Math 4100 Notes

Pete L. Clark

Contents

CHAPTER 1

Topology of Euclidean Space

1. A Little Review of Math 3100

The name of the course is **Real Analysis**, so let us begin with a check-in on the real numbers $\mathbb{R}$, perhaps the most important single mathematical object. Intuitively we view $\mathbb{R}$ as being the points on a number line, with the origin marked as 0 and with an orientation so that we may distinguish positive from negative. We may *represent* every real number via an infinite decimal expansion, and while this is certainly an excellent way to think about and work with real numbers, it works poorly as a definition.

The modern approach is to lean on a certain collection of axioms for $\mathbb{R}$:

I. The field axioms: $\mathbb{R}$ is endowed with two binary operations $+$ and $\cdot$, satisfying many familiar properties like commutativity, associativity and so forth.

II. The order axioms: $\mathbb{R}$ is endowed with a total order relation $\leq$.

III. The ordered field axioms, which give compatibility between the field operations and the order structure:
(OF1) For all $x \in \mathbb{R}$, exactly one of the following holds: $x = 0$, $x > 0$, $-x > 0$.
(OF2) For all $x, y, z \in \mathbb{R}$, if $x \leq y$ then $x + z \leq y + z$.
(OF3) For all $x, y \in \mathbb{R}$, if $x \geq 0$ and $y \geq 0$, then $x \cdot y \geq 0$.

A structure that satisfies all of the properties so far is called an **ordered field**. There are in fact an enormous number of ordered fields: the rational numbers, $\mathbb{Q}$, is one. A **subfield** of $\mathbb{R}$ is a subset $F \subseteq \mathbb{R}$ that contains 0 and 1 and is closed under the field operations: if $x, y \in F$ then $x + y, x - y, x \cdot y \in F$ and $\frac{x}{y} \in F$ if $y \neq 0$. If we take any subfield of $\mathbb{R}$ and restrict the relation $\leq$ to F, then we get an ordered field. This builds *a lot* (infinitely many, to say the least!) of subfields of $\mathbb{R}$. And there are also more exotic ordered fields that do not arise in this way, although you will probably never meet any in an undergraduate course (including this one).

The real numbers is characterized among ordered fields by satisfying:

IV. The completeness axiom, which can be stated in several equivalent forms.

But first let me nail down what "characterized" means: first of all, the real numbers satisfy these completeness axioms. Second, an ordered field F that satisfies one of these completeness axioms is "essentially" the real numbers, which means that one can find a bijective function $f : F \to \mathbb{R}$ that preserves all of the structures: $+$, $\cdot$ and $\leq$. (More formally, f is an **isomorphism of ordered fields**.)

Now back to the completeness axioms. The most useful formulation is:

Dedekind's Completeness: If X is a subset of $\mathbb{R}$ that is nonempty and bounded above, then X has a least upper bound, or supremum, in $\mathbb{R}$.

Another version is **O'Connor's Completeness / Monotone Sequence Lemma**: Every bounded monotone sequence $\{x_n\}_{n=1}^{\infty}$ in $\mathbb{R}$ converges to a real number.

What do I mean by "versions" of completeness? I mean that it can be shown that if an ordered field satisfies Dedekind's Completeness than it also satisfies O'Connor's Completeness (this was an important result in Math 3100) and also conversely an ordered field that satisfies O'Connor's Completeness also satisfies Dedekind's Completeness (this was probably *not* covered in class in Math 3100 but see [**SS**]).

In any ordered field satisfying the completeness axioms, the following properties also hold:

Archimedean Property: For every real number x, there is a positive integer n with $n > x$.

Cauchy's Completeness: Every Cauchy sequence in $\mathbb{R}$ converges.

The rational numbers $\mathbb{Q}$ are an example of an ordered field that do not satisfy Cauchy's completeness. It turns out that an ordered field satisfies the Archimedean property iff it is (isomorphic to) a subfield of $\mathbb{R}$, so non-Archimedean ordered fields are precisely the ones we called "exotic" above. Moreover:

PROPOSITION 1.1. *An ordered field that satisfies the Achimedean Property* and *Cauchy's completeness is Dedekind complete – and thus isomorphic to* $\mathbb{R}$.

PROOF. See [**SS**, Proposition 2.6.7 and Theorem 2.6.13b)]. □

Before we move on to the material of our course, one remark / reminder: one *does* need to show that there is a Dedekind complete ordered field that is unique up to isomorphism; that is, we still need to "construct the real numbers $\mathbb{R}$." At least, someone does. The first such construction was given by Dedekind in the late 1800's. The truth of it is that no such construction is particularly simple, so that one needs a certain amount of mathematical sophistication to understand it...at which point it seems to be a better use of any instructor's time to cover something else. So it is extremely rare to encounter the construction $\mathbb{R}$ in a course. This course will be no exception. But if by chance you do want to see a construction of $\mathbb{R}$, it is written up in [**HC**, Chapter 16].

2. Convergence in Euclidean Space

Let $N \in \mathbb{Z}^+$. By $\mathbb{R}^N$ we mean the set of ordered N-tuples of real numbers

$$\mathbf{x} = (x_1, \ldots, x_N).$$

This is a familiar object from linear algebra, as a vector space over $\mathbb{R}$. This means that elements of $\mathbb{R}^N$ can be added to each other, and it also makes sense to "scale" an element $\mathbf{x}$ by a real number α:

$$\alpha(x_1, \ldots, x_N) \coloneqq (\alpha x_1, \ldots, \alpha x_n).$$

However, we are interested in $\mathbb{R}^N$ not just as a real vector space, but endowed with the **Euclidean norm**, which is a function from $\mathbb{R}^N$ to $[0, \infty)$, the non-negative real numbers. Specifically:

$$\forall\ \mathbf{x} = (x_1, \ldots, x_N) \in \mathbb{R}^N,\ ||\mathbf{x}|| \coloneqq \sqrt{x_1^2 + \ldots + x_N^2}.$$

We recall a *very* basic fact: for elements $x_1, \ldots, x_N$ in any ordered field F, we have

$$x_1^2 + \ldots + x_N^2 \geq 0,$$

and

$$x_1^2 + \ldots + x_N^2 = 0 \iff x_1 = \ldots = x_N = 0.$$

That is: a sum of squares is never negative, and is 0 iff every term is 0.

From this we deduce:

$$\forall\ \mathbf{x} \in \mathbb{R}^N,\ \mathbf{x} = 0 \iff ||\mathbf{x}|| = 0.$$

Here is another easy property of the Euclidean norm:

Proposition 1.2. *For all $\mathbf{x} \in \mathbb{R}^N$ and all $\alpha \in \mathbb{R}$, we have*

$$||\alpha\mathbf{x}|| = |\alpha|||\mathbf{x}||.$$

The proof of Proposition 1.2 is left as an exercise.

By **Euclidean N-space** I mean $\mathbb{R}^N$ equipped with its Euclidean norm. By the way, the Euclidean norm itself can be defined in terms of an **inner product** operation

$$\cdot : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R},$$

$$(x_1, \ldots, x_N) \cdot (y_1, \ldots, y_N) \coloneqq x_1 y_1 + \ldots + X_N Y_N.$$

Then:

$$\forall\ \mathbf{x} \in \mathbb{R}^N,\ ||\mathbf{x}|| = \sqrt{\mathbf{x} \cdot \mathbf{x}}.$$

Inner products are *extremely* important in certain branches of analysis, but I think they will only make a brief appearance in this course.

We use the Euclidean norm to measure distance between points in $\mathbb{R}^N$, namely: for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, we define the **Euclidean distance**

$$d(\mathbf{x}, \mathbf{y}) \coloneqq ||\mathbf{x} - \mathbf{y}||.$$

In mathematics, if we have a set X and a function $d : X \times X \to \mathbb{R}$, then in order to call d a "distance function" we usually required that it satisfy the following three properties:

(D1) (Positive Definiteness) For all $x, y \in X$ we have $d(x, y) \geq 0$, with equality iff $x = y$.
(D2) (Symmetry) For all $x, y \in X$ we have $d(x, y) = d(y, x)$.
(D3) (Triangle Inequality) For all $x, y, z \in X$ we have $d(x, z) \leq d(x, y) + d(y, z)$.

Shall we try to show that our Euclidean distance satisfies these three properties?

It starts out easily:

(D1): For $\mathbf{x},\mathbf{y} \in \mathbb{R}^N$, we have $d(\mathbf{x},\mathbf{y}) = ||\mathbf{x}-\mathbf{y}|| \geq 0$ because the norm of anything is at least 0, and moreover $||\mathbf{x}-\mathbf{y}|| = 0$ iff $\mathbf{x}-\mathbf{y} = 0$ iff $\mathbf{x} = \mathbf{y}$. No problem!

(D2) For $\mathbf{x},\mathbf{y} \in \mathbb{R}^N$, we have $d(\mathbf{x},\mathbf{y}) = ||\mathbf{x}-\mathbf{y}|| = ||-(\mathbf{y}-\mathbf{x})||$. Using Proposition 1.2 we have

$$||-(\mathbf{y}-\mathbf{x})|| = |-1|||\mathbf{y}-\mathbf{x}|| = ||\mathbf{y}-\mathbf{x}|| = d(\mathbf{y},\mathbf{x}).$$

Again, no problem.
(D3) We want to show:

$$\forall\ \mathbf{x},\mathbf{y},\mathbf{z} \in \mathbb{R}^N,\ ||\mathbf{x}-\mathbf{z}|| \leq ||\mathbf{x}-\mathbf{y}|| + ||\mathbf{y}-\mathbf{z}||. \tag{1}$$

Hmm. Well, I notice that $\mathbf{x}-\mathbf{z} = (\mathbf{x}-\mathbf{y}) + (\mathbf{y}-\mathbf{z})$, so if we put

$$\mathbf{A} := \mathbf{x}-\mathbf{y},\ \mathbf{B} := \mathbf{y}-\mathbf{z},$$

then we have $\mathbf{A},\mathbf{B} \in \mathbb{R}^N$ and we want to show

$$||\mathbf{A}+\mathbf{B}|| \leq ||\mathbf{A}|| + ||\mathbf{B}||.$$

In other words, we see that in order to show (D3) it suffices to show the slightly simpler statement:

$$\forall\ \mathbf{x},\mathbf{y} \in \mathbb{R}^N, ||\mathbf{x}+\mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||. \tag{2}$$

To show (2) we really need to do something, although there is more than one "something" that will work. The following approach is a good one in that it uses very little. The main step is to establish the following closely related result.

Theorem 1.3 (Cacuhy-Schwarz in $\mathbb{R}^N$). *For all $\mathbf{x},\mathbf{y} \in \mathbb{R}^N$, we have*

$$|\mathbf{x}\cdot\mathbf{y}| \leq ||\mathbf{x}||||\mathbf{y}||.$$

Proof. Write $\mathbf{x} = (x_1,\ldots,x_N)$, $\mathbf{y} = (y_1,\ldots,y_N)$. For non-negative real numbers X, Y we have $X \leq Y$ iff $X^2 \leq Y^2$, so it is equivalent to show

$$|\mathbf{x}\cdot\mathbf{y}|^2 \leq ||\mathbf{x}||^2||\mathbf{y}||^2.$$

Without any vector notation, what we want to show is:

$$(x_1y_1 + \ldots + x_ny_n)^2 \leq (x_1^2 + \ldots + x_n^2)(y_1^2 + \ldots + y_n^2).$$

Put

$$L := (x_1y_1 + \ldots + x_ny_n)^2$$

and

$$R := (x_1^2 + \ldots + x_n^2)(y_1^2 + \ldots + y_n^2),$$

so we want to show that $L \leq R$; it will certainly suffice to show $R - L \geq 0$. Now:

$$R = \sum_{i=1}^n x_i^2y_i^2 + \sum_{1\leq i\neq j\leq n} x_i^2y_j^2 = \sum_i x_i^2y_i^2 + \sum_{i<j} x_i^2y_j^2 + \sum_{i<j} x_j^2y_i^2,$$

while

$$L = \sum_{i=1}^n x_i^2y_i^2 + \sum_{1\leq i\neq j\leq n} x_iy_ix_jy_j = \sum_i x_i^2y_i^2 + 2\sum_{i<j} x_iy_ix_jy_j,$$

so

$$R - L = \sum_{i<j} x_i^2 y_j^2 - 2\sum_{i<j} x_i y_j x_j y_i + \sum_{i<j} x_j^2 y_i^2 = \sum_{i<j} (x_i y_j - x_j y_i)^2 \geq 0. \qquad \square$$

Using Theorem 1.3, it easy to prove (2), especially if we allow ourselves to use simple properties of inner products from Exercise 1.2. Indeed, let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$. We want to show that $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$. Again it suffices to show this after squaring both sides, so equivalently we want to show:

$$||\mathbf{x} + \mathbf{y}||^2 \leq (||\mathbf{x}|| + ||\mathbf{y}||)^2.$$

Now we have

$$||\mathbf{x} + \mathbf{y}||^2 = (\mathbf{x} + \mathbf{y}) \cdot (\mathbf{x} + \mathbf{y}) = (\mathbf{x} \cdot \mathbf{x}) + (\mathbf{x} \cdot \mathbf{y}) + (\mathbf{y} \cdot \mathbf{x}) + (\mathbf{y} \cdot \mathbf{y})$$

$$= ||\mathbf{x}||^2 + 2(\mathbf{x} \cdot \mathbf{y}) + ||\mathbf{y}||^2 \leq ||\mathbf{x}||^2 + 2|\mathbf{x} \cdot \mathbf{y}| + ||\mathbf{y}||^2 \overset{\text{CS}}{\leq} ||\mathbf{x}||^2 + 2||\mathbf{x}||||\mathbf{y}|| + ||\mathbf{y}||^2$$
$$= (||\mathbf{x}|| + ||\mathbf{y}||)^2.$$

It is important to know when equality holds in Cauchy-Schwarz or (this is very closely related) in the Triangle Inequality.

Corollary 1.4. *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$. The following are equivalent:*

(i) *The vectors $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent: that is, either $\mathbf{x} = 0$ or there is $\alpha \in \mathbb{R}$ such that $\mathbf{y} = \alpha\mathbf{x}$.*
(ii) *We have $|\mathbf{x} \cdot \mathbf{y}| = ||\mathbf{x}|| \cdot ||\mathbf{y}||$.*

Your are asked to prove Corollary 1.4 in Exercise 1.4.

2.1. Exercises.

General Comment: Most exercises will refer to $\mathbb{R}^N$. Here it should be understood that N is an *arbitrary* positive integer. That is, unless you are asked for an example, your solution should apply no matter what the value of N is.

Exercise 1.1. *Show: for all $\mathbf{x} \in \mathbb{R}^N$ and all $\alpha \in \mathbb{R}$, we have*

$$||\alpha\mathbf{x}|| = |\alpha|||\mathbf{x}||.$$

Exercise 1.2. *Let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^N$ and $\alpha \in \mathbb{R}$.*

a) *Show:* $\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x}$.
b) *Show:* $(\alpha\mathbf{x}) \cdot \mathbf{y} = \alpha(\mathbf{x} \cdot \mathbf{y})$.
c) *Show:* $(\mathbf{x} + \mathbf{y}) \cdot \mathbf{z} = (\mathbf{x} \cdot \mathbf{z}) + (\mathbf{y} \cdot \mathbf{z})$.

Exercise 1.3. *We showed that (2) implies (1). Show that conversely, (1) implies (2). Explicitly, suppose that:*

$$\forall\ \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^N,\ ||\mathbf{x} - \mathbf{z}|| \leq ||\mathbf{x} - \mathbf{y}|| + ||\mathbf{y} - \mathbf{z}||.$$

Show:

$$\forall\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^N,\ ||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||.$$

Exercise 1.4. *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$. Show that the following are equivalent:*

(i) *The vectors $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent: that is, either $\mathbf{x} = 0$ or there is $\alpha \in \mathbb{R}$ such that $\mathbf{y} = \alpha\mathbf{x}$.*
(ii) *We have $|\mathbf{x} \cdot \mathbf{y}| = ||\mathbf{x}|| \cdot ||\mathbf{y}||$.*

Exercise 1.5. *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$.*

a) *Suppose that* $||\mathbf{x}+\mathbf{y}|| = ||\mathbf{x}||+||\mathbf{y}||$. *Show:* $\mathbf{x}$ *and* $\mathbf{y}$ *are linearly dependent.*
b) *Find necessary and sufficient conditions for* $||\mathbf{x}+\mathbf{y}|| = ||\mathbf{x}|| + ||\mathbf{y}|$.

EXERCISE 1.6 (Reverse Triangle Inequality). *Show: for all* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$*, we have*

$$\big| ||\mathbf{x}|| - ||\mathbf{y}|| \big| \leq ||\mathbf{x} - \mathbf{y}||.$$

3. Sequences in $\mathbb{R}^N$

3.1. Sequences in a Set. Let X be any set. We have the notion of a sequence in X: informally, this is an infinite ordered list of elements of X:

$$x_1, x_2, \ldots, x_n, \ldots \quad \text{with } x_n \in X \ \forall n \in \mathbb{Z}^+.$$

This is formalized as a function $x_\bullet : \mathbb{Z}^+ \to X$; then we have $x_\bullet(n) = x_n$. For instance we could consider sequences in the set of real-or-made-up English words (such a thing consists of a finite string of letters from our alphabet, whether it is a valid English word or not), and then

$$b,\ bo,\ boo,\ booo,\ boooo \ldots \tag{3}$$

defines a sequence.

In this level of generality we can consider **subsequences**: to form a subsequence, we choose an infinite, strictly increasing sequence of positive integers

$$n_1 < n_2 < \ldots < n_k < \ldots$$

and then form the new sequence

$$x_{n_1}, x_{n_2}, \ldots, x_{n_k}, \ldots.$$

Again we can be a bit more formal: a strictly increasing sequence of positive integers corresponds to a strictly increasing function $n_\bullet : \mathbb{Z}^+ \to \mathbb{Z}^+$, and then to pass from the sequence $x_\bullet : \mathbb{Z}^+ \to X$ to the corresponding subsequence we form the composition of functions

$$x_\bullet \circ n_\bullet : k \mapsto n_k \mapsto x_{n_k}.$$

So for instance if we take $n_k = k^2$ for all k then in our above weird example we get the subsequence

$$b,\ booo, booooooooo, boooooooooooooooo, \ldots. \tag{4}$$

But we're not really cooking with gas here. Rather we'd like a notion of *convergence* of sequences, and for this one needs some kind of extra structure on our set: for our weird sequence (3) above, if you asked me whether it converges I can only look at you quizzically: we just haven't set up enough for that question to be meaningful.

3.2. Sequences in $\mathbb{R}^N$. Let me give the definition for convergence of sequences in $\mathbb{R}^N$, after a little motivation. First, for $N = 1$ we have seen this definition already: it is the single most important definition of Math 3100. If we have a sequence $\{x_n\}$ of real numbers, we say the sequence **converges** to a real number L if

$$\forall\ \epsilon > 0,\ \exists K \in \mathbb{Z}^+ \text{ such that } \forall n > K,\ |x_n - L| < \epsilon.$$

A sequence **converges** if it converges to some $L \in \mathbb{R}$; otherwise it **diverges**. One of the first things one shows is that if a sequence converges then its limit is *unique.*

Now let me rephrase this definition slightly. First, when $N = 1$ the Euclidean norm is precisely the absolute value, and thus $|x_n - L|$ is nothing else than the distance $d(x_n, L)$ between x_n and L...as is certainly familiar from Math 3100. Now I make the following observation:

- The sequence $\{x_n\}$ converges to L iff the sequence $d(x_n, L)$ converges to 0.

Indeed, if we write out the latter convergence statement, it is: for all $\epsilon > 0$, there is $N \in \mathbb{Z}^+$ such that for all $n > N$ we have $\big||x_n - L| - 0\big| < \epsilon$. But

$$\big||x_n - L| - 0\big| = |x_n - L|,$$

so this is the same as saying that $x_n \to L$.

Aha. So if $\{x_n\}$ is a sequence in $\mathbb{R}^N$ and $L \in \mathbb{R}^N$, we can (and do!) say that x_n converges to L – and write $x_n \to L$ – if $d(x_n, L) \to 0$. This is still an (ϵ, K) definition: spelling it out, we get that $x_n \to L$ means: for all $\epsilon > 0$, there is $K \in \mathbb{Z}^+$ such that for all $n > K$ we have $||x_n - L|| < \epsilon$.[1]

EXAMPLE 3.1. *Consider the sequence* $\mathbf{x}_n = (\frac{1}{n^2}, \frac{n+3}{n+4})$ *in* $\mathbb{R}^2$*. We will show that* $\mathbf{x}_n \to (0,1)$*.*

From Math 3100 we know how to show that $\frac{1}{n^2} \to 0$ *and* $\frac{n+3}{n+4} \to 1$*. Let's put those together to show that* $\mathbf{x}_n \to (0,1)$*. Let* $\epsilon > 0$*, and put*

$$K := \left\lceil \frac{\sqrt{2}}{\epsilon} \right\rceil.$$

Step 1: If $n > K$ *we have*

$$\left|\frac{1}{n^2} - 0\right| = \left|\frac{1}{n^2}\right| = \frac{1}{n^2} \leq \frac{1}{n} < \frac{1}{K_1} \leq \frac{1}{\lceil \frac{\sqrt{2}}{\epsilon} \rceil} < \frac{1}{\frac{\sqrt{2}}{\epsilon}} \leq \frac{\epsilon}{\sqrt{2}}.$$

Step 2: If $n > K$*, we also have*

$$\left|\frac{n+3}{n+4} - 1\right| = \frac{1}{n+4} < \frac{1}{n} \leq \frac{\epsilon}{\sqrt{2}}.$$

Step 3:Thus, if $n > K$ *then*

$$||\mathbf{x}_n - (0,1)|| = \sqrt{(\frac{1}{n^2} - 0)^2 + (\frac{n+3}{n+4} - 1)^2} < \sqrt{\left(\frac{\epsilon}{\sqrt{2}}\right)^2 + \left(\frac{\epsilon}{\sqrt{2}}\right)^2} = \epsilon.$$

So $\mathbf{x}_n \to (0,1)$*.*

There is a general moral to extract here. We will get back to this shortly.

THEOREM 1.5 (Familiar Facts About Convergence). *Let* $\{\mathbf{x}_n\}$ *and* $\{\mathbf{y}_n\}$ *be sequences in* $\mathbb{R}^N$*. Suppose that* $\mathbf{x}_n \to \mathbf{L} \in \mathbb{R}^N$ *and* $\mathbf{y}_n \to \mathbf{M} \in \mathbb{R}^N$*.*

a) *Let* $\alpha \in \mathbb{R}$*. Then* $\alpha\mathbf{x}_n \to \alpha\mathbf{L}$*.*
b) *We have* $\mathbf{x}_n + \mathbf{y}_n \to \mathbf{L} + \mathbf{M}$*.*
c) *Every subsequence* $\{\mathbf{x}_{n_k}\}$ *of* $\mathbf{x}_n$ *also converges to* $\mathbf{L}$*.*
d) *If* $\mathcal{P} \in \mathbb{R}^N$ *is such that* $\mathbf{x}_n \to \mathcal{P}$*, then* $\mathbf{L} = \mathcal{P}$*.*

[1] So you see that I could have just said that we're replacing the absolute value $|\cdot|$ by the Euclidean norm $||\cdot||$. But I have my reasons for phrasing it in terms of the distance function, as will become clear at the end of the course.

You are asked to prove each of these facts as exercises. Of course, this is mostly to get you to look back at the corresponding proofs for real sequences.

A subset $S \subseteq \mathbb{R}^N$ is **bounded** if there is $M \geq 0$ such that for all $x \in S$ we have $||x|| \leq M$. In other words, a subset is bounded if the distances of its elements from the origin are bounded above by a fixed real number. (Soon enough we will rephrase this by saying that S is contained in some closed ball centered at 0.)

We say that a sequence $\{\mathbf{x}_n\}$ in $\mathbb{R}^N$ is **bounded** if the set of terms $\{\mathbf{x}_n \mid n \in \mathbb{Z}^+\}$ is a bounded subset of $\mathbb{R}^N$. Here is one more familiar fact:

THEOREM 1.6. *Convergent sequences in $\mathbb{R}^N$ are bounded.*

PROOF. Suppose $\mathbf{x}_n \to L$. Then there is $K \in \mathbb{Z}^+$ such that for all $n > K$ we have $||\mathbf{x}_n - L|| \leq 1$. By the Reverse Triangle Inequality, we get:

$$\forall n > K, \ \Big| ||\mathbf{x}_n|| - ||L|| \Big| \leq ||\mathbf{x}_n - L|| \leq 1,$$

so

$$\forall n > K, ||\mathbf{x}_n|| \leq ||L|| + 1.$$

Now put

$$M := \max(||x_1||, \ldots, ||x_K||, ||L|| + 1).$$

Then for all $n \in \mathbb{Z}^+$ we have $||\mathbf{x}_n|| \leq M$, so $\{\mathbf{x}_n\}$ is bounded. □

There are some things that we did with real sequences that *do not* make sense for sequences in $\mathbb{R}^N$ for $N > 1$, namely:

• In $\mathbb{R}$ we can multiply sequences and show the analogue of Theorem 1.5 for products. In $\mathbb{R}^N$ we cannot in general multiply two vectors so as to get another vector. However, there are a few loopholes here:
(i) We can multiply vectors in $\mathbb{R}^2$. Indeed we can identify $\mathbb{R}^2$ with the complex numbers $\mathbb{C}$ and use the given multiplication.
(ii) We can multiply vectors in $\mathbb{R}^3$, using the cross product. This is a kind of weird multiplication operation (neither commutative nor associative), but it still exists
(iii) For all $N \in \mathbb{Z}^+$ we can multiply two elements of $\mathbb{R}^N$ to get an element of $\mathbb{R}$, using the scalar product.

It happens to be true that in all three cases, these products preserve convergence of sequences. The first two of these are explored in the exercises; we will prove the third a little later on.

• Whereas $\mathbb{R}$ comes equipped with an ordering, for $N > 1$ we do not have any (natural, useful) total ordering on $\mathbb{R}^N$. Thus the important notion of monotone sequence in $\mathbb{R}^N$ has no analogue in $\mathbb{R}^N$, although we could speak of monotonicity of the sequence of norms.

• In $\mathbb{R}$ we have the notion of diverging to $+\infty$ and also the notion of diverging to $-\infty$. For $N > 1$ we have something similar but less precise. Namely, a sequence $\mathbf{x}$ in $\mathbb{R}^N$ **diverges to infinity** if the real sequence $||\mathbf{x}_N||$ diverges to ∞.

