - b| < \epsilon/2$. Thus if

$n \geq \max(N_2, N_2)$, we have

$$
\begin{aligned}
|(a_n + b_n) - (a + b)| &= |(a_n - a) + (b_n - b)| \\
&\leq |a_n - a| + |b_n - b| \\
&< \epsilon/2 + \epsilon/2 \\
&= \epsilon
\end{aligned}
$$

In other words, we use the triangle inequality to *split* the distance between $a_n + b_n$ and $a + b$ into two portions, each of which we can bound.

(b) This is similar but easier; we just use an $\epsilon/|k|$ argument. Given ϵ, we can find N such that for all $n \geq N$ we have $|a_n - a| < \epsilon/|k|$. It follows that $|ka_n - ka| < \epsilon$, as desired.

(c) Given ϵ, we wish to show that eventually we have $|a_n b_n - ab| < \epsilon$. To see what we will need, we calculate

$$
\begin{aligned}
|a_n b_n - ab| &\leq |a_n b_n - ab_n + ab_n - ab| \\
&\leq |a_n b_n - ab_n| + |ab_n - ab| \\
&= |a_n - a||b_n| + |a||b_n - b|
\end{aligned}
$$

Since b_n converges, we can find a bound M such that $|b_n| \leq M$ for all n (see below). Now since a_n and b_n converge, we can find an N large enough so that for $n \geq N$ we have both $|a_n - a| < \epsilon/2M$ and $|b_n - b| < \epsilon/2|a|$. Together with the calculation above, we thus conclude that $|a_n b_n - ab| < \epsilon$, as desired.

(d),(e) The proof of (d) is similar to (c), and the proof of (e) is left as an exercise.

$\square$

8.2. EXAMPLE. We can now verify that $\frac{n}{n+1} \to 1$ as follows:

$$
\begin{aligned}
\lim \frac{n}{n+1} &= \frac{1}{\lim \frac{n+1}{n}} \\
&= \frac{1}{\lim 1 + 1/n} \\
&= \frac{1}{1 + \lim 1/n} \\
&= 1
\end{aligned}
$$

Before we continue, we owe

8.3. DEFINITION. The sequence a_n is *bounded* if there exists some $M \in \mathbb{R}$ such that for all n we have $|a_n| \leq M$.

It is equivalent to say that a_n is bounded if there exists an interval $[a, b]$ such that $\{a_n\} \subset [a, b]$.

8.4. PROPOSITION. *If a_n is convergent, then a_n is bounded.*

PROOF. Suppose $a_n \to L$, so that for all ϵ we can find N such that $|a_n - L| < \epsilon$ whenever $n > N$. In particular whenever $n \geq N$ we have that $|a_n| \leq |a_n - L| + |L| < \epsilon + |L|$. Thus we may take $M_0 = \epsilon + |L|$ as a bound for these terms.

Meanwhile all that remains of the sequence is the finite initial segment $a_1, \ldots, a_N$. Letting $M_1 = \max(|a_1|, \ldots, |a_N|)$, we of course have for all $n \leq N$ that $|a_n| \leq M_1$. It follows that the full sequence is bounded by the constant $M = \max(M_0, M_1)$. $\qquad \square$

The converse of this Proposition is false, with a simple example being the sequence with terms $+1, -1, +1, -1, \ldots$.

The limit calculus theorem above provides a method of establishing convergence using algebraic properties of the sequence. Another key method of establishing convergence is using the order properties of sequence. The result is sometimes known as the *squeeze theorem*.

8.5. THEOREM (Squeeze theorem).

- *If $a_n \geq 0$ and $\lim a_n$ exists then $\lim a_n \geq 0$.*
- *If $a_n \leq b_n \leq c_n$ and $\lim a_n = \lim c_n = L$ then $\lim b_n$ exists and $= L$.*

PROOF. For the first statement, suppose towards a contradiction that $a = \lim a_n < 0$. Then applying the definition of convergence to $\epsilon = -a$ we see that eventually $|a_n - a| < -a$. It follows that $a_n - a < -a$ and hence $a_n < 0$, which is a contradiction.

For the second statement, first note that by the limit laws $\lim(c_n - a_n) = 0$. Thus given ϵ we eventually have both $c_n - a_n < \epsilon/2$ and also $|a_n - L| < \epsilon/2$. Now we calculate that

$$
\begin{aligned}
|b_n - L| &= |b_n - a_n + a_n - L| \\
&\leq b_n - a_n + |a_n - L| \\
&\leq c_n - a_n + |a_n - L| \\
&< \epsilon
\end{aligned}
$$

and thus $\lim b_n = L$, as desired. $\qquad \square$

8.6. EXAMPLE. We can now show that $\lim \frac{1}{n^2} e^{1/n} \to 0$ by writing $0 \leq \frac{1}{n^2} e^{1/n} \leq \frac{e}{n^2}$ and noting the left and right sides converge to 0.

Activity for §8.

1. Use the limit calculus and/or the squeeze theorem to evaluate the limits. Carefully justify all of your steps.

   (a) $\lim(2 + \frac{1}{n})^2$

   (b) $\lim \frac{1}{3n^2 - n}$

   (c) $\lim \left(\frac{2n^2 - 3n + 2}{3n^2 + 4n + 1} \right)^{100}$

   (d) $\lim \frac{\sqrt{n^2 + 1}}{n - 2}$

   (e) $\lim \frac{\sin(n)}{n}$

   (f) $\lim \frac{2n + \cos(n)}{n + 2}$

READING FOR NEXT TIME. Abbott, §2.4

§9. Convergence and summations

READING. Abbott, §2.4

In this section we begin to look at convergence in the context of series summations. We will define what it means for an infinite series to converge, and give a first look into the so-called convergence tests. We will return to convergence tests later on.

9.1. DEFINITION. If a_n is any sequence then the *sequence of partial sums* of the a_n is the sequence S_N with terms

$$S_n = a_1 + \cdots + a_n = \sum_1^n a_k$$

We say that the series $\sum a_n$ *converges* and equals L if the sequence of partial sums S_n converges to L in the ordinary sense.

This definition answers the question of how the terms of an infinite sum are supposed to be grouped. When we write $\sum a_n$, we mean $\cdots (((a_1 + a_2) + a_3) + a_4) + \cdots$

It is important to remark that the sequence of terms a_n and the sequence of partial sums S_n are both just ordinary sequences. The precise relation ship between the two is given by the two equations:

$$S_n = a_1 + \cdots + a_n$$
$$a_n = S_n - S_{n-1}$$

It is natural to ask about the relationship between convergence of a_n and convergence of S_n. We will see later on that if S_n converges then a_n converges too. On the other hand the converse is false, that is, a_n may converge while S_n diverges. For a simple example, if a_n is the sequence $1, 1, 1, \ldots$ then S_n is the sequences $1, 2, 3, \ldots$, and this clearly diverges.

The definition of S_n as a sum of a variable number of terms frequently makes it difficult to compute its values exactly. Still, sometimes one can find tricks to do so, such as with the following "geometric" sequence.

9.2. EXAMPLE. The series with terms $a_n = 1/2^n$, beginning at $n = 1$, has partial sums $S_1 = 1/2, S_2 = 3/4, S_3 = 7/8, S_4 = 15/16$, etc. It is easy to prove by induction that

$$S_n = \frac{2^n - 1}{2^n}$$

This calculates to $S_n = 1 - 1/2^n$ which makes it clear that $S_n \to 1$. In series notation, we write $\sum 1/2^n = 1$.

It is not usually so easy to calculate the exact value of S_n and of $\sum a_n$. Thus when given a series, we usually begin with the simpler question of whether it converges at all, or whether it diverges.

9.3. EXAMPLE. The series $\sum 1/n^2$ converges. In fact the value is $\pi^2/6$, but this fact rests on some serious calculus! Rather than calculate the partial sums exactly, we will

show only that the sequence S_n of partial sums is bounded as follows.

$$S_{2^n} = 1 + \left(\frac{1}{4} + \frac{1}{9}\right) + \left(\frac{1}{16} + \frac{1}{25} + \frac{1}{36} + \frac{1}{49}\right) + \cdots + \frac{1}{(2^n)^2}$$

$$\leq 1 + 2\left(\frac{1}{4}\right) + 4\left(\frac{1}{16}\right) + \cdots + 2^{n-1}\left(\frac{1}{(2^{n-1})^2}\right) + \frac{1}{(2^n)^2}$$

$$\leq 1 + \sum \frac{1}{2^n} = 2$$

This shows that S_{2^n} is bounded by 2, and since the sequence S_n is increasing, clearly the whole sequence S_n is bounded by 2. The next result implies that we can conclude $\sum 1/n^2$ converges even without knowing its value.

9.4. THEOREM (Monotone convergence theorem). *Suppose that b_n is a monotone increasing sequence, that is, for all n we have $b_n \leq b_{n+1}$. Suppose further that b_n is bounded. Then b_n converges to the limit $L = \sup(b_n)$.*

PROOF. Let $\epsilon > 0$ be given. Since L is an upper bound for the b_n, we have $b_n \leq L$ for all n. Since L is the least upper bound for the b_n, we can find a particular element b_N of the sequence such that $b_N > L - \epsilon$. Since the sequence is monotone increasing, $n > N \implies b_n > L - \epsilon$ as well. Thus for all $n > N$ we have $L - \epsilon < b_n < L + \epsilon$, and the convergence is established. □

To complete Example 9.3, we know that the sequence of partial sums S_n is increasing and bounded above by 2, therefore it is convergent. Hence the series $\sum 1/n^2$ converges to some unknown value ≤ 2.

We conclude our examples with a divergent series.

9.5. EXAMPLE. The series $\sum 1/n$ looks similar to the previous one, but in fact it diverges. To prove this, we show that the sequence of partial sums is unbounded using a similar grouping trick.

$$S_{2^n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \cdots + \frac{1}{2^n}$$

$$\geq 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right) + \cdots + 2^{n-1}\frac{1}{2^n}$$

$$= 1 + n\left(\frac{1}{2}\right)$$

It follows that the sequence of partial sums S_n is unbounded and therefore divergent.

The method of this last two examples can be used in a wide variety of situations. We can package this method of proof in the following convergence test.

9.6. THEOREM (Condensation test). *If a_n is a positive, decreasing sequence, then $\sum a_n$ converges if and only if $\sum 2^k a_{2^k}$ converges.*

PROOF. First suppose that $\sum 2^k a_{2^k}$ converges and let S_n denote the partial sums of the sequence a_n. Then

$$S_{2^n} = a_1 + (a_2 + a_3) + \cdots + (a_{2^{n-1}} + \cdots + a_{2^n-1}) + a_{2^n}$$
$$\leq a_1 + 2a_2 + \cdots + 2^{n-1}a_{2^{n-1}} + 2^n a_{2^n}$$
$$\leq \sum 2^k a_{2^k}$$

Thus the monotone convergence theorem implies that S_n converges, in other words that $\sum a_n$ converges.

Conversely suppose that $\sum a_n$ converges and let T_k be the partial sums of the sequence $2^k a_{2^k}$. Then

$$T_k = a_1 + 2a_2 + 4a_4 + \cdots + 2^k a_{2^k}$$
$$\leq 2a_1 + 2a_2 + 2(a_3 + a_4) + \cdots + 2(a_{2^{k-1}+1} + \cdots + a_{2^k})$$
$$\leq 2 \sum a_n$$

Thus again the monotone convergence theorem implies that T_k converges, in other words that $\sum 2^k a_{2^k}$ converges. $\qquad\square$

The condensation test may be applied to our earlier examples as follows.

- To test whether $\sum 1/n^2$ converges, we may instead test $\sum 2^k/(2^k)^2 = \sum 1/2^k$. This latter series converges by Example 9.2, so $\sum 1/n^2$ converges as well.
- To test whether $\sum 1/n$ converges, we may instead test $\sum 2^k/2^k = \sum 1$. This latter series clearly diverges, so $\sum 1/n$ diverges as well.

Activity for §9.

1. Use the condensation test to determine whether the series converge.

    (a) $\sum \frac{1}{n \ln(n)}$

    (b) $\sum \frac{1}{n(\ln n)^2}$

2. Let p be any real number. Prove that $\sum \frac{1}{n^p}$ converges if and only if $p > 1$. (This is called the "p-test.")

READING FOR NEXT TIME. Abbott, §2.5

§10. Subsequences

READING. Abbott, §2.5

Given a sequence a_n, a *subsequence* of a_n is a selection of infinitely many terms of a_n. More formally, if $n_1 < n_2 < n_3 < \ldots$ is a properly increasing sequence of indices, then the corresponding subsequence a_{n_k} consists of the temrs $a_{n_1}, a_{n_2}, a_{n_3}, \ldots$.

10.1. EXAMPLE. Starting with the sequence $a_n = 1/n$, let's make some subsequences:

$$a_{2k} : \frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \frac{1}{8}, \ldots$$

$$a_{k^2} : 1, \frac{1}{4}, \frac{1}{9}, \frac{1}{16}, \ldots$$

One could also have a selection of terms with no apparent pattern whatsoever:

$$a_{n_k} : \frac{1}{4}, \frac{1}{15}, \frac{1}{37}, \ldots$$

One must be careful to point out that it is not allowed to reorder or repeat the terms. Thus in the example of $a_n = 1/n$ the terms $1, \frac{1}{3}, \frac{1}{2}, \frac{1}{4}, \ldots$ cannot be part of a subsequence of a_n, and neither can the terms $1, \frac{1}{2}, \frac{1}{2}, \frac{1}{3}, \ldots$.