3.3. The Secret to Convergence in $\mathbb{R}^N$. Look back at Example 3.1 of a convergent sequence in $\mathbb{R}^2$:

$$\mathbf{x}_n = \left(\frac{1}{n^2}, \frac{n+3}{n+4}\right).$$

Put $x_n = \frac{1}{n^2}$ and $y_n = \frac{n+3}{n+4}$. Then the sequence of x-components converges to 0 and the sequence of y-components converges to 1; having established this it took us only one more line to show that $(x_n, y_n) \to (0,1)$.

In fact this is a general phenomenon of convergence in $\mathbb{R}^N$! Namely, let $\{\mathbf{x}_n\}$ be a sequence in $\mathbb{R}^N$. For each $1 \leq i \leq N$, let $\mathbf{x}_{n,i}$ be the ith component of $\mathbf{x}_n$. Then the vector sequence $\{\mathbf{x}_n\}$ can be traded in for N different real sequences: $\{\mathbf{x}_{n,1}\}, \ldots, \{\mathbf{x}_{n,N}\}$. It turns out that the convergence of the vector sequence is equivalent to the convergence of all of the scalar sequences:

THEOREM 1.7. *Let $\{\mathbf{x}_n\}_{n=1}^{\infty}$ be a sequence in $\mathbb{R}^N$, and let $\mathbf{L} = (L_1, \ldots, L_N) \in \mathbb{R}^N$. Then the following are equivalent:*

(i) *The vector sequence $\mathbf{x}_n$ converges to $\mathbf{L}$.*
(ii) *For each $1 \leq i \leq N$, the real sequence $\{\mathbf{x}_{n,i}\}$ of ith components converges to L_i.*

PROOF. The key to this is the following relatively simple observation: let $\mathbf{x} = (x_1, \ldots, x_N) \in \mathbb{R}^N$. There is at least one $1 \leq I \leq N$ such that

$$\forall 1 \leq i \leq N, \ |x_i| \leq |x_I|.$$

Fix such an I. Then for each $1 \leq i \leq N$ we have

$$|x_i| = \sqrt{x_i^2} \leq \sqrt{x_1^2 + \ldots + x_N^2} = ||\mathbf{x}|| \leq \sqrt{x_I^2 + \ldots + x_I^2} = \sqrt{N x_I^2} = \sqrt{N}|x_I|.$$

This shows: if $||\mathbf{x}||$ is small, then so is the absolute value of each coordinate of $\mathbf{x}$ – in fact, each is no larger than $||\mathbf{x}||$ – and conversely, if all of the absolute values of the coordinates are small, then $||\mathbf{x}||$ is also small: at most $\sqrt{N}$ times as large as the largest coordinate absolute value. These inequalities imply that for any sequence $\{\mathbf{x}_n\}$ in $\mathbb{R}^N$ we have $\mathbf{x}_n \to 0$ if and only if $\mathbf{x}_{n,i} \to 0$ for all $1 \leq i \leq N$. The general case follows from this special case applied to the sequence $\{\mathbf{x}_n - \mathbf{L}\}$. $\square$

We extend the notion of **Cauchy sequence** to $\mathbb{R}^N$ in a straightforward way: a sequence $\{\mathbf{x}_n\}$ in $\mathbb{R}^N$ is Cauchy if for all $\epsilon > 0$, there is $K \in \mathbb{Z}^+$ such that for all $m, n \geq K$ we have $||\mathbf{x}_m - \mathbf{x}_n|| < \epsilon$. The same simple inequalities used in the proof of Theorem 1.7 also work to show:

THEOREM 1.8. *A sequence $\{\mathbf{x}_n\}$ in $\mathbb{R}^N$ is Cauchy if and only if for all $1 \leq i \leq N$, the real sequence $\{\mathbf{x}_{n,i}\}$ is Cauchy.*

We leave the details of this as an exercise. It follows that:

COROLLARY 1.9. *A sequence in $\mathbb{R}^N$ is convergent if and only if it is Cauchy.*

PROOF. We know the result for $N = 1$ from Math 3100. So by what we have just seen, the vector sequence is convergent iff each of its component scalar sequences is convergent iff each of its component scalar sequences is Cauchy iff the vector sequence is Cauchy. $\square$

This has the same advantage of knowing the equivalence of Cauchy sequences and convergent sequences in $\mathbb{R}$: it allows us to decouple the question of convergence of a sequence from the question of knowing the limit of the sequence; often the former questions is much easier than the latter.

3.4. Bolzano-Weierstrass in $\mathbb{R}^N$. The celebrated Bolzano-Weierstrass Theorem says that every bounded real sequence has a convergent subsequence. This extends verbatim to sequences in $\mathbb{R}^N$, as we will now show. Let us first give an equivalent formulation of boundedness of subsets of Euclidean N-space. Suppose we are given real numbers $a_1 \leq b_1, a_2 \leq b_2, \ldots, a_N \leq b_N$. To this data we associate the set

$$B(a_1, b_1, \ldots, a_N, b_N) \coloneqq \{\mathbf{x} = (x_1, \ldots, x_N) \coloneqq \forall 1 \leq i \leq N, a_i \leq x_i \leq b_i\}.$$

A set $B(a_1, b_1, \ldots, a_N, b_N)$ is called a **closed box**. A closed box in $\mathbb{R}$ is simply a closed bounded interval. A closed box in $\mathbb{R}^2$ is a rectangle (together with its interior) whose edges are parallel to the coordinate axes. And so forth. Now:

LEMMA 1.10. *A subset $X \subseteq \mathbb{R}^N$ is bounded if and only if X is contained in some closed box. In particular, all closed boxes are bounded subsets of $\mathbb{R}^N$.*

We leave the proof of Lemma 1.10 as an exercise.

THEOREM 1.11 (Bolzano-Weierstrass in $\mathbb{R}^N$). *Every bounded sequence in $\mathbb{R}^N$ has a convergent subsequence.*

PROOF. Let $\{\mathbf{x}_n\}$ be a bounded sequence in $\mathbb{R}^N$. By Lemma1.10 there are real numbers $a_1 \leq b_1, \ldots, a_N \leq b_N$ such that every term $\mathbf{x}_n$ of the sequence lies in the box $B(a_1, b_1, \ldots, a_N, b_N)$.
Step 1: The sequence $\{\mathbf{x}_{n,1}\}$ of first coordinates lies in the interval $[a_1, b_1]$, so by Bolzano-Weierstrass in $\mathbb{R}$ it has a subsequence that converges to $L_1 \in \mathbb{R}$.

Interregnum: We now have a purely notational pitfall to avoid: we are going to be passing to subsequences quite a lot of times, so if we actually write this out using double index notation then in Step 2 we are going to get triple indices, in Step 3 quadruple indices, and so forth: it will be a terrible mess. So we will just *remember* that we passed to a subsequence so as to make the sequence of first coordinates converge.
Step 2: The sequence (which is actually a subsequence of our original sequence) $\mathbf{x}_{n,2}$ of second coordinates lies in the interval $[a_2, b_2]$, so by Bolzano-Weierstrass in $\mathbb{R}$ it has a subsequence that converge to $L_2 \in \mathbb{R}$. What happens with the sequence of first coordinates when we do this? Fortunately, if a sequence converges to a limit then every subsequence converges to the same limit, so passing to this second subsequence does not screw up what we did in Step 1: after two steps we have passed to a subsubsequence – which is still a subsequence! – of the original sequence so as to make each of the first two component real sequences converge.
Steps 3 to N: We move on to the bounded sequence of third components, apply Bolzano-Weierstrass again, and so forth. After N steps we have passed to a subsequence N times altogether to get a sequence in which each of the component sequences converge, hence by Theorem 1.7 the subsub.....subsequence converges. Passing from a sequence to a subsequence any finite number of times still yields a subsequence of the original sequence, so...we're done. $\square$

A point $L \in \mathbb{R}^N$ is a **partial limit** of a sequence $\{\mathbf{x}_n\}$ in $\mathbb{R}^N$ if there is some subsequence $\mathbf{x}_{n_k} \to L$. Thus Theorem 1.11 can be rephrased as: every bounded sequence in $\mathbb{R}^N$ has at least one partial limit. An unbounded sequence may or may not have a partial limit; a divergent sequence may have more than one partial limit, and it can be interesting to contemplate the set of all partial limits of a sequence. This is pursued in the exercises.

3.5. Exercises.

EXERCISE 1.7. *Let $\{\mathbf{x}_n\}$ be a sequence in $\mathbb{R}^N$ such that $\mathbf{x}_n \to L$, and let $\alpha \in \mathbb{R}$. Show: $\alpha\mathbf{x}_n \to \alpha L$.*

EXERCISE 1.8. *Let $\{\mathbf{x}_n\}$, $\{\mathbf{y}_n\}$ be sequences in $\mathbb{R}^N$. Suppose that $\mathbf{x}_n \to L$ and $\mathbf{y}_n \to M$. Show: $\mathbf{x}_n + \mathbf{y}_n \to L + M$.*

EXERCISE 1.9. *Let $\{\mathbf{x}_n\}$ be a sequence in $\mathbb{R}^N$. Show: if $\mathbf{x}_n \to \mathbf{L}$, then every subsequence $\{\mathbf{x}_{n_k}\}$ also converges to $\mathbf{L}$.*

EXERCISE 1.10. *Let $\{\mathbf{x}_n\}$ be a sequence in $\mathbb{R}^N$, and let $\mathbf{L}, \mathcal{P} \in \mathbb{R}^N$. Suppose that $\mathbf{x}_n \to \mathbf{L}$ and $\mathbf{x}_n \to \mathcal{P}$. Show: $\mathbf{L} = \mathcal{P}$.*

EXERCISE 1.11. *Show: every finite subset of $\mathbb{R}^N$ is bounded.*

EXERCISE 1.12. *The set $\mathbb{C} = \{x + iy \mid x, y \in \mathbb{R}\}$ has a nice multiplication operation:*

$$(x + iy)(z + iw) = (xz - yw) + (xw + zy)i.$$

If we identify the vector (x, y) with the complex number $x + iy$, this gives a multiplication operation on $\mathbb{R}^2$:

$$(x, y) \cdot (z, w) := (xz - yw, xw + zy).$$

Show: if $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$ are two sequences in $\mathbb{R}^2$ such that $\mathbf{x}_n \to L$ and $\mathbf{y}_n \to M$, then $\mathbf{x}_n \cdot \mathbf{y}_n \to L \cdot M$.

EXERCISE 1.13. *Let $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^3$, and let $\mathbf{x} \times \mathbf{y} \in \mathbb{R}^3$ be the cross product. Show: if $\mathbf{x}_N \to L$ and $\mathbf{y}_N \to M$ then $\mathbf{x}_n \times \mathbf{y}_n \to L \times M$.*

EXERCISE 1.14. *Let $\{\mathbf{x}_n\}$ be a sequence in $\mathbb{R}^N$.*

a) *Show: if $\{\mathbf{x}_n\}$ is bounded, so is every subsequence.*
b) *Show: $\{\mathbf{x}_n\}$ is unbounded iff some subsequence of $\{\mathbf{x}_n\}$ diverges to ∞.*

EXERCISE 1.15. *A sequence $\{\mathbf{x}_n\}$ in $\mathbb{R}^N$ is Cauchy if and only if for all $1 \leq i \leq N$, the real sequence $\{\mathbf{x}_{n,i}\}$ is Cauchy.*

EXERCISE 1.16. *Suppose we are given real numbers $a_1 \leq b_1, a_2 \leq b_2, \ldots, a_N \leq b_N$. To this data we associate the set*

$$B(a_1, b_1, \ldots, a_N, b_N) := \{\mathbf{x} = (x_1, \ldots, x_N) := \forall 1 \leq i \leq N, a_i \leq x_i \leq b_i\}.$$

*A set of the form $B(a_1, b_1, \ldots, a_N, B_N)$ is called a **closed box**.*
Show: a subset $X \subseteq \mathbb{R}^N$ is bounded iff X is contained in some closed box.

EXERCISE 1.17. *Let $1 \leq i \leq N$. We define the **ith coordinate projection map***

$$\pi_i : \mathbb{R}^N \to \mathbb{R},\ (x_1, \ldots, x_N) \mapsto x_i.$$

Show: a subset $X \subseteq \mathbb{R}^N$ is bounded iff for all $1 \leq i \leq N$, the subset $\pi_i(X)$ is a bounded subset of $\mathbb{R}$.

EXERCISE 1.18. *For a sequence $\{\mathbf{x}_n\}$ in $\mathbb{R}^N$, show that the following are equivalent:*

(i) *The sequence diverges to ∞ (recall this means that $||\mathbf{x}_n|| \to +\infty$).*
(ii) *Every subsequence of $\mathbf{x}_n$ is unbounded.*
(iii) *The sequence $\{\mathbf{x}_n\}$ has no partial limit.*

COMMENT: If we wished, we could define the "extended Euclidean space"

$$\widetilde{\mathbb{R}^N} := \mathbb{R}^N \cup \{\infty\}$$

and say that ∞ is a partial limit of a sequence $\{\mathbf{x}_n\}$ in $\mathbb{R}^N$ iff some subsequence diverges to ∞. With this convention, combining Exercises 1.14 and 1.18 we would have that *every* sequence in $\mathbb{R}^N$ has at least one partial limit in $\widetilde{\mathbb{R}^N}$. We will not adopt this definition simply because we will not use it further in this course.

EXERCISE 1.19. *By Bolzano-Weierstrass, every bounded sequence in $\mathbb{R}^N$ has at least one partial limit. Show that a bounded sequence converges iff it has* exactly *one partial limit.*

EXERCISE 1.20. *Find a sequence in $\mathbb{R}^N$ that has* every *element of $\mathbb{R}^N$ as a partial limit.*

4. Topology of $\mathbb{R}^N$

4.1. Open and Closed Sets. Let $\mathbf{x} \in \mathbb{R}^N$ and $r > 0$. We define the **open ball centered at x with radius** r to be

$$B^\circ(\mathbf{x}, r) := \{\mathbf{y} \in \mathbb{R}^N \mid ||\mathbf{x} - \mathbf{y}|| < r\}$$

and the **closed ball centered at x with radius** r to be

$$B^\bullet(\mathbf{x}, r) := \{\mathbf{y} \in \mathbb{R}^N \mid ||\mathbf{x} - \mathbf{y}|| \le r\}.$$

Thus $B^\circ(\mathbf{x}, r)$ consists of all points of $\mathbb{R}^N$ whose distance from $\mathbf{x}$ – called the **center** of the ball – is less than r, and the same goes for $B^\bullet(\mathbf{x}, r)$ except that now the distance is less than *or equal to* r.

The latter concept has actually arisen already: to see this, notice that our definition of a subset of $\mathbb{R}^N$ being bounded is precisely that it is contained in $B^\bullet(0, M)$ for some M. Notice also that open and closed balls are always bounded sets: indeed, for any $\mathbf{x} \in \mathbb{R}^N$ and $r > 0$, the Triangle Inequality gives:

$$B^\circ(\mathbf{x}, r) \subseteq B^\bullet(\mathbf{x}, r) \subseteq B^\bullet(0, ||\mathbf{x}|| + r).$$

In $\mathbb{R}^1$ balls are not very interesting: you are asked to show as an exercise that a subset of $\mathbb{R}$ is an open ball if and only if it is a bounded open interval and that a subset of $\mathbb{R}$ is a closed ball if and only if it is a bounded closed interval. (This is actually one reason why we want to work in $\mathbb{R}^N$ at the beginning of the course: the topological concepts we want to deal with are not trivial when we restrict to the one variable case, but they are "geometrically degenerate" in a way that may hamper intuition.)

We say that a subset $U \subseteq \mathbb{R}^N$ is **open** if for every $\mathbf{x} \in U$, there is $\epsilon > 0$ such that $B^\circ(\mathbf{x}, \epsilon) \subseteq U$. In other words, a set is open if whenever it contains a point $\mathbf{x}$ it also contains all points of $\mathbb{R}^N$ that are sufficiently close to $\mathbf{x}$.

The terminology suggests than an open ball should itself be an open set, but we had better prove that.

PROPOSITION 1.12. *Every open ball is an open subset of* $\mathbb{R}^N$.

PROOF. Consider the open ball $B^\circ(\mathbf{x}, r)$ and a point $\mathbf{y}$ in it. We need to find a (smaller!) ball centered at $\mathbf{y}$ that is entirely contained in the first ball. The question really is: how do we choose $\epsilon > 0$ such that

$$B^\circ(\mathbf{y}, \epsilon) \subseteq B^\circ(\mathbf{x}, r)?$$

Let's try to work it out: if $\mathbf{z} \in B^\circ(\mathbf{y}, \epsilon)$, then

$$d(\mathbf{y}, \mathbf{z})) = ||\mathbf{z} - \mathbf{y}|| < \epsilon,$$

so by the Triangle Inequality we have

$$d(\mathbf{x}, \mathbf{z}) < \epsilon + d(\mathbf{x}, \mathbf{y}).$$

So the point $\mathbf{z}$ will lie in $B^\circ(\mathbf{x}, r)$ provided that $\epsilon + d(\mathbf{x}, \mathbf{y}) \leq r$. Thus we can take

$$\epsilon = r - d(\mathbf{x}, \mathbf{y}),$$

which is indeed positive: since $\mathbf{y} \in B^\circ(\mathbf{x}, r)$, we have $d(\mathbf{x}, \mathbf{y}) < r$. □

Non-examples are as helpful as understanding new concepts as examples, so:

PROPOSITION 1.13. *No closed ball is an open subset of* $\mathbb{R}^N$.

PROOF. Consider $B := B^\bullet(\mathbf{x}, r)$. Let $p := \mathbf{x} + (r, 0, \ldots, 0)$ be the rightmost point on the ball. Then any open ball $B^\circ(p, \epsilon)$ contains the point $p + (\frac{\epsilon}{2}, 0, \ldots, 0) = \mathbf{x} + (r + \frac{\epsilon}{2}, 0, \ldots, 0)$. This point has distance $r + \frac{\epsilon}{2}$ from $\mathbf{x}$ so does not lie in B. □

Let A be a subset of $\mathbb{R}^N$. We say that $L \in \mathbb{R}^N$ is a **limit point** of A if there is a sequence $\{\mathbf{x}_n\}$ in A such that $\mathbf{x}_n \to L$.

Every point $L \in A$ is a a limit point of A, because we can take the constant sequence $L, L, L, \ldots$. (If this feels like cheating...good! You are probably grasping for the related concept of **accumulation point**, which is coming up soon.)

EXAMPLE 4.1. *We claim that every point of the closed ball* $B^\bullet(\mathbf{x}, r)$ *is a limit point of the corresponding open ball* $B^\circ(\mathbf{x}, r)$. *Indeed, we need only look at points* $p \in B^\bullet(\mathbf{x}, r) \setminus B^\circ(\mathbf{x}, r)$, *i.e., points* p *whose distance from* $\mathbf{x}$ *is exactly* r. *Then take*

$$\mathbf{x}_n = \mathbf{x} + (1 - \frac{1}{n})(p - \mathbf{x}).$$

Then

$$d(\mathbf{x}_n, x) = (1 - \frac{1}{n})||p - \mathbf{x}|| = (1 - \frac{1}{n})r < r,$$

so $\mathbf{x}_n \in B^\circ(\mathbf{x}, r)$. *And* $\lim_{n \to \infty} \mathbf{x}_n = \mathbf{x} + (p - \mathbf{x}) = p$.

This example motivates the second key definition of this section: a subset A of $\mathbb{R}^N$ is **closed** if every limit point of A is an element of A. Another way of saying this is that A is *closed* under taking limits of convergent sequences.

A basic fact in Math 3100 is that limits of sequences preserve non-strict inequalities: that is, if we every term of a convergent sequence is at least a, then the limit is also at least a, and if every term of a convergent sequence is at most b, then the limit is also at most b. This means precisely that the closed interval $[a, b]$ is a closed subset of $\mathbb{R}$. Recalling that these are precisely the closed balls in $\mathbb{R}$, we get that every closed ball in $\mathbb{R}^1$ is a closed subset of $\mathbb{R}^1$.

We would like to extend this to $\mathbb{R}^N$: let's try. Consider a closed ball $B^\bullet(\mathbf{x}, r)$ in $\mathbb{R}^N$. Seeking a contradiction, suppose that there is some $L \in \mathbb{R}^N \setminus B^\bullet(\mathbf{x}, r)$ and a sequence $\{\mathbf{x}_n\}$ in $B^\bullet(\mathbf{x}, r)$ such that $\mathbf{x}_N \to L$. Let

$$d := d(\mathbf{x}, L)$$

be the distance from the limit point to the center of the ball. Our assumption is that $d > r$. Take $\epsilon := d - r$. I claim that

$$B^\circ(L, \epsilon) \cap B^\bullet(\mathbf{x}, r) = \varnothing.$$

Indeed, if $y \in B^\circ(L, \epsilon) \cap B^\bullet(\mathbf{x}, r)$ then

$$d = d(\mathbf{x}, L) \leq d(\mathbf{x}, y) + d(y, L) < r + \epsilon = d.$$

That's a contradiction. But if we had a sequence in $B^\bullet(\mathbf{x}, r)$ converging to L then sufficiently large terms of the sequence will give elements of $B^\bullet(\mathbf{x}, r)$ that are less than ϵ away from L, so there is no such sequence. Therefore closed balls are closed.

If we look back this proof, we really showed that for every point L of the complement $\mathbb{R}^N \setminus B^\bullet(\mathbf{x}, r)$, there is an open ball centered at L and contained in the complement. In other words, we showed that $B^\bullet(\mathbf{x}, r)$ is closed essentially by showing that its complement was open. This is true in general, very important, and not so difficult to prove.

Theorem 1.14. *A subset $A \subseteq \mathbb{R}^N$ is closed iff its complement $\mathbb{R}^N \setminus A$ is open.*

You are asked to prove Theorem 1.14 as an exercise.

4.2. Continuous Functions. Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is continuous at a point $c \in \mathbb{R}$ if: for all $\epsilon > 0$, there is $\delta > 0$ such that for all $x \in \mathbb{R}$, if $|x - c| < \delta$ then $|f(x) - f(c)| < \epsilon$. A function $f : \mathbb{R} \to \mathbb{R}$ is continuous if it is continuous at every $c \in \mathbb{R}$.

If we have a function f defined not on all of $\mathbb{R}$ but only on some subset A, then we used the same definition as above but with one reasonable change: $f : A \to \mathbb{R}$ is continuous at $c \in A$ if for all $\epsilon > 0$ there is $\delta > 0$ such that for all $x \in A$, if $|x - c| < \delta$ then $|f(x) - f(c)| < \epsilon$.

If we observe that $|f(x) - f(c)| < \epsilon$ means that $d(f(x), f(c)) < \epsilon$ and $|x - c| < \delta$ means $d(x, c) < \delta$, it should be pretty clear how to generalize this definition to maps between Euclidean spaces. Again, let's do it in two steps. First suppose that M, N are positive integers and we have

$$f : \mathbb{R}^N \to \mathbb{R}^M.$$

We say that f is continuous at $c \in \mathbb{R}^N$ if for all $\epsilon > 0$, there is $\delta > 0$ such that for all $\mathbf{x} \in \mathbb{R}^N$, if $d(\mathbf{x}, c) = ||\mathbf{x} - c|| < \delta$ then $d(f(x), f(c)) = ||f(\mathbf{x}) - f(c)|| < \epsilon$. We say f is continuous if it is continuous at every $c \in \mathbb{R}^N$.

And again it is no problem to make a more general definition: if A is a subset of $\mathbb{R}^N$ and $f : A \to \mathbb{R}^M$ is a function, then f is continuous at $c \in \mathbb{R}^N$ if for all $\epsilon > 0$ there is $\delta > 0$ such that for all $\mathbf{x} \in A$, if $d(\mathbf{x}, c) = ||\mathbf{x} - c|| < \delta$ then $||f(\mathbf{x}) - f(c)|| < \epsilon$.

Let us rephrase this definition in terms of open balls. A function $f : A \to \mathbb{R}^N$ is continuous at $\mathbf{c} \in A$ if for all $\epsilon > 0$, there is some $\delta > 0$ such that f maps

$A \cap B^\circ(\mathbf{c}, \delta)$ into $B^\circ(f(\mathbf{c}), \epsilon)$.

The following is an extension of an important result from Math 3100: continuous functions are characterized by their preservation of limits of convergent sequences.

THEOREM 1.15. *Let $X \subseteq \mathbb{R}^N$, and let $f : X \to \mathbb{R}^M$ be a function. Let $\mathbf{c} \in X$. The following are equivalent:*

(i) *f is continuous at $\mathbf{c}$.*
(ii) *For every sequence $\{\mathbf{x}_n\}$ in X such that $\mathbf{x}_n \to \mathbf{c}$, we have $f(\mathbf{x}_n) \to f(\mathbf{c})$.*

PROOF. (Compare this to [**SS**, Theorem 2.7.5]: it's virtually identical.)
(i) $\implies$ (ii): Fix $\epsilon > 0$. Because f is continuous at $\mathbf{c}$, there is $\delta > 0$ such that for all $\mathbf{x} \in X$ with $||\mathbf{x} - \mathbf{c}|| < \delta$, we have $||f(\mathbf{x}) - f(\mathbf{c})|| < \epsilon$. Because $\mathbf{x}_n \to c$, there is $K \in \mathbb{Z}^+$ such that for all $n > K$ we have $||\mathbf{x}_n - \mathbf{c}|| < \delta$. Thus for all $n > K$ we have $||f(\mathbf{x}_n) - f(\mathbf{c})|| < \epsilon$.
(ii) $\implies$ (i): We will prove the contrapositive: suppose f is *not* continuous at $\mathbf{c}$. Then there is $\epsilon > 0$ such that for all $\delta > 0$ there is $\mathbf{x} \in X$ with $||\mathbf{x} - \mathbf{c}|| < \delta$ and $||f(\mathbf{x}) - f(\mathbf{c})| \geq \epsilon$. For $n \in \mathbb{Z}^+$, taking $\delta = \frac{1}{n}$ gives $\mathbf{x}_n \in X$ such that $||\mathbf{x}_n - \mathbf{c}|| < \frac{1}{n}$ and $||f(\mathbf{x}_n) - f(\mathbf{c})|| \geq \epsilon$. Thus $\mathbf{x}_n \to \mathbf{c}$, but $f(\mathbf{x}_n)$ *does not converge* to $f(\mathbf{c})$. $\square$

4.3. New Continuous Functions From Old. Let us now discuss some ways of building new continuous functions out of old continuous functions. We can start with the real ground floor:

PROPOSITION 1.16. *Let $X \subset \mathbb{R}^N$, and let $f : X \to \mathbb{R}^M$ be a constant function: for all $\mathbf{x}, \mathbf{y} \in X$ we have $f(\mathbf{x}) = f(\mathbf{y})$. Then f is continuous.*

PROOF. Indeed for any $\epsilon > 0$ we may take *any* positive value of δ we like, since in fact for *any* $\mathbf{x}, \mathbf{y} \in X$ we have $||f(\mathbf{x}) - f(\mathbf{y})|| = ||0|| = 0 < \epsilon$. $\square$

After constant functions, perhaps the simplest functions $f : \mathbb{R}^N \to \mathbb{R}$ are the **coordinate functions** or **coordinate projections**: for $1 \leq i \leq N$, put

$$\pi_i : \mathbb{R}^N \to \mathbb{R} \text{ by } (x_1, \ldots, x_N) \mapsto x_i.$$

This is pretty fancy/careful notation. In practice we will often speak of "the function x_i". It is quite easy to see that these functions are continuous: indeed, let $\mathbf{x} \in \mathbb{R}^N$, and fix $\epsilon > 0$. Then for $\mathbf{y} \in \mathbb{R}^N$, we have

$$|x_i - y_i| \leq \sqrt{(x_1 - y_1)^2 + \ldots + (x_N - y_N)^2} = ||\mathbf{x} - \mathbf{y}||,$$

so if $||\mathbf{x} - \mathbf{y}|| < \epsilon$ then also $|x_i - y_i| < \epsilon$, so we may take $\delta = \epsilon$.