Not surprisingly, if a sequence converges, so do its subsequences.

10.2. PROPOSITION. *If a_n is a sequence which converges to L, and a_{n_k} is a subsequence of a_n, then a_{n_k} converges to L too.*

PROOF. Since $a_n \to L$, if $\epsilon > 0$ is given we can find N such that $n > N$ implies $|a_n - L| < \epsilon$. Since the n_k are strictly increasing, we always have $n_k \geq k$. Hence if $k \geq N$ then we have $n_k \geq k \geq N$ and therefore $|a_{n_k} - L| < \epsilon$ too.   □

Of course the converse is false; for a simple example the sequence $1, -1, +1, -1, \ldots$ has the convergent subsequence $1, 1, 1, 1, \ldots$.

Although it is very simple, Proposition 10.2 has several useful applications. The first application gives a method of determining that a sequence is divergent. In the past, we have checked that a given sequence fails to converge to a proposed limit. The following result lets you check that a sequence fails to converge to any/every limit.

10.3. COROLLARY. *Let a_n be a sequence, and suppose the subsequences a_{n_k} and a_{n_l} both converge but to different limits. Then a_n does not converge.*

For example if $a_n = (-1)^n(1 + 1/n)$, then a_n has terms

$$-2, \frac{3}{2}, -\frac{4}{3}, \frac{5}{4}, -\frac{6}{5}, \ldots$$

The subsequence $a_{2n} = 1 + 1/(2n)$ converges to 1, and the subsequence $a_{2n+1} = -1 - 1/(2n+1)$ converges to -1. Therefore the original sequence a_n diverges.

A second application of Proposition 10.2 allows us to settle the question posed about the associative law for series way back in Section 6.

10.4. COROLLARY. *The associative law holds for convergent series. That is, if $\sum a_n$ converges then any association of its terms converges.*

So for example if $\sum a_n$ converges then the following series converges:

$$(a_1 + a_2 + a_3) + (a_4 + a_5 + a_6) + \cdots$$

The proof is simple: the partial sums of the re-associated series form a subsequence of the partial sums of the original series. Therefore the re-associated series converges by Proposition 10.2.

We conclude this section with a key theorem about subsequences. Recall that we saw every convergent sequence is bounded, but not every bounded sequence is convergent. The following theorem gives a consolation prize instead: even if a bounded sequence doesn't converge, at least it has a convergent subsequence.

10.5. THEOREM (Bolzano–Weierstraß). *If a_n is a bounded sequence, then a_n has a convergent subsequence.*

PROOF. Since a_n is bounded we may assume it is a subset of some large interval $[-M, M]$. We go in search of the limit by iteratively subdividing the interval. To begin, at least one of $[-M, 0]$ or $[0, M]$ must contain infinitely many terms of the sequence. Whichever of these intervals it is, call it I_1, and let a_{n_1} be a term of the sequence lying in I_1.

Next, one of the two halves of I_1 must contain infinitely many terms of the sequence a_n, call it I_2, and let a_{n_2} be a term of the sequence lying in I_2. (Let us also ensure that $n_1 < n_2$.)

Continuing in this manner, we find intervals $I_k \subset I_{k-1}$ with length $M/2^{k-1}$, and terms of the sequence $a_{n_k} \in I_k$. By the NIP there exists a real number $L \in \bigcap I_k$. We claim a_{n_k} is convergent and has limit L.

For this, let $\epsilon > 0$ be given. Choose K large enough so that $M/2^{K-1} < \epsilon$. Then for all $k \geq K$ we have $a_{n_k} \in I_k \subset I_K$ and therefore $|a_{n_k} - L| < \epsilon$, as desired.   □

The Bolzano–Weierstraß theorem has numerous theoretical applications, some of which we will see later in the course. To give one small example, we can establish the following partial converse to Corollary 10.3.

10.6. PROPOSITION. *Let a_n be a bounded sequence. If a_n does not converge, then it has subsequences a_{n_k} and a_{n_l} which both converge but to different limits.*

PROOF. By the Bolzano–Weierstraß theorem, there is a convergent subsequence $a_{n_k} \to L$. Since a_n itself does not converge to L, there exists ϵ such that infinitely many terms a_n of the sequence satisfy $|a_n - L| \geq \epsilon$. Enumerate these terms in a subsequence a_{n_l}. Again by the Bolzano–Weierstraß theorem, this subsequence a_{n_l} has a further subsequence $a_{n_{l_m}}$ which converges to a limit M. Then we have $|M - L| \geq \epsilon$, in particular $M \neq L$.   □

Activity for §10.

1. (Abbott, ex 2.5.3) Give an example of each of the following, or prove that no such example exists.

   (a) A sequence that does not have 0 or 1 as a term, but has a subsequence which converges to 0 and a susequence which converges to 1

   (b) A bounded divergent sequence with a convergent subsequence

   (c) An unbounded sequence with a convergent subsequence

   (d) An unbounded monotone sequence with a convergent subsequence

   (e) A sequence that has subsequences which converge to each of the following: $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$

   (f) A sequence that has a subsequence that is bounded, but no subsequence that converges

READING FOR NEXT TIME. Abbott, §2.6

§11. Completeness

READING. Abbott, §2.6

Recall that the the real number system satisfies the Axiom of Completeness, which says that every bounded set has a supremum. We have said that this axiom should be interpreted as saying that the real number system does not have any gaps. But this leads to the question of what it means for other spaces, besides the real numbers, to be complete? For example it is intuitively clear that $\mathbb{R}^2$ or the sphere $\mathbb{S}^2$ do not have any gaps. To capture this, we need a more general way to define completeness.

11.1. DEFINITION. The sequence a_n is *Cauchy* if for all $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $m, n \geq N$ we have $|a_n - a_m| < \epsilon$.

The Cauchy sequences are the ones which really, really want to converge, but don't know what their limit should be!

11.2. EXAMPLE. The sequence $a_n = 1/n$ is Cauchy. As scratch work, we seek N large enuogh that whenever $n, m \geq N$ we have $|1/n - 1/m| < \epsilon$. Since the variables n, m are symmetric, assume for convenience that $n < m$. This way $|1/n - 1/m|$ is simply $1/n - 1/m$. Now $1/n - 1/m < \epsilon$ would follow from $1/n < \epsilon$, or $n > 1/\epsilon$.

To complete the proof, let $N > 1/\epsilon$ and without loss of generaltiy let $m > n \geq N$. Then $|1/n - 1/m| = 1/n - 1/m < 1/n \leq 1/N < \epsilon$, as desired.

The proof that the above sequence is Cauchy is not too different than the proof that it is convergent. In fact the notions of Cauchy and convergent sequences turn out to be exactly the same.

11.3. THEOREM (Cauchy criterion). *A sequence of real numbers is Cauchy if and only if it is convergent.*

PROOF. Let us first suppose that a_n is convergent with limit L and show that it is Cauchy. Given ϵ we can choose N large enough that $n > N$ implies $|a_n - L| < \epsilon/2$. Then if $m, n > N$ we have

$$\begin{aligned} |a_n - a_m| &= |a_n - L + L - a_m| \\ &\leq |a_n - L| + |L - a_m| \\ &< \epsilon \end{aligned}$$

and hence a_n is Cauchy.

Conversely suppose that a_n is Cauchy. Before showing it is convergent, we will first show it is bounded. For this, we can choose N large enough that $m, n > N$ implies $|a_n - a_m| < 1$. This shows that a tail of the sequence is bounded, and since finite sets are always bounded, the whole sequence must be bounded too.

Now since a_n is bounded, it follows from the Bolzano–Weierstraß theorem that a_n has a convergent subsequence a_{n_k} with limit L. We claim that the whole sequence a_n

converges to L as well. Indeed given ϵ we can choose N large enough that $m, n > N$ implies $|a_n - a_m| < \epsilon/2$. Moreover we can find k large enough that $n_k > N$ and $|a_{n_k} - L| < \epsilon/2$. Therefore

$$|a_n - L| \leq |a_n - a_{n_k}| + |a_{n_k} - L| < \epsilon$$

and so a_n converges to L, as desired. $\qquad\square$

The above theorem says that in the real number system, every sequence that wants to converge (beacuse it is Cauchy) really does converge. This statement is false if you restrict consider the rational number system only. Instead, the statement can be regarded as another way of saying that the real number system is complete. In general, we say that a space is *Cauchy complete* if all of its Cauchy sequences converge. The spaces $\mathbb{R}$, $\mathbb{R}^2$, $\mathbb{S}^2$ are all Cauchy complete.

It turns out that many of the theorems we have investigated so far can be regarded as statements of completeness. Just now we proved the Cauchy criterion (CC) using the Bolzano–Weierstraß theorem (BWT). Earlier we proved the Bolzano–Weierstraß theorem using the NIP. And we proved the NIP using our original Axiom of Completeness (bounded sets have a least upper bound). We therefore see that:

$$AOC \implies NIP \implies BWT \implies CC$$

The following result states that in fact all four of these statements are equivalent to one another!

11.4. THEOREM. *All of the statements AOC, NIP, BWT, CC are equivalent.*

PROOF OUTLINE. It remains only to show that CC implies AOC. Suppose we know Cauchy sequences converge and let A be a bounded subset of $\mathbb{R}$. Then A is contained in some interval $[a, b]$. For each n, divide $[a, b]$ into cells of width $(b - a)/2^n$ and let a_n be the right endpoint of the right-most cell meeting A. Then for any N, the sequence a_n eventually lies in an interval of with $(b - a)/2^N$ and hence a_n is Cauchy.

Since Cauchy sequences converge we may let L be the limit of a_n. We claim that L is the least upper bound for A. Indeed since for all n we have that a_n is an upper bound for A, it is clear that L is an upper bound for A. Moreover given any n we can find $a \in A$ such that $a \geq a_n - 1/2^n$. Since the latter is $\geq L - 1/2^n$, this implies $L = \sup(A)$. Thus AOC holds, as desired. $\qquad\square$

In the exercises you will be asked to prove that the monotone convergence theorem (MCT) is also equivalent to AOC.

Activity for §11.

1. Show directly from the definition that the sequence $x_n = \frac{n}{n+1}$ is Cauchy.

2. (Abbott, ex 2.6.2) Give an example of each of the following, or prove that no such example exists.

   (a) A Cauchy sequence that is not monotone

   (b) A monotone sequence that is not Cauchy

   (c) A Cauchy sequence with a divergent subsequence

   (d) An unbounded sequence containing a subsequence that is Cauchy

READING FOR NEXT TIME. Abbott, §2.7

§12. Series convergence tests

READING. Abbott, §2.7

In this section we return to series and the so-called tests for series convergence. First, we note that just as limits obey calculus laws, series obey certain rules as well.

12.1. THEOREM (Series calculus). *Suppose $\sum a_n$ and $\sum b_n$ both converge. Then $\sum c a_n = c \sum a_n$ and $\sum (a_n + b_n) = \sum a_n + \sum b_n$.*

PROOF. This follows almost directly from the limit calculus. For example, to show that $\sum a_n + b_n = \sum a_n + \sum b_n$, we let S_n denote the sequence of partial sums of the a_n and T_n the sequence of partial sums of b_n.

$$\sum (a_n + b_n) = \lim \sum_{1}^{N} (a_n + b_n)$$
$$= \lim (S_N + T_N)$$
$$= \lim S_N + \lim T_N$$
$$= \sum a_n + \sum b_n$$

Here, the third equality is an instance of the limit summation law.   □

Before proceeding to the convergence tests, we state formally one simple test for divergence which should be kept in mind when considering practical examples.

12.2. THEOREM (Divergence criterion). *If $\sum a_n$ is a series such that the terms a_n do not converge to zero, then $\sum a_n$ diverges.*

PROOF. We use the contrapositive. Suppose that $\sum a_n = S$. Then the sequence partial sums S_n converges to S. It follows from the addition law for limits that the terms $a_n = S_n - S_{n-1}$ converge to $S - S = 0$.   □

12.3. EXAMPLE. Consider the series $\sum (-1)^n \frac{3n^2 + 1}{2n^2 - 5}$. The terms are alternating in sign and one might think they could conceivably cancel with one another to yield a finite summation. However we know from the limit calculus that the subsequence of positive terms converges to $3/2$ and the subsequence of negative terms to $-3/2$. Hence the terms of the series do not converge to 0, so the series diverges.

The next two tests summarize some familiar series, and help give us a library of known series convergence and divergence for future use.

12.4. PROPOSITION (Series library).   ○ $\sum r^n$ *converges if and only if* $|r| < 1$
         ○ $\sum \frac{1}{n^p}$ *converges if and only if* $p > 1$

The geometric series test is standard in calculus and can be proved with a calculation similar to that of Example 9.2. The p-series test was proved in an earlier exercise as an application of the condensation test.

12.5. EXAMPLE. Consider the series $\sum 9^{-n+2}4^{n+1}$. We can rewrite it as $4 \cdot 81 \sum (4/9)^n$. Since $4/9 < 1$ the series converges and the summation is $4 \cdot 81/(1 - 4/9)$.

12.6. THEOREM (Comparison test). *Let a_n and b_n be nonnegative sequences of terms, and suppose $\sum b_n$ converges. If for all n we have $a_n \leq b_n$, then $\sum a_n$ converges too.*

PROOF. Let S_n denote the partial sums of a_n and let T_n denote the partial sums of b_n. By assumption we have that T_n converges to some value, let's call it T. Meanwhile since $a_n \leq b_n$ we clearly have $S_n \leq T_n \leq T$. Thus we have shown S_n is monotone and bounded above, and it follows from the MCT that the sequence S_n converges. This means by definition that the series $\sum a_n$ converges. $\square$

12.7. EXAMPLE. Consider the series $\sum \frac{n^2-3n-1}{n^4+2n^3+3}$. Intuitively this series isn't very different than the simpler series $\sum \frac{1}{n^2}$, but it would be a mess to apply the condensation test to this one directly. Instead, we can easily calculate

$$\frac{n^2 - 3n - 1}{n^4 + 2n^3 + 3} \leq \frac{n^2}{n^4} = \frac{1}{n^2}$$

Since $\sum \frac{1}{n^2}$ converges by the p-series test, the comparison test implies that $\sum \frac{n^2-3n-1}{n^4+2n^3+3}$ converges too.

12.8. EXAMPLE. The comparison test may also be applied in contrapositive as follows. Consider the series $\sum \frac{n^2+2n+1}{n^3-2n-1}$. Intuitively this series isn't very different than the simpler series $\sum \frac{1}{n}$, and we can again bear this out:

$$\frac{n^2 + 2n + 1}{n^3 - 2n - 1} \geq \frac{n^2}{n^3} = \frac{1}{n}$$

Since $\sum \frac{1}{n}$ diverges, the comparison test implies that $\sum \frac{n^2+2n+1}{n^3-2n-1}$ diverges too.

Although the comparison test is very simple, it may be leveraged to yield several more powerful tests. Perhaps the most famous applications of the comparison test are the ratio and root tests.

12.9. THEOREM (Root and ratio test). *Let a_n be a nonnegative sequence of terms.*
- *If the sequence $(a_n)^{1/n}$ converges to a limit $r < 1$, then the series $\sum a_n$ converges.*
- *If the sequence a_{n+1}/a_n converges to a limit $r < 1$, then the series $\sum a_n$ converges.*

PROOF. We prove only the root test. Let r' be any number such that $r < r' < 1$. Then there exists N such that $n > N$ implies $a_n^{1/n} < r'$. Since we can ignore initial segments for the purposes of series convergence, we can suppose without loss of generality that $a_n^{1/n} < r'$ for all n. It follows that $a_n < (r')^n$, and since we know $\sum (r')^n$ converges, the comparison test implies that $\sum a_n$ converges too. $\square$

12.10. EXAMPLE. Consider the series $\sum (1/2 + 1/n)^n$. Intuitively it is similar to the series $\sum 1/2^n$, but it is hard to apply either the geometric series test or the comparison test

directly. Instead, we calculate

$$\lim[(1/2 + 1/n)^n]^{1/n} = \lim(1/2 + 1/n) = 1/2$$

Since this value is less than 1, the root test implies the series converges.

Activity for §12.

1. Decide whether the given series converge or diverge. Prove your answers using the series tests we have discussed in class so far.

   (a) $\sum \dfrac{1}{n^2 + 250n + 37}$

   (b) $\sum \dfrac{1}{n^2 - 37}$

   (c) $\sum \dfrac{1}{3n - 2}$

   (d) $\sum \dfrac{1}{3n + 2}$

   (e) $\sum \dfrac{e^n}{\pi^n}$

   (f) $\sum \dfrac{n!}{n^n}$

§13. Conditional convergence

The convergence tests we studied in the previous section applied primarily to series with nonnegative terms. But many series have terms of mixed signs, and as we saw previously these series can exhibit some of the most confusing behavior (such as a lack of commutativity).

Of course, if $\sum a_n$ is a series with mixed signs, one can always take absolute values of the terms and study $\sum |a_n|$ instead. Then all of the tests in the previous section may be applied to this latter series. This idea leads to the question of what is the relationship between convergence of $\sum a_n$ and convergence of $\sum |a_n|$.

13.1. DEFINITION. A series $\sum a_n$ is said to be *absolutely convergent* if $\sum |a_n|$ is convergent.

Thus if $\sum a_n$ is a series with nonnegative terms and $\sum a_n$ converges, then $\sum a_n$ is absolutely convergent by definition.

13.2. THEOREM. *If $\sum a_n$ is absolutely convergent, then $\sum a_n$ is convergent.*

PROOF. Consider the new series $\sum(a_n + |a_n|)$ and note that the terms $a_n + |a_n|$ are nonnegative. Since $a_n + |a_n| \leq 2|a_n|$, the comparison test implies that $\sum(a_n + |a_n|)$ converges. Now the series calculus implies that $\sum a_n = \sum(a_n + |a_n|) - \sum |a_n|$, and in particular $\sum a_n$ converges. $\square$

13.3. EXAMPLE. The series $\sum \sin(n)/n^2$ is absolutely convergent. Indeed, its absolute variant is $\sum |\sin(n)|/n^2$, and this can be compared with the convergent series $\sum 1/n^2$.

The converse to the Theorem 13.2 is very false, that is, it is possible for a series to be convergent but not absolutely convergent.

13.4. EXAMPLE. The series $\sum(-1)^{n+1}/n$ is convergent. Although we will prove this shortly, for now recall that in the introductory section we estimated its value at 0.69. On the other hand, the absolute variant of this series is $\sum 1/n$, which we know to be divergent.

13.5. DEFINITION. A series $\sum a_n$ is said to be *conditionally convergent* if it is convergent but not absolutely convergent.

The possibilities for convergence are shown in Figure 13.f1.

	$\sum a_n$ conv	$\sum a_n$ div
$\sum |a_n|$ conv	AC	X
$\sum |a_n|$ div	CC	Div

FIGURE 13.F1. Possibilities for convergence. The "X" denotes an impossible situation

All of the tests in the previous section can now be regarded as tests for absolute convergence. To apply them to a series $\sum a_n$ with mixed signs, one simply considers the series

$\sum |a_n|$ and applies some test to conclude that $\sum |a_n|$ converges and hence $\sum a_n$ converges absolutely. Tests for conditional convergence are somewhat harder to come by, but there are several. The simplest and most famous is the following.

13.6. THEOREM (Alternating series test). *Assume that a_n is a decreasing sequence with limit 0. Then the series $\sum (-1)^n a_n$ converges.*

PROOF. Observe that $S_{2n+2} = S_{2n} - (a_{2n+1} - a_{2n+2})$ and that $S_{2n+3} = S_{2n+1} + (a_{2n+2} - a_{2n+3})$. Since the terms a_n are decreasing, it follows that the even-indexed subsequence S_{2n} is decreasing and the odd-indexed subsequence S_{2n+1} is increasing. Moreover S_{2n} is bounded below by $-a_1$ and S_{2n+1} is bounded above by $-a_1 + a_2$.

Therefore by the monotone convergence theorem, both subsequences converge, say to S and T respectively. Now it follows that $S_{2n+1} - S_{2n} \to S - T$. But we also have $S_{2n+1} - S_{2n} = a_{2n+1} \to 0$. Therefore $S = T$ and it is easy to see this implies $S_n \to S$. Hence $\sum a_n$ converges. $\qquad\square$

13.7. EXAMPLE. The series $\sum (-1)^n \frac{n^2}{n^3+9}$ is conditionally convergent. To see that it is not absolutely convergent one can compare $\sum \frac{n^2}{n^3+9}$ against a variant of the series $\sum 1/n$. To see that it is convergent, we apply the alternating series test. The only tricky thing to check is that $n^2/(n^3+9)$ is (eventually) decreasing. For this one can simply check that the derivative of $x^2/(x^3+9)$ is (eventually) negative.

We close our discussion of series by answering our initial question about commutativity in infinite summations. Recall that we calculated two different values for series with terms $1, -1/2, +1/3, -1/4, \ldots$. As we shall now see, this phenomenon occurs precisely when a series is conditionally convergent.

13.8. THEOREM.
  ◦ *If $\sum a_n$ is absolutely convergent then any rearrangement of its terms converges to the same limit.*
  ◦ *If $\sum a_n$ is conditionally convergent then any value can be obtained as the sum of a rearrangement of its terms.*

PROOF SKETCH. Beginning with the first statement, suppose that $\sum a_n$ is absolutely convergent and that $\sum a_n = A$. Let $\sum b_n$ be a given rearrangement of $\sum a_n$. Since $\sum a_n$ is absolutely convergent, given ϵ we can find N so large that $\sum_N^\infty |a_n| < \epsilon/2$. Next choose N' large enough that all of the terms $a_1, \ldots, a_N$ appear in the list $b_1, \ldots, b_{N'}$. Then

$$\left| \sum_1^{N'} b_n - A \right| \leq \left| \sum_1^{N'} b_n - \sum_1^N a_n \right| + \left| \sum_1^N a_n - A \right|$$

$$\leq \sum_{N+1}^\infty |a_n| + \left| \sum_{N+1}^\infty a_n \right|$$

$$\leq 2 \sum_{N+1}^\infty |a_n|$$

$$< \epsilon$$

This shows that the partial sums of the b_n converge to A, as desired.

For the second statement, note that the sum of just the positive terms must be divergent, and the sum of just the negative terms must also be divergent. Thus given a target summation S, we can add together enough positive terms until we just overshoot S. We can then add together enough negative terms until we just undershoot S. Continuing in like fashion, we straddle S infinitely many times. Since the terms $a_n \to 0$, these attempts get arbitrarily close to the value S. □

Activity for §13.

EXERCISE 13.1. Decide whether the given series converge absolutely, converge conditionally, or diverge.

(a) $\sum \dfrac{(-2)^n}{3^n}$

(b) $\sum \dfrac{(-3)^n}{2^n}$

(c) $\sum \dfrac{(-1)^n \ln n}{n}$

(d) $\sum \dfrac{(-1)^n}{\sqrt[3]{n^2 + 7}}$

Topology

§14. The Cantor set

READING. Abbott, §3.1

In this chapter we study subsets of the real line using *topology*, which is a notion of abstract shape. In the context of our class, topology will let us explore the shape of subsets of $\mathbb{R}$. Before commencing with the appropriate definitions, we wish to give a small taste of just how complicated subsets of $\mathbb{R}$ can be.

To motivate our discussion, let us begin with the notoriously difficult problem of measuring the *size* of a subset. We have already introduced the notion of *cardinality*, which gives us some idea of the size of a set (countable vs uncountable). But this notion is very course since it has just two outcomes. In this section we will discuss two more notions: the *length* and the *dimension* of a set. After we have introduced topology we will briefly touch on a fourth notion called *category*.

We will now introduce the *Cantor set*, which will help us challenge our ideas about what cardinality, length, and dimension mean, and how they interact with one another. The Cantor set set has many unusual properties not seen in our familiar examples of sets, that is, the intervals, cuts, sequences, and so forth.

The Cantor set is constructed as follows. Begin with the unit interval $[0, 1]$, and remove its open middle third to obtain the set $C_1 = [0, \frac{1}{3}] \cup [2/3, 1]$. Next remove the open middle third of each of these intervals to obtain the set $C_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$. Continuing in like fashion, we obtain a decreasing sequence of closed sets C_n, each of which is the union of several disjoint closed intervals. The Cantor set C is then defined to be $\bigcap C_n$.

FIGURE 14.F1. The first few steps in the construction of the Cantor set, and a rough image of the Cantor set itself.

Initially it is hard to see whether C has any elements at all. But it does, for example both 0 and 1 lie in C. In fact C is infinite because the endpoints of every interval making

up C_n lie in C (for example $\frac{1}{3}$, $\frac{1}{9}$, etc). The following result states that the Cantor set is in fact very large.

14.1. PROPOSITION. *The Cantor set is uncountable.*

PROOF. Let S denote the set of infinite binary sequences. We have already seen using the diagonalization argument that S is uncountable. Thus it suffices to find a one-to-one function from S to the Cantor set. Indeed, given any element $s \in S$ we define a sequence of intervals I_n from C_n as follows. If $s(1) = 0$ we let $I_1 =$ the left interval of C_1 and if $s(1) = 1$ we let $I_1 =$ the right such. Similarly if $s(2) = 0$ we let $I_2 =$ the left interval of C_2 contained in I_1, and if $s(2) = 1$ we let $I_2 =$ the right such.

Continuing in this fashion, s determines a nested sequence of closed intervals I_n. By the NIP there exists a point x_s lying in $\bigcap I_n$. Moreover the function $x \mapsto x_s$ is one-to-one. Indeed, if $s \neq s'$ then there is some first index n such that $s(n) \neq s(n')$. Then the interval I_n determined by s and the interval I'_n determined by s' are disjoint, and this implies $x_s \neq x_{s'}$. $\qquad\square$

We next consider the *length* of the Cantor set. Although we do not give the definition of length in full generality, we need only keep in mind the following very simple properties. First, the length of an interval $[a, b]$ or (a, b) is equal to $b - a$. Second, the length of a finite or countable union of disjoint sets is equal to the sum of the lengths of the factors.