Here is one use of the coordinate projections: let $X \subseteq \mathbb{R}^N$ and let $f : X \to \mathbb{R}^M$ be a function. Then for all $x \in X$, we have

$$f(x) = (\pi_1(f(x)), \ldots, \pi_M(f(x)).$$

The notation may momentarily obscure this unprofound identity: we are just reassembling the components of the vector-valued function f. Now we have:

PROPOSITION 1.17. *For a function $f : X \subseteq \mathbb{R}^N \to \mathbb{R}^M$ and $\mathbf{x} \in X$, the following are equivalent:*

(i) *The function f is continuous at $\mathbf{x}$.*
(ii) *Each of the functions $f_1, \ldots, f_M$ is continuous at $\mathbf{x}$.*

The proof uses the same idea as Theorem 1.7 – a vector has small norm if and only if each of its components has small absolute value – and is left as an exercise.

PROPOSITION 1.18. *Let $f_1, \ldots, f_M : X \subseteq \mathbb{R}^N \to \mathbb{R}$, and let $\mathbf{x} \in X$. If each of $f_1, \ldots, f_M$ are continuous at $\mathbf{x}$, then so are $\sum_{i=1}^M f_i$ and $\prod_{i=1}^M f_i$.*

PROOF. Let's use Theorem 1.15: let $\{\mathbf{x}_n\}$ be a sequence in X that converges to $\mathbf{x}$. Since each $f_1, \ldots, f_M$ is continuous at $\mathbf{x}$ we have $f_i(\mathbf{x}_n) \to f_i(\mathbf{x})$ as sequences in $\mathbb{R}$. By the extension Theorem 1.5b) from 2 sequences to M sequences (a completely routine induction argument does this) we know that $f_1(\mathbf{x}_n) + \ldots + f_M(\mathbf{x}_n) \to f_1(\mathbf{x}) + \ldots + f_M(\mathbf{x})$, and applying Theorem 1.15 once more we get that $\sum_{i=1}^M f_i$ is continuous at $\mathbf{x}$. The argument for $\prod_{i=1}^M f_i$ except we use the fact that for real sequences we have $x_n \to L$ and $y_n \to M$ implies $x_n y_n \to LM$ [**SS**, Theorem 2.5.4b)] (and again, its evident extension from 2 sequences to M sequences). $\square$

A function $f : \mathbb{R}^N \to \mathbb{R}$ is a **polynomial** if it is built up out of constant functions and coordinate functions by (finitely!) repeated addition and multiplication. Thus for instance $xyz + 17y^5 - \pi x^2 y^2 z^2$ is a polynomial function. It follows from Propositions 1.16, 1.17 and 1.18 that polynomial functions are continuous. In particular:

COROLLARY 1.19. *The inner product map $\mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$, $(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{x} \cdot \mathbf{y}$ is continuous.*

PROOF. We may identify $\mathbb{R}^N \times \mathbb{R}^N$ with $\mathbb{R}^{2N}$ and then the inner product map is $(x_1, \ldots, x_{2N}) \mapsto x_1 x_{N+1} + x_2 x_{N+2} + \ldots + x_N x_{2N}$. This is a polynomial function, so it is continuous. $\square$

PROPOSITION 1.20. *Let $M, N, P \in \mathbb{Z}^+$. Let $X \subseteq \mathbb{R}^N$ and $Y \subseteq \mathbb{R}^M$. Let $f : X \to \mathbb{R}^M$ and $g : Y \to \mathbb{R}^P$ be functions. Suppose that $f(X) \subseteq Y$, so that the composition $g \circ f$ is defined.*

a) *Let $\mathbf{x} \in X$. If f is continuous at $\mathbf{x}$ and g is continuous at $f(\mathbf{x})$, then $g \circ f$ is continuous at $\mathbf{x}$.*
b) *If f and g are both continuous, so is $g \circ f$.*

PROOF. a) Let $\epsilon > 0$. Since g is continuous at $f(\mathbf{x})$ there is $D > 0$ such that if $\mathbf{w} \in Y$ is such that $||\mathbf{w} - f(\mathbf{x})|| < D$, then $||g(\mathbf{w}) - g(f(\mathbf{x}))|| < \epsilon$. Since f is continuous at $\mathbf{x}$, there is $\delta > 0$ such that if $\mathbf{z} \in X$ is such that $||\mathbf{z} - \mathbf{x}| < \delta$, then $||f(\mathbf{z}) - f(\mathbf{x})|| < D$. So altogether, if $\mathbf{z} \in X$ is such that $||\mathbf{z} - \mathbf{x}|| < \delta$, then $||f(\mathbf{z}) - f(\mathbf{x})|| < D$, so $||g(f(\mathbf{z})) - g(f(\mathbf{x}))|| < \epsilon$, so $g \circ f$ is continuous at $\mathbf{x}$.
b) This follows immediately. $\square$

If we assume as known that the function $\sqrt{x} : [0, \infty) \to \mathbb{R}$ is continuous (actually we will discuss inverses of continuous functions of one variable later on in the course), then we can also prove that the norm function

$$|| \cdot || : \mathbb{R}^N \to \mathbb{R}$$

is continuous: indeed, it is the composition of the polynomial function $x_1^2 + \ldots x_N^2$ with the square root function. Similarly, the Euclidean distance function

$$d : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}, \ (x_1, \ldots, x_N, y_1, \ldots, y_N) \mapsto \sqrt{(x_1 - y_1)^2 + \ldots + (x_N - y_N)^2}$$

is the composition of the polynomial function $(\mathbf{x} - \mathbf{y}) \cdot (\mathbf{x} - \mathbf{y})$ with the square root function, hence is continuous.

As one more application of these ideas, we will prove:

PROPOSITION 1.21. *The addition function* $+ : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^N$ *is continuous.*

PROOF. Again, we may identify $\mathbb{R}^N \times \mathbb{R}^N$ with $\mathbb{R}^{2N}$, and then we are trying to show that the function

$$(x_1, \ldots, x_N, x_{N+1}, \ldots, x_{2N}) \mapsto (x_1 + x_{N+1}, \ldots, x_N + x_{2N})$$

is continuous. By Proposition 1.17 it's enough to show that each component is continuous. But the ith component is $x_i + x_{N+i}$, which is a polynomial. $\square$

4.4. Sequential Compactness in $\mathbb{R}^N$. A subset $X \subseteq \mathbb{R}^N$ is **sequentially compact** if every sequence in X has a subsequence that converges to some $L \in X$.

PROPOSITION 1.22. *Let* $X \subset \mathbb{R}^N$ *be a sequentially compact, and let* $f : X \to \mathbb{R}^M$ *be a continuous function. Then the image* $f(X)$ *is sequentially compact.*

PROOF. Let $\{\mathbf{y}_n\}$ be a sequence in $f(X)$. By definition of the image, every element of $f(X)$ is of the form $f(x)$ for some $x \in X$, so for each $n \in \mathbb{Z}^+$ we may choose $\mathbf{x}_n \in X$ such that $f(\mathbf{x}_n) = \mathbf{y}_n$. Because X is sequentially compact, there is some subsequence $\{\mathbf{x}_{n_k}\}$ that converges to an element $\mathbf{x}$ of X. By Theorem 1.15 we have

$$y_{n_k} = f(\mathbf{x}_{n_k}) \to f(\mathbf{x}).$$

Since $f(\mathbf{x}) \in f(X)$, this shows that $f(X)$ is sequentially compact. $\square$

The following is actually quite a big theorem.

THEOREM 1.23. *A subset of* $\mathbb{R}^N$ *is sequentially compact iff it is closed and bounded.*

PROOF. Step 1: We show that sequentially compact sets are both closed and bounded. We do this contrapositively.

First suppose that X is not closed. Then there is a sequence $\{\mathbf{x}_n\}$ in X that converges to an element $L \in \mathbb{R}^N \setminus X$. Because every subsequence of a convergent sequence converges *to the same limit*, whatever subsequence we take will still be convergent but the limit will lie outside of X, so X is not sequentially compact.

Now suppose that X is not bounded. We will produce a sequence in X no subsequence of which is convergent. Indeed, since X is not bounded, for all $n \in \mathbb{Z}^+$ there is $\mathbf{x}_n \in X$ with $||x_n|| \geq n$. Such a sequence is unbounded, hence divergent. Moreover, passing to a subsequence $\{x_{n_k}\}$ is no help: $||x_{n_k}|| \geq n_k \geq k$, so every subsequence is unbounded. (In other words, this sequence diverges to ∞, hence so does every subsequence.) So X is not sequentially compact.

Step 2: Suppose X is closed and bounded. Let $\{\mathbf{x}_n\}$ be a sequence in X. Since X is bounded, by Bolzano-Weierstrass, there is a subsequence that converges to some $L \in \mathbb{R}^N$. Since X is closed, we have $L \in X$. So X is sequentially compact. $\square$

At this point you're probably thinking: "Hey, I'm not impressed with sequential compactness because it turns out to be a fancy way to say closed and bounded." Let me try to debunk this. First, even if you want to think of it that way, we have learned something very important about closed and bounded subsets of Euclidean spaces. Namely, putting together the last two results, we (immediately!) get:

COROLLARY 1.24. *Let $X \subseteq \mathbb{R}^N$ be closed and bounded, and let $f : X \to \mathbb{R}^M$ be a continuous function. Then the image $f(X)$ is a closed and bounded subset of $\mathbb{R}^M$.*

On the other hand, this does not work for either closedness or boundedness alone.

EXAMPLE 4.2.

a) *Consider $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = \frac{1}{x^2+1}$. Then f is continuous and $f(\mathbb{R}) = (0,1)$, so f takes the closed set $\mathbb{R}$ to the not-closed set $(0,1)$.*
b) *Consider $f : (0,1) \to \mathbb{R}$ by $f(x) = \frac{1}{x}$. Then f is cotinuous and $f((0,1)) = (1,\infty)$, so f takes the bounded set $(0,1)$ to the unbounded set $(1,\infty)$.*

The actual answer is a bit more complicated though. All of the concepts that we have introduced in the course so far can in fact be studied in much more generality: namely in any metric space (we will learn a bit about metric spaces at the end of the course). If we have a subset X of any metric space, then it will turn out that if it is sequentially compact then it must also be closed and bounded, but in a general metric space a closed, bounded subset does not need to be sequentially compact. To get a glimpse of this, imagine we were working in $\mathbb{Q}$ instead of $\mathbb{R}$, with all the rest of the definitions being the same. Then

$$[0,2]_{\mathbb{Q}} \coloneqq \{x \in \mathbb{Q} \mid 0 \leq x \leq 2\}$$

is a closed, bounded subset of $\mathbb{Q}$, but it is *not* sequentially compact: there is a sequence in $[0,2]_{\mathbb{Q}}$ that converges to the irrational real number $\sqrt{2}$, hence so does every subsequence, hence no subsequence converges to an element *of* $[0,2]_{\mathbb{Q}}$.

Just as in Math 3100 we used Bolzano-Weiersrtrass in $\mathbb{R}$ to show that Cauchy sequences in $\mathbb{R}$ must converge, pretty much the same argument will show that in a sequantially compact metric space, every Cauchy sequence must converge. So sequential compactness has something to do with *completeness*, but it is even stronger, since Cauchy sequences in $\mathbb{R}^N$ converge but $\mathbb{R}^N$ is not sequentially compact.

Coming back to earth: from Corollary 1.24 we deduce:

COROLLARY 1.25 (Multivariable Extreme Value Theorem). *Let X be a subset of $\mathbb{R}^N$ that is nonempty, closed and bounded, and let $f : X \to \mathbb{R}$ be a continuous function. Then f assumes its maximum and minimum values.*

PROOF. By the previous corollary, $f(X)$ is a subset of $\mathbb{R}$ that is nonempty, closed and bounded. By Exercise 1.25 it follows that $\sup f(X)$ and $\inf f(X)$ both lie in $f(X)$, so $f(X)$ has a largest and smallest element. $\square$

4.5. Exercises.

EXERCISE 1.21.

a) *Show: a subset of $\mathbb{R}$ is an open ball if and only if it is a bounded open interval (a,b).*
b) *Show: a subset of $\mathbb{R}$ is a closed ball if and only if it is a bounded closed interval $[a,b]$.*

EXERCISE 1.22. *Let I be a nonempty set, and let $\{U_i\}_{i\in I}$ be an indexed family of open subsets of $\mathbb{R}^N$.*

a) *Show: $\bigcup_{i\in I} U_i$ is also an open subset of $\mathbb{R}^N$.*

b) *Show: if I is finite, then $\bigcap_{i \in I} U_i$ is also an open subset of $\mathbb{R}^N$.*
c) *Give an example in which I is infinite and $\bigcap_{i \in I} U_i$ is* not *an open subset of $\mathbb{R}^N$.*

EXERCISE 1.23. *Show: A subset $A \subseteq \mathbb{R}^N$ is closed iff its complement $\mathbb{R}^N \setminus A$ is open.*

EXERCISE 1.24. *Let I be a nonempty set, and let $\{A_i\}_{i \in I}$ be an indexed family of closed subsets of $\mathbb{R}^N$.*

a) *Show: $\bigcap_{i \in I} A_i$ is also a closed subset of $\mathbb{R}^N$.*
b) *Show: if I is finite, then $\bigcup_{i \in I} A_i$ is also a closed subset of $\mathbb{R}^N$.*
c) *Give an example in which I is infinite and $\bigcup_{i \in I} A_i$ is* not *an open subset of $\mathbb{R}^N$.*

(Comment: if you remember DeMorgan's Laws, you can immediately deduce this from Exercises 1.22 and 1.23. But if not, you can still solve this exercise directly, and even if you, you might want to try it that way as well.)

EXERCISE 1.25. *Let $A \subseteq \mathbb{R}$ be a nonempty subset.*

a) *Suppose that A is bounded above. Show that the supremum* $\sup(A)$ *is a limit point of A.*
b) *Suppose that A is bounded below. Show that the infimum* $\inf(A)$ *is a limit point of A.*
c) *Deduce: if A is closed and bounded, then A has a maximum element (i.e., an element larger than any other element of A and a minimum element (i.e., an element smaller than any other element of A).*

EXERCISE 1.26. *Let $X \subseteq \mathbb{R}^N$, and let $f : X \to R^M$. For $x \in X$, we may write $f(x)$ as $(f_1(x), \ldots, f_M(x))$; this defines functions $f_1, \ldots, f_M : X \to \mathbb{R}$.*

a) *For $1 \leq i \leq M$, let $\pi_i : \mathbb{R}^M \to R$ be the coordinate projections of Exercise 1.17. Show: for all $1 \leq i \leq M$, we have $f_i(x) = \pi_i \circ f$.*
b) *Let $\mathbf{x} \in X$. Show: f is continuous at $\mathbf{x}$ iff f_i is continuous at $\mathbf{x}$ for all $1 \leq i \leq M$.*

EXERCISE 1.27. *Show that the scalar multiplication operation $\alpha \cdot \mathbf{x} \mapsto \alpha x$ defines a continuous function $\mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$.*

EXERCISE 1.28. *For a sequence $\{\mathbf{x}_n\}$ in $\mathbb{R}^N$, let $\mathcal{L}(\mathbf{x}_n)$ be the set of all partial limits of the sequence.*

a) *Show: $\mathcal{L}(\mathbf{x}_n)$ is a closed subset of $\mathbb{R}^N$.*
b) *Suppose that the sequence $\{\mathbf{x}_n\}$ is injective (i.e., for all $m \neq n$ we have $\mathbf{x}_m \neq \mathbf{x}_n$). Let $X \coloneqq \mathbf{x}_\bullet(\mathbb{Z}^+) = \{\mathbf{x}_n \mid n \in \mathbb{Z}^+\}$ be the set of terms of the sequence. Show that $\mathcal{L}(\mathbf{x}_n)$ is the set of accumulation points of X.*

EXERCISE 1.29. *Show: for every closed subset $X \subseteq \mathbb{R}^N$ there is a sequence $\{\mathbf{x}_n\}$ in $\mathbb{R}^N$ such that the set $\mathcal{L}(\mathbf{x}_n)$ of partial limits is X.*

5. Uniform Continuity

Let $A \subseteq \mathbb{R}^N$. A function $f : A \to \mathbb{R}^M$ is **uniformly continuous** if for all $\epsilon > 0$ there is $\delta > 0$ such that for all $x, y \in A$, if $d(x,y) = ||x - y|| < \delta$, then $d(f(x), f(y)) = ||f(x) - f(y)|| < \epsilon$.

The point of this definition is that ordinary continuity applies to one point at a domain at a time, so for each fixed $\epsilon > 0$, the δ that works for one point may not work for another point. Uniform continuity means precisely that we may choose the same δ to work for all points at once. Thus uniformly continuous functions are continuous. The converse is not always true.

EXAMPLE 5.1. *Consider the continuous function $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^2$. We claim that f is* not *uniformly continuous. Indeed, for a positive integer n, take $x \in \mathbb{R}$ and $y = x + \delta$. Then*

$$|f(x) - f(y)| = |(x+\delta)^2 - x^2| = |2x\delta + \delta^2|.$$

No matter how small δ is, this quantity will still be large if $|x|$ is sufficiently large, so in fact for no $\epsilon > 0$ is there a $\delta > 0$ such that $|x-y| < \delta$ implies $|f(x)-f(y)| < \epsilon$.

EXAMPLE 5.2. *Consider the continuous function $f : (0,1) \to \mathbb{R}$ by $f(x) = \frac{1}{x}$. We claim that f is not uniformly continuous. Take $x \in (0,1)$ and $y = x + \delta$. Then*

$$|f(x) - f(y)| = \left|\frac{1}{x} - \frac{1}{x+\delta}\right| = \left|\frac{\delta}{x(x+\delta)}\right|.$$

For each fixed δ, as $x \to 0$ the above expression tends to ∞, so for no $\epsilon > 0$ is there a $\delta > 0$ such that $|x - y| < \delta$ implies $|f(x) - f(y)| < \epsilon$.

In the first example the domain is closed but not bounded. In the second example the domain is bounded but not closed.

Here is a sequential characterization of uniform continuity:

PROPOSITION 1.26. *Let $X \subseteq \mathbb{R}^N$ be a subset, and let $f : X \to \mathbb{R}^M$ be a function. The following are equivalent:*

(i) *f is uniformly continuous.*
(ii) *For all pairs of sequences $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$ in X such that $d(\mathbf{x}_n, \mathbf{y}_n) \to 0$, we have $d(f(\mathbf{x}_n), f(\mathbf{y}_n)) \to 0$.*

PROOF. First suppose that (i) fails: then there is some $\epsilon > 0$ such that for all $\delta > 0$ there are $\mathbf{x}_\delta, \mathbf{y}_\delta \in X$ such that $||\mathbf{x}_\delta - \mathbf{y}_\delta|| < \delta$ and $||f(\mathbf{x}_\delta) - f(\mathbf{y}_\delta)|| \geq \epsilon$. In particular, for each $n \in \mathbb{Z}^+$ this holds for $\delta = \frac{1}{n}$. Let's write $\mathbf{x}_n$ and $\mathbf{y}_n$ in place of $\mathbf{x}_{\frac{1}{n}}$ and $\mathbf{y}_{\frac{1}{n}}$: then for all $n \in \mathbb{Z}^+$ we have $d(\mathbf{x}_n, \mathbf{y}_n) < \frac{1}{n}$ and $d(f(\mathbf{x}_n), f(\mathbf{y}_n)) \geq \epsilon$. In particular $d(f(\mathbf{x}_n), f(\mathbf{y}_n))$ fails to converge to 0, so condition (ii) fails.

Now suppose that condition (ii) fails: then we have sequences $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$ in X such that $\mathbf{x}_n - \mathbf{y}_n$ converges to 0 but $f(\mathbf{x}_n) - f(\mathbf{y}_n)$ fails to converge to 0. The latter means that there is some $\epsilon > 0$ and infinitely many positive integers n such that $||f(\mathbf{x}_n) - f(\mathbf{y}_n)|| \geq \epsilon$. This infinite set of positive integers defines subsequences $\mathbf{x}_{n_k}$ and $\mathbf{y}_{n_k}$. Because passing to a subsequence preserves convergence we have

$$\lim_{k\to\infty} \mathbf{x}_{n_k} - \mathbf{y}_{n_k} \to 0,$$

and now $||f(\mathbf{x}_{n_k}) - f(\mathbf{y}_{n_k})|| \geq \epsilon$ for all positive integers k. Beause $\mathbf{x}_{n_k} - \mathbf{y}_{n_k}$ converges to 0, for all $\delta > 0$ there is $k \in \mathbb{Z}^+$ such that $||\mathbf{x}_{n_k} - \mathbf{y}_{n_k}|| < \delta$, and we still have $||f(\mathbf{x}_{n_k}) - f(\mathbf{y}_{n_k})|| \geq \epsilon$. So condition (i) fails. $\square$

LEMMA 1.27. *Let $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$ be sequences in $\mathbb{R}^N$ with $d(\mathbf{x}_n, \mathbf{y}_n) \to 0$. If $\mathbf{x}_n \to L \in \mathbb{R}^N$, then also $\mathbf{y}_n \to L$.*

You are asked to prove Lemma 1.27 in Exercise 1.32.

THEOREM 1.28 (Uniform Continuity Theorem). *Let $X \subseteq \mathbb{R}^N$ be sequentially compact. Then every continuous function $f : X \to \mathbb{R}^M$ is uniformly continuous.*

PROOF. Seeking a contradiction, we suppose that f is *not* uniformly continuous. Then by Proposition 1.26 there are sequences $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$ in X with $d(\mathbf{x}_n, \mathbf{y}_n) \to 0$ and $d(f(\mathbf{x}_n), f(\mathbf{y}_n))$ *not* converging to 0. As we saw in the proof of Proposition 1.26, this means that there is $\epsilon > 0$ and subsequences $\{\mathbf{x}_{n_k}\}$ and $\{\mathbf{y}_{n_k}\}$ such that $d(\mathbf{x}_{n_k}, \mathbf{y}_{n_k}) \geq \epsilon$ for all $k \in \mathbb{Z}^+$. Thus if X is *not* uniformly continuous then there are sequences $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$ in X such that $d(\mathbf{x}_n, \mathbf{y}_n) \to 0$ and $\epsilon > 0$ such that $d(f(\mathbf{x}_n), f(\mathbf{y}_n)) \geq \epsilon$ for all $n \in \mathbb{Z}^+$.

We will use the sequential compactness of X to get a contradiction. Indeed, since X is sequential compact, there is a subsequence $\{\mathbf{x}_{n_k}\}$ that converges to some element of X, say L. Since $\mathbf{x}_n - \mathbf{y}_n \to 0$, also $\mathbf{x}_{n_k} - \mathbf{y}_{n_k} \to 0$, so by Lemma 1.27 the sequence $\{\mathbf{y}_{n_k}\}$ also converges to L. Because f is continuous, we have

$$f(\mathbf{x}_{n_k}) \to f(L) \text{ and } f(\mathbf{y}_{n_k}) \to f(L),$$

from which it follows that $f(\mathbf{x}_{n_k}) - f(\mathbf{y}_{n_k}) \to 0$ and thus for all sufficiently large k we have $d(\mathbf{x}_{n_k}, \mathbf{y}_{n_k}) < \epsilon$. Contradiction! $\square$

So here is a question that is so much more than fair: **why** uniform continuity? What we have established up to this point is that uniform continuity is a variant of continuity that is in general subtly stronger, still has a sequential characterization, and that the two concepts coincide on closed, bounded subsets of Euclidean space. But...what's the point?

One thing that makes the study of theoretical mathematics challenging is that key definitions emerge after years (centuries, here) of work on specific problems. When the mathematics is presented however it is much more efficient to present the definitions first and the application later on. Indeed, later on in this course we will absolutely want to know that every continuous function $f : [a, b] \to \mathbb{R}$ is uniformly continuous: this will be the key to showing that every such function is Riemann integrable. However, I would like to show an application of uniform continuity *now*, so in the next section we consider the extension problem for continuous functions.