14.2. PROPOSITION. *The Cantor set has zero length.*

PROOF. Going from C_0 to C_1 we remove one interval of length $1/3$. Going from C_1 to C_2 we remove two intervals of length $1/9$. Going from C_2 to C_3 we remove four intervals of length $1/27$. In general, going from stage n to stage $n + 1$ we remove a total length of $2^{n-1}/3^n$. Therefore in the whole construction we remove a total of

$$\sum \frac{2^{n-1}}{3^n} = \frac{1}{3} \sum \left(\frac{2}{3}\right)^n$$
$$= \frac{1}{3} \cdot \frac{1}{1 - 2/3}$$
$$= 1$$

We therefore conclude that the total length remaining in the Cantor set is 0. $\qquad\square$

We close with a brief discussion of the *dimension* of the Cantor set. Although we do not give a precise definition of dimension, we motivate it by observing the following property. When you dilate a square by a factor of two, the result contains $4 = 2^2$ copies of the original square. When you dilate a cube by a factor of two, the result contains $8 = 2^3$ copies of the original cube. In each case the number appearing in the exponent matches our intuitive idea of the dimension of the object. This leads to the following formula, which is used whenever it is applicable:

$$\text{number of copies} = (\text{scaling factor})^{\text{dimension}}$$

The most general definitions of dimension are somewhat complicated to describe, but all of them reduce to the above formula when it applies.

14.3. PROPOSITION. *The Cantor set has dimension strictly between* 0 *and* 1.

PROOF. The trick is to dilate the Cantor set by a factor of 3. Looking at the result applied to the first step C_1, we see it becomes two unit intervals. Thus if the construction is completed on the dilated copy, we will obtain two full copies of the Cantor set. The dimension formula therefore yields:

$$2 = 3^{\text{dimension}}$$

Hence the dimension of C is equal to $\log(2)/\log(3) \approx 0.63$. □

The three propositions taken together yield an interesting conclusion: the Cantor set is large from the point of view of cardinality, small from the point of view of length, and intermediate from the point of view of dimension!

Activity for §14. For $0 < \alpha < 1$, the *modified Cantor set* C^α is constructed by letting $B_0 = [0,1]$ and inductively letting B_{n+1} be the set obtained by removing the middle α from each interval of B_n. For example, if $\alpha = 1/3$ then we obtain the set $C^{1/3}$, which is precisely the ordinary Cantor set C.

For a new example, if $\alpha = \frac{1}{4}$ then we construct the modified Cantor set $C^{1/4}$ as follows. First B_1 is obtained by removing the middle $1/4$ of $[0,1]$. Next B_2 is obtained by removing the middle $1/4$ of each interval of B_1, for a total removal of _____.

1. (a) In the construction of $C^{1/4}$ shown above, for each n find the total length removed when going from B_n to B_{n+1}.
   (b) Find the sum of all the lengths removed, that is, the numbers calculated in part (a). Then take 1 minus this number to find the length of the modified Cantor set $C^{1/4}$.

2. Find the fractal dimension of $C^{1/4}$. [Hint: you need to use a scaling factor that grows both intervals in B_1 to a length of 1.]

3. Repeat Question 1 for the set C^α, where $0 < \alpha < 1$ is arbitrary.

4. Repeat Question 2 for the set C^α, where $0 < \alpha < 1$ is arbitrary.

5. Given $\ell \in (0,1)$, can you always choose α to make the length of C^α equal to ℓ? Given $d \in (0,1)$, can you always choose α to make the fractal dimension of C^α equal to d? Why or why not?

§15. The boundary of a set

READING. Abbott, §3.2

We can think of a subset A of some universe set X as a black-and-white division: every point is either in A or in its complement A^c. However this division is purely set-theoretic in that it doesn't take into account the relative placement of the points. If one the other hand we have some idea about the location or ordering of the points of X, then we can imagine there is a grey area consisting of points that are "on the boundary" of A.

Before we formally define the boundary of a set, recall the neighborhood notation that we introduced alongside the definition of convergence. The *ϵ-neighborhood of x* consists of all points that are less than ϵ away from x. In symbols:

$$N_\epsilon(x) = \{\, y \in \mathbb{R} \mid |y - x| < \epsilon \,\}$$

In other words, $N_\epsilon(x)$ is an interval centered at x, that is, $(x - \epsilon, x + \epsilon)$. Note that if we were working in $X = \mathbb{R}^2$ we would define $N_\epsilon(x)$ to be a disk centered at x, in $X = \mathbb{R}^3$ we would define $N_\epsilon(x)$ to be a ball centered at x, and so on.

15.1. DEFINITION. Let A be a subset of $\mathbb{R}$. A point x is said to be a *boundary point* of A if for every $\epsilon > 0$, we have that $N_\epsilon(x) \cap A \neq \emptyset$ and $N_\epsilon(x) \cap A^c \neq \emptyset$.

The set of all boundary points of A is denoted ∂A.

Intuitively, if we think of points of A as colored black and points of A^c as colored white, the boundary of A consists of those points such that no matter how far you zoom in on them, you still see both the colors black and white.

15.2. EXAMPLE. If A is the closed interval $[0, 1]$ then the boundary of A is the set $\{0, 1\}$. Similarly, if A is the open interval $(0, 1)$ then the boundary is once again the set $\{0, 1\}$.

Notice that the boundary points of a set A may be in either A or A^c. In fact following the example above, it is easy to see that a closed interval will always contain its boundary, while an open interval will always omit its boundary. This observation leads us to make the following general definition of closed and open sets.

15.3. DEFINITION.   ○ The set A is *open* if $\partial A \cap A = \emptyset$.
   ○ The set $A \subset \mathbb{R}$ is *closed* if $\partial A \subset A$.

15.4. EXAMPLE.   ○ Open intervals are open, and so are unions of open intervals such as $(0, 1) \cup (2, 3)$.
   ○ Closed intervals are closed, and so are unions of (finitely many) closed intervals such as $[0, 1] \cup [2, 3]$.
   ○ Single points $\{x\}$ are closed, as is the set $\mathbb{Z}$.
   ○ The set $\{\, 1/n \mid n \in \mathbb{N} \,\}$ is neither open nor closed: 1 is a boundary point which lies in the set, and 0 is a boundary point which does not lie in the set.
   ○ The empty set $\emptyset$ is both open and closed.

As we have just seen, open and closed sets are not opposites, since it is possible for a set to be both or neither. The following fundamental result states rather that the two notions are complementary.

15.5. THEOREM. *A is open if and only if A^c is closed.*

PROOF. By the symmetry of the definition of boundary point, it is clear that $\partial A = \partial(A^c)$. Hence A omits its boundary if and only if A^c contains its boundary. □

We next investigate the relationship between the boolean operations of union and intersection, and the concepts of open and closed sets.

15.6. THEOREM. *The union of any collection of open sets is open. The intersection of finitely many open sets is open.*

PROOF. For the first statement let $\{\, A_\alpha \,\}$ be a collection of open sets; we use the index α to allow the collection to be finite, countable, or even uncountable. Let $S = \bigcup_\alpha A_\alpha$ and let $x \in S$ be given. We wish to show that x is not in the boundary of S. Since $x \in S$ there exists an index α such that $x \in A_\alpha$. Since A_α is open, x is not in the boundary of A_α. Hence there exists a neighborhood $N_\epsilon(x) \subset A_\alpha$. It follows that also $N_\epsilon(x) \subset S$, which means x is not in the boundary of S.

For the second statement let $A_1, \ldots, A_n$ be open sets, let $T = A_1 \cap \cdots \cap A_n$, and let $x \in T$ be given. Again we wish to show that x is not in the boundary of T. Since A_i are open for each $i \le n$, we can find ϵ_i such that $N_{\epsilon_i}(x) \subset A_i$. Letting $\epsilon = \min(\epsilon_1, \ldots, \epsilon_n)$ we clearly have that $N_\epsilon(x) \subset T$. Thus x is not in the boundary of T, as desired. □

We also have an analogous statement regarding closed sets. It follows directly from the previous theorem using complements.

15.7. THEOREM. *The intersection of any collection of closed sets is closed. The union of finitely many closed sets is closed.*

15.8. EXAMPLE. The Cantor set is closed. Indeed, all of the sets removed in the construction of the Cantor set are open. This includes the initial removal of $(-\infty, 0)$ and $(1, \infty)$, and then all the open middle thirds. Thus the complement of the Cantor set is a union of open sets and hence open. Hence the Cantor set is closed.

We conclude this section with a justification for the term "closed". The next result says that closed sets are closed under the operation of taking limits of sequences.

15.9. THEOREM. *Suppose a_n is a sequence of A and $a_n \to x$. If A is closed then $x \in A$.*

PROOF. Using contraposition, suppose $x \notin A$. Since $a_n \to x$, for every ϵ there exists n such that $a_n \in N_\epsilon(x)$. Thus $N_\epsilon(x)$ meets A (at a_n) and meets A^c (at x). This means by definition that $x \in \partial A$. Thus x is a boundary point of A that is not in A, so A is not closed. □

Activity for §15.

1. For each of the following sets A, find the boundary ∂A. Then decide whether the following sets are open, closed, both, or neither.

   (a) $A = \{0, 1, \ldots, 10\}$

   (b) $A = \{1/n \mid n \in \mathbb{N}\} \cup \{0\}$

   (c) $A = (-1, 1) \cup \{2\}$

   (d) $A = [-1, 1] \cup \{2\}$

   (e) $A = \mathbb{Q}$

   (f) $A = \left\{ x \mid (\exists n \in \mathbb{N}) \; \frac{1}{n+1} < x < \frac{1}{n} \right\}$

§16. Compact sets

READING. Abbott, §3.3

Recall the Nested Interval Property, which states that the intersection of a family of nested closed intervals I_n is nonempty. We have seen that this theorem may fail if the intervals are not assumed to be closed, or not assumed to be bounded. However, a version of the NIP does hold even if the intervals are replaced by arbitrary sets which are both closed and bounded.

For subsets of $\mathbb{R}$, being both closed and bounded is very strong. To understand why, notice that if A is bounded then by the Bolzano–Weierstraß theorem every sequence of elements of A has a convergent subsequence, and if A is also closed then the limit of the subsequence lies in A. Sets with this compound property are given a special name.

16.1. DEFINITION. A subset $A \subset \mathbb{R}$ is called *compact* if every sequence from A has a convergent subsequence whose limit lies in A.

16.2. EXAMPLE. Any set which is closed and bounded is an example of a compact set. Thus the interval $[0, 1]$ is an example, as is the Cantor set C. The open interval $(0, 1)$ is not compact, since the sequence $a_n = 1/n$ has no subsequence that converges to a point in $(0, 1)$. The unbounded interval $[0, \infty)$ is not compact, since the sequence $a_n = n$ has no subsequence which converges at all.

The following result gives what one might call the "nested compact sets property."

16.3. THEOREM. *Let A_n be a sequence of nonempty compact sets such that $A_{n+1} \subset A_n$ for all n. Then $\bigcap A_n \neq \varnothing$.*

PROOF. For each n choose any element $a_n \in A_n$. In particular the entire sequence lies in A_1, and since A_1 is compact, a_n has a convergent subsequence a_{n_k} with some limit $L \in A_1$. Moreover, since the A_n are nested, for every n some tail of the subsequence a_{n_k} lies entirely within A_n. Since A_n is compact, a_{n_k} has a subsequence whose limit is in A_n. Since a_{n_k} converges to L, this limit must be L and therefore $L \in A_n$. Thus we have shown that $L \in \bigcap A_n$. $\qquad\square$

Compact sets play a role in almost every area of mathematics. As just a small example consider the extreme value theorem from calculus. The way it is usually stated, the EVT says that a continuous function on a closed, bounded interval attains its minimum and maximum value. It turns out that the domain of the function can be taken to be any compact set. Since we have not yet introduced continuous functions, we give a sketch of the proof that will be made more formal in the future.

16.4. THEOREM (Extreme Value Theorem). *Any continuous function with a compact domain attains a minimum and maximum value.*

PROOF SKETCH. Let A be a compact set and let $f\colon A \to \mathbb{R}$ be a continuous function. Let $R = \{f(x) \mid x \in A\}$ be the range of f, and let $\alpha = \sup(R)$. We wish to show there is some $x \in A$ such that $f(x) = \alpha$. Since $\alpha = \sup(R)$ there exists a sequence y_n contained within R such that $y_n \to \alpha$. For each n we may then let $x_n \in A$ be such that $f(x_n) = y_n$. Since A is compact, there exists a convergent subsequence x_{n_k} with some limit x. Then since f is continuous $f(x) = f(\lim x_{n_k}) = \lim f(x_{n_k}) = \lim y_{n_k} = \alpha$, which means f attains a maximum value. [It remains for us to justify why we can pass the limit from the inside to the outside of the function f.]   □

As we have outlined above, any closed and bounded subset of $\mathbb{R}$ is compact. In fact for subsets of $\mathbb{R}$ (or even $\mathbb{R}^n$), this turns out to be an equivalence.

16.5. THEOREM (Heine–Borel). *If $A \subset \mathbb{R}$ then A is compact if and only if A is closed and bounded.*

PROOF. We have already seen that the Bolzano–Weierstraß theorem implies that every set which is closed and bounded is compact. For the converse we will show that if A is either not closed or unbounded, then A is not compact.

First suppose that A is unbounded. Then for all n there exists some point $a_n \in A$ such that $|a_n| > n$. It follows that every subsequence of a_n does not converge at all, and therefore that A is not compact.

Next suppose that A is not closed. Then by there exists some $a \in \partial A \smallsetminus A$. That is, a is in the boundary of A but not in A. From the definition of boundary, it is not difficult to verify there exists a sequence a_n of elements of A such that $a_n \to a$. Then a_n has the property that none of its subsequences converges to a point of A, and so A is again not compact.   □

The equivalence in the last theorem turns out to be something of a coincidence. Indeed when one studies the topology of more general types of spaces, the equivalence can fail. For example, if one studies topology on $\mathbb{R}$ with the usual distance $|x - y|$ replaced by the function $d(x, y) = |x - y|/(1 + |x - y|)$, then the entire space $\mathbb{R}$ is closed and bounded, and fails to be compact. In general topology one doesn't use a distance function at all, so "bounded" doesn't make sense. In this setting, even the subsequence definiton of compactness fails to be adequate.

Activity for §16.

1. (See Abbott, ex 3.3.5) Decide whether each of the following sets A is compact. If it is not compact, give an example of a sequence in A with no subsequential limit in A.

   (a) $A = \{0, 1, \ldots, 10\}$

   (b) $A = \mathbb{Z}$

   (c) $A = \{1/n \mid n \in \mathbb{N}\} \cup \{0\}$

   (d) $A = \{q \in \mathbb{Q} \mid 0 \leq q \leq 2\}$

   (e) $A = \{x \mid (\exists n \in \mathbb{N}) \, |x - n| \leq .1\}$

§17. Connected sets

<small>READING.</small> Abbott, §3.4

Thinking about sets in the plane, it is easy to visualize the idea that some sets come all in one blob, while others are split up into multiple blobs. But if a set is just a collection of points, how can we make precise the distinction between one and multiple blobs? This task turns out to be somewhat difficult to achieve using only naïve intuition.

For open sets, the following is a natural and good definition of connected: an open set is connected if it cannot be expressed as a union of two nonempty disjoint open subsets. Thus $(0,1) \cup (1,2)$ is disconnected, and one can verify that $(0,3)$ is connected. However this definition is problematic for more general types of sets. For example $[0,1] \cup [2,3]$ should be disconnected but it can't be expressed as a union of two disjoint open subsets. At this point one might suggest dropping the condition that the two disjoint subsets be open. But consider that $(0,1) \cup [1,2)$ is actually connected, being equal to $(0,2)$.

To begin to resolve this mess, let us return to the example of $[0,1] \cup [2,3]$. Although it cannot be expressed as a union of two disjoint open subsets, its two components are *contained* in disjoint open subsets. That is, $[0,1] \subset (-0.1, 1.1)$ and $[2,3] \subset (1.9, 3.1)$. The following correct definition is the end result of this line of thinking.

17.1. <small>DEFINITION.</small> A subset $A \subset \mathbb{R}^n$ is said to be *disconnected* if there exist disjoint open sets V, V', both meeting A, and such that $A \subset V \cup V'$. A subset $A \subset \mathbb{R}^n$ is said to be *connected* if it is not disconnected.

Observe that it was easier to define disconnectedness first, and connectedness as its opposite. Thus we begin with examples of disconnected sets.

17.2. <small>EXAMPLE.</small> The union $(0,1) \cup (3,4)$ is clearly disconnected. The Cantor set is disconnected, being contained in $(-1, \frac{1}{2}) \cup (\frac{1}{2}, 2)$. The set $\mathbb{Q}$ is disconnected, being contained in $(-\infty, \sqrt{2}) \cup (\sqrt{2}, \infty)$.

In fact, as the next proposition shows, any subset of $\mathbb{R}$ which is not an interval must be disconnected. In the past we have mentioned open and closed intervals, but when we say just *interval*, we mean an interval of any type. Thus the sets (a,b), $(a,b]$, $[a,b)$, and $[a,b]$ are all intervals, where a,b may be real numbers or even $\pm\infty$ when it makes sense. It is easy to see that a subset $A \subset \mathbb{R}$ is an interval if and only if whenever $a,b \in A$ and $a < c < b$ we have $c \in A$ too.

17.3. <small>PROPOSITION.</small> *If A is a subset of $\mathbb{R}$ and A is connected, then A is an interval.*

<small>PROOF.</small> Let $a,b \in A$ and $a < c < b$; we wish to show that $c \in A$ too. Indeed if $c \notin A$, then the sets $V = (-\infty, c)$ and $V' = (c, \infty)$ would witness that A is disconnected, which is contrary to our hypothesis. □

At this point we should stress that for subsets of $\mathbb{R}^2$ or higher, the situation is much more complex. In fact, in these higher dimensions there are connected figures of any shape you can imagine.

So far we have not given any examples of connected sets. Intuitively speaking all intervals should be connected, that is, the converse to the above proposition should be true. Thus if our definition of connectedness is a good one, it will allow us to verify that intervals are connected. Because the definition was so complex, this verification turns out to be slightly tricky.

17.4. THEOREM. *If A is a subset of $\mathbb{R}$, then A is connected if and only if A is an interval.*

PROOF. We have already shown that connected subsets of $\mathbb{R}$ must be intervals. Conversely, let I be an interval, and suppose towards a contradiction that $I \subset V \cup V'$ where V, V' are disjoint open sets meeting I. Let us also choose specific elements $a_1 \in I \cap V$ and $b_1 \in I \cap V'$, and assume without loss of generality that $a_1 < b_1$.

In the next part of the proof, we will use a_1, b_1 as the basis for sequences a_n and b_n. The sequence a_n will be monotone increasing and lie in V, while the sequence b_n will be monotone decreasing and lie in V'. Moreover the two sequences will be getting closer and closer to one another, that is, $a_n - b_n \to 0$.

Assuming this has been done, we can complete the proof as follows. By the monotone convergence theorem a_n and b_n both converge, and since $a_n - b_n \to 0$ they both converge to the same limit L. Since I is an interval and $a_1 \le L \le b_1$, we clearly have $L \in I$. Since $I \subset V \cup V'$, we have either $L \in V$ or V'. But L cannot be in V since V is open and L is a limit of points outside V. Similarly, L cannot be in V', resulting in a contradiction.

To construct the sequences a_n and b_n, assume the terms a_n, b_n have been defined and define a_{n+1}, b_{n+1} as follows. Let c be the midpoint of a_n, b_n. Since I is an interval, c lies in I too. Since $I \subset V \cup V'$, we either have $c \in V$ or V'. If c lies in V then let $a_{n+1} = c$ and $b_{n+1} = b_n$; if c lies in V' then let $a_{n+1} = a_n$ and $b_{n+1} = c$. In either case we have everything we want: $a_{n+1} \in V, b_{n+1} \in V', a_n \le a_{n+1} < b_{n+1} \le b_n$ and the distance between a_{n+1}, b_{n+1} has decreased by half. $\square$

Recall that the compact sets will play a key role in the Extreme Value Theorem from calculus. Similarly, the connected sets will play a key role in the Intermediate Value Theorem from calculus. Once again, we provide only a preview.

17.5. THEOREM (Intermediate Value Theorem). *If f is a continuous function with a connected domain, then $f(a) < 0 < f(b)$ implies there exists c such that $f(c) = 0$.*

SKETCH OF PROOF. If there was no such c, then the sets $V = \{x \mid f(x) < 0\}$ and $V' = \{x \mid f(x) > 0\}$ would witness that the domain of f is disconnected, a contradiction. [It remains to justify why V and V' are open.] $\square$

Athough it took some work to show the intervals are connected, this fact can be leveraged to show that a variety of subsets of $\mathbb{R}^n$ are connected. For this we will need to see how one can piece together several connected sets to make a larger one.

17.6. PROPOSITION. *If $\mathcal{F}$ is a family of connected sets which overlap pairwise, then $S = \bigcup \mathcal{F}$ is connected.*

PROOF. Suppose, towards a contradiction, that there exist disjoint open sets V, V' meeting S such that $S \subset V \cup V'$. Since each set $A \in \mathcal{F}$ is connected, it must be contained in either V or V'. Indeed, otherwise V, V' would witness that A is disconnected. Since both V, V' meet S, we can find $A, B \in \mathcal{F}$ such that $A \subset V$ and $B \subset V'$. But we are assuming that A, B overlap, while V, V' are disjoint, a contradiction. $\square$

As a simple application, this result can be used to show that disks in $\mathbb{R}^2$ are connected.

17.7. COROLLARY. *Let D be a disk in $\mathbb{R}^2$, possibly containing some or all of its boundary. Then D is connected.*

PROOF. D can be written as a union of straight line segments from $\mathbf{0}$ to the boundary of D. The proof that intervals are connected may also be used to show that line segments are connected. Since these line segments all overlap at $\mathbf{0}$, the previous proposition implies that D is connected. $\square$

Activity for §17.

1. Decide whether each of the following sets A is connected. If it is not connected, give an example of disjoint open sets V, V' meeting A such that $A \subset V \cup V'$.

    (a) $\{\, x \mid x^2 < 5 \,\}$

    (b) $\{\, x \mid 2 < x^2 < 5 \,\}$

    (c) $\{\, x \mid x^3 - 3x^2 + x = 0 \,\}$

    (d) $\{\, x \mid x^3 + x^2 + x = 0 \,\}$

    (e) A countably infinite set A

2. How many different kinds of interval are there?

3. Explain why A is an interval if and only if whenever $a, b \in A$ and $a < c < b$ we have $c \in A$ too.

§18. Metric space

READING. Abbott, §8.2

So far we have studied the topology of the real line $\mathbb{R}$ and to some extent the Euclidean spaces $\mathbb{R}^n$. Almost everything we have said about these spaces applies in a vastly more general context

18.1. DEFINITION. A *metric space* consists of a set of points X together with a distance function $d\colon X \times X \to \mathbb{R}$ satisfying the following properties for all $x, y, z \in X$:

  (a) $d(x,y) \geq 0$, with $d(x,y) = 0$ iff $x = y$;
  (b) $d(x,y) = d(y,x)$; and
  (c) $d(x,z) \leq d(x,y) + d(y,z)$.

Parts (a) and (b) of the definition should seem quite natural. Part (c) is motivated by the triangle inequality for real numbers, which we have often used in the form $|x - z| \leq |x - y| + |y - z|$.

18.2. EXAMPLE. The real line $\mathbb{R}$ and Euclidean spaces $\mathbb{R}^n$ are the quintessential examples of metric spaces. The usual distance function on $\mathbb{R}$ is $d(x,y) = |x - y|$ and the usual distance function on $\mathbb{R}^n$ is $d(x,y) = \|x - y\|$.

18.3. EXAMPLE. Any subset A of a given metric space X can be thought of as a metric space in its own right. The restriction of the metric d to $A \times A$ still satisfies properties (a)–(c). Thus for examlpe the Cantor set is a metric space.

18.4. EXAMPLE. The space $C[0,1]$ consists of all continuous functions $f\colon [0,1] \to \mathbb{R}$. (Recall that we will discuss continuity more rigorously in the next chapter.) The distance function is the *uniform distance* $d(f,g) = \sup_{t \in [0,1]} |f(t) - g(t)|$.

To do topology in metric space, we first have to have neighborhoods. Recall that in $\mathbb{R}$ the neighborhood notation $N_\epsilon(x)$ denotes the interval $(x - \epsilon, x + \epsilon)$. In $\mathbb{R}^2$ the neighborhood notation $N_\epsilon(\mathbf{x})$ denotes the disk centered at x of radius ϵ, and so forth. By analogy, if X is a metric space with distance function d, we can define the neighborhoods $N_\epsilon(x) = \{y \mid d(x,y) < \epsilon\}$.

We can use these neighborhoods to repeat the definitions of all the basic topological concepts.

18.5. DEFINITION. Let X be a metric space with distance function d and let $A \subset X$.

  ○ A point $x \in X$ is on the boundary of A if every neighborhood $N_\epsilon(x)$ meets both A and $X \smallsetminus A$
  ○ A is open if it omits its boundary, A is closed if it contains its boundary
  ○ A is compact if every sequence of A has a subsequence converging to a limit in A
  ○ A is connected if one cannot find disjoint open sets V, V' meeting A such that $A \subset V \cup V'$.

We can also define convergent and Cauchy sequences in metric space as follows. The sequence x_n is convergent with limit x if and only if for all ϵ there exists N such that $n > N$ implies $d(x_n, x) < \epsilon$. The sequence x_n is Cauchy if and only if for all ϵ there exists N such that $m, n > N$ implies $d(x_m, x_n) < \epsilon$. At the level of generality of metric spaces, the statement we called the Cauchy criterion now becomes a definition.

18.6. DEFINITION. A metric space X is *complete* if every Cauchy sequence is convergent.

The space $\mathbb{R}$ is complete by the Cauchy criterion itself, and the spaces $\mathbb{R}^n$ are complete as well. On the other hand, the space $X = (0, 1)$, that is the open unit interval considered as a subspace of $\mathbb{R}$, is not complete. For example the sequence $1/n$ is a Cauchy sequence of $(0, 1)$ but has no limit in $(0, 1)$.

18.7. PROPOSITION. *A subspace $A \subset \mathbb{R}$ is complete if and only if A is closed.*

PROOF. First suppose that A is closed and let a_n be a sequence of A. By the Cauchy criterion, a_n has a limit a, and since A is closed, $a \in A$.

Conversely, suppose that A is complete and let a be a limit point of A. Then there exists a sequence a_n of A with limit a. Since a_n is convergent, it is Cauchy and therefore has a subsequence with limit in A. Since subsequences of a convergent sequence converge to the same limit, it follows that $a \in A$ and therefore that A is closed. $\square$

18.8. THEOREM. *The space $C[0, 1]$ is complete.*

PROOF. Let f_n be a sequence of functions which is Cauchy in the metric of $C[0, 1]$. This means that for all ϵ there exists N such that $n > N$ implies $\sup_{t \in [0,1]} |f_n(t) - f_m(t)| < \epsilon$. We seek to find a function f which is the limit of the sequence. To define the function f, first note that since f_n is Cauchy, it follows that for each $t \in [0, 1]$ the sequence $f_n(t)$ is a Cauchy sequence of real numbers. Now by the Cauchy criterion, for each t the sequence $f_n(t)$ has a limit. Thus we can define a function $f(t) = \lim f_n(t)$.