5.1. Exercises.

EXERCISE 1.30. *Use Proposition 1.26 to show that the function of Example 5.1 is not uniformly continuous.*

EXERCISE 1.31. *Use Proposition 1.26 to show that the function of Example 5.2 is not uniformly continuous.*

EXERCISE 1.32. *Let $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$ be sequences in $\mathbb{R}^N$ with $d(\mathbf{x}_n, \mathbf{y}_n) \to 0$. If $\mathbf{x}_n \to L \in \mathbb{R}^N$, show that also $\mathbf{y}_n \to L$.*

EXERCISE 1.33. *State and prove an analogue of Exercise 1.26 for uniform continuity.*

EXERCISE 1.34. *Let $M, N, P \in \mathbb{Z}^+$. Let $X \subseteq \mathbb{R}^N$ and let $Y \subseteq \mathbb{R}^M$. Let $f : X \to R^M$ and let $g : Y \to \mathbb{R}^P$ be functions. Suppose that $f(X) \subseteq Y$, so that the composition $g \circ f$ is defined. Show: if f and g are both uniformly continuous, so is $g \circ f$.*

EXERCISE 1.35. *A function $f : \mathbb{R} \to \mathbb{R}$ is called* ***periodic*** *if there is $a \in \mathbb{R} \setminus \{0\}$ such that for all $x \in \mathbb{R}$ we have $f(x+a) = f(a)$. Show: a function that is continuous and periodic is uniformly continuous.*

EXERCISE 1.36. *Let $f : \mathbb{R} \to \mathbb{R}$ be a polynomial function, say $f(x) = a_n x^n + \ldots + a_1 x + a_0$ with $n \in \mathbb{Z}^{\geq 0}$ and $a_n \neq 0$.*

a) *Suppose that f is a linear function (plus a constant, the linear algebraists would add): i.e., that $n \leq 1$. Show: f is uniformly continuous.*
b) *Suppose that $n \geq 2$. Show: f is* not *uniformly continuous.*
c) *Can you generalize this to polynomials $f : \mathbb{R}^N \to \mathbb{R}$?*

EXERCISE 1.37. *We say that a function $f : \mathbb{R}^N \to \mathbb{R}^M$* ***vanishes at infinity*** *if for all $\epsilon > 0$ there is $R > 0$ such that for all $x \in \mathbb{R}^N$, if $||x|| > R$ then $||f(x)|| < \epsilon$. Show: if f is continuous and vanishes at infinity, then f is uniformly continuous.*

EXERCISE 1.38. *Let $X \subseteq \mathbb{R}^N$ and let $f : X \to \mathbb{R}^M$ be uniformly continuous. Show: if X is bounded, then $f(X)$ is bounded.*

6. Extending Continuous Functions

Suppose X is a subset of $\mathbb{R}^N$ and $f : \mathbb{R}^N \to \mathbb{R}^M$ is a continuous function. It is natural to ask: can f be extended to a continuous function on all of $\mathbb{R}^N$?

EXAMPLE 6.1. *Let $f : [a, b] \to \mathbb{R}$ be continuous. Then f extends continuously to all of $\mathbb{R}$: indeed, we can put $f(x) = f(a)$ for all $x < a$ and $f(x) = f(b)$ for all $x > b$. This works!*

EXAMPLE 6.2. *The function $f(x) = \frac{1}{x}$ is a continuous function on $(0, 1]$ that does* not *extend continuously to all of $\mathbb{R}$. In the language of calculus, we would say that $\lim_{x \to 0^+} f(x) = \infty$, which prevents such an extension. This is correct, but here is an explanation using the language and concepts we have been developing: f is continuous at a point $\mathbf{x}$ if for all $\epsilon > 0$ f maps some ball $B^\circ(\mathbf{x}, \delta)$ into the ball $B^\circ(f(\mathbf{x}), \epsilon)$. In particular f must be bounded in some small ball around $\mathbf{x}$. Since $f((0, \delta)) = (\frac{1}{\delta}, \infty)$, no matter how we define f at 0, the function will be unbounded in any δ-ball around 0, so it does not have any continuous extension to $[0, 1]$.*

One moral one can extract from this is:

PROPOSITION 1.29. *Let $f : X \subseteq \mathbb{R}^N \to \mathbb{R}^M$. If f has a continuous extension to $\mathbb{R}^N$, then for all bounded subsets $Y \subseteq X$, the image $f(Y)$ is bounded.*

PROOF. It is enough to see that for $f : \mathbb{R}^N \to \mathbb{R}^M$, if $Y \subseteq \mathbb{R}^N$ is bounded then so is $f(Y)$. If Y is bounded, then it is contained in a closed ball B, which is closed and bounded, so $f(B)$ is closed and bounded, so $f(Y) \subseteq f(B)$ is bounded. □

This criterion is however not sufficient.

EXAMPLE 6.3. *The function $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ defined by $f(x) = \sin(1/x)$ is not only bounded on every bounded subset; it is just bounded. Nevertheless, it does not extend continuously to $\mathbb{R}$, as is left as an exercise.*

Let us worry about extending a continuous function one point at a time. There are two cases of this; one is trivial, and the other is not.

EXAMPLE 6.4. *Let* $f : (0, 1) \to \mathbb{R}$ *be a continuous function. Suppose we want to extend* f *to a continuous function on* $(0, 1) \cup \{2\}$*. There is precisely no problem here: for any* $L \in \mathbb{R}$ *we can put* $f(2) \coloneqq L$*, and the function* $f : (0, 1) \cup \{2\}$ *will be continuous. Why? Because for any* $\epsilon > 0$*, take* $\delta = 1$*: we need to check that if* $|x - 2| < 1$ *then* $|f(x) - L| < \epsilon$*. But the only* $x \in (0, 1) \cup \{2\}$ *with* $|x - 2| < 1$ *is* $x = 2$ *itself, and* $|f(2) - L| = 0$.

This example motivates the following definition: let A be a subset of $\mathbb{R}^N$. An **isolated point** of A is a point $L \in A$ such that for some $\delta > 0$ we have $B^\circ(L, \delta) \cap A = \{L\}$. In other words, a point of A is isolated if for some $\delta > 0$ the only point of A that is within δ of L is L itself.

PROPOSITION 1.30. *Let* $A \subseteq \mathbb{R}^N$*, and let* $L \in A$ *be an isolated point. Then* every *function* $f : A \to \mathbb{R}^M$ *is continuous at* L.

PROOF. This is the same argument as in Example 6.4: if $\delta > 0$ is such that $B^\circ(L, \delta) \cap A = \{L\}$, then for any $\epsilon > 0$, we have that for all $\mathbf{x} \in A$, $d(\mathbf{x}, L) < \delta \implies d(f(\mathbf{x}), f(L) < \epsilon$....because the only $\mathbf{x}$ that satisfies the first inequality is $\mathbf{x} = L$! □

In terms of the extension problem, this means: if L is an isolated point of A, then every continuous function f on $A \setminus \{L\}$ extends continuously to A, and we can do so by defining $f(L)$ to be whatever we want!

Okay, that was indeed a trivial case. Let's move on to the other case, which involves a variant of the notion of limit point that was alluded to before. By an **injective sequence** in a set X, we mean a sequence $\{\mathbf{x}_n\}$ in X for which the defining function $\mathbf{x}_\bullet : \mathbb{Z}^+ \to X$ is injective. In plainer language, an injective sequence is a sequence in which every term is a different element of X. Now for a subset $A \subseteq \mathbb{R}^N$, an **accumulation point** is a point $L \in \mathbb{R}^N$ for which there is an injective sequence $\{\mathbf{x}_n\}$ in A converging to L.

Compare with the definition of a limit point: the only difference is that we have added the word "injective." Thus every limit point of A is an accumulation point of A. The converse is not true in general: for instance if A is finite there are no injective sequences in A, so A has no accumulation points, but as always, every element of A is a limit point of A. In general:

PROPOSITION 1.31. *Let* $A \subseteq \mathbb{R}^N$*, and let* $L \in \mathbb{R}^N$ *be a limit point of* A*. Then exactly one of the following holds:*

(i) L *is an accumulation point of* A.
(ii) L *is an isolated point of* A.

PROOF. Step 1: If L is an isolated point of A, then a sequence $\{\mathbf{x}_n\}$ in A converges to L if and only if we have $\mathbf{x}_n = L$ for all sufficiently large n. We leave this as an exericse (Exercise 1.41). From this it follows that if L is an isolated point of A then L is not the limit of any injective sequence in A, so L is not an accumulation point of A. Thus we have shown that conditions (i) and (ii) are mutually exclusive.
Step 2: Suppose that L is a limit point of A that is not an isolated point of A. This means that *either* $L \notin A$ *or* $L \in A$ but for all $\delta > 0$ there is $\mathbf{x}_\delta \in A$ with $0 < ||\mathbf{x}_\delta - L|| < \delta$. In each of these two cases we will produce an injective sequence

in A that converges to L.
Case 1: $L \notin A$. Because L is a limit point of A there is a sequence $\{\mathbf{x}_n\}$ in A that converges to A. The problem is that is may not be injective: i.e., terms may repeat. However, any element of $p \in A$ can show up only finitely many times in the sequence: indeed, since $p \in A$ and $L \notin A$, we have $p \neq L$, so $d = d(p, L) > 0$, and because the sequence converges to L, we have $d(\mathbf{x}_n, L) < d$ for all sufficiently large n. Therefore we can form a subsequence simply by omitting every term that is a repetition: i.e., for which the same element of A has alerady occurred earlier in the sequence. This builds an injective subsequence, which must still converge to L.
Case 2: $L \in A$. Because we have elements of A arbitrarily close to L but different from L, we can build a sequence the nth term of which has distance less than $\frac{1}{n}$ from L and is also closer than any previous term. In other words, let $\mathbf{x}_1$ be an element of $A \setminus \{L\}$ with $d(\mathbf{x}_1, L) < 1$. Let $\mathbf{x}_2$ be an element of $A \setminus \{L\}$ with $d(\mathbf{x}_2, L) < \min(\frac{1}{2}, d(\mathbf{x}_1, L))$. Let x_3 be an element of $A \setminus \{L\}$ with $d(\mathbf{x}_3, L) < \min(\frac{1}{3}, d(\mathbf{x}_2, L))$. And so forth. This gives an injective sequence in A converging to L. $\square$

Note that an isolated point of A is necessarily a point of A, but an accumulation point of A may or may not be a point of A. For instance, every point of an open or closed ball is an accumulation point.

Since a set $X \subseteq \mathbb{R}^N$ is closed if it contains all its limit points, but every limit point of X is either an element of X or an accumulation point (again, both are possible!), it follows that a set is closed iff it contains its accumulation points.

So now let's consider the nontrivial case of the "one point extension problem": let $A \subseteq \mathbb{R}^N$, let $L \in \mathbb{R}^N \setminus A$ be an accumulation point of A, and let $f : A \to \mathbb{R}^M$ be a continuous function. The question is whether we can extend f to a continuous function on $A \cup \{L\}$. First we observe that there is at most one way to do this: indeed, suppose that $g : A \cup \{L\} \to \mathbb{R}^M$ is a continuous extension of f. Because L is an accumulation point of A, there is a sequence $\{\mathbf{x}_n\}$ in A such that $\mathbf{x}_n \to L$. By Theorem 1.15 we have

$$g(L) = g(\lim_{n\to\infty} \mathbf{x}_n) = \lim_{n\to\infty} f(\mathbf{x}_n).$$

This tells us how to define $g(L)$, so its value must indeed be unique.

THEOREM 1.32. *Let $A \subseteq \mathbb{R}^N$, let $L \in \mathbb{R}^N \setminus A$ be an accumulation point of A, and let $f : A \to \mathbb{R}^M$ be a continuous function. Suppose there is some $r > 0$ such that*

$$f|_{B^\bullet(L,r)\cap A} : B^\bullet(L, r) \cap A \to \mathbb{R}^M$$

is uniformly continuous. Then f admits a continuous extension to $A \cup \{L\}$.

PROOF. Step 1: Above we assumed that the continuous extension g existed and gave a formula for it: namely, choose a sequence $\{\mathbf{x}_n\}$ in A such that $\mathbf{x}_n \to L$; then $g(L) = \lim_{n\to\infty} f(\mathbf{x}_n)$. So we want to *define*

$$f(L) := \lim_{n\to\infty} f(\mathbf{x}_n).$$

In fact, so as to make use of the assumed uniform continuity, we want the sequence $\mathbf{x}_n$ to lie in $B^\bullet(L, r)$. Because the original sequence converges to L, we can attain this just by removing finitely many terms, so let's do so. Now we need to show first that this limit actually exists and second that it does not depend upon the

sequence $\{\mathbf{x}_n\}$ we chose.
Step 1a): Because we know that Cauchy sequences in $\mathbb{R}^M$ converge, it is enough to show that $\{f(\mathbf{x}_n)\}$ is Cauchy. For this, we know that the sequence $\{\mathbf{x}_n\}$ is convergent in $\mathbb{R}^N$, so it is Cauchy. Happily, it is easy to show that *uniformly* continuous maps send Cauchy sequences to Cauchy sequences: let $\epsilon > 0$. Because of the uniform continuity of f, there is $\delta > 0$ such that for all $y, z \in B^\bullet(L, r) \cap A$, we have $d(y, z) < \delta \implies d(f(y), f(z)) \leq \epsilon$. Since $\{\mathbf{x}_n\}$ lies in $B^\bullet(L, r)$ and is Cuachy, there is $K \in \mathbb{Z}^+$ such that if $m, n \geq K$ then $d(\mathbf{x}_m, \mathbf{x}_n) \leq \delta$, and thus

$$\forall m, n \geq K, \ d(f(\mathbf{x}_m), f(\mathbf{x}_n)) < \epsilon.$$

This shows that the sequence $f(\mathbf{x}_n)$ converges.
Step 1b): Let $\{\mathbf{y}_n\}$ be another sequence in $B^\bullet(L, r) \cap A$ such that $\mathbf{y}_n \to L$. Then $d(\mathbf{x}_n, \mathbf{y}_n) \to 0$, so by Proposition 1.26, we have $d(f(\mathbf{x}_n), f(\mathbf{y}_n)) \to 0$. Since both sequences are convergent, it follows from Lemma 1.27 their limits are equal.
Step 2: It remains to show that our extended function is continuous at L. But in fact this follows from our Sequential Characterization of Uniform Continuity, since we have just shown that if if $\mathbf{x}_n \to L$ then $f(\mathbf{x}_n) \to f(L)$. Strictly speaking, we showed this only for sequences each of whose terms lie in $B^\bullet(L, r) \cap A$, but again any sequence that converges to L becomes such a sequence after removing finitely many terms; so any such sequence converges after removing finitely many of its terms...so any such sequence converges. □

Let me quickly discuss some further developments of these ideas.

For a subset A of $\mathbb{R}^N$, we can define its **closure** $\overline{A}$ to be A together with all of its limit points (equivalently, with all of its accumulation points). As the name implies, $\overline{A}$ is then a closed set (this is not completely obvious: it comes down to showing a limit point of limit points of A is still a limit point of A). In fact $\overline{A}$ is the smallest closed set containing A. It follows from our discussion that every continuous function had *at most one* continuous extension to $\overline{A}$. Such a continuous extension need not exist, but it will if f is *uniformly continuous* on A. But in fact this condition is a little too strong, and the precise result is the following.

THEOREM 1.33. *Let $A \subseteq \mathbb{R}^N$, and let $f : A \to R^M$ be continuous. The following are equivalent:*

(i) *f admits a continuous extension to $\overline{A}$.*
(ii) *The restriction of f to each bounded subset of A is uniformly continuous.*

We are not so terribly far away from a proof of this important result; it will be developed in some exercises.

A subset $X \subset \mathbb{R}^N$ is called **dense** if its closure is all of $\mathbb{R}^N$. This means: for every $\mathbf{y} \in \mathbb{R}^N$ and every $\epsilon > 0$, there is $\mathbf{x} \in X$ with $d(\mathbf{x}, \mathbf{y}) < \epsilon$. For instance $\mathbb{Q}$ is dense in $\mathbb{R}$. Theorem 1.33 therefore shows that if you have a continuous function on a dense subset of $\mathbb{R}^N$ then it extends continuously to all of $\mathbb{R}^N$ if and only if it is uniformly continuous on each bounded subset. As an example, consider an exponential function a^x. If you think about it, we can make good sense of a^x using the methods of precalculus when x is any rational number, but what does $a^{\sqrt{2}}$ mean? In order to make sense of it we need to use some limiting process. One way to define a^x as a function on all of $\mathbb{R}$ is to show that $a^x : \mathbb{Q} \to \mathbb{R}$ is uniformly continuous on bounded subsets. (It is not uniformly continuous on all of $\mathbb{Q}$.)

What if A is not dense, so $\overline{A} \subsetneq R^N$ is closed? It turns out that if $X \subseteq \mathbb{R}^N$ is a closed subset and $f : X \to \mathbb{R}^M$ is continuous, then there is always a continuous extension of f to all of $\mathbb{R}^N$: in fact there are always lots and lots of such extensions. This is a special case of an important result called the **Tietze Extension Theorem**, which you might learn about in Math 4200: see [**GT**, Theorem 2.89].

Before we depart this topic, let us observe that we have essentially rediscovered the notion of **limit**. Namely, let $X \subseteq \mathbb{R}^N$, let $\mathbf{c}$ be an accumulation point of X, and let $f : X \setminus \{\mathbf{c}\} \to \mathbb{R}^M$ be a function. Then we define

$$\lim_{\mathbf{x} \to \mathbf{c}} f(x) = L$$

to mean: if we extend f to X by setting $f(\mathbf{c}) \coloneqq L$, then f is continuous at $\mathbf{c}$. Again, the value L is then the common value $\lim_{n\to\infty} f(\mathbf{x}_n)$ for all sequences $\{\mathbf{x}_n\}$ in $A\setminus\{L\}$ that converge to c, so it is uniquely determined, if it exists. It is immediate to see that the limit is L iff: forall $\epsilon > 0$ there is $\delta > 0$ such that for all $\mathbf{x} \in A \setminus \{\mathbf{c}\}$, we have $d(\mathbf{x}, \mathbf{c}) < \delta \implies d(f(\mathbf{x}), L) < \epsilon$.

6.1. Exercises.

Exercise 1.39. *Let $A \subseteq \mathbb{R}^N$ and let $\mathbf{c}$ be an isolated point of A. Show: every function $f : A \to \mathbb{R}^M$ is continuous at $\mathbf{c}$.*

Exercise 1.40. *Show that the function $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ given by $f(x) = \sin(\frac{1}{x})$ has no continuous extension to 0.*

Exercise 1.41. *Let $A \subseteq \mathbb{R}^N$, and let $L \in A$ be an isolated point. Let $\{\mathbf{x}_n\}$ be a sequence in A. Show that $\mathbf{x}_n \to L$ if and only if there is $K \in \mathbb{Z}^+$ such that $\mathbf{x}_n = L$ for all $n > K$.*

Exercise 1.42. *Let $A \subseteq \mathbb{R}^N$. Let $\overline{A}$ be the union of A and the limit points of A.*

- a) *Show: $\overline{A}$ is closed.*
- b) *Show: $\overline{A}$ is the intersection of all closed subsets of $\mathbb{R}^N$ containing A.*
- c) *Show: A is bounded iff $\overline{A}$ is sequentially compact.*

Exercise 1.43. *Let $X \subseteq \mathbb{R}^N$. Use Bolzano-Weierstrass (Theorem 1.11) to show that the following are equivalent:*

- (i) *X has an accumulation point in $\mathbb{R}^N$.*
- (ii) *There is some bounded subset $B \subseteq \mathbb{R}^N$ such that $X \cap B$ is infinite.*

Exercise 1.44. *Let $A \subseteq \mathbb{R}^N$ be bounded. Let $f : A \to \mathbb{R}^M$ be continuous.*

- a) *Show that the following are equivalent:*
  - (i) *f is uniformly continuous.*
  - (ii) *f admits a continuous extension to $\overline{A}$.*
- b) *Show that under the equivalent condiitons of part a), the continuous extension of f to $g : \overline{A} \to \mathbb{R}^M$ is unique and uniformly continuous.*

Exercise 1.45. *Let $A \subseteq \mathbb{R}^N$, and let $f : A \to \mathbb{R}^M$.*

- a) *Show that f is continuous iff its restriction to each bounded subset of f is continuous.*

b) *Show: f admits a continuous extension to $\overline{A}$ iff its restriction to each bounded subset of A is uniformly continuous.*

The remaining exercises make use of the following definitions: let $X \subseteq \mathbb{R}^N$. We say that X is **discrete** if every point of X is an isolated point. We say that X is **uniformly discrete** if there is $\delta > 0$ such that for all $x_1, x_2 \in X$, if $||x_1 - x_2|| < \delta$ then $x_1 = x_2$.

EXERCISE 1.46. a) *Show: if X is uniformly discrete, then X is discrete.*

b) *Show $\{\frac{1}{n} \mid n \in \mathbb{Z}^+\}$ is a subset of $\mathbb{R}$ that is discrete, not uniformly discrete and not closed.*
c) *Show: if X is uniformly discrete, then X is closed.*
d) *Find a closed subset $X \subseteq \mathbb{R}$ that is discrete but not uniformly discrete.*

EXERCISE 1.47. *Let $X \to \mathbb{R}^N$ be a subset.*

a) *Show that the following are equivalent:*
  (i) *X is discrete.*
  (ii) *Every function $f : X \to \mathbb{R}^M$ is continuous.*
  (iii) *Every function $f : X \to \mathbb{R}$ is continuous.*
b) *Show that the following are equivalent:*
  (i) *X is uniformly discrete.*
  (ii) *Every function $f : X \to \mathbb{R}^M$ is uniformly continuous.*
  (iii) *Every function $f : X \to \mathbb{R}$ is uniformly continuous.*

EXERCISE 1.48. *Let $X \subseteq \mathbb{R}^N$. Show that the following are equivalent:*

(i) *Every continuous function $f : X \to \mathbb{R}^M$ is uniformly continuous.*
(ii) *Every continuous function $f : X \to \mathbb{R}$ is uniformly continuous.*
(iii) *X is either sequentially compact or uniformly discrete.*

CHAPTER 2

Derivatives and Inverse Functions

1. Functional Limits

Let $X \subseteq \mathbb{R}^N$ be a nonempty subset, and let $c \in X$ be a nonisolated point. For a function $f : X \setminus \{c\} \to \mathbb{R}^M$, recall that we say that $\lim_{x\to c} f(x) = L$ if defining $f(c) := L$ makes f continuous at c. Spelling out, this means: for all $\epsilon > 0$, there is $\delta > 0$ such that for all $x \in X$, if $0 < ||x - c|| < \delta$ then $||f(x) - L|| < \epsilon$. If the limit exists, then its value is unique.

The following is a variation on the fact that compositions of continuous functions are continuous.[1]

Proposition 2.1. *Let $X \subseteq \mathbb{R}^N$ and $Y \subseteq \mathbb{R}^M$, and let $f : X \setminus \{c\} \to \mathbb{R}^M$ and $g : Y \to \mathbb{R}^P$. Suppose that $f(X) \subseteq Y$. Let $c \in X$ be a nonisolated point. Then: if $\lim_{x\to c} f(x) = L$ and $\lim_{y\to L} g(y) = M$. Then*

$$\lim_{x\to c} g(f(x)) = M = g(\lim_{x\to c} f(x)).$$

Proof. Define f at c by $f(c) := L$; then f is continuous at c. There is a sequence $\{x_n\}$ in X converging to c, hence $f(x_n)$ converges to $f(c) = L$, so L is a limit point of Y. If $L \notin Y$, we put $g(L) := M$; if $L \in Y$, then we *redefine* $g(L) := M$. Either way this makes g continuous at M. By Proposition 1.20, the composition $g \circ f$ is then continuous at c, so $\lim_{x\to c} g(f(x)) = g(f(c)) = g(L) = M$. $\square$

1.1. Exercises.

Exercise 2.1 (Squeeze Theorem). *Let $X \subseteq \mathbb{R}^N$, and let $c \in X^\circ$. Let $f : X \setminus \{c\} \to \mathbb{R}$ be a function. Let $\delta > 0$ be such that $B^\circ(c,\delta) \subseteq X$, and suppose there are functions*

$$m, M : B^\circ(c,\delta) \setminus \{c\} \to \mathbb{R}$$

such that

$$\forall x \in B^\circ(c,\delta) \setminus \{c\},\ m(x) \le f(x) \le M(x)$$

and

$$\lim_{x\to c} m(x) = L = \lim_{x\to c} M(x).$$

Show: $\lim_{x\to c} f(x) = L$.

In the above statement of the Squeeze Theorem, it is in fact not critical that c be an interior point of X: we could have worked with any accumulation point c of X and in place of $B^\circ(c,\delta)$ used $B^\circ(c,\delta) \cap X$. We just wanted a relatively clean statement.

[1] The annoying complicatedness of the statement is again a sign that continuity is the fundamental concept on which limits should be based, not the other way around.

Exercise 2.2. *Let $X \subseteq \mathbb{R}$, and let a be an accumulation point of X. We say that the **right-handed limit** $\lim_{x \to a^+} f(x) = L$ exists if a is still an accumulation point of $X \cap [a, \infty)$ and upon restricting the domain from X to $X \cap [a, \infty)$, the limit exists and is equal to L. We say that the **left-handed limit** $\lim_{x \to a^-} f(x) = L$ exists if a is still an accumulation point of $X \cap (-\infty, a]$ and upon restricting the domain from X to $X \cap (-\infty, a])$, the limit exists and is equal to L. Show: $\lim_{x \to a} f(x)$ exists if and only if $\lim_{x \to a^+} f(x)$ and $\lim_{x \to a^-} f(x)$ both exists and are equal, in which case the common value is $\lim_{x \to a} f(x)$.*

2. The Derivative

Suppose now that $X \subseteq \mathbb{R}$, that $x \in X$ is an accumulation point of X, and we have $f : X \to \mathbb{R}$. (Thus with respect to the notation of the previous paragraph we have taken $N = M = 1$.) We say that f is **differentiable at c** if $\lim_{h \to 0} \frac{f(x+h)-f(x)}{h}$ exists, and if so we denote this value by $f'(x)$ and call it the **derivative** of f at x. We say that $f : X \to \mathbb{R}$ is **differentiable** if it is differentiable at each point of X, and if so we get a new function

$$f' : X \to \mathbb{R}, \; x \mapsto f'(x),$$

which is, of course, called the **derivative** of f.

An equivalent form of the limit defining $f'(a)$ is

$$\lim_{x \to a} \frac{f(x) - f(a)}{x - a}.$$

Remark. a) *In practice, we will almost always apply this definition in the case when $X = I$ is an* interval.

b) *Our definition of derivative makes sense for functions $f : X \subseteq \mathbb{R} \to \mathbb{R}^M$. In fact the same thing happens as usually happens for vector-valued functions: we may write $f = (f_1, \ldots, f_M)$ with $f_j : X \to R$ and then f is differentiable at c if and only if each $f_j(x)$ is differentiable at x, in which case we have $f'(x) = (f_1'(x), \ldots, f_r'(x))$. Thus in a sense the analysis of derivatives of a vector-valued function is a bit banal. The geometry of such functions is still interesting – essentially these are parameterized curves in $\mathbb{R}^M$ – but we will not be studying it in this course.*

c) *On the other hand, it does* not *make sense to extend this definition (at least, not verbatim) to functions of a vector variable: i.e., to $f : X \subseteq \mathbb{R}^N \to R$ for $N \geq 1$, because then division of $f(x + h) - f(x)$ by the vector h doesn't make sense. This is certainly not the end of the story: there are several remedies. The one you have probably seen before is to consider* partial derivatives *by holding all but one of the inputs fixed. As the word "partial" suggests, there is also a "total' derivative, which has a more elaborate definition: the derivative of a function $f : X \subseteq \mathbb{R}^N \to R^M$ is a linear map $L : \mathbb{R}^N \to \mathbb{R}^M$ such that*

$$\lim_{h \to 0} \frac{||f(x + h) - f(x) - L(h)|||}{||h||} = 0.$$

We will not be studying this definition in our course.