While it may seem like we are done, we actually have to verify that f really lies in $C[0, 1]$, in other words, that f is a continuous function. We also have to verify that f_n converges to f in the metric of $C[0, 1]$. The proof that f is continuous will be postponed until we have studied continuity more formally.

To see that f_n converges to f in the uniform metric, let ϵ be given and choose N large enough such that $m, n > N$ implies $\sup_{t \in [0,1]} |f_m(t) - f_n(t)| < \epsilon/2$. Fixing $n > N$ and $t \in [0, 1]$, find $m > N$ large enough that $|f_m(t) - f(t)| < \epsilon/2$. Now we have

$$|f_n(t) - f(t)| \leq |f_n(t) - f_m(t)| + |f_m(t) - f(t)| < \epsilon/2 + \epsilon/2 = \epsilon \, .$$

Since t was arbitrary, we have $\sup_{t \in [0,1]} |f_n(t) - f(t)| \leq \epsilon$, which shows that $f_n \to f$ in the uniform metric. $\square$

When $f_n \to f$ in the uniform metric, we say that the sequence f_n *converges uniformly* to f. Our text addresses the missing component in the above proof by showing that whenever a sequence of continuous functions converges uniformly to a limit, the limit must be continuous.

PART IV

Continuity and calculus

§19. Functions and limits

READING. Abbott, §4.2

So far in this course we have investigated real numbers and sets of real numbers. In this part we will study functions because they are fundamental to applications such as modeling. Many of the most important functions turn out to be continuous. For example if you go for a jog, the distance you have traveled varies continuously over time. Moreover many common expressions such as x^2, $\ln(x)$, or $3\sin(x) - 2e^x$ define continuous functions. However it is important to note that not all natural functions are continuous. For example the charge for a phone call does not vary continuously over time because it jumps at each minute.

What exactly is a continuous function? We often say intuitively that a continuous function has no breaks in its graph. But what does this mean in terms of formal definitions? A simple example of a discontinuous function is one with a jump:
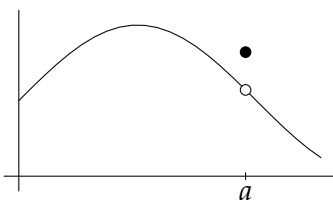
What makes this function discontinuous is that the value $f(a)$ (the filled dot) is nowhere near the limit of $f(x)$ as x approaches a (the unfilled dot). In order to follow through on this idea, we need to give a formal definition for "the limit of $f(x)$ as x approaches a."

19.1. DEFINITION. We say that $\lim_{x \to a} f(x) = L$ if for all $\epsilon > 0$ there exists $\delta > 0$ such that $0 < |x - a| < \delta$ implies $|f(x) - L| < \epsilon$.

Visualizing the definition on the graph of f, it means that whenever we draw a band of radius ϵ around the line $y = L$, there exists a small enough δ such that on $N_\delta(a)$ the function lies entirely within the band.
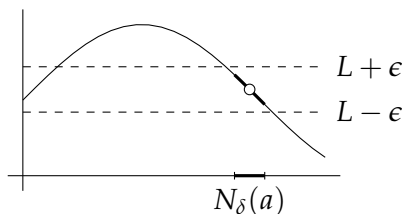
19.2. EXAMPLE. We can verify by hand that $\lim_{x \to 2}(5x + 3) = 13$. As scratch work, let ϵ be given and let us work backwards as follows:

$$|5x + 3 - 13| < \epsilon \iff |5x - 10| < \epsilon$$
$$\iff 5|x - 2| < \epsilon$$
$$\iff |x - 2| < \epsilon/5$$

This calculation suggests we should let $\delta = \epsilon/5$. To write the proof forwards, we do indeed let $\delta = \epsilon/5$ and then observe that the calculation above shows that $|x - 2| < \delta \implies |5x + 3 - 13| < \epsilon$, as desired.

The linear example above was perhaps a little too simple; the following quadratic example gives one an idea of the kinds of manipulations that can arise generally.

19.3. EXAMPLE. Let us verify by hand that $\lim_{x \to 2} x^2 = 4$. As scratch work, we once again work backwards from the desired conclusion:

$$|x^2 - 4| < \epsilon \iff |x - 2| \cdot |x + 2| < \epsilon$$
$$\iff |x - 2| < \epsilon/|x + 2|$$

Now unfurtunately this last quantity is not a constant so it does not tell us directly what number to choose for δ. However for values of x near 2 we will certainly have $|x + 2| < 10$. Thus we choose $\delta = \epsilon/10$, promise ourselves to consider values of x near 2, and write the proof forwards as follows:

$$|x - 2| < \delta \implies |x - 2| < \epsilon/10$$
$$\implies |x - 2| < \epsilon/|x + 2|$$
$$\implies |x^2 - 4| < \epsilon$$

which concludes this example.

It is also possible to use sequential limits to describe functional limits, as the next result shows.

19.4. THEOREM. *The following are equivalent.*

○ *for all $\epsilon > 0$ there exists $\delta > 0$ such that $0 < |x - a| < \delta$ implies $|f(x) - L| < \epsilon$;*
○ *for all sequences $x_n \neq a$ if $x_n \to a$ then $f(x_n) \to L$.*

PROOF. We first assume the ϵ, δ statement holds. Letting x_n be a given sequence such that $x_n \neq a$ and $x_n \to a$, we wish to show that $f(x_n) \to L$. To this end let ϵ be arbitrary, and choose δ as in the the hypothesis. Since $x_n \to a$ we can find N such that $n > N$ implies $|x_n - a| < \delta$. The hypothesis now gives that $|f(x_n) - L| < \epsilon$, and therefore that $f(x_n) \to L$.

For the converse, we proceed by contrapositive. Assume that the negation of the ϵ, δ statement holds, that is, there exists ϵ such that for al δ there exists $x \neq a$ such that $|x -$

$a| < \delta$ but $|f(x) - L| \geq \epsilon$. Fix this special value of ϵ, and apply the statement repeatedly to δ-values $1/n$. We thus obtain a sequence of $x_n \neq a$ such that $|x_n - a| < 1/n$ but $f|(x_n) - L| \geq \epsilon$. It follows that $x_n \to a$, but $f(x_n)$ does not converge to L. This establishes the negation of the latter statement, as desired.   □

This theorem is nice to have because the two equivalent clauses have complementary uses. The sequential definition is really handy for refuting functional limit equalities, since one can use a single sequence as a counterexample. On the other hand, the definition is cumbersome for establishing functional limit equalities because of the "for all sequences" quantifier. The ϵ, δ definition is often one's only resort for establishing a functional limit equality.

We close this section with the analog of the limit calculus for functional limits.

19.5. THEOREM (Functional limit calculus). *Suppose that* $\lim_{x \to a} f(x) = L$ *and* $\lim_{x \to a} g(x) = M$. *Then we have*

(a) $\lim_{x \to a} [f(x) + g(x)] = L + M$

(b) $\lim_{x \to a} kf(x) = kM$ *for any* $k \in \mathbb{R}$

(c) $\lim_{x \to a} f(x)g(x) = LM$

(d) $\lim_{x \to a} f(x)/g(x) = L/M$ *provided* $M \neq 0$

(e) $\lim_{x \to a} \sqrt{f(x)} = \sqrt{L}$ *provided* $f(x) \geq 0$

PROOF. Each of these statements can be proved using the corresponding statement in the ordinary limit calculus. For example let us prove part (a). For this let x_n be an arbitrary sequence such that $x_n \to a$. Then $\lim[f(x_n) + g(x_n)] = \lim f(x_n) + \lim g(x_n)$ by the limit calculus. By our hypothesis and the limit characterization of functional limits, this latter quantity equals $L + M$. Since x_n was arbitrary, we can conclude that $\lim_{x \to a}[f(x) + g(x)] = L + M$.   □

19.6. THEOREM (Functional squeeze theorem).

○ *If* $f(x) \geq 0$ *in a neighborhood of a and* $\lim_{x \to a} f(x)$ *exists then* $\lim_{x \to a} f(x) \geq 0$.

○ *If* $f(x) \leq g(x) \leq h(x)$ *and* $\lim_{x \to a} f(x) = \lim_{x \to a} h(x) = L$ *then* $\lim_{x \to a} g(x)$ *exists and* $= L$.

The proof is requested as an exercise.

Activity for §19.

1. Prove the following using the ϵ, δ-definition of functional limits.
    (a) $\lim_{x \to 4} \sqrt{x} = 2$
    (b) $\lim_{x \to 3} x^2 = 9$

2. (Abbott, ex 4.2.6(a)) Suppose $\delta > 0$ has been constructed as a suitable response to a particular ϵ challenge. Which is correct and why: any larger δ will also suffice, or; any smaller δ will also suffice.

§20. Continuous functions

READING. Abbott, §4.3

Thanks to our rigorous treatment of functional limits, we are now ready to define continuous fucntions. Recall that intuitively speaking, we said that f is continuous at a point a if whenever values of x get near to a, values of $f(x)$ get near to $f(a)$. Formally, this can be stated as follows.

20.1. DEFINITION. Let f be a real-valued function defined on an interval of $\mathbb{R}$. We say f is *continuous* at a point a in its domain if $\lim_{x\to a} f(x)$ exists and equals $f(a)$. We say that f is *continuous* if it is continuous on every point of its domain.

Our theorems regarding functional limits are now ready to bear fruit. First of all, since functional limits admit both an ϵ, δ and a sequential characterization, so too does continuity.

20.2. PROPOSITION. *The function f is continuous at a if and only if either of the two following conditions holds:*

(a) *for all ϵ there exists a δ such that for all x, $|x - a| < \delta$ implies $|f(x) - f(a)| < \epsilon$; or*
(b) *for all sequences x_n, if $x_n \to a$ then $f(x_n) \to f(a)$*

PROOF. The equivalence of continuity with (a) is clear from the ϵ, δ definition of functional limit. The equivalence of continuity with (b) is clear from the sequential characterization of functional limit. □

20.3. THEOREM (Continuity calculus). *Suppose that f, g are continuous functions with the same domain. Then we have:*

(a) $f + g$ *is continuous*
(b) kf *is continuous for any $k \in \mathbb{R}$*
(c) fg *is continuous*
(d) f/g *is continuous (whenever $g(x) \neq 0$)*
(e) $\sqrt{f}$ *is continuous (provided $f \geq 0$)*

These properties follow trivially from the functional limit calculus, together with the definition of continuity. To this list of laws we can add one additional operation, namely composition of functions.

20.4. THEOREM. *If f, g are continuous and $\mathrm{rng}(f) \subset \mathrm{dom}(g)$ then the composition $g \circ f$ is continuous.*

PROOF. This follows trivially from the sequential characterization of continuity. Let a be a point of $\mathrm{dom}(g \circ f) = \mathrm{dom}(f)$ and let $x_n \to a$. Then since f is continuous $f(x_n) \to f(a)$, and since g is continuous $g \circ f(x_n) \to g \circ f(a)$. This shows $g \circ f$ is continuous at a. □

The continuity calculus allows us to establish that a variety of functions are continuous.

20.5. EXAMPLE. Since it is clear that the identity function $f(x) = x$ is continuous, we can use the arithmetic laws to conclude that polynomial functions such as $3x^5 - 17x^2 + 12$ are continuous. We can also conclude that rational functions such as $(3x^3 + 17x)/(2x^2 - 13)$ are continuous whenever the denominator is not zero.

20.6. EXAMPLE. The function $f(x) = \sin(x)$ is continuous. Again we must check this by hand using the definition of the sin function. Without going into details, it is not difficult to see geometrically that we always have $|\sin(x) - \sin(a)| \leq |x - a|$, so we may take $\delta = \epsilon$ in this proof.

20.7. EXAMPLE. The function $\sqrt{x^2 + 17\sin(1/x)}$ is continuous where it is defined. Here we are using the composition law two times.

We close with a beautiful characterization which expresses continuity in purely topological terms. First we need to define forward and inverse images of a set with respect to a function.

20.8. DEFINITION. Let $f : X \to Y$ be any function. If A is a subset of X then the *forward image* of A with respect to f is the set $f(A) = \{ f(a) \mid a \in A \}$. And if B is a subset of Y then the *inverse image* of B with respect to f is $f^{-1}(B) = \{ x \in X \mid f(x) \in B \}$.

Notice that if f is a continuous function and V is an open set, it is not necessarily the case that $f(V)$ is open too. For example if $f(x) = x^2$ and $V = (-1, 1)$, then $f[V] = [0, 1)$ which is not open. On the other hand, the following result states that the continuous functions are precisely those which preserve the open sets under inverse images.

20.9. THEOREM. *Let $f : \mathbb{R} \to \mathbb{R}$. Then f is continuous everywhere if and only if for every open set V we have $f^{-1}(V)$ is open too.*

PROOF. First suppose f is continuous, and let V be open and $a \in f^{-1}(V)$. Since V is open, $f(a)$ has a neighborhood $N_\epsilon(f(a)) \subset V$. Since f is continuous at a there exists δ such that $|x - a| < \delta$ implies $|f(x) - f(a)| < \epsilon$. This is equivalent to the set containment $N_\delta(a) \subset f^{-1}(N_\epsilon(f(a)))$. It follows that $N_\delta(a) \subset f^{-1}(V)$, so $f^{-1}(V)$ is open.
Conversely, suppose V open implies $f^{-1}(V)$ is open. Given a and ϵ we apply this to the open set $N_\epsilon(f(a))$ to conclude $f^{-1}(N_\epsilon(a))$ is open. This implies there exists δ such that $N_\delta(a) \subset N_\epsilon(f(a))$. This means precisely that $|x - a| < \delta$ implies $|f(x) - f(a)| < \epsilon$, that is, f is continuous at a. $\qquad\square$

We remark that this characterization is true in the world of metric spaces too: if $f : X \to Y$ then f is continuous (preserves limits) if and only if for every open set V in Y we have $f^{-1}(V)$ is open in X.

Activity for §20.

1. Write an expression for a function f with the given property.
   (a) f is defined and continuous everywhere except $x = 1$
   (b) f is defined and continuous everywhere except $x = 1, 2$
   (c) f is defined and continuous everywhere except the integers
   (d) f is defined and continuous everywhere except $(-1, 1)$

2. Give an example of a discontinuous function f such that $|f|$ is continuous.

3. (Abbott, ex 4.3.9) Let f be a continuous function and let $Z = \{\, x \mid f(x) = 0 \,\}$ (the zeroes or roots of the function). Prove that Z is closed.

§21. Continuity and preservation

READING. Abbott, §4.4, 4.5

In the previous section we concluded that continuous functions preserve open sets under inverse images but not necessarily under forward images. In this section we address the same questions for closed, compact, and connected sets. In doing so we will finally be able to explain two major theorems from calculus: the extreme value theorem and the intermediate value theorem.

Before beginning our investigation, we should point out some further properties of forward and inverse images. It is not hard to check that for any function f, the inverse image mapping f^{-1} preserves unions, intersections, and differences. That is, we have

- $f^{-1}(A \cup B) = f^{-1}(A) \cup f^{-1}(B)$
- $f^{-1}(A \cap B) = f^{-1}(A) \cap f^{-1}(B)$
- $f^{-1}(A \smallsetminus B) = f^{-1}(A) \smallsetminus f^{-1}(B)$

The forward image mapping $f(\cdot)$ also preserves unions, but fails to preserve intersections or differences. For counterexamples, look to the function $f(x) = x^2$.

21.1. PROPOSITION. *If f is continuous and C is closed, then $f^{-1}(C)$ is closed.*

PROOF. We have shown that if f is continuous and V is open, then $f^{-1}(V)$ is open. Since C is closed, its complement $\mathbb{R} \smallsetminus C$ is open. It follows from the pervious discussion that $f^{-1}(\mathbb{R}) \smallsetminus f^{-1}(C)$ is open since it is equal to $f^{-1}(\mathbb{R} \smallsetminus C)$. We therefore conclude that $f^{-1}(C)$ is closed. $\qquad \square$

Just as continuous functions need not preserve open sets under forward images, neither must they preserve closed sets under forward images. For example if $f(x) = 1/(1 + x^2)$ and $A = \mathbb{R}$ then $f(A) = (0,1]$. Observe that for this counterexample we chose to go with an unbounded set A. The following result implies that it was necessary to do so.

21.2. THEOREM. *If f is continuous and K is compact, then $f(K)$ is compact.*

PROOF. Let y_n be a seqence of $f(K)$, and let x_n be such that $y_n = f(x_n)$ for all n. Since K is compact, x_n has a convergent subsequence x_{n_k} with limit $x \in K$. Since f is continuous we have that $y_{n_k} = f(x_{n_k})$ converges to the limit $f(x)$ which is in $f(K)$. Thus $f(K)$ is compact. $\qquad \square$

We have mentioned the extreme value theorem before as an important property of compact sets. At the time we did not have the framework to give a formal proof, but now we do.

21.3. COROLLARY (Extreme Value Theorem). *If K is compact and $f: K \to \mathbb{R}$ is continuous, then f achieves a minimum and maximum value on K.*

PROOF. By the theorem, $f(K)$ is a compact subset of $\mathbb{R}$. Since compact sets are bounded, $f(K)$ has an infemum α and a supremum β. Since compact sets are closed, and we have

seen that α, β lie on the boundary of $f(K)$, it follows that $\alpha, \beta \in f(K)$. Thus there exist $a, b \in K$ such that $f(a) = \alpha$ and $f(b) = \beta$, as claimed. $\qquad \square$

The extreme value theorem is the key to any kind of optimization. If one is seeking an optimum value (usually an infemum or supremum), the exterem value theorem gives conditions under which one can be assured that such a value exists.

Finally we turn to preservation of connected sets. To motivate this, observe intuitively that if f is a continuous real-valued function defined on an interval, then the range of f had better be an interval too. That's because if there were a gap in the range then f would have to have some kind of jump, which means a discontinutiy.

21.4. THEOREM. *If f is continuous and A is connected then $f(A)$ is connected.*

PROOF. Let us prove the contrapositive, that if $f(A)$ is disconnected then A is disconnected. So suppose that there exist disjoint open sets V, V' both meeting $f(A)$ and such that $f(A) \subset V \cup V'$. Observe that the inverse image mapping f^{-1} preserves open sets, disjointness, and unions. It follows that $f^{-1}(V), f^{-1}(V')$ are disjoint open sets, both meeting A, and such that $A \subset f^{-1}(V) \cup f^{-1}(V')$. This means that A is disconnected, as desired. $\qquad \square$

Once again, this simple preservation property has an important consequence in calculus.

21.5. COROLLARY (Intermediate Value Theorem). *If f is continuous on $[a, b]$ and $f(a) < L < f(b)$ then there exists c such that $a < c < b$ and $f(c) = L$.*

PROOF. Since $[a, b]$ is connected, the previous theorem gives that $f([a, b])$ is connected. As we have shown that connected subsets of $\mathbb{R}$ are intervals, we have that $f([a, b])$ is an interval. Recall our characterization that I is an interval iff $x, y \in I$ and $x < z < y$ implies $z \in I$. We currently have $f(a), f(b) \in f([a, b])$ and $f(a) < L < f(b)$. Thus $L \in f([a, b])$, which means there exists $c \in [a, b]$ such that $f(c) = L$. Finally note that c clearly cannot equal a or b itself. $\qquad \square$

The intermediate value theorem has many practical consequences, just one of which is the following.

21.6. COROLLARY. *If $f \colon [0, 1] \to [0, 1]$ is a continuous function, then f has a fixed point.*

PROOF. Consider the function $g(x) = x - f(x)$; we are looking for some point c such that $g(c) = 0$. If either $g(0)$ or $g(1) = 0$ then we would be done. Otherwise, we clearly have that $g(0) < 0 < g(1)$. By the intermediate value theorem it follows that there exists c such that $g(c) = 0$, as desired. $\qquad \square$

Activity for §21.

1. Give an example of each of the following, or argue that no such example exists.

   (a) A continuous function f and a compact set K such that $f^{-1}(K)$ is not compact.

   (b) A continuous function f and a connected set C such that $f^{-1}(C)$ is disconnected.

   (c) A continuous function f and a bounded set A such that $f(A)$ is not bounded.

   (d) A continuous function f, whose domain is $\mathbb{R}$, and a bounded set A such that $f(A)$ is not bounded.

   (e) a continuous function f and a bounded set B such that $f^{-1}(B)$ is not bounded.

§22. Continuity and derivatives

READING. Abbott, §5.2

The derivative is one of the most important developments of modern mathematics because it captures the intuitive notion of "rate of change". As simple as this intuition may seem, one must be careful since not all functions have a well-defined derivative. Our study of continuous functions and limits leaves us prepared to explore the concept rigorously.

22.1. DEFINITION. If f is a real-valued function defined on an interval and x is in the domain of f, then f is said to be *differentiable* at x if the limit exists:

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

In this case the value of the limit is called the *derivative* of f at x and denoted $f'(x)$.

As one often observes in a calculus class, the derivative is a limit of slopes between $(x, f(x))$ and nearby points $(x+h, f(x+h))$. This means that $f'(x)$ represents the slope of the line tangent to f at x. If a function is differentiable on its domain, we sometimes say that it is "locally linear".

22.2. EXAMPLE. The function $f(x) = x^2$ is differentiable at every point and we can use the functional limit calculus to find $f'(x)$ as follows:

$$f'(x) = \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h} = \lim_{h \to 0} \frac{x^2 + 2xh + h^2 - x^2}{h} = \lim_{h \to 0} 2x + h = 2x$$

22.3. EXAMPLE. Any function with a corner or pointy cusp is not differentiable at that point. For example $f(x) = |x|$ is not differentiable at $x = 0$ because $\lim_{h \to 0} |h|/h$ does not exist.

The primary development of differential calculus now proceeds in two pieces. First, one uses the definition of the derivative to calculate a library of *formulas* for the derivatives of common functions. Second, one uses the definition of the derivative to prove the algebraic *laws* of the derivative operation itself.

22.4. THEOREM (Derivative formulas).   ○ *The derivative of x^α is $\alpha x^{\alpha-1}$*
   ○ *The derivative of $\ln(x)$ is $1/x$*
   ○ *The derivative of e^x is e^x*
   ○ *The derivative of $\sin(x)$ is $\cos(x)$*
   ○ *The derivative of $\cos(x)$ is $-\sin(x)$*

22.5. THEOREM (Derivative calculus). *Asusme f, g are differentiable functions with the same domain.*
   ○ $(f+g)' = f' + g'$
   ○ $(kf)' = kf'$

○ $(fg)' = f'g + fg'$
○ $(f/g)' = (f'g - g'f)/g^2$
○ *chain rule:* $(f \circ g)' = (f' \circ g)g'$
○ *inverse rule:* $(f^{-1})' = 1/(f' \circ f^{-1})$ *if f is invertible*

We leave the proof of this theorem to your old calculus textbook.

In the rest of the section, we investigate the subtle relationship between continuity and differentiability. Our first result in this direction implies that differentiability is stronger than continuity. Phrased another way, any function which is discontinuous will also be non-differentiable.

22.6. PROPOSITION. *If f is differentiable at a point x then f is continuous at x.*

PROOF. For this proof it is convenient to use the common substitution $y = x + h$ and write $f'(x) = \lim_{y \to x}(f(y) - f(x))/(y - x)$. It follows from this and the functional limit calculus that

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

This implies that $\lim_{y \to x} f(y) = f(x)$, which means by definition that f is continuous at x. □

Now we know that every differentiable function is continuous, but not every continuous function is differentiable. It is natural to ask whether f being differentiable implies f' is continuous. It turns out that the answer is "no".

22.7. EXAMPLE. Let $f(x) = x^2 \sin(1/x)$ and $f(0) = 0$. The function is clearly differentiable at any point $x \neq 0$, and it is actually differentiable at $x = 0$ too. To see this, observe that

$$f'(0) = \lim_{h \to 0}\frac{h^2 \sin(1/h) - 0}{h} = \lim_{h \to 0} h \sin(1/h)$$

Since $\sin(1/h)$ is bounded and $h \to 0$, it follows that $f'(0) = 0$.

Next, the derivative of f is $f'(x) = 2x \sin(1/x) - \sin(1/x)$ and $f'(0) = 0$. It is easy to see that the first term $x \sin(1/x)$ is continuous, but the second term $\sin(1/x)$ is not continuous at $x = 0$. It follows that $f'(x)$ is not continuous at $x = 0$.

In fact there is a whole hierarchy of differentiability properties. We say that a function f is in the class C^0 if f is continuous, in the class C^1 if f is differentiable f' is continuous, in the class C^2 if f' is differentiable and f'' is continuous, and so forth. We say that f is in the class C^∞ or *smooth* if it is differentiable infinitely many times. Functions such as polynomials and $\sin(x)$ are all smooth. But the function $x^n \sin(1/x)$ lies in the class C^{n-2}.

In the example above, the discontinuous derivative $f'(x)$ did not just have a simple jump discontinuity, but something much uglier. In our final result of the section, we will show that this ugliness can't be avoided. The result states that while derivatives need not be continuous, derivatives still share the *intermediate value property* with continuous

functions. That is, the conclusion of the intermediate value theorem is true for derivatives. First we need the following familiar lemma.

22.8. LEMMA (Interior Extremum Theorem). *If f is differentiable on the interval (a, b) and f attains a maximum or minimum value at $c \in (a, b)$, then $f'(c) = 0$.*

PROOF. Let us treat the case when f attains a maximum value at c. Then for any h we have $f(c + h) - f(c) \leq 0$. It follows that for $h > 0$ we have $(f(c + h) - f(c))/h \leq 0$, which implies that $f'(c) \leq 0$. Similarly for $h < 0$ we have $(f(c + h) - f(c))/h \geq 0$, which implies that $f'(c) \geq 0$. This leaves only the possibility that $f'(c) = 0$.   □

The interior extremum theorem has applications to optimization, because it says that to find a minimum or maximum a good place to look is at the critical points.

22.9. THEOREM (Darboux's theorem). *If f is differentiable on $[a, b]$ and $f'(a) < L < f'(b)$, then there exists $c \in (a, b)$ such that $f'(c) = L$.*

PROOF. We may assume without loss of generality that $L = 0$. Since $[a, b]$ is compact, the extreme value thorem implies that f attains its minimum value at some point $c \in [a, b]$. We claim that c is not equal to a or b. Admitting the claim for the moment, this implies f attains its minimum at a point $c \in (a, b)$. Thus the Interior Extremum Theorem implies that $f'(c) = 0$, as desired.