Proposition 2.2. *Let $X \subseteq \mathbb{R}$, and let $a \in X$ be an accumulation point. Let $f : X \to \mathbb{R}$ be a function. If f is differentiable at a, then f is continuous at a.*

PROOF. As we know, f is continuous at a if and only if $\lim_{x \to a} f(x)$ exists and is equal to $f(a)$. So:

$$\lim_{x \to a} (f(x) - f(a)) = \lim_{x \to a} \left(\frac{f(x) - f(a)}{x - a} \right) \cdot (x - a)$$

$=$

$$\left(\lim_{x \to a} \frac{f(x) - f(a)}{x - a} \right) \left(\lim_{x \to a} x - a \right) = f'(a) \cdot 0 = 0.$$

Thus

$$0 = \lim_{x \to a} (f(x) - f(a)) = \left(\lim_{x \to a} f(x) \right) - f(a),$$

so

$$\lim_{x \to a} f(x) = f(a). \qquad \square$$

Thus differentiable functions are continuous. There are two statements that sound vaguely similar to this but are not the same.

The first is "Continuous functions are differentiable." Of course this assertion is the logical *converse* of Proposition 2.2 so we should not necessarily expect it to be true but rather inquire whether it is true. It isn't.

EXAMPLE 2.1. *Let $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = |x|$. This function is continuous: in fact it is uniformly continuous with $\delta = \epsilon$ for all $\epsilon > 0$ (a "short map"). But it is not differentiable at $a = 0$. Indeed,*

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

The function $\frac{|h|}{h}$ is $\begin{cases} 1 & h > 0 \\ -1 & h < 0 \end{cases}$, so it has no continuous extension to $h = 0$. So f is not differentiable at 0.

There are a couple of followup remarks to make here:

a) *If we restricted f to $[0, \infty)$, then it would be differentiable at 0. For then in $\lim_{h \to 0} \frac{|h|}{h}$ we are only conisdering positive values of h, so the limit is 1. In general, if we have a subset $X \subseteq \mathbb{R}$, a function $f : X \to \mathbb{R}$ and an accumulation point $a \in X$, we define the* ***right handed limit*** *$\lim_{x \to a^+} f(x)$ to exist if (i) a is still an accumulation point of $X \cap [a, \infty)$ and (ii) upon restricting the domain from X to $X \cap [a, \infty)$, the limit at a exists.*[2] *When the right-handed limit defining the derivative exists, we say that f is* ***right-handed differentiable at a*** *and write the $f'_+(a)$. There is a corresponding definition of left-handed limit and left-handed derivative, denoted $f'_-(a)$.*

   In our case of $f(x) = |x|$ we have $f'_+(0) = 1$ and $f'_-(0) = -1$, and the differing values of these left- and right-handed derivatives are what cause the derivative overall not to exist.

[2]This shows by the way that our approach of allowing the domain of a limit to be a rather general subset of $\mathbb{R}^N$ is a good one: it allows us to handle certain variations on the limiting process that one encounters in calculus in routine ways.

b) *The function f is differentiable at every $a \neq 0$. Indeed, if $a > 0$, then there is $\delta > 0$ – indeed, one may take $\delta = a$ – such that upon restriction to $(a - \delta, a + \delta)$ the function is just $f(x) = x$, which we surely know is differentiable, with derivative 1. Because the derivative is defined by a limit, it is a* local *property of f: the value of f' at a point a depends only on the values of f in an arbtirarily small ball centered at a. So $f'(a) = 1$ for all $a > 0$. Similarly, if $a < 0$, then there is $\delta > 0$ – indeed, one may take $\delta = |a|$ – such that upon restriction to $(a - \delta, a + \delta)$ the function is just $f(x) = -x$, which we surely know is differentiable, with derivative -1. By the same discussion we have $f'(a) = -1$ for all $a < 0$.*

By the way, this is just about the mildest example of a continuous, nondifferentiable function. Weierstrass used uniform convergence to construct functions $f : \mathbb{R} \to \mathbb{R}$ that are continuous and differentiable *at no point* of $\mathbb{R}$.

The second statement that sounds like "differentiable functions are continuous" is "derivatives are continuous." Since f is differentiable, f is continuous: but the derivative f' is a different function. It does *not* have to be continuous:

EXAMPLE 2.2. *Let $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = \begin{cases} x^2 \sin(1/x) & x \neq 0 \\ 0 & x = 0 \end{cases}$. Then f is certainly differentiable away from 0 (we will shortly nail down the differentiation rules from calculus that lead to this): indeed, for $a \neq 0$ we have*

$$f'(a) = 2x \sin(1/x) + x^2 \cdot \cos(1/x) \cdot (-1/x^2) = 2x \sin(1/x) - \cos(1/x).$$

We have

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

because $h \to 0$ and $\sin(1/h)$ is bounded. Therefore f is differentiable, and its derivative is

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

Since $\lim_{x \to 0} f'(x)$ does not exist, f' is not *continuous at 0.*

The fact that derivatives need not be continuous is often something of a nuisance. A function $f : X \to \mathbb{R}$ is called **continuously differentiable, or C^1** is $f' : X \to \mathbb{R}$ exists *and is continuous.* There are many, many theorems in analysis that have the hypothesis that f is C^1 rather than just that f is differentiable. On the other hand, from the Fundamental Theorem of Calculus it will follow that every continuous function *is a derivative* (again, not the same as "every continuous function is differentiable"), which leads to the question "Which discontinuous functions are derivatives?" We will not answer this question in this course, but we will see an important property that every derivative must satisfy, continuous or not.

When it comes to the differentiation rules, we will mostly leave the task of confirming them to the reader.

PROPOSITION 2.3. *Let X be a subset of $\mathbb{R}$, let $a \in X$ be an accumulation point, and let $f, g : X \to \mathbb{R}$ be functions that are both differentiable at a.*

a) *If* $C \in \mathbb{R}$ *then* Cf *is differentiable at* a *and* $(Cf)'(a) = Cf'(a)$.
b) *The function* $f+g$ *is differentiable at* a *and* $(f+g)'(a) = f'(a) + g'(a)$.
c) *The function* fg *is differentiable at* a *and* $(fg)'(a) = f'(a)g(a) + f(a)g'(a)$.
d) *If* $g(a) \neq 0$*, then* $\frac{1}{g}$ *is differentiable at* a *and* $(1/g)'(a) = \frac{-g'(a)}{g(a)^2}$.
e) *If* $g(a) \neq 0$*, then* $\frac{f}{g}$ *is differentiable at* a *and* $(f/g)'(a) = \frac{g(a)f'(a) - f(a)g'(a)}{g(a)^2}$.

You are asked to prove Proposition 2.3 over the course of several exercises.

Now we want to prove the Chain Rule, which takes a bit more doing. In the proof we will use the following result:

LEMMA 2.4. *Let* $X \subseteq \mathbb{R}$ *and let* $a \in X$ *be an accumulation point. We suppose that:*
(i) *The limit* $\lim_{x \to a} f(x)$ *exists, and*
(ii) *There is* $L \in \mathbb{R}$ *such that for all* $\delta > 0$*, there is* $x \in \mathbb{R}$ *such that* $0 < |x-a| < \delta$ *and* $f(x) = L$.

Then: $\lim_{x \to a} f(x) = L$.

You are asked to prove Lemma 2.4 as an exercise.

THEOREM 2.5 (Chain Rule). *Let* f *and* g *be functions, each defined on a subset of* $\mathbb{R}$ *and taking values in* $\mathbb{R}$ *Suppose that* $g \circ f$ *is defined, that* f *is defined and differentiable at* $a \in \mathbb{R}$ *and that* g *is differentiable at* $f(a)$*. Then* $g \circ f$ *is differentiable at* a *and*

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

PROOF. We have not used Leibniz notation for the derivative here and we will probably never do so, but nevertheless we have all seen it and it motivates the following attempted proof:

$$(g \circ f)'(a) = \lim_{x \to a} \frac{g(f(x)) - g(f(a))}{x-a} = \left(\lim_{x \to a} \frac{g(f(x)) - g(f(a))}{f(x) - f(a)} \right) \cdot \left(\lim_{x \to a} \frac{f(x) - f(a)}{x-a} \right).$$

Since f is continuous at a, as $x \to a$ we have $f(x) \to f(a)$, and thus we may rewrite the above as

$$\left(\lim_{f(x) \to f(a)} \frac{g(f(x)) - g(f(a))}{f(x) - f(a)} \right) \cdot \left(\lim_{x \to a} \frac{f(x) - f(a)}{x-a} \right) = g'(f(a))f'(a).$$

There is however a **gap** in this argument: we multiplied and divided by $f(x) - f(a)$: what if that quantity is 0? Well, it could be, so we need to think carefully about what happens if that is the case. More precisely, the above argument works so long as $f(x) - f(a)$ is nonzero for all $x \neq a$ sufficiently close to a (at which f is defined). So the other possibility is that for all $\delta > 0$ there is x in the domain of f such that $0 < |x-a| < \delta$ such that $f(x) = f(a)$. In this case, by Lemma 2.4 we have

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

and also

$$(g \circ f)'(a) = \lim_{x \to a} \frac{g(f(x)) - g(f(a))}{x-a} = 0,$$

$$(g \circ f)'(a) = 0 = g'(f(a))f'(a).$$

This completes the proof. □

2.1. Exercises.

Exercise 2.3. *We revisit the notation and assumptions of Exercise 2.2.*
Show: $f'(x)$ *exists if and only if* $f'_+(x)$ *and* $f'_-(x)$ *both exist and are equal, in which case* $f'(x) = f'_+(x) = f'_-(x)$.

Exercise 2.4. *Let* $X \subseteq \mathbb{R}$, *and let* a *be an accumulation point of* X. *Let* $f : X \to \mathbb{R}$.

a) *Show: if* f *is continuous at* a, *then* $|f|$ *is continuous at* a.
b) *Show by example that* $|f|$ *can be continuous at* a *when* f *is not.*
c) *Show that the following are equivalent:*
   (i) f *is differentiable at* a, *and either* $f(a) \neq 0$ *or* $f(a) = f'(a) = 0$.
   (ii) f' *is differentiable at* a.

Exercise 2.5. *Prove parts a) and b) of Proposition 2.3.*

Exercise 2.6. *Prove part c) of Proposition 2.3.*

Exercise 2.7. *Prove parts d) and e) of Proposition 2.3.*

Exercise 2.8. *Almost everyone thinks at first that the product rule should read* $(fg)'(a) = f'(a)g'(a)$. *But actually this is a bad guess. Use* ***dimensional analysis*** *to explain why this cannot be correct. (That is, assign dimensional units to the input and output variables, figure out what the dimensional units of* f' *should be, and show that the dimensional units of* $(fg)'$ *are not the same as those of* $f'g'$*.)*

Exercise 2.9 (Generalized Product Rule). *Let* $X \subseteq \mathbb{R}$ *and let* $a \in X$ *be an accumulation point. Let* $f_1, \ldots, f_n : X \to \mathbb{R}$ *be functions that are each differentiable at* a. *Show:* $f_1 \cdots f_n$ *is differentiable at* a, *and*

$$(f_1 \cdots f_n)'(a) = f_1'(a)f_2(a) \cdots f_n(a) + \ldots + f_1(a) \cdots f_{n-1}(a)f_n'(a).$$

Exercise 2.10. *(Generalized Leibniz Rule) Recall that* $f^{(n)}$ *denotes the* n*th derivative of* f. *Show: for all* $n \in \mathbb{Z}^+$, *if* f *and* g *are both* n *times differentiable then so is* fg, *and*

$$(fg)^n(a) = \sum_{k=0}^{n} \binom{n}{k} f^{(k)} g^{(n-k)}.$$

Exercise 2.11. *Use the methods of Math 3100 to show that* $(\sin x)' = \cos x$ *and* $(\cos x)' = -\sin x$. *That is, define* $\sin x$ *by* $\sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}$ *and* $\cos x$ *by* $\sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!}$ *and check the identity by termwise differentation. Please cite a result from* [**SS**] *that justifies this termwise differentiation!*

Exercise 2.12. *Prove Lemma 2.4.*

3. The Mean Value Theorem

3.1. Functions increasing or decreasing at a point. Let $I \subset \mathbb{R}$ be an interval, and let a be an interior point of I. We say that a function $f : I \to \mathbb{R}$ is **increasing at a** if there is $\delta > 0$ such $(a - \delta, a + \delta) \subseteq I$ and for all $x \in (a - \delta, a)$ and $y \in (a, a + \delta)$ we have

$$f(x) \leq f(a) \leq f(y).$$

The definition of **strictly increasing at a** is the same as for increasing at a except that the conclusion is

$$f(x) < f(a) < f(y).$$

Thus the idea of being increasing (resp. strictly increasing) at a is that all points x slightly to the left of a have a function value $f(x)$ that is no larger (resp. smaller) than the value $f(a)$ at a, which is no larger (resp. smaller) than the value $f(y)$ at all points y slightly to the right of a.

We have similar definitions for **decreasing at a** and **strictly decreasing at a** where the conclusions are respectively

$$f(x) \leq f(a) \leq f(y)$$

and

$$f(x) > f(a) > f(y).$$

THEOREM 2.6. *Let I be an interval, and let $a \in I^\circ$ (that is, a is an interior point of I). Let $f : I \to \mathbb{R}$ be differentiable at a.*

a) *If $f'(a) > 0$, then f is strictly increasing at a.*
b) *If $f'(a) < 0$, then f is strictly decreasing at a.*

PROOF. In the limit definition of the derivative, take $\epsilon = f'(a)$. Then there is $\delta > 0$ such that for all $x \in I$ with $0 < |x - a| < \delta$, we have

$$\left|\frac{f(x) - f(a)}{x - a} - f'(a)\right| < f'(a),$$

or equivalently:

$$0 < \frac{f(x) - f(a)}{x - a} < 2f'(a).$$

So: • If $x \in (a-\delta, a)$, then $\frac{f(x)-f(a)}{x-a}$ is positive and $x-a$ is negative, so $f(x)-f(a)$ is negative: that is, $f(x) < f(a)$.
• If $x \in (a, a+\delta)$, then $\frac{f(x)-f(a)}{x-a}$ is positive and $x - a$ is positive, so $f(x) - f(a)$ is positive: that is, $f(x) > f(a)$.
b) This is similar enough to part a) that we leave it to the reader. One method is simply to repeat the argument of part a) with some flipped inequalities. Those who don't like direct repetition may wish to reduce to part a) using Exercise 2.13. □

We note that when $f'(a) = 0$, no conclusion can be drawn. For instance, if $f(x) = x^3$, then $f'(0) = 0$ and f is strictly increasing at 0. If $f(x) = -x^3$, then $f'(0) = 0$ and f is strictly decreasing at 0. If $f(x) = x^2$, then $f'(0) = 0$ and f is neither increasing at 0 nor decreasing at 0.

If $X \subset \mathbb{R}^N$ and $f : X \to \mathbb{R}$ is a function, we say f has a **local minimum** at $a \in X$ if there is $\delta > 0$ such that $f(a)$ is the minimum value of the restriction of f to $X \cap B^\circ(a, \delta)$: that is, for all $x \in X$ with $||x - a|| < \delta$, we have $f(x) \geq f(a)$. Similarly, we say f has a **local maximum** at $a \in X$ if there is $\delta > 0$ such that $f(a)$ is the maximum value of the restriction of X to $X \cap B^\circ(a, \delta)$: that is, for all $x \in X$ with $||x - a|| < \delta$, we have $f(x) \leq f(a)$.

Theorem 2.6 is mainly of interest to us because of the following consequence that should be familiar from calculus:

COROLLARY 2.7. *Let I be an interval, let $a \in I^\circ$. If $f : I \to \mathbb{R}$ has a local minimum or local maximum at a, then $f'(a) = 0$.*

PROOF. We will show the contrapositive: if $f'(a) \neq 0$, then f has neither a local minimum nor local maximum at a. Indeed, if $f'(a) > 0$, then by Theorem 2.6a), the function f is strictly increasing at a: thus it takes smaller values slightly to the left of a – so does not have a local minimum at a – and larger values slightly to the right of a – so does not have a local maximum at a. Similarly, if $f'(a) < 0$, then by Theorem 2.6b) the function f is strictly decreasing at a so again takes both larger and smaller values in any small interval around a, so again has neither a local minimum nor local maximum at a. $\square$

Now here comes an intriguing result showing that derivatives have the Intermediate Value Property. Let's firm up some terminology: for real numbers x, y, z, we say that y **lies between** x and z if either $x < z < y$ or $y < z < x$.

THEOREM 2.8 (Darboux). *Let $f : [a, b] \to \mathbb{R}$ be a differentiable function. If $L \in \mathbb{R}$ lies in between $f'(a)$ and $f'(b)$, then there is $c \in (a, b)$ such that $f'(c) = L$.*

PROOF. We will handle the case in which $f'(a) < L < f'(b)$; the case in which $f'(b) < L < f'(a)$ will be left to the reader.
Step 1: First suppose that $L = 0$: so $f'(a) < 0 < f'(b)$. By the Extreme Value Theorem, f attains a minimum at some point $c \in [a, b]$. Because $f'(a) < 0$, the function f is strictly decreasing at a, so we cannot have $c = a$; similarly, because $f'(b) > 0$, the function f is strictly increasing at b, so we cannot have $c = b$. Therefore $c \in (a, b)$, so by Corollary 2.7 we have $f'(c) = 0$.
Step 2: In general, we have $f'(a) < L < f'(b)$. To reduce to Step 1, we define $g : [a, b] \to \mathbb{R}$ by $g(x) := f(x) - L(x)$. Then g is again differentiable, and moreover $g'(a) = f'(a) - L < 0$ and $g'(b) = f'(b) - L > 0$, so by Step 1 there is $c \in (a, b)$ such that $0 = g'(c) = f'(c) - L$. Thus $f'(c) = L$. $\square$

This curious similarity between continuous functions and derivatives motivates us to formalize the Intermediate Value Property. Here is a clean way to do so: if I is an interval, we say a function $f : I \to \mathbb{R}$ is **Darboux** if for all $a < b$ in I, every value of $\mathbb{R}$ in between $f(a)$ and $f(b)$ lies in the image $f(I)$. In other words, if $f(a) < f(b)$ then the image $f(I)$ contains the interval $[f(a), f(b)]$, while if $f(a) > f(b)$ then the image $f(I)$ contains the interval $[f(b), f(a)]$.

REMARK. *The Intermediate Value Theorem tells us that continuous functions $f : I \to \mathbb{R}$ are Darboux: for $a < b$ in I, the Intermediate Value Theorem applied to $f|_{[a,b]} : [a, b] \to \mathbb{R}$ tells us that every value in between $f(a)$ and $f(b)$ lies in $f(I)$; exactly the same restriction argument holds if $f = g'$ is a derivative.*

Thus we know two classes of Darboux functions: continuous functions and derivatives. Actually the second class includes the first: later, as an application of the Fundamental Theorem of Calculus, we will see that every continuous function is a derivative. On the other hand, not every Darboux function is a derivative (if this were true, the problem of characterizing derivatives would have an easier answer than it actually does).

Suppose that I is an interval, $c \in I^\circ$ and $f : I \to \mathbb{R}$ is a function. We say that f has a **simple discontinuity** at c if each of the one-sided limits $\lim_{x \to c^+} f(x)$

and $\lim_{x\to c^-} f(x)$ exist but f is *not* continuous at c: this means that either $\lim_{x\to c^+} \neq f(c)$ or $\lim_{x\to c^-} f(c)$ (or both). (Alternate terminology: **jump discontinuity**.) In Exercise 2.15 you are asked to show that Darboux functions cannot have simple discontinuities. What is interesting about this result is that it does *not* say that Darboux functions must be continuous: again, the function $f'(x)$ of Example 2.2 is a dicontinuous derivative. So when Darboux functions do have discontinuities, these functions must be "complicated."

3.2. The Mean Value Theorem.

THEOREM 2.9 (Rolle's Theorem). *Let $f : [a, b] \to \mathbb{R}$ be differentiable. If $f(a) = f(b)$, then there is $c \in (a, b)$ such that $f'(c) = 0$.*

PROOF. Since f is continuous and $[a, b]$ is closed and bounded, by the Extreme Value Theorem f attains a minimum value m and a maximum value M. Suppose first that the minimum is attained at an interior point: there is $c \in (a, b)$ such that $f(c) = m$. By Corollary 2.7 we must then have $f'(c) = 0$, completing the argument in this case. Now suppose that the maximum is attained at an interior point: there is $c \in (a, b)$ such that $f(c) = M$. Again we may apply Corollary 2.7 to conclude that $f'(c) = 0$, completing the argument in this case.

So we are left to deal with the case in which the minimum and maximum both occur at the endpoints of $[a, b]$: that is, we either have $f(a) = m$ and $f(b) = M$ or we have $f(a) = M$ and $f(b) = m$. But we've assumed that $f(a) = f(b)$ so either way we have $m = M$. That is, the minimum value is the same as the maximum value, so f is constant and thus its derivative is 0 *at every point* of (a, b). $\square$

THEOREM 2.10 (Mean Value Theorem). *Let $f : [a, b] \to \mathbb{R}$ be continuous and differentiable at every $c \in (a, b)$. Then there is $c \in (a, b)$ such that*

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

PROOF. We'll deduce this from Rolle's Theorem (Theorem 2.9).

Namely, there is a unique line in $\mathbb{R}^2$ joining the points $(a, f(a))$ and $(b, f(b))$, and let $L(x)$ be this line, viewed as a function of x. (We could certainly write down a formula for $L(x)$, but we don't need to.) We put

$$g : [a, b] \to \mathbb{R} \text{ by } g(x) := f(x) - L(x).$$

Since f and L are differentiable, so is g. Moreover we have

$$g(a) = f(a) - L(a) = f(a) - f(a) = 0 = f(b) - f(b) = f(b) - L(b) = g(b).$$

Applying Rolle's Theorem to g, there is $c \in (a, b)$ such that

$$0 = g'(c) = f'(c) - L'(c),$$

so

$$f'(c) = L'(c).$$

But $L'(c)$ is the slope of L, which is $\frac{f(b)-f(a)}{b-a}$. We're done. $\square$

There is a strengthening of the Mean Value Theorem that is sometimes useful, though perhaps only for things that we will not actually cover in our course, like l'Hôpital's Rule and Taylor's Theorem. Here it is:

THEOREM 2.11 (Cauchy Mean Value Theorem). *Let $f, g : [a,b] \to \mathbb{R}$ be continuous and moreover differentiable on (a,b). Then there is $c \in (a,b)$ such that*

$$(f(b) - f(a))g'(c) = (g(b) - g(a))f'(c). \tag{5}$$

PROOF. Case 1: Suppose $g(a) = g(b)$. By Rolle's Theorem, there is $c \in (a,b)$ such that $g'(c) = 0$. For this value of c, both sides of (5) are 0 so (5) holds.
Case 2: Suppose $g(a) \neq g(b)$, and define

$$h(x) = f(x) - \left(\frac{f(b) - f(a)}{g(b) - g(a)}\right) g(x).$$

Then h is continuous on $[a,b]$, differentiable on (a,b), and

$$h(a) = \frac{f(a)(g(b) - g(a)) - g(a)(f(b) - f(a))}{g(b) - g(a)} = \frac{f(a)g(b) - f(b)g(a)}{g(b) - g(a)},$$

$$h(b) = \frac{f(b)(g(b) - g(a)) - g(b)(f(b) - f(a))}{g(b) - g(a)} = \frac{f(a)g(b) - f(b)g(a)}{g(b) - g(a)} = h(a).$$

Applying Rolle's Theorem to h, we get a $c \in (a,b)$ such that

$$0 = h'(c) = f'(c) - \left(\frac{f(b) - f(a)}{g(b) - g(a)}\right) g'(c),$$

or equivalently,

$$(f(b) - f(a))g'(c) = (g(b) - g(a))f'(c). \qquad \square$$

3.3. Mean Value Inequalities. The following result is a consequence of the Mean Value Theorem:

COROLLARY 2.12 (Mean Value Inequality). *Let $f : [a,b] \to \mathbb{R}$ be continuous on $[a,b]$ and differentiable on (a,b). Suppose there are real numbers $m \leq M$ such that*

$$\forall x \in (a,b),\ m \leq f'(x) \leq M.$$

Then we have

$$m(b - a) \leq f(b) - f(a) \leq M(b - a).$$

PROOF. By the Mean Value *Theorem*, we have $\frac{f(b)-f(a)}{b-a} = f'(c)$ for some $c \in (a,b)$. Since $m \leq f'(c) \leq M$, we have

$$m(b - a) \leq f'(c)(b - a) = f(b) - f(a) = f'(c)(b - a) \leq M(b - a). \qquad \square$$

Here is a slightly less precise consequence of Corollary 2.12:

COROLLARY 2.13 (Norm Mean Value Inequality). *Let $f : [a,b] \to \mathbb{R}$ be continuous and differentiable on (a,b). Define*

$$||f'|| := \sup_{c \in (a,b)} ||f'(c)||,$$

and assume that $||f'|| < \infty$: that is, we assume that f' is bounded on (a,b). Then:

$$||f(b) - f(a)|| \leq ||f'||(b - a).$$

PROOF. Indeed, for all $x \in (a,b)$, we have $-||f'|| \leq f'(c) \leq ||f'||$. Applying Corollary 2.12 with $m = -||f'||$ and $M = ||f'||$ gives the result. $\square$

The Norm Mean Value Inequality is arguably more intuitive than the Mean Value Theorem. If we interpret the derivative of a function as its instantaneous velocity, MVT says that there is always at least one interior point at which the instantaneous velocity at that point is equal to the average velocity over the interval $[a,b]$. While the statement is crisp and appealing, I wouldn't find it to be "obvious" if I didn't know the proof. On the other hand, the Norm Mean Value Inequality says that for $f : [a,b] \to \mathbb{R}$, an upper bound on the displacement between time a and time b is $(b-a)$ times the maximum speed. *This* seems intuitively clear, since if you are given a speed limit $L > 0$ and asked to maximize your displacement, then it seems clear that the way to do this is to either move to the right at speed L at all times, giving a displacement of $L(b-a)$ or to move to the left at speed L at all times, giving a displacement of $-L(b-a)$.

In fact this latter argument makes sense for vector-valued functions $f : [a,b] \to \mathbb{R}^M$: if a prisoner escapes their prison at time a and has a maximum speed of L, then by time b they cannot have gone any farther than $L(b-a)$ from where they started. Indeed this is mathematically correct:

THEOREM 2.14 (Vector-Valued Norm Mean Value Inequality). *Let $f : [a,b] \to \mathbb{R}^M$ be continuous and differentiable on (a,b). (Recall that the latter simply means that if $f = (f_1, \ldots, f_M)$, then each $f_j : (a,b) \to \mathbb{R}$ is differentiable.) Suppose*

$$||f'|| \coloneqq \sup_{c \in (a,b)} ||f'(c)|| < \infty.$$

Then we have

$$||f(b) - f(a)|| \leq ||f'||(b-a).$$

PROOF. [**R**, p. 113–114] Define $g : [a,b] \to \mathbb{R}$ by

$$g(t) \coloneqq (f(b) - f(a)) \cdot f(t).$$

(Note that we have taken the dot product of two vector-valued functions to get a scalar-valued function.) Simply by writing out components, one sees that since f is differentiable on (a,b), so is g, so the Mean Value theorem applies: there is $c \in (a,b)$ such that

$$g(b) - g(a) = (b-a)g'(c).$$

Now

$$\begin{aligned} g(b) - g(a) &= (f(b) - f(a)) \cdot f(b) - (f(b) - f(a)) \cdot f(a) \\ &= (f(b) - f(a)) \cdot (f(b) - f(a)) = ||f(b) - f(a)||^2, \end{aligned}$$

while

$$\begin{aligned} (b-a)g'(c) = (b-a)\Big((f(b) - f(a)) \cdot f'(c)\Big) &\leq (b-a)||f(b) - f(a)|| \cdot ||f'(c)|| \\ &\leq (b-a)||f(b) - f(a)|| \cdot ||f'||; \end{aligned}$$

we used Cauchy-Schwarz to get the first inequality and the definition of $||f'||$ to get the second. So

$$||f(b) - f(a)||^2 = g(b) - g(a) = (b-a)g'(c) \leq (b-a)||f(b) - f(a)|| \cdot ||f'||.$$

The result holds trivially if $f(b) = f(a)$; otherwise, we may divide through by $||f(b) - f(a)||$ to get

$$||f(b) - f(a)|| \leq ||f'||(b-a). \qquad \square$$

EXAMPLE 3.1. *Let* $f : [0, 2\pi] \to \mathbb{R}$ *by* $f(x) = (\cos x, \sin x)$. *Then for all* x *we have* $f'(x) = (-\sin x, \cos x)$, *so* $||f'(x)|| = 1$. *On the other hand, we have* $f(2\pi) - f(0) = (1, 0) - (1, 0) = (0, 0)$. *Thus whereas in* 1 *dimension, knowing the instantaneous speed (i.e.,* $|f'(x)|$*) at all points determines the absolute value of the displacement, in more than* 1 *dimension this is no longer the case...e.g. because you can travel in a circle at constant speed! This example seems to show that there is no higher analogue of the Mean Value* Theorem *– and indeed, no such result is known to me – whereas there is a higher dimensional analogue of the Norm Mean Value* Inequality*. In some circles this is a convincing argument for preferring MVI over MVT...but I don't see why we can't have both.*

Let me end this section with an observation that I made rather recently...in 2022, after hearing five short lectures on MVT in the same week. MVT is often motivated as a theorem that can be used by the highway patrol: suppose that a (linear!) highway has cameras set up at various checkpoints. Suppose there are cameras at checkpoints $A < B$ and that these cameras observe a car passing the first checkpoint at time a and the second checkpoint at time b. Then the average velocity is $\frac{B-A}{b-a}$. It follows from the Norm Mean Value Inequality that for at least one point the car had speed at least $\frac{B-A}{b-a}$: for if not, the displacement would be less than $B - A$. So the highway patrol can award a ticket even if they don't witness the speeding happening in between the checkpoints.

After hearing this argument three or four times in rapid succession, finally something occurred to me: how do we know that the car's position is a differentiable function of time? Honestly, I think we don't. (For what it's worth, I mentioned this to my PhD student Freddy Saia and he did not agree. It's not so much a math question as a question about how we use mathematics to model physical reality.) The differentiability on (a, b) is a hypothesis of all of these results...so maybe a savvy driver could evade the ticket by claiming to have driven nondifferentiably?

It turns out that the answer is no if we enlarge our definition of "speeding." Namely, for $M > 0$, say that a function $f : [a, b] \to \mathbb{R}$ is an **M-speeder** if for all $\delta > 0$ there are $a \leq x < y \leq b$ with $0 < y - x < \delta$ and $\frac{|f(y)-f(x)|}{y-x} > M$. In other words, you are an M-speeder if there are arbitrarily short intervals on which your average speed exceeds M. If I say so myself, I find this definition rather reasonable: covering too much distance in too small a time is exactly the unsafe behavior that speed limits are trying to prevent.

I proved the following result:

THEOREM 2.15. *Let* $f : [a, b] \to \mathbb{R}$ *be* any *function. If for some* $M > 0$ *we have* $\frac{|f(b)-f(a)|}{b-a} > M$, *then* f *is an* M*-speeder. More precisely, there is* $c \in [a, b]$ *such that for all* $\delta > 0$ *there is* $x \in [a, b]$ *with* $0 < |c - x| < \delta$ *and* $\frac{|f(c)-f(x)|}{|c-x|} > M$. *If* f *is continuous, we may take* $c \in (a, b)$.

For the proof, please see [**Cl22**]. The level of the argument *is* appropriate for our course; it's just that at our level it will probably take some time to understand it.

Theorem 2.15 generalizes the Norm Mean Value Inequality, since if f is differentiable then at the point c (which can be taken in (a,b) since differentiable functions are continuous) the given conditions imply that $|f'(c)| > M$. In fact the result holds verbatim for functions $f : [a,b] \to \mathbb{R}^M$, taking norms instead of absolute values when appropriate. Finally, in the one-dimensional case I also give a normless version with two inequalities; this generalizes the Mean Value Inequality.

3.4. Exercises.

EXERCISE 2.13. *Let $I \subseteq \mathbb{R}$ be an interval, and let a be an interior point of I.*

a) *Show: f is increasing at a if and only if $-f$ is decreasing at a.*
b) *Show: f is strictly increasing at a if and only if $-f$ is strictly decreasing at a.*

EXERCISE 2.14. *Treat the case of Theorem 2.8 in which $f'(a) > L > f'(b)$.*

EXERCISE 2.15. *Let I be an interval, let $c \in I$, and suppose $f : I \to \mathbb{R}$ has a simple discontinuity at c. Show: f is* not *a Darboux function.*

EXERCISE 2.16. *Let I be an interval, and let $f : I \to \mathbb{R}$ be differentiable.*

a) *Show: if $f'(x) \geq 0$ for all $x \in I$, then f is increasing.*
b) *Show: if $f'(x) > 0$ for all $x \in I$, then f is strictly increasing.*
c) *Show: if $f'(x) \leq 0$ for all $x \in I$, then f is decreasing.*
d) *Show: if $f'(x) < 0$ for all $x \in I$, then f is strictly decreasing.*

EXERCISE 2.17. *Let I be an interval, and let $f : I \to \mathbb{R}$ be differentiable.*

a) *Suppose that $f'(x) \geq 0$ for all $x \in I$, so by Exercise 2.16, f is increasing. If we moreover had that $f'(x) > 0$ for all $x \in I$, then by Exercise 2.16b) f would be strictly increasing. However, f may be strictly increasing even when $f'(x) = 0$ for some $x \in I$. Show that the following are equivalent:*
   (i) *f is* not *strictly increasing.*
   (ii) *There are $a < b$ in I such that $f|_{[a,b]} : [a,b] \to \mathbb{R}$ is constant.*
b) *Use the criterion of part a) to show that for all odd integers $n \geq 1$, the function $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^n$ is strictly increasing.*
c) *State an analogous criterion for a function to be strictly decreasing. (You need not prove it.)*

EXERCISE 2.18. *Let $f : I \to \mathbb{R}$ be a differentiable function.*

a) *[Zero Velocity Theorem]*
   Show: if $f'(x) = 0$ for all $x \in I$, then f is constant.
b) *[(Almost) Uniqueness of Antiderivatives] Suppose that $f, g : I \to \mathbb{R}$ are differentiable functions and that $f' = g'$. Show: there is $C \in \mathbb{R}$ such that $g = f + C$: i.e., for all $x \in I$, we have $g(x) = f(x) + C$.*

EXERCISE 2.19. *Let $k \in \mathbb{Z}^+$. Suppose that the kth derivative $f^{(k)}$ of f exists and is identically 0: $f^{(k)}(x) = 0$ for all $x \in I$. Show: f is a polynomial function of degree at most k. (Suggestion: use induction on k.)*

EXERCISE 2.20. *Let I be an interval, let $c \in I^\circ$, and let $f : I \to \mathbb{R}$. We suppose:*

(i) *f is continuous.*
(ii) *For all $x \in I \setminus \{c\}$, f is differentiable at x.*
(iii) $\lim_{x \to c} f'(x) = L$ *exists (as a real number).*

Show: f is differentiable at c and $f'(c) = L$.

4. Inverse Functions

4.1. The Interval Image Theorem. Let us extend the notion of "betweenness" from $\mathbb{R}$ to $\mathbb{R}^N$. If x and y are distinct points of $\mathbb{R}^N$, we say that $z \in \mathbb{R}^N$ **lies between x and y** if z lies on the line segment

$$\overline{xy} := \{(1-t)x + ty \mid t \in [0,1]\}.$$

A subset X of $\mathbb{R}^N$ is **convex** if whenever it contains distinct points x and y it also contains all points in between, i.e., it also contains the line segment $\overline{xy}$.

You are asked to show in an Exercise that if for all i in a set I we have a convex subset X_i of $\mathbb{R}^N$, then the intersection $\bigcap_{i \in I} X_i$ is convex, but the union $\bigcup_{i \in I} X_i$ need not be.

It is easy to see that intervals in $\mathbb{R}$ are convex. In fact:

Proposition 2.16. *For a nonempty subset $X \subseteq \mathbb{R}$, the following are equivalent:*
(i) *X is an interval.*
(ii) *X is convex.*

Proof. (i) $\implies$ (ii): this is not hard but is somewhat tedious because there are so many different kinds of intervals. The first kind is $\mathbb{R}$ itself, which is certainly convex. The next are (a,∞), $[a,\infty)$, $(-\infty,b)$ and $(-\infty,b]$. If $x < y$ are both in (a,∞), then $a < x$, so if $x < z < y$ then $a < x < z$ and thus $z \in (a,\infty)$. Variations of this same easy argument handle $[a,\infty)$, $(-\infty,b)$ and $(-\infty,b]$. The remaining intervals are the bounded intervals $[a,b]$, $[a,b)$, $(a,b]$ and (a,b). Each of these is the intersection of two unbounded intervals: e.g. $[a,b] = [a,\infty) \cap (-\infty,b]$, so they are convex by Exercise 2.26.
(ii) $\implies$ (i): Let $X \subseteq \mathbb{R}$ be a nonempty convex subset.
Case 1: Suppose that X is unbounded above and below. Then for all $N \in \mathbb{Z}^+$ there is $x \in X$ with $x \leq -N$ and $y \in X$ with $y \geq N$, so by convexity $[-N,N] \subseteq X$. Since this holds for all N, we have $X = \mathbb{R}$.
Case 2: Suppose that X is unbounded above and bounded below; let a be the infimum of X.
Case 2a): Suppose that $a \in X$. We claim that $X = [a,\infty)$. Indeed, since X contains its infimum, a is the minimum element of X, so $X \subseteq [a,\infty)$. Conversely, if $x \in \mathbb{R}$ is greater than a, then because X is unbounded above there is $b \in X$ such that $x < b$. Thus $a,b \in X$ and $x \in [a,b]$, so by convexity we have $x \in X$. So $X = [a,\infty)$.
Case 2b): Suppose that $a \notin X$. In a very similar manner we can show that $X = (a,\infty)$.
Case 3: Suppose that X is unbounded below and bounded above. If b is the supremum of X, then as in Case 2 we can show that $X = (-\infty,b]$ (if $b \in X$) or $X = (-\infty,-b)$ (if $b \notin X$).
Case 4: Suppose that X is bounded above and below: let $a = \inf X$ and $b = \sup X$. Then again, we can show that X is one of $[a,b]$, $[a,b)$, $(a,b]$ or (a,b) depending on whether it contains a and on whether it contains b. $\square$

Proposition 2.16 gives a simple description of all nonempty convex subsets of $\mathbb{R}$. Nothing like this holds for $\mathbb{R}^N$ for any $N \geq 2$: convex subsets are too complicated to be "classified" in this way, although there is a rich theory concerning them.

In Exercise 2.27 you are asked to deduce from Proposition 2.16 that boxes in $\mathbb{R}^N$ are convex. More generally, if for $N_1, \ldots, N_k$ we are given a convex subset X_k of $\mathbb{R}^{N_k}$, then the Cartesian product

$$X = \prod_{i=1}^{k} X_i$$

is a convex subset of $\prod_{i=1}^{k} \mathbb{R}^{N_k} = \mathbb{R}^{N_1 + \ldots + N_k}$: Exercise 2.28.

PROPOSITION 2.17. *Closed balls in $\mathbb{R}^N$ are convex.*

PROOF. Being convex is a translation-invariant property of a subset of $\mathbb{R}^N$: that is, if $X \subseteq \mathbb{R}^N$, and $a \in \mathbb{R}^N$, then X is convex if and only if

$$a + X := \{a + x \mid x \in X\}$$

is convex: this is true because the translate of a line segment is another line segment. So it suffices to show that for $r > 0$ the closed ball $B^\bullet(0,r)$ is convex. So, let $P, Q \in B^\bullet(0,r)$, and let $t \in [0,1]$. Then we need to show that $||(1-t)P + tQ|| \leq r$. Thank goodness for the Triangle Inequality:

$$||(1-t)P + tQ|| \leq ||(1-t)P|| + ||tQ||$$
$$= (1-t)||P|| + (t)||Q|| \leq (1-t)r + tr = r. \qquad \square$$

REMARK. *Using Proposition 2.17 it is not hard to show that for any $a \in \mathbb{R}^N$ and $r > 0$, if $X \subseteq \mathbb{R}^N$ satisfies:*

$$B^\circ(a,r) \subseteq X \subseteq B^\bullet(a,r),$$

then X is convex. This suggests that convexity of a subset of $\mathbb{R}^N$ is unaffected by what happens "on the boundary." And this suggestion is correct...though it is not important enough to us to stop and formalize it, beyond noting that a convex subset certainly need not *be either open or closed.*

Although our interest in convexity in Euclidean space is sincere, at the present time our goal is something very specific:

THEOREM 2.18 (Subinterval Image Theorem). *Let I be an interval. For a function $f : I \to \mathbb{R}$, the following are equivalent:*

(i) *The function f is Darboux.*
(ii) *For every subinterval $J \subseteq I$, the image $f(J)$ is an interval.*

PROOF. (i) $\implies$ (ii): Suppose f is Darboux. Let c and d be distinct points of $f(J)$, so there are $a, b \in J$ with $f(a) = c$ and $f(b) = d$; interchanging c and d if necessary, we may assume that $a < b$. Because f is Darboux, $f([a,b])$ contains the interval between c and d, hence so does $f(J)$, since $f(J) \supseteq f([a,b])$. We have shown that $f(J)$ is convex, which by Proposition 2.16 implies that $f(J)$ is an interval.
(ii) $\implies$ (i): Let $a < b$ be points of I. By assumption $f([a,b])$ is an interval, hence $f([a,b])$ is a convex subset that contains $f(a)$ and $f(b)$, hence $f([a,b])$ contains the interval between $f(a)$ and $f(b)$. This is the definition of a Darboux function. $\square$

Thus Darboux functions are precisely the functions $f : I \to \mathbb{R}$ that map each subinterval of I onto a subinterval of $\mathbb{R}$. However, just having $f(I)$ be an interval is **not enough** for f to be Darboux: Exercise 2.29 gives a counterexample.

COROLLARY 2.19. *If I is an interval and $f : I \to \mathbb{R}$ is either continuous or the derivative of some other function, then $f(I)$ is an interval.*

PROOF. The Intermediate Value Theorem proved in the previous course [**SS**, Thm. 2.7.6] says precisely that continuous functions are Darboux, whereas Darboux's Theorem (Theorem 2.8) and Remark 3.1 tells us that derivatives are Darboux. Either way, Theorem 2.18 applies.[3] $\square$

It follows almost immediately from Corollary 2.19 and the Extreme Value Theorem that if $f : [a, b] \to \mathbb{R}$ is continuous, then $f([a, b])$ is a closed bounded interval: this is Exercise 2.23a). (This is precisely the conjunction of the Intermediate and Extreme Value Theorems from calculus, but for some reason stating it this way does not seem to be standard.) The case where I is *not* both closed and bounded is more interesting: in Exercise 2.23b) you are asked to show that if I and J are two intervals in $\mathbb{R}$ such that $I \neq [a, b]$ for any $a \leq b$ then there is a continuous function $f : I \to \mathbb{R}$ scuh that $f(I) = J$.

4.2. Inverses of Continuous Functions. Our main goal of this section is to prove the following result.

THEOREM 2.20. *Let I be an interval, let $f : I \to \mathbb{R}$ be an injective continuous function, and put $J := f(I)$. Then:*

a) *$f : I \to J$ is a bijection, so it has an inverse function $f^{-1} : J \to I$.*
b) *The set J is also an interval.*
c) *The inverse function $f^{-1} : J \to \mathbb{R}$ is also continuous.*

Regarding Theorem 2.20, part a) is a general fact: if $f : X \to Y$ is injective, then f also defines a surjective function from X to $f(X)$. (Every function can be made surjective by replacing its codomain with its image.) Thus $f : X \to f(X)$ is a bijection, so it has a unique inverse function $f^{-1} : f(X) \to X$. Part b) is precisely Theorem 2.18. What remains is to show that f^{-1} is also continuous: this is the crux.

We will come at this rather indirectly.

LEMMA 2.21. *Let X and Y be subsets of $\mathbb{R}$, and let $f : X \to Y$ be a bijection, with inverse function $f^{-1} : Y \to X$.*

a) *If f is strictly increasing, then f^{-1} is strictly increasing.*
b) *If f is strictly decreasing, then f^{-1} is strictly decreasing.*

PROOF. a) Let $y_1 < y_2$ in Y. We want to show that $f^{-1}(y_1) < f^{-1}(y_2)$. Seeking a contradiction, we assume otherwise: $f^{-1}(y_1) \geq f^{-1}(y_2)$. Since f^{-1} is a bijection, we cannot have equality, so we must have $f^{-1}(y_2) < f^{-1}(y_1)$. Applying the strictly increasing function f to both sides, we get

$$y_2 = f(f^{-1}(y_2)) < f(f^{-1}(y_1)) = y_1,$$

so $y_2 < y_1$, a contradiction.
b) Let $y_1 < y_2$ in Y. We want to show that $f^{-1}(y_1) > f^{-1}(y_2)$. Again, suppose not;

[3]In fact for every subinterval J of I we know that $f(J)$ is a subinterval. But restricting a continuous function or a derivative to a subinterval, we still get a continuous function or a derivative, so this more complicated conclusion doesn't carry any more content.

then we must have $f^{-1}(y_1) < f^{-1}(y_2)$. Applying the strictly decreasing function f to both sides, we get

$$y_1 = f(f^{-1}(y_1)) > f(f^{-1}(y_2)) = y_2,$$

so $y_1 > y_2$, a contradiction. $\square$

LEMMA 2.22 (Λ-V Lemma). *Let I be an interval, and let $f : I \to \mathbb{R}$. The following are equivalent:*

(i) *The function f is neither strictly increasing nor strictly decreasing.*
(ii) *At least one of the following holds:*
  (a) *f is* not *injective: there are $a < b$ in I with $f(a) = f(b)$.*
  (b) *f admits a Λ-**configuration**: there are $a < b < c$ in I with $f(a) < f(b) > f(c)$.*
  (c) *f admits a V-**configuration**: there are $a < b < c$ in I with $f(a) > f(b) < f(c)$.*

Lemma 2.22 is called the "Λ-V Lemma" because of the shapes of the letters Λ and V: if a function defined on an interval is injective and not monotone, then we can find three points $a < b < c$ in I such that value of the function goes up from a to b and then goes down from b to c – like the letter Λ – *or* we can find three points $a < b < c$ in I such that the value of the function goes down from a to b and then goes up from b to c – like the letter V. The idea of the proof is straightforward: if f is injective and neither increasing nor decreasing then there are points $a < b$ such that $f(a) > f(b)$ (if not, f would be increasing) and there are also points $c < d$ such that $f(c) < f(d)$ (if not, f would be decreasing).

Now there are various cases depending upon the ordering of the points a, b, c, d and of the points $f(a), f(b), f(c), f(d)$. For instance: suppose that $a < b < c < d$. If $f(b) < f(c)$, then going from a to b to c gives a V-configuration. On the other hand, if $f(b) > f(c)$, then going from b to c to d gives a V-configuration. For one more case, suppose that $a < c < b < d$. If $f(a) < f(c)$, then $f(a) < f(c) > f(a) > f(b)$, so going from a to c to b gives a Λ-configuration. If on the other hand $f(a) > f(c)$, then $f(a) > f(c) < f(d)$, so going from a to c to d gives a V-configuration. In Exercise 2.21 you are asked to work out the remaining cases.

THEOREM 2.23. *Let I be an interval, and let $f : I \to \mathbb{R}$ be an injective Darboux function. Then f is monotone.*

PROOF. It suffces to show that if f is injective and not monotone, then f is not Darboux. By Lemma 2.22, f admits either a Λ-configuration or a V-configuration.

• Suppose first that f admits a Λ-configuation: there are $a < b < c$ in I with $f(a) < f(b) > f(c)$. Then $\max(f(a), f(c)) < f(b)$, and if $L \in (\max(f(a), f(c)), f(b))$, then we have

$$f(a) < L < f(b) > L > f(c).$$

Since f is Darboux, there is $d \in (a, b)$ such that $f(d) = L$ and also $e \in (b, c)$ such that $f(e) = L$. But then $f(d) = L = f(e)$, contradicting the injectivity of f.

• Now suppose that f admits a V-configuration: there are $a < b < c$ with $f(a) > f(b) < f(c)$. Then a very similar argument shows that for every $L \in (f(b), \min(f(a), f(c))$ then there is $d \in (a, b)$ with $f(d) = L$ and also $e \in (b, c)$ with $f(e) = L$, so $f(d) = L = f(e)$, contradicting the injectivity of f. $\square$

LEMMA 2.24. *Let I be an interval, and let $f : I \to \mathbb{R}$ be a monotone function.*

(a) *Suppose that* $c \in I^{\circ}$. *Then* $\lim_{x \to c^+} f(x)$ *and* $\lim_{x \to c^-} f(x)$ *both exist. Moreover, if* f *is increasing, then*

$$\lim_{x \to c^-} f(x) = \sup\{f(x) \mid x < c\} \leq f(c) \leq \inf\{f(x) \mid x > c\} = \lim_{x \to c^+} f(x),$$

while if f *is decreasing, then*

$$\lim_{x \to c^-} f(x) = \inf\{f(x) \mid x < c\} \geq f(c) \geq \sup\{f(x) \mid x > c\} = \lim_{x \to c^+} f(x).$$

b) *Suppose that* I *has a left endpoint* a. *Then* $\lim_{x \to a^+} f(x)$ *exists. Moreover, if* f *is increasing then*

$$f(a) \leq \lim_{x \to a^+} f(x),$$

while if f *is decreasing then*

$$f(a) \geq \lim_{x \to a^+} f(x).$$

c) *Suppose that* I *has a right endpoint* b. *Then* $\lim_{x \to b^-} f(x)$ *exists. Moreover, if* f *is increasing then*

$$f(b) \geq \lim_{x \to b^-} f(x),$$

while if f *is decreasing then*

$$f(b) \leq \lim_{x \to b^-} f(x).$$

You are asked to prove Lemma 2.24 in Exercise 2.24.

Corollary 2.25. *Let* I *be an interval. For a monotone function* $f : I \to \mathbb{R}$, *the following are equivalent:*

(i) f *is continuous.*
(ii) f *is Darboux.*
(iii) $f(I)$ *is an interval.*

Proof. Corollary 2.19 gives (i) $\Longrightarrow$ (ii) and (i) $\Longrightarrow$ (iii) even without the assumption that f is monotone.
(ii) $\Longrightarrow$ (i): If f is Darboux, then by Exercise 2.15 it cannot have simple discontinuities, while by Lemma 2.24 since f is monotone, it can only have simple discontinuities. So f is continuous!
(iii) $\Longrightarrow$ (i): We will treat the case in which f is increasing; the case in which f is decreasing is very similar and left to the reader. We will also argue by contrapositive, so suppose that f fails to be continuous at some point $c \in I$. Suppose first that c is an interior point of I. By Lemma 2.24 we must have

$$m := \lim_{x \to c^-} f(x) < \lim_{x \to c^+} f(x) =: M.$$

Thus we must have either $m < f(c)$ or $f(c) < M$ (or both). If $m < f(c)$, choose a point a of I that is less than c and choose L such that $m < L < f(c)$. Then $f(a) \leq m < L < f(c)$, but L is not a value of f: since f is increasing, the value L would have to be taken on the subinterval $[a, c)$, but for all $x \in [a, c)$ we have $f(x) \leq m < L$. If $f(c) < M$, choose a point b of I that is greater than c and choose L such that $f(c) < L < M$. Then similarly $f(c) < L < M \leq f(b)$ but L is not a value of f. Either way, $f(I)$ is not convex and thus is not an interval.