To see that f doesn't attain its minimum value at a or b, first note that since $f'(a) < 0$ we have $\lim_{h \to 0}(f(a + h) - f(a))/h < 0$. It follows that there exists $h > 0$ such that $f(a + h) < f(a)$, that is, f cannot attain its minimum value at a. Similarly since $f'(b) > 0$ we have $\lim_{h \to 0}(f(b + h) - f(b))/h > 0$. It follows that there exists $h < 0$ such that $f(b + h) < f(b)$, so once eagain f cannot attain its minimum value at b either.   □

Activity for §22.

1. Find the derivative of each function directly from the definition.

    (a) $f(x) = x^3$

    (b) $f(x) = 1/x$

    (c) $f(x) = \dfrac{1}{\sqrt{x}}$

2. Decide, with justification, whether each of the following functions is differentiable.

    (a) $f(x) = |x| + |x - 1|$

    (b) $f(x) = x\,|x|$

    (c) $f(x) = |\sin(x)|$

§23. Continuity and integrals

READING. Abbott, §7.2, 7.4

Along with the derivative, the integral is one of the highlights of modern mathematics because it helps capture the concept of area. Just as the idea of "slope of a function" is intuitively simple but required rigorous definition, the idea of "area under a curve" is easy to visualize but must be defined carefully. Once again, our knowledge of functions, infemums, and supremums leaves us prepared to define and explore integrals.

23.1. DEFINITION. Let f be a bounded function defined on an interval $[a, b]$. Given a finite sequence P of endpoints $a = x_0 < x_1 < \cdots < x_n = b$, we let $m_i = \inf f([x_{i-1}, x_i])$ and $M_i = \sup f([x_{i-1}, x_i])$, and:

$$L(f, P) = \sum_{i=1}^{n} m_i(x_i - x_{i-1})$$

$$U(f, P) = \sum_{i=1}^{n} M_i(x_i - x_{i-1})$$

We then define the *lower integral* of f as $L(f) = \sup_P\{L(f, P)\}$ and the *upper integral* of f as $U(f) = \inf_P\{U(f, P)\}$. One can verify that we always have $L(f) \leq U(f)$. If in fact $L(f) = U(f)$, then we say f is *integrable* and write

$$\int_a^b f\, dx = \text{the common value of } L(f) \text{ and } U(f)$$

23.2. EXAMPLE. The function $f(x) = x^2$ on the interval $[0, b]$ is integrable. To see why, given any n we define the regularly-spaced partition P_n by $x_i = bi/n$. Since f is a monotone increasing function, maximum values occur at right endpoints, which means we always have $M_i = (bi/n)^2$. Now we can calculate:

$$
\begin{aligned}
U(f, P_n) &= \sum_{i=1}^{n} \left(\frac{bi}{n}\right)^2 \left(\frac{b}{n}\right) \\
&= \frac{b^3}{n^3} \sum_{i=1}^{n} i^2 \\
&= \frac{b^3}{n^3} \frac{n(n+1)(2n+1)}{6}
\end{aligned}
$$

Here we are using a classic formula for the sum of the first n squares. It is easy to calculate that $\lim U(f, P_n) = b^3/3$, and since $U(f)$ is the infemum over all possible $U(f, P)$, we conclude that $U(f) \leq b^3/3$. An analogous calculation shows that $L(f) \geq b^3/3$, and we conclude from these inequalities that $\int_0^b f\, dx$ exists and equals $b^3/3$.

23.3. EXAMPLE. Let f be the Dirichlet function on $[0, 1]$, which is defined by

$$f(x) = \begin{cases} 1 & x \in [0, 1] \cap \mathbb{Q} \\ 0 & x \in [0, 1] \setminus \mathbb{Q} \end{cases}$$

Then f is not integrable. In fact using rational and irrational density, for any P we have $L(f,P) = 0$ and $U(f,P) = 1$.

As was the case with limits and derivatives, the integral operation obeys some useful laws.

23.4. THEOREM (Integral calculus). *Assume f,g are integrable functions on $[a,b]$.*

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

(b) $\int_a^b kf(x)\,dx = k\int_a^b f(x)\,dx$

(c) *If $f \le g$ then $\int_a^b f(x)\,dx \le \int_a^b g(x)\,dx$*

(d) *If $a < c < b$ then $\int_a^c f(x)\,dx + \int_c^b f(x)\,dx = \int_a^b f(x)\,dx$*

PROOF. As these results are standard facts of calculus, we prove only (c). For this let $h(x) = g(x) - f(x)$, so that we have $h \ge 0$. Now using parts (a) and (b), h is integrable and our desired conclusion is equivalent to $\int_a^b h(x)\,dx \ge 0$. Since $h \ge 0$ we have that for any P, all values of $m_i \ge 0$ and thus $L(f,p) \ge 0$. It follows that $L(f) \ge 0$ too, and hence that $\int_a^b h(x)\,dx \ge 0$. $\square$

We close this section by establishing the first link between continuity and integration. We show that any continuous function is integrable. Since we have seen that every differentiable function is continuous and not every continuous function is differentiable, we can conclude that the integral operation is properly *more general* than the derivative operation!

23.5. THEOREM. *If f is a continuous function on $[a,b]$ then f is integrable. If f has finitely many discontinuities on $[a,b]$ then f is integrable.*

PROOF. We will use Abbott, Theorem 4.4.7 as a black box. It states that if f is a continuous function on a compact domain, then it satisfies the following strong form of continuity:
$$(\forall \epsilon > 0)(\exists \delta > 0)(\forall x,y)\ |x - y| < \delta \implies |f(x) - f(y)| < \epsilon$$
(The difference is only that δ does not depend on x.) Now given any $\epsilon > 0$, we create a partition P such that $x_i - x_{i-1} < \delta$ for all i. This implies that for all i, we have $M_i - m_i \le \epsilon$. Thus
$$U(f,P) - L(f,P) = \sum_{i=1}^n (M_i - m_i)(x_i - x_{i-1}) \le \sum_{i=1}^n \epsilon(x_i - x_{i-1}) = \epsilon(b - a)$$
Since ϵ was arbitrary, we conclude that $U(f,P) - L(f,P) = 0$.

The second statement follows from this together with law (d). $\square$

Activity for §23.

1. Use the definition of integral to calculate the value of

$$\int_0^b x^3 \, dx$$

You will need the sum of cubes formula.

$$\sum_1^n i^3 = \frac{n^2(n+1)^2}{4}$$

§24. Derivatives and integrals

READING. Abbott, §7.5

We have introduced derivatives as slopes or rates of change, and integrals as areas or accumulations. How are the two concepts related? Since we have already calculated that $\int_0^b x^2\,dx = b^3/3$ and $\int_0^b x^3\,dx = b^4/4$, it is not difficult to guess that integration is related to antiderivatives. Intuitively, it is also not hard to accept that the rate of change of an accumulation of some quantity should at least be related to the size of the original quantity. This is all confirmed in surprisingly great generality by the Fundamental Theorem of Calculus, which says in brief that derivatives and integrals are inverse operations.

Before stating and proving the FTC, we need to visit the Mean Value Theorem, a cornerstone result about derivatives that is closely related to the FTC.

24.1. THEOREM (Mean Value Theorem). *Let f be differentiable on $[a,b]$. Then there exists $c \in [a,b]$ such that*

$$\frac{f(b) - f(a)}{b - a} = f'(c)$$

PROOF. We can add a linear term to f to assume without loss of generality that $f(a) = f(b)$. (Note that passing from $f(x)$ to $f(x) + mx + b$ has the effect of adding an m to both sides of the desired equation.) Now we can use the Extreme Value Thoerem together with the Interior Extremum Theorem to conclude that there must exist $c \in [a,b]$ such that $f'(c) = 0$, the desired result.   □

The MVT says that the "average rate of change" of a function over an interval is equal to the instantaneous rate of change at some point in the interval. On the other hand, the conclusion of the MVT can also be written as follows:

$$f(b) - f(a) = f'(c)(b - a)$$

This version foreshadows the FTC because the left side involves a change in f, and the right side involves a derivative times a change in x.

The MVT has a number of important applications. For example, it can be used to show that antiderivatives are unique up to an additive constant.

24.2. COROLLARY. *Suppose that $f'(x) = g'(x)$. Then $f - g$ is a constant function.*

PROOF. Let $h = f - g$ and suppose h is not a constant function. Then there exist a, b in the domain of h such that $(h(b) - h(a))/(b - a) \neq 0$. It follows from the MVT that there exists c such that $h'(c) \neq 0$. Thus $f'(c) \neq g'(c)$, contradicting our hypothesis.   □

We are now ready to dive into the FTC, which we recall says that derivatives and indegrals are inverse operations. This is actually two separate statements: the integral of a derivative is (related to) the original function, and the derivative of an integral is the original function. We begin with the first version, sometimes called the net change theorem.

24.3. THEOREM (FTC, net change theorem). *Let f be differentiable on $[a, b]$ and f' be integrable on $[a, b]$. Then*

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

PROOF. Let P be a partition of $[a, b]$ consisting of endpoints $0 = x_0 < x_1 < \cdots < x_n = b$. Then for each i, the MVT gives us $c_i \in [x_{i-1}, x_i]$ such that

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

Summing both sides over $i \leq n$, the left-hand side telescopes and becomes $f(b) - f(a)$. Therefore we have

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

Since we clearly have $m_i \leq f'(c_i) \leq M_i$, we obtain

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

Since the partition P was arbitrary, we conclude that $L(f') \leq f(b) - f(a) \leq U(f')$. Since we are assuming $L(f') = U(f') = \int_a^b f'(x)\,dx$, we have $f(b) - f(a) = \int_a^b f'(x)\,dx$ as desired. $\square$

Dividing both sides of the net change theorem by $b - a$, we obtain a statement which says that the "average rate of change" of a function over an interval is equal to the average of its instantaneous rates of change over the interval.

$$\frac{f(b) - f(a)}{b - a} = \frac{1}{b - a}\int_a^b f'(x)\,dx$$

We now continue with the second part of the FTC. First, given any integrable function f we define its *accumulation function* (or *area function*) as $\int_0^x f(t)\,dt$. The second part of the FTC concerns the derivative of the accumulation function. To understand the statement intuitively, imagine extending a $0.5 \times x$ rectangle in the x direction. The area grows at a rate of 0.5. In a similar way, if we extend the region under f in the x direction the area should grow at a rate of $f(x)$. Said another way, the derivative of the accumulation of f is f, or, the accumulation of f is an antiderivative of f.

24.4. THEOREM (FTC, accumulation-is-antiderivative theorem). *Let $f(x)$ be a continuous function and let $F(x) = \int_a^x f(t)\,dt$. Then F is differentiable, and*

$$F'(x) = f(x)$$

PROOF. Substituting the definition of F into the difference quotient for F, we have:

$$F'(x) = \frac{1}{h}\int_x^{x+h} f(t)\,dt$$

By the continuity of f, given $\epsilon > 0$ we can find $\delta > 0$ such that $|t - x| < \delta$ implies $|f(t) - f(x)| < \epsilon$. Letting $h < \delta$, it follows that

$$h(f(x) - \epsilon) \leq \int_x^{x+h} f(t)\, dt \leq h(f(x) + \epsilon)$$

Dividing all sides by h, we have:

$$f(x) - \epsilon \leq F'(x) \leq f(x) + \epsilon$$

Since ϵ was arbitrary, we conclude that $F'(x) = f(x)$. $\qquad\square$

Recall that when we introduced the derivative calculus we had both formulas for specific derivatives as well as laws for the derivative operation. On the other hand for the integral calculus we have only introduced laws, and still owe the formulas. The accumulation-is-antiderivative theorem implies that every derivative formula can be reversed to give an integral formula, and therefore completes our basic picture of calculus.

24.5. THEOREM (Integral calculus formulas). *For each function $f(x)$, we have $\int_a^b f(t)\, dt = F(b) - F(a)$ where F is the function shown.*

   (a) *If $f(x) = x^\alpha$ then $F(x) = \frac{1}{\alpha+1} x^{\alpha+1} + C$, for $\alpha \neq -1$*
   (b) *If $f(x) = 1/x$ then $F(x) = \ln(x) + C$*
   (c) *If $f(x) = e^x$ then $F(x) = e^x + C$*
   (d) *If $f(x) = \sin(x)$ then $F(x) = -\cos(x) + C$*
   (e) *If $f(x) = \cos(x)$ then $F(x) = \sin(x) + C$*

Activity for §24.

1. Find the derivative of the function:
   (a) $F(x) = \int_0^x \cos(3t)\, dt$
   (b) $F(x) = \int_0^x \cos(t^2)\, dt$
   (c) $F(x) = \int_0^{\ln(x)} \cos(t^2)\, dt$

2. Suppose you know $f(2) = 5$ and $-.5 \leq f'(x) \leq .5$ for all x. What is the biggest $f(8)$ can be? What is the smallest? How do you prove it?

3. (Abbott, ex 7.5.4) Let f be a continuous function on $[a, b]$ and assume that $\int_a^x f(t)\, dt = 0$ for all x. Show that $f = 0$.

   Give a counterexample to show that the statement is false if f is not assumed to be continuous.