The cases in which c is an endpoint are similar. For instance, if c is the left endpoint of I, discontinuity at c means $f(c) < M := \inf\{f(x) \mid c < x\}$. Choose L

such that $f(c) < L < M$, and let $b \in I \setminus \{c\}$. Then $f(c) < L < M \leq f(b)$ and L is not a value of f, so f is not convex and thus not an interval. $\square$

And now we are ready to prove Theorem 2.20c)! Indeed, let I be an interval, let $f : I \to \mathbb{R}$ be an injective continuous function, and let J be the interval $f(I)$. Then $f : I \to J$ is a continuous bijection, with inverse function $f^{-1} : J \to I$. By Theorem 2.23 f is a monotone injection, hence either strictly increasing or strictly decreasing. By Lemma 2.21 it follows that f^{-1} is also a monotone injection. Finally, we know that $f^{-1}(J) = I$ is an interval, so applying Corollary 2.25 to $f^{-1} : J \to \mathbb{R}$, we get that f^{-1} is continuous.

4.3. Inverses of Differentiable Functions.

THEOREM 2.26. *Let I and J, be intervals, and let $f : I \to J$ be a continuous bijection. Let $b \in J^\circ$, and put $a := f^{-1}(b)$. Suppose that f is differentiable at a and $f'(a) \neq 0$. Then f^{-1} is differentiable at b and*

$$(f^{-1})'(b) = \frac{1}{f'(a)} = \frac{1}{f'(f^{-1}(b))}. \tag{6}$$

Before giving the proof we remark that most of the content lies in the assertion that f^{-1} is differentiable at b. If we assume that f^{-1} is differentiable, then differentiating $f(f^{-1}(x)) = x$ gives $f'(f^{-1}(x))(f^{-1})'(x) = 1$, so $(f^{-1})'(x) = \frac{1}{f'(f^{-1}(x))}$.

PROOF. We have

$$(f^{-1})'(b) = \lim_{h \to 0} \frac{f^{-1}(b+h) - f^{-1}(b)}{h} = \lim_{h \to 0} \frac{f^{-1}(b+h) - a}{h}.$$

Since $J = f(I)$, for each $b + h \in J$ (we assume h is sufficiently small so that $b + h$ does lie in J; we can assume this since $b \in J^\circ$) there is a unique $k_h \in \mathbb{R}$ such that

$$b + h = f(a + k_h).$$

Then

$$f^{-1}(b + h) = a + k_h \text{ and } h = f(a + k_h) - f(a).$$

Making these substitutions, we get

$$(f^{-1})'(b) = \lim_{h \to 0} \frac{a + k_h - a}{f(a + k_h) - f(a)} = \lim_{h \to 0} \left(\frac{f(a + k_h) - f(a)}{k_h} \right)^{-1}.$$

So we are clearly on the right track: if we can show that $\lim_{h \to 0} k_h = 0$, then the above limit is $\frac{1}{f'(a)}$ and we will be done.

Now we have $k_h = f^{-1}(b + h) - a$. Moreover, by Theorem 2.20, since f is continuous so is f^{-1}, and thus

$$\lim_{h \to 0} k_h = \lim_{h \to 0} f^{-1}(b + h) - a = f^{-1}(b) - a = a - a = 0,$$

and we're done. $\square$

4.4. Several Variables. As I have hinted before, the concept of a continuous function is actually one of the most general in all of mathematics: whenever we have sets X and Y each endowed with classes of subsets called "open" (and required to satisfy several familiar properties), one can define the notion of a continuous function $f : X \to Y$.[4] Much what we said about continuous functions applies in this level of generality: e.g. continuous functions preserve limits of sequences and compositions of continuous functions are continuous. However, it is **not** a general fact that a continuous bijection must have a continuous inverse – not even close. In fact this does not even hold when X and Y are subsets of Euclidean space.

Example 4.1. *Let $X = [0, 2\pi)$ and let $Y = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$ be the unit circle. Let $f : X \to Y$ by $f(x) = (\cos x, \sin x)$. This is a continuous function because cosine and sine are both continuous functions. It is also a bijection: the inverse function g maps a point on the unit circle to its angle, taken to lie in $[0, 2\pi)$. However, g is* not *continuous at the point $Q = (1, 0)$, since points with angle slightly less than 2π are very close to Q, but under g these points get mapped almost 2π units away from $g(Q) = 0$.*

It turns out that if $X \subseteq \mathbb{R}^N$ is closed and bounded and $f : X \to \mathbb{R}^M$ is a continuous injection, so $f : X \to f(X)$ is a continuous bijection, then the inverse function $f^{-1} : f(X) \to X$ is necessarily continuous. This uses the sequential compactness of X and is actually not terribly difficult: Exercises 2.24 and 2.25 cover this.

The following result *is* terribly difficult:

Theorem 2.27 (Brouwer, 1912). *Let $U \subseteq \mathbb{R}^N$ be an open subset, and let $f : U \to \mathbb{R}^N$ be a continuous injection. Then $V \coloneqq f(U)$ is open in $\mathbb{R}^N$, and the inverse function $f^{-1} : V \to U$ is also continuous.*

It would be possible to prove this result in Math 8200 (Algebraic Topology), as an application of singular homology. If you ever take this course, ask for it!

Finally, there is an inverse function theorem for differentiable functions $f : U \to \mathbb{R}^N$, where $U \subseteq \mathbb{R}^N$ is an open subset. Although we barely mentioned the total derivative, we can still state the result. Write $f = (f_1, \ldots, f_N)$, and let $P \in U$. Suppose that on some smaller open subset containing P, each of the partial derivatives $\frac{\partial f_i}{\partial x_j}$ exists and is continuous. (With the extra hypothesis of *continuity* of the partials, this implies that f is differentiable in the "total" sense we mentioned above.) We can therefore form the **Jacobian matrix** $J(f)(P)$, whose i, j entry is $\frac{\partial f_i}{\partial x_j}(P)$. We suppose that this matrix is invertible (this is the higher-dimensional analogue of assuming that $f'(a) \neq 0$). Then there is an open subset V of $\mathbb{R}^N$ containing P such that $W \coloneqq f(V)$ is an open subset of $\mathbb{R}^N$ containing $f(P)$ – so $f : V \to W$ is a bijection – and moreover $f^{-1} : W \to V$ has continuously differentiable partials for all $y \in W$, and indeed if $y = f(x)$ is in W, the Jacobian matrix $J(f^{-1})(y)$ is the inverse of the Jacobian matrix $J(f)(x)$. For a proof, see [**R**, Thm. 9.24].

[4] It is actually only a little trouble to be much less vague: a **topological space** is a set X endowed with a family τ of subsets of X. This family is required to satisfy: (i) $\varnothing, X \in \tau$; (ii) if for all $i \in I$, U_i lies in τ then so does $\bigcup_{i \in I} U_i$; (iii) if $U_1, \ldots, U_n$ lie in τ, then so does $\bigcap_{i=1}^n U_i$. The elements of τ are called **open subsets** of X. Taking $X = \mathbb{R}^N$ and τ to be the open sets as we defined them at the beginning of the course is a – very, very important! – example of a topological space.

4.5. Exercises.

Exercise 2.21. *Prove the* Λ-V *Lemma* [**HC**, Lemma 5.38].

Exercise 2.22. *Prove Lemma 2.24.*

Exercise 2.23. *Let* $I, J \subset \mathbb{R}$ *be intervals*[5].

a) *Suppose* $I = [a,b]$ *and* $f : I \to \mathbb{R}$ *is a continuous function. Show: there are real numbers* $m \leq M$ *such that* $f(I) = [m, M]$.
b) *Suppose that* I *is* not *a closed, bounded interval. Show: there is a continuous function* $f : I \to \mathbb{R}$ *with* $f(I) = J$.

Exercise 2.24. *Let* $X \subseteq \mathbb{R}^N$ *and let* $f : X \to \mathbb{R}^M$ *be a function. Show that the following are equivalent:*

(i) f *is continuous.*
(ii) *For all open subsets* V *of* $\mathbb{R}^M$, $f^{-1}(V) := \{x \in X \mid f(x) \in V\}$ *is of the form* $U \cap X$ *for an open subset* U *of* $\mathbb{R}^N$.
(iii) *For all closed subets* V *of* $\mathbb{R}^M$, $f^{-1}(V)$ *is of the form* $A \cap X$ *for a closed subset* A *of* $\mathbb{R}^N$.

Exercise 2.25. *Let* $X \subseteq \mathbb{R}^N$ *be closed and bounded, and let* $f : X \to \mathbb{R}^M$ *be a continuous injection. Let* $Y := f(X)$, *so* $f : X \to Y$ *is a continuous bijection. Show that* Y *is closed and bounded and* $f^{-1} : Y \to X$ *is continuous.*
(Suggestion: if $Z \subseteq X$ *is closed, then* $f(Z)$ *is closed; apply Exercise 2.24.)*

Exercise 2.26.

a) *Let* I *be a nonempty set, and let* $\{X_i\}_{i \in I}$ *be any family of convex subsets of* $\mathbb{R}^N$. *Show:* $\bigcap_{i \in I} X_i$ *is also convex.*
b) *Show by example that the union of two convex subsets of* $\mathbb{R}^N$ *need* not *be convex.*

Exercise 2.27. *Show the closed box* $B = \prod_{i=1}^N [a_i, b_i]$ *is a convex subset of* $\mathbb{R}^N$.

Exercise 2.28. *Let* $N_1, \ldots, N_k$ *be positive integers, and put* $N := N_1 + \ldots N_k$. *We may view* $\mathbb{R}^N$ *as the Cartesian product* $\prod_{i=1}^k \mathbb{R}^{N_i}$. *(Just concatenate the coordinates of the various vectors into one vector.) For* $1 \leq i \leq k$, *let* X_i *be a convex subset of* $\mathbb{R}^{N_i}$. *Show that the Cartesian product* $\prod_{i=1}^k X_i$ *is a convex subset of* $\mathbb{R}^N$.

Exercise 2.29. *Consider the following function* $f : \mathbb{R} \to \mathbb{R}$:

$$f(x) = \begin{cases} -5 & x < 0 \\ x \sin x & x \geq 0 \end{cases}.$$

a) *Show* $f(\mathbb{R}) = \mathbb{R}$.
b) *Show:* f *is* not *a Darboux function. This shows that being Darboux is stronger than having the image be an interval.*

[5] Our convention is that the empty set is *not* an interval.

CHAPTER 3

The Riemann Integral

1. Abstract Integrals and the Fundamental Theorem of Calculus

We now begin our study of "the integral calculus." The basic idea here is as follows: for a function $f : [a, b] \to \mathbb{R}$ we wish to associate a real number $\int_a^b f$, the **definite integral**. When f is non-negative, our *intutition* is that $\int_a^b f$ should represent the area under the curve $y = f(x)$ — more precisely the area of the region bounded above by $y = f(x)$, below by $y = 0$, on the left by $x = a$ and on the right by $x = b$. For general functions f, the integral $\int_a^b f$ is supposed to represent the **signed area** — more on this later.

The above sentiment is roughly analogous to the intuition that a continuous function is one whose graph is a "nice, unbroken" curve. Namely, it is a geometric idea that must be analytically formalized, and whose analytic formalization requires further ideas. The above gives a precise description of a subset of the plane associated to $f : [a, b] \to [0, \infty)$, namely the set

$$S_f := \{(x, y) \in \mathbb{R}^2 \mid a \leq x \leq b \text{ and } 0 \leq y \leq f(x)\}.$$

It is easy to see that S_f is bounded if and only if f is bounded (Exercise 3.1). So if we knew how to assign an area to every bounded subset of $\mathbb{R}^2$, then this would work as a definition. The issue is that this "assigning areas" problem is itself a very challenging one: the part of mathematics that deals with this in a satisfactory way is called **measure theory**, which is part of *graduate* real analysis.

So our main task here is to define a new limiting process telling us how to assign the real number $\int_a^b f$ to the function $f : [a, b] \to \mathbb{R}$. Just as for all previous limiting processes (limits of sequences and series, functional limits at a point, continuity, differentiability) the limit need not exist for all functions, and indeed there are some functions $f : [a, b] \to \mathbb{R}$ for which $\int_a^b f$ is not defined. (This is true both for the particular limiting processes that we will study but also, for certain choices of f, for *any* reasonable limiting process.) Just as we call a function differentiable if the limiting process defining the limit exists, we will call a function **integrable** if it lies in the class of functions for which the limiting process works to assign a number $\int_a^b f$. (This is not yet a definition since we haven't said what the process is!)

Before we plunge into the details of a particular limiting process, it will be helpful to consider some properties that we want our integral to study. If the integral is supposed to be a signed area, it should surely satisfy the following properties:

(I1) If $f = C$ is constant, then $\int_a^b C = C(b - a)$.

Indeed, when $C > 0$ the set S_f is just a rectangle with base length $b - a$ and with height C: that's an area that we know. When $C < 0$, the set S_f is a rectangle with the same description, but now it is bounded *above* by the x-axis and *below* by $y = C$, so our convention is that this counts as "negative area." When $C = 0$, the set S_f is just the line segment $[a, b]$, which indeed should have area 0.

Comment: Until further notice, we will "explain" our properties only for non-negative functions f. This case is simpler and easier to explain. Once we sufficiently develop the theory we will be able to understand how to recover the general case from this (essentially we add a sufficiently large constant to make f non-negative).

(I2) If $f_1, f_2 : [a, b] \to \mathbb{R}$ satisfy $f_1 \leq f_2$ — that is, for all $x \in [a, b]$ we have $f_1(x) \leq f_2(x)$ — then $\int_a^b f_1 \leq \int_a^b f_2$.

Under our running "explanatory assumption" that f_1 and f_2 are non-negative, if $f_1 \leq f_2$ then S_{f_1} is a subset of S_{f_2}, and certainly the area of a subset should be less than or equal to the area of the entire set.

(I3) If $f : [a, b] \to \mathbb{R}$ and $a \leq c \leq b$, then $\int_a^b f = \int_a^c f + \int_c^b f$.

To explain this, again under the additional assumption that $f \geq 0$, we will add to our notation by writing $S_{f,[a,b]}$ for what we above wrote as S_f, taking the interval $[a, b]$ as known. Then we have

$$S_{f,[a,b]} = S_{f,[a,c]} \cup S_{f,[c,b]}$$

and $S_{f,[a,c]} \cap S_{f,[c,b]}$ is just the vertical line segment from $(c, 0)$ to $(c, f(c))$, which should have area 0. The way we think areas should work is that the area of the union should be the sum of the areas minus the area of the intersection, so this explains (I3).

Again, let me emphasize: I am not *proving* (I1), (I2) and (I3). I couldn't possibly do that until I tell you what $\int_a^b f$ means. I am just writing down some desired consequences of any reasonable definition of $\int_a^b f$. Or, if you like, we are writing down **axioms** that our integration process should satisfy.

In fact, I do like – I find the axiomatic approach to be a clean way to come at this problem. To make it work completely, I want to add one more ingredient: what is the "domain." Namely, suppose we are given a subset $\mathcal{R}[a, b]$ of the set of all functions $f : [a, b] \to \mathbb{R}$ that we call the **integrable functions**.

(There is a little fine print here: first of all, we actually mean to define $\mathcal{R}[a, b]$ for each pair of real numbers (a, b) with $a \leq b$. Second of all, if $a \leq c \leq b$ and $f \in \mathcal{R}[a, b]$, we want $f|_{[a,c]} : [a, c] \to \mathbb{R}$ to lie in $\mathcal{R}[a, c]$ and $f|_{[c,b]} : [c, b] \to \mathbb{R}$ to lie in $\mathcal{R}[c, b]$. This is necessary to make sense of Axiom (I3), for instance.)

Having done this, an **integral** is, for each $a \leq b$, a function

$$\int : \mathcal{R}[a, b] \to \mathbb{R}, \ f \mapsto \int_a^b f$$

that should satisfy the above axioms. This means that we want constant functions to be integrable and that we require (I2) and (I3) to hold for functions $f_1, f_2, f \in \mathcal{R}[a, b]$.

However, we need one more thing in order to be sure we are doing something nontrivial. That is...we need to say something about $\mathcal{R}[a, b]$, the set of integrable functions. The only functions that our axioms ensure lie in $\mathcal{R}[a, b]$ are the constant functions. So we could take $\mathcal{R}[a, b]$ to consist of constant functions and then we are only talking about signed areas of rectangles. One step away would be to take $\mathcal{R}[a, b]$ to be all polynomial functions. In this case, verfiying the axioms corresponds roughly to the amount of understanding posssessed by a B-level calculus student: we just need to know to reverse the power rule for differentiation.

So let us sneak in one more axiom to ensure that there is some content here:

(I0) For all real numbers $a < b$ we have that:
(I0(a)) Every continuous function $f : [a, b] \to \mathbb{R}$ lies in $\mathcal{R}[a, b]$; and
(I0(b)) Every function $f \in \mathcal{R}[a, b]$ is bounded.

Concerning this last axiom: we start with part b) and *do not* give a justification but rather admit that it is there to simplify the situation. However we observe that parts a) and b) are compatible because of the Extreme Value Theorem: every continuous function is bounded. Therefore because of axiom (IO) we can — at the least – integrate every continuous function $f : [a, b] \to \mathbb{R}$. Such an integral is guaranteed to have real content: because of the close connection to the area problem, such an integral gives a rigorous mathematical meaning to "the area under a non-negative continuous curve $y = f(x)$."

Now something remarkable happens: if we *assume* that we have an integral $\int : \mathcal{R}[a, b] \to \mathbb{R}$ satisfying axioms (I0) through (I3), then without knowing anything about how this function is actually defined, we can use it to prove the Fundamental Theorem of Calculus!

THEOREM 3.1 (Fundamental Theorem of Calculus). *Let $\int : \mathcal{R}[a, b] \to \mathbb{R}$ satisfy (IO), (I1), (I2) and (I3). Let $f \in \mathcal{R}[a, b]$. For $x \in [a, b]$, we define*

$$\mathcal{F}(x) := \int_a^x f.$$

Then:

a) *The function $\mathcal{F} : [a, b] \to \mathbb{R}$ is continuous.*
b) *If f is continuous at c, then $\mathcal{F}$ is differentiable at c, and $\mathcal{F}'(c) = f(c)$.*
c) *If f is continuous and $F : [a, b] \to \mathbb{R}$ is any antiderivative of f – i.e., $F' = f$ – then*

$$\int_a^b f = F(b) - F(a).$$

PROOF. a) By (I0(b)), there is $M > 0$ such that $|f(x)| \leq M$ for all $x \in [a, b]$. Let $\epsilon > 0$, and take $\delta := \frac{\epsilon}{M}$. For any $a \leq c \leq d \leq b$, because $-M \leq f \leq M$, applying (I2) and (I1) we get

$$-M(d-c) = \int_c^d (-M) \leq \int_c^d f \leq \int_c^d M = M(d-c),$$

so using (I3) we get

$$\left|\mathcal{F}(d) - \mathcal{F}(c)\right| = \left|\int_a^d f - \int_a^c f\right| = \left|\int_c^d f\right| \leq M(d-c), \tag{7}$$

which shows that $\mathcal{F}$ is uniformly continuous with $\delta = \frac{\epsilon}{M}$.
b) Since f is continuous at c, for all $\epsilon > 0$, there is δ such that $|x-c| < \delta$ implies

$$f(c) - \epsilon < f(x) < f(c) + \epsilon.$$

Thus:

$$f(c) - \epsilon = \frac{\int_c^x (f(c) - \epsilon)}{x-c} \leq \frac{\int_c^x f}{x-c} \leq \frac{\int_c^x (f(c)+\epsilon)}{x-c} = f(c) + \epsilon,$$

which we may rewrite as

$$\left|\frac{\mathcal{F}(x) - \mathcal{F}(c)}{x-c} - f(c)\right| = \left|\frac{\int_c^x f}{x-c} - f(c)\right| \leq \epsilon,$$

which shows that

$$\mathcal{F}'(c) = \lim_{x\to c} \frac{\mathcal{F}(x) - \mathcal{F}(c)}{x-c} = f(c).$$

c) Suppose f is continuous. By part b), we know that $\mathcal{F}(x) = \int_a^x f$ is an antiderivative of f. By Exercise 2.18 we know that antiderivatives are unique up to the addition of a constant, which means that if F is any antiderivative of f there is $C \in \mathbb{R}$ such that

$$\forall x \in [a,b],\ F(x) = \mathcal{F}(x) + C,$$

and thus,

$$\begin{aligned} F(b) - F(a) &= (\mathcal{F}(b) + C) - (\mathcal{F}(a) + C) \\ &= \mathcal{F}(b) - \mathcal{F}(a) = \int_a^b f - \int_a^a f = \int_a^b f; \end{aligned}$$

above we used Exercise 3.2 to get $\int_a^a f = 0$. □

We now have several important remarks to make.

First, as discussed above, any integral $\int : \mathcal{R}[a,b] \to \mathbb{R}$ restricts to an integral $\int : \mathcal{C}[a,b] \to \mathbb{R}$ on the set of all continuous functions $f : [a,b] \to \mathbb{R}$. But part c) of the Fundamental Theorem of Calculus tells us that in this case there is no need for axiomatics: the integral of any continuous function is *necessarily* given as $F(b) - F(a)$ for any antiderivative F of f. In other words, the function $\int : \mathcal{C}[a,b] \to \mathbb{R}$ is **unique**.

Second: I must observe that the proof of Theorem 3.1 was...quite easy. Admittedly the statement was a bit technical, but the proof of each part took only a few lines. Our proofs that our fancy-looking function $\mathcal{F}$ is always continuous and is differentiable when f is continuous each came out right away: earlier in our course we worked harder to prove the continuity/differentiability of very *specific* functions.

Why is the proof of the Fundamental Theorem of Calculus so easy? This is a question I thought a lot about the first time I taught undergraduate real analysis, in

2004 at McGill University. The proof of FTC is usually given in freshman calculus courses, but the theory of the Riemann integral is much more intricate (um, wait for it; you'll see). How is it possible that the theory is hard but its main theorem is easy?

The answer is that as we've stated it, the Fundamental Theorem of Calculus is *not* the crux that we might think it is. Do you see why? The answer is that our statement of the Fundamental Theorem *assumes* that we have an integral satisfying the axioms (I0) through (I3) and defined on the class of continuous functions (so (I0) is satisfied). After we prove the theorem, it turns out that on the class of continuous functions, this integral is *unique*. But how do we know that the integral exists? Answer: we don't, yet. That's where the real work lies.

Third: Theorem 3.1 has the following very important consequence:

COROLLARY 3.2. *Each continuous function $f : [a, b] \to \mathbb{R}$ has an antiderivative.*

Indeed, the Fundamental Theorem supplies us with the particular antiderivative $\mathcal{F}(x) = \int_a^x f$. (I emphasize that at the moment we know this conditionally on the assumption that the integral exists.) Once again we know, as a consequence of the Mean Value Theorem, that antiderivatives are unique up to an additive constant. As we saw in the proof, we have $\mathcal{F}(a) = \int_a^a f = 0$, so that tells us which antiderivative we're getting: the unique one that is 0 at the left endpoint.

It may be interesting to ask *how much* of the content of the Fundamental Theorem of Calculus is carried by Corollary 3.2: that is, suppose that we know, somehow, that every continuous function has an antiderivative. Can we then use this to show the existence of an integral on $\mathcal{C}[a, b]$? The answer is **yes**: if F is antiderivative of f, then you can show directly that $\int_a^b f := F(b) - F(a)$ defines an integral $\int : \mathcal{C}[a, b] \to \mathbb{R}$. This is an amusing exercise: Exercise 3.3. On the other hand, although there are several ways to go about constructing this integral $\int : \mathcal{C}[a, b] \to \mathbb{R}$ that we have been talking about, I believe that I do not know any way to prove Corollary 3.2 that does not involve constructing the integral in some way and then differentiating $\int_a^x f$ to get $f(x)$.

Let me now give a small preview of what's coming next: we will define a certain **process** that can be applied to any function $f : [a, b] \to \mathbb{R}$. This process returns two different extended real numbers – i.e., either real numbers, ∞ of $-\infty$. These are called the **upper Darboux integral** $\overline{\int}_a^b f$ and the **lower Darboux integral** $\underline{\int}_a^b f$. It will turn out that in all cases we have

$$\underline{\int}_a^b f \leq \overline{\int}_a^b f.$$

We say that the function f is **Darboux integrable** if the two are equal and the common value is a real number (and not $\pm\infty$).

We will study the Darboux integration process and show that it satisfies all our axioms: that is, if we define $\mathcal{R}_D[a, b]$ to be the set of Darboux integrable functions, then these functions satisfy (I1), (I2), (I3), and most importantly, (I0): every Darboux integrable function is bounded (indeed boundedness is equivalent to the upper and lower integrals both being finite) and every continuous function $f : [a, b] \to \mathbb{R}$. It is this last statement that carries most of the content of the Fundamental Theorem of Calculus. We will also show some further useful properties of the Darboux integral: for instance, we will see that $\mathcal{R}_D[a, b]$ is a vector space over the real numbers and the integral $\int : \mathcal{R}_D[a, b] \to \mathbb{R}$ is a linear map.

At this point, we will know that $\mathcal{R}_D[a, b]$ contains all the continuous functions, and it will not be hard to see that it contains many other functions as well – e.g. all bounded functions that are either monotone or have finitely many discontinuities. So it is natural to ask: can we determine exactly which functions are Darboux integrable?

Leaving that question hang in the air for now, here is a very different question: why have we not said "Riemann" yet? After all, in calculus one speaks of the Riemann integral and after all that is the title of this chapter. Well, what we called the "Darboux integral" above is what many would call the Rieman integral. However we have a distinction to make: Riemann himself defined a different *process* from Darboux's: in other words, Riemann's actual technical definition of the limit is different from Darboux's. Rather we should say that Darboux's definition is different from Riemann's, since Riemann's came first: Darboux's is actually easier to understand and easier to work with in many respects. The main advantage of Riemann's definition is that it is indeed a (rather complicated!) limit of **Riemann sums**, which means that certain sequential limits can be evaluated by interpreting them as Riemann sums of a Riemann integrable function.

What is the relationship between the integrals of Darboux and Riemann? Although their descriptions are different, we have already shown that as functions $\mathcal{C}[a, b] \to \mathbb{R}$ they must be equal, i.e., the real number $\int_a^b$ assigned to each continuous $f : [a, b] \to \mathbb{R}$ must be the same, because both satisfy the axioms and there is a unique integral on the continuous functions satisfying the axioms. In fact their relationship is closer still: if we let $\mathcal{R}_R[a, b]$ denote the set of Riemann integrable functions, then in fact

$$\mathcal{R}_D[a, b] = \mathcal{R}_R[a, b]$$

— that is, a function is Riemann integrable if and only if it is Darboux integrable – and moreover when a function $f : [a, b] \to \mathbb{R}$ is integrable according to either definition the assigned values $\int_a^b f$ agree. So at the end of the day, although Riemann and Darboux are different *processes*, they yield exactly the same *integral.* In other words, they are ultimately two different descriptions of the same thing.

1.1. Exercises.

EXERCISE 3.1. *Let $f : [a, b] \to [0, \infty)$ be a function. Show that the subset*

$$S_f := \{(x, y) \in \mathbb{R}^2 \mid a \leq x \leq b \text{ and } 0 \leq y \leq f(x)\}$$

is bounded if and only if f is bounded.

EXERCISE 3.2. *Show that the axioms (I1), (I2) and (I3) imply that for any integrable $f : [a, b] \to \mathbb{R}$ and any $c \in [a, b]$, we have $\int_c^c f = 0$.*

EXERCISE 3.3. *Suppose that you happen to know that every continuous function has an antiderivative. Show that, defining, for every continuous function* $f : [a, b] \to \mathbb{R}$, $\int_a^b f$ *to be* $F(b) - F(a)$ *where* F *is any antiderivative of* f, *defines an integral* $\int : \mathcal{C}[a, b] \to \mathbb{R}$: *in other words, check that the axioms (I1) through (I3) are satisfied.*

2. Darboux's Riemann Integral

In this section we present Darboux's approach to the Riemann integral. Throughout this section $a < b$ are real numbers.

2.1. Upper and lower sums, upper and lower integrals. Partitions: A **partition** of $[a, b]$ is a finite subset $\mathcal{P}$ of $[a, b]$ containing a and b. Thus we may write $\mathcal{P}$ as $\{x_0, x_1, \ldots, x_{n-1}, x_n\}$ with $a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$. Notice that the positive integer n is one less than the number of elements of $\mathcal{P}$; we think of a partition $\mathcal{P}$ as subdividing the interval $[a, b]$ into subintervals $[a, x_1], [x_1, x_2], \ldots, [x_{n-1}, b]$, and thus n is the number of subintervals into which we subdivided $[a, b]$. The telescoping sum

$$\sum_{i=0}^{n-1}(x_{i+1} - x_i) = (x_1 - a) + (x_2 - x_1) + \ldots + (b - x_{n-1}) = b - a$$

shows that the length of the interval $[a, b]$ is the sum of the lengths of the subintervals into which we divided it using $\mathcal{P}$.

Because $[a, b]$ is infinite, there are certainly infinitely many partitions of it. We introduce a relation among them: we say that a partition $\mathcal{P}_2$ of $[a, b]$ **refines** a partition $\mathcal{P}_1$ of $[a, b]$ if $\mathcal{P}_1 \subseteq \mathcal{P}_2$: thus, $\mathcal{P}_2$ contains all the points of $\mathcal{P}_1$ and (if $\mathcal{P}_2 \neq \mathcal{P}_1$) some others. We can think of $\mathcal{P}_2$ as being obtained from $\mathcal{P}_1$ by repeatedly choosing one of the subintervals $[x_i, x_{i+1}]$ given by $\mathcal{P}_1$ and subdividing it by adding an addition point $z \in (x_i, x_{i+1})$. This refinement relation is a partial ordering on the set of partitions of $[a, b]$: this just means that every partition refines itself; if each of two partitions refines the other than they are equal; and if $\mathcal{P}_3$ refines $\mathcal{P}_2$ and $\mathcal{P}_2$ refines $\mathcal{P}_1$ then $\mathcal{P}_3$ refinee $\mathcal{P}_1$.

Now let $f : [a, b] \to \mathbb{R}$ be a *bounded* function. To every partition $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ of $[a, b]$, we will define an **upper sum** $L(f, \mathcal{P}) \in \mathbb{R}$ and a **lower sum** $U(f, \mathcal{P}) \in \mathbb{R}$. To this we first define:

• For all $0 \leq i \leq n - 1$, let $M_i(f)$ be the supremum of $f([x_i, x_{i+1}])$, and
• For all $0 \leq i \leq n - 1$, let $m_i(f)$ be the infimum of $f[(x_i, x_{i+1}])$.

Now we put

$$U(f, \mathcal{P}) \coloneqq \sum_{i=0}^{n-1} M_i(f)\,(x_{i+1} - x_i)$$

and

$$L(f, \mathcal{P}) \coloneqq \sum_{i=0}^{n-1} m_i(f)\,(x_{i+1} - x_i)\,.$$

Some remarks are in order.

REMARK. a) *Since for any nonempty subset X of $\mathbb{R}$ we have* $\inf X \leq \sup X$*, for any $f : [a,b] \to \mathbb{R}$ and all $0 \leq i \leq n-1$ we have $m_i(f) \leq M_i(f)$, from which it follows that*

$$L(f, \mathcal{P}) \leq U(f, \mathcal{P}).$$

b) *For all $0 \leq i \leq n-1$, we have $M_i(f) \in \mathbb{R}$ because f is bounded above. Suppose on the other hand that f were not bounded above. Then by Exercise 3.4b), there is at least one i such that f is not bounded above on the subinterval $[x_i, x_{i+1}]$, so the supremum of $f([x_i, x_{i+1}])$ is ∞. It is a standard convention the arithmetic of extended real numbers that $\infty + \infty = \infty$ and for $a \in (0, \infty]$ we have $a \cdot \infty = \infty$. Using these conventions we find that if f is unbounded above we can make sense of the upper sum $U(f, \mathcal{P})$: it will always be ∞.*

c) *Similarly, for all $0 \leq i \leq n-1$, we have $m_i(f) \in \mathbb{R}$ because f is bounded below. If were unbounded below then by the same reasoning as part a) we find that we can make sense of the lower sum $L(f, \mathcal{P})$ but it will always be $-\infty$. Thus for* any *function f we have $U(f, \mathcal{P}) \in \mathbb{R} \cup \{\infty\}$ and $L(f, \mathcal{P}) \in \mathbb{R} \cup \{-\infty\}$.*

d) *Suppose that $f : [a,b] \to \mathbb{R}$ is continuous. Then f is bounded, and for all $1 \leq i \leq n$, by the Extreme Value Theorem we get that $m_i(f)$ is the **minimum** of f on $[x_i, x_{i+1}]$ and $M_i(f)$ is the **maximum** of f on $[x_i, x_{i+1}]$. Though we will not define Riemann sums until the next section – the extra complication of choosing a "sample point" in each subinterval is part of what Darboux's approach manages to avoid – nevertheless we remark now that when f is continuous the upper sum and lower sum are both Riemann sums for f.*

EXAMPLE 2.1. *Consider $f : [0,1] \to \mathbb{R}$ by $f(x) = x^2$.*

a) *Suppose we take the smallest possible partition: $\mathcal{P}_1 = \{0, 1\}$. The minimum of f on $[0,1]$ is 0 and the maximum of f is 1, so*

$$L(x^2, \mathcal{P}_1) = 0 < 1 = U(x^2, \mathcal{P}_1).$$

We can interpret this geometrically: consider the $S_{x^2} = \{(x,y) \in \mathbb{R}^2 \mid 0 \leq x \leq 1,\ 0 \leq y \leq x^2\}$, whose area we are trying to define via some limiting process. This set contains the line segment $S_0 = [0,1] \times \{0\}$, that has area 0, and it is contained in the unit square $S_1 = [0,1] \times [0,1]$, that has area 1. So although we haven't defined the integral yet, the idea is that we have learned from $\mathcal{P}_1$ is that we want $\int_0^1 x^2$ to be some real number such that

$$L(x^2, \mathcal{P}_1) = 0 < \int_0^1 x^2 \leq 1 = U(x^2, \mathcal{P}_1).$$

b) *Even a vague memory of definite integrals from calculus should suggest that we try something else: for $n \in \mathbb{Z}^+$ let*

$$\mathcal{P}_n := \{0 < \frac{1}{n} < \frac{2}{n} < \ldots \frac{n-1}{n} < 1\}$$

be the partition that subdivides $[0,1]$ into n equally spaced subintervals. Because $f(x) = x^2$ is increasing, the supremum it takes on any subinterval

$[c,d]$ of $[0,1]$ is just $f(d)$ and the infimum it takes on any subinterval $[c,d]$ of $[0,1]$ is just $f(c)$. So:

$$U(x^2,\mathcal{P}_n) = \sum_{i=0}^{n-1} M_i(x^2)\left(\frac{i+1}{n} - \frac{i}{n}\right) = \sum_{i=0}^{n-1}(\frac{i+1}{n})^2 \cdot \frac{1}{n}$$
$$= \frac{1}{n^3}\sum_{i=0}^{n-1}(i+1)^2 = \frac{1}{n^3}\sum_{i=1}^{n} i^2.$$

Oh, thank goodness that in a previous course (Math 3200) we practiced induction with sums like these: we happen to remember that

$$\forall n \in \mathbb{Z}^+, \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6},$$

so we get

$$U(x^2,\mathcal{P}_n) = \frac{n(n+1)(2n+1)}{6n^3}.$$

The computation for the lower sums is very similar: for all i we have $m_i(x^2) = \left(\frac{i}{n}\right)^2$, which leads to

$$L(x^2,\mathcal{P}_n) = \frac{1}{n^3}\sum_{i=0}^{n-1} i^2 = \frac{1}{n^3}\sum_{i=1}^{n-1} i^2 = \frac{(n-1)n(2n-1)}{6n^3}.$$

Now we observe that

$$\lim_{n\to\infty} U(x^2,\mathcal{P}_n) = \frac{1}{3} = \lim_{x\to\infty} L(x^2,\mathcal{P}_n).$$

So we found a sequence of partitions along which the lower sums converged to $\frac{1}{3}$ and along which the upper sums also converged to $\frac{1}{3}$. This makes us strongly suspect that we want $\int_0^1 x^2 = \frac{1}{3}$. In calculus we would probably be happy to take either one of these limits *as sufficient to give the answer, but now we are trying to find our way to a principled definition of an integrable function. We can reason as follows: for* any *partition $\mathcal{P}$, we can interpret $U(f,\mathcal{P})$ as the area enclosed by a piecewise constant function that is always greater than or equal to f and we can interpret $L(f,\mathcal{P})$ as the area enclosed by a piecewise constant function that is always less than or equal to f, so we should have*

(8) $$\forall \text{ partitions } \mathcal{P} \text{ of } [a,b],\ L(f,\mathcal{P}) \le \int_a^b f \le U(f,\mathcal{P}).$$

So in our case we want $\int_a^b f$ to satisfy

$$\forall n \in \mathbb{Z}^+,\ \frac{(n-1)n(2n-1)}{6n^3} \le \int_0^1 x^2 \le \frac{n(n+1)(2n+1)}{6n^3}.$$

Limits of sequences preserve lax inequalities ($\le$ and $\ge$, not $<$ and $>$), so

$$\frac{1}{3} = \lim_{n\to\infty} L(f,\mathcal{P}) \le \int_0^1 x^2 \le \lim_{n\to\infty} U(f,\mathcal{P}) = \frac{1}{3}.$$

This tells us that $\int_0^1 x^2 = \frac{1}{3}$!

In other words, our one idea about $\int_a^b f$ is that it should lie in between

$L(f,\mathcal{P})$ and $U(f,\mathcal{P})$ for any partition $\mathcal{P}$ of $[a,b]$. In this particular example, just by looking at the sequence of partitions $\{\mathcal{P}_n\}_{n=1}^{\infty}$ we see that the only real number that could possibly satisfy this is $\frac{1}{3}$.

This leads us to our first definition of Darboux integrability: a function $f : [a,b] \to \mathbb{R}$ is **Darboux integrable** if there is exactly one real number S such that for all partitions $\mathcal{P}$ of $[a,b]$ we have

$$L(f,\mathcal{P}) \leq S \leq U(f,\mathcal{P});$$

for a Darboux integrable function f we put $\int_a^b f$ to be this unique real number S.

PROPOSITION 3.3. *Let $f : [a,b] \to \mathbb{R}$ be a function. If f is Darboux integrable, then it is bounded.*

PROOF. We will show the contrapositive: suppose f is unbounded; we claim that f is *not* Darboux integrable.
Case 1: If f is unbounded both above and below then for all partitions $\mathcal{P}$ of $[a,b]$ we have $U(f,\mathcal{P}) = \infty$ and $L(f,\mathcal{P}) = -\infty$. So every real number lies in between every lower sum and upper sum: thus the uniqueness of I fails.
Case 2: Suppose f is unbounded above but bounded below. Then for all partitions $\mathcal{P}$ of $[a,b]$ we have that $U(f,\mathcal{P}) = \infty$ but $L(f,\mathcal{P}) \in \mathbb{R}$. Every real number is at most ∞, so in order to be Darboux integrable there would have to be a unique real number I greater than or equal to $L(f,\mathcal{P})$ for all partitions $\mathcal{P}$. In other words, the set $\{L(f,\mathcal{P}) \mid \mathcal{P} \text{ is a partition of [a,b]}\}$ would need to have a *unique* upper bound. That's not how upper bounds in $\mathbb{R}$ work: if $I \in \mathbb{R}$ is an upper bound for *any* subset X of $\mathbb{R}$ then so is $I+1$, so X cannot have a unique, finite upper bound.
Case 3: If f is bounded above but unbounded below, the reasoning of Case 2 applies: there is no *unique* real number less than or equal $U(f,\mathcal{P})$ for all $\mathcal{P}$. □

However, this definition of Darboux integrability is not so easy to work with: one can see this by observing that in Example 2.1 we did *not* show that x^2 is Darboux integrable on $[0,1]$: all we showed was that *if it is*, then the integral is $\frac{1}{3}$.

The awkwardness in our definition of Darboux integrability is characteristic of many definitions in theoretical mathematics: the definition involves a universal quantifier over an infinite set and for each element of that set asserts something nontrivial. Here that set is the set of all partitions of $[a,b]$. In our above example, showing Darboux integrability apparently asks us to compute $U(f,\mathcal{P})$ and $L(f,\mathcal{P})$ for *every* partition of $[0,1]$ and check that $L(f,\mathcal{P}) \leq \frac{1}{3} \leq U(f,\mathcal{P})$. Are we really supposed to perform *infinitely many* computations to check that x^2 is integrable?!?

No, not really. This definition is too hard to check directly, so we need a result that tells us that it is sufficient to do something easier. The result that we are going for here is as follows: a bounded function $f : [a,b] \to \mathbb{R}$ is Darboux integrable if and only if: for all $\epsilon > 0$, there is a partition $\mathcal{P}$ of $[a,b]$ such that $U(f,\mathcal{P}) - L(f,\mathcal{P}) < \epsilon$. Once we establish this, we don't need to look at *all* partitions; we just need to exhibit a *sequence* of partitions along which the gap between the upper and lower sums tends to 0. That is much easier: in Example 2.1, the sequence $\{\mathcal{P}_n\}$ works. Because $\lim_{n\to\infty} U(f,\mathcal{P}_n) = \frac{1}{3} = \lim_{n\to\infty} L(f,\mathcal{P}_n)$, it follows that $\lim_{n\to\infty} U(f,\mathcal{P}_n) - L(f,\mathcal{P}_n) = 0$, so for any $\epsilon > 0$, just taking $\mathcal{P}_n$ for large enough n does what we want.

In order to show this, we need a few preliminaries. They involve refinements of partitions, which you may notice that we defined but have not yet used for anything whatsoever. Well, now is the time.

LEMMA 3.4. *Let $f : [a,b] \to \mathbb{R}$ be a function.*

a) *Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be partitions of $[a,b]$, with $\mathcal{P}_2$ refining $\mathcal{P}_1$ (that is, $\mathcal{P}_2 \supseteq \mathcal{P}_1$). Then we have*

$$L(f,\mathcal{P}_1) \leq L(f,\mathcal{P}_2) \leq U(f,\mathcal{P}_2) \leq U(f,\mathcal{P}_1).$$

b) *Let $\mathcal{P}$ and $\mathcal{Q}$ be any partitions of $[a,b]$. Then we have*

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

PROOF. We get from $\mathcal{P}_1$ to $\mathcal{P}_2$ by adding finitely many more points. So it suffices to treat the case in which $\mathcal{P}_2$ is obtained from $\mathcal{P}_1$ by adding a single additional point $c \in (x_i, x_{i+1})$ for some $0 \leq i \leq n-1$ and then show that $L(f,\mathcal{P}_2) \leq L(f,\mathcal{P}_1)$ and $U(f,\mathcal{P}_2) \leq U(f,\mathcal{P}_1)$. (Notice that we already know the middle inequality $L(f,\mathcal{P}_2) \leq U(f,\mathcal{P}_2)$; it is just there to make everything look nice.) This is actually quite easy: most of the terms in the sums $U(f,\mathcal{P}_1)$ and $U(f,\mathcal{P}_2)$ are the same; the only change is that we replace the ith term $\sup(f[x_i,x_{i+1}]) \cdot (x_{i+1}-x_i)$ of $U(f,\mathcal{P}_1)$ with the two terms $\sup(f[x_i,c]) \cdot (c-x_i) + \sup(f[c,x_{i+1}]) \cdot (x_{i+1}-x_c)$. If $A \subseteq B \subset \mathbb{R}$ then $\sup A \leq \sup B$, so we have

$$\sup(f[x_i,c]) \cdot (c-x_i) + \sup(f[c,x_{i+1}]) \cdot (x_{i+1}-x_c)$$

$$\leq \sup(f[x_i,x_{i+1}]) \cdot (c-x_i) + \sup(f[x_i,x_{i+1}]) \cdot (x_{i+1}-c) \leq \sup(f[x_i,x_{i+1}]) \cdot (x_{i+1}-x_i).$$

Thus $U(f,\mathcal{P}_2) \leq U(f,\mathcal{P}_1)$. The same reasoning works for the lower sums: the infimum of f on $[x_i,x_{i+1}]$ is less than or equal to its infimum on $[x_i,c]$ and its infimum on $[c,x_{i+1}]$.

b) Let $\mathcal{R} \coloneqq \mathcal{P} \cup \mathcal{Q}$; this is a partition of $[a,b]$ that is a *common refinement* of $\mathcal{P}$ and $\mathcal{Q}$. Applying part a) twice, we get

$$L(f,\mathcal{P}) \leq L(f,\mathcal{R}) \leq U(f,\mathcal{R}) \leq U(f,\mathcal{Q}). \qquad \square$$

Now one final definition that we hinted at in the last section: for any function $f : [a,b] \to \mathbb{R}$, we define the **upper Darboux integral**

$$\overline{\int_a^b} f \coloneqq \inf U(f,\mathcal{P}) \in [-\infty,\infty]$$

and the **lower Darboux integral**

$$\underline{\int_a^b} f \coloneqq L(f,\mathcal{P}) \in [-\infty,\infty];$$

in each case we are ranging over all partitions $\mathcal{P}$ of $[a,b]$. Notice that $\overline{\int_a^b} f$ is a "minimax": for each partition we maximized f (actually we took suprema, but people don't say "infysup"), collected these values over all partitions and then took the minimum (actually the infimum). Simlarly, $\underline{\int_a^b} f$ is a "maximin." When it makes sense to do, it's often a surprisingly good idea to take minimaxes and maximins and to compare them: see e.g. `https://en.wikipedia.org/wiki/Minimax_theorem`, which is the foundational result in Game Theory. Anyway, there is a clear geometric idea: the upper integral $\overline{\int_a^b} f$ is the best upper bound one can get on the area of the region S_f using upper rectangles, while the lower integral $\underline{\int}_a^b f$ is the best lower bound one can get on the same area using lower rectangles. So if we want there to be a unique real number lying between all the areas of lower rectangles and all the areas of upper rectangles, then presumably we want the upper and lower integrals to be equal, finite numbers. We are about to show this, but first one "sanity check":

LEMMA 3.5. *Let $f : [a, b] \to \mathbb{R}$ be a function.*

a) *We have*

$$\underline{\int_a^b} f \leq \overline{\int_a^b} f.$$

b) *If f is bounded, then $\overline{\int}_a^b f, \underline{\int}_a^b f \in \mathbb{R}$.*

PROOF. a) The lower integral $\underline{\int}_a^b f$ is the supremum of the set

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

while the upper integral $\overline{\int}_a^b f$ is the infimum of the set

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

But Lemma 3.4b) says that for all $x \in X$ and all $y \in Y$ we have $x \leq y$. Thus every $x \in X$ is a lower bound for Y, so $x \leq \inf Y$, and since this holds for all $x \in X$ we have $\sup X \leq \inf Y$.
b) Suppose that f is bounded: there is $M > 0$ such that $|f| \leq M$. Then for any partition $\mathcal{P}$ of $[a, b]$ we have

$$-M(b - a) \leq L(f, \mathcal{P}) \leq U(f, \mathcal{P}) \leq M(b - a),$$

and it follows that $\overline{\int}_a^b f$, $\underline{\int}_a^b f \in [-M(b - a), M(b - a)]$. $\square$

THEOREM 3.6 (Darboux's Integrability Criterion). *For a function $f : [a, b] \to \mathbb{R}$, the following are equivalent:*

(i) *There is a unique real number I such that for all partitions $\mathcal{P}$ of $[a, b]$ we have $L(f, \mathcal{P}) \leq I \leq U(f, \mathcal{P})$.*
(ii) *We have $\underline{\int}_a^b f = \overline{\int}_a^b f \in \mathbb{R}$.*
(iii) *For all $\epsilon > 0$, there is a partition $\mathcal{P}$ such that $U(f, \mathcal{P})$ and $L(f, \mathcal{P})$ are real numbers and $U(f, \mathcal{P}) - L(f, \mathcal{P}) < \epsilon$.*

Henceforth we call a function satisfying these conditions ***Darboux integrable****.*

PROOF. Step 1: Suppose first that f is unbounded. By Proposition 3.3, condition (i) fails. Moreover either f is unbounded above — in which case $U(f, \mathcal{P}) = \infty$ for all $\mathcal{P}$ hence $\overline{\int}_a^b f = \infty$, so (ii) and (iii) fail — or f is unbounded below – in which case $L(f, \mathcal{P}) = -\infty$ for all $\mathcal{P}$ hence $\underline{\int}_a^b f = -\infty$, so again (ii) and (iii) fail.

So it suffices to consider the case in which f is bounded. In this case, by Exercise 3.5 we know that $\overline{\int}_a^b f$ and $\underline{\int}_a^b f$ are both finite.

Step 2: We show that (i) $\iff$ (ii). For a real number S, we have $S \leq \overline{\int}_a^b f$ if and only if I is less than or equal to every upper sum of f, and we have $S \geq \underline{\int}_a^b f$ if and only if S is greater than or equal to every lower sum of f, so a real number S

lies in the interval $[\underline{\int}_a^b f, \overline{\int}_a^b f]$ if and only if it lies in between every lower sum of f and every upper sum of f. So the upper and lower integrals are equal if and only if there is a unique S in between every lower sum of f and every upper sum of f.
Step 3: We show that (ii) $\iff$ (iii): If (ii) holds, let $\epsilon > 0$. Then there is a partition $\mathcal{P}$ of $[a,b]$ such that

$$L(f,\mathcal{P}) > \underline{\int}_a^b f - \frac{\epsilon}{2}$$

and another partition $\mathcal{Q}$ of $[a,b]$ such that

$$U(f,\mathcal{Q}) < \overline{\int}_a^b f + \frac{\epsilon}{2},$$

so

$$U(f,\mathcal{Q}) - L(f,\mathcal{P}) < \overline{\int}_a^b f - \underline{\int}_a^b f + \epsilon = \epsilon.$$

Now let $\mathcal{R} := \mathcal{P} \cup \mathcal{Q}$. Since $\mathcal{R}$ refines both $\mathcal{P}$ and $\mathcal{Q}$, we have

$$U(f,\mathcal{R}) \le U(f,\mathcal{P}) \text{ and } L(f,\mathcal{R}) \ge L(f,\mathcal{P}),$$

so

$$U(f,\mathcal{R}) - L(f,\mathcal{R}) \le U(f,\mathcal{Q}) - L(f,\mathcal{P}) < \epsilon.$$

If (iii) holds, then let $\epsilon > 0$, and choose a partition $\mathcal{P}$ such that $U(f,\mathcal{P}) - L(f,\mathcal{P}) < \epsilon$. Then, since $\overline{\int}_a^b f \le U(f,\mathcal{P})$ and $\underline{\int}_a^b f \ge L(f,\mathcal{P})$, we have

$$\overline{\int}_a^b f - \underline{\int}_a^b f \le U(f,\mathcal{P}) - L(f,\mathcal{P}) < \epsilon.$$

Since this holds for all $\epsilon > 0$, we have $\overline{\int}_a^b f = \underline{\int}_a^b f$. □

2.2. Verification of the Axioms. Our next order of business is to check that the Darboux integral that we have defined satisfies Axioms (I0), (I1), (I2) and (I3) from §3.1. "Checking axioms" doesn't sound so exciting, but we get quite a payoff: the Fundamental Theorem of Calculus, which includes the fact that every continuous function $f : [a,b] \to \mathbb{R}$ is a derivative.

The verification of (I1) is left as Exercise 3.6 and the verification of (I2) is left as Exercise 3.7. Checking the third axiom is less straightforward:

PROPOSITION 3.7. *Let $f : [a,b] \to \mathbb{R}$ be a Darboux integrable function.*

a) *For any $a \le c \le d \le b$, the function $f|_{[c,d]} : [c,d] \to \mathbb{R}$ is Darboux integrable.*
b) *For any $a \le c \le b$, we have $\int_a^b f = \int_a^c f + \int_c^b f$.*

PROOF. a) Let $\epsilon > 0$. Since f is Darboux integrable, there is a partition $\mathcal{P}$ of $[a,b]$ such that $U(f,\mathcal{P}) - L(f,\mathcal{P}) < \epsilon$. Let $\mathcal{P}' := \mathcal{P} \cup \{c,d\}$, and write

$$\mathcal{P}' = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}.$$

Since $\mathcal{P}'$ refines $\mathcal{P}$, we have $U(f,\mathcal{P}') \le U(f,\mathcal{P})$ and $L(f,\mathcal{P}') \ge L(f,\mathcal{P})$, so

$$U(f,\mathcal{P}') - L(f,\mathcal{P}') < \epsilon.$$

Now $\mathfrak{p} := \mathcal{P}' \cap [c,d]$ is a partition of $[c,d]$: it contains c and d and is a suset of a finite set, hence finite. To be specific, suppose that $c = x_I$ and $d = x_J$. Then

$$U(f,\mathfrak{p}) - L(f,\mathfrak{p}) = \sum_{i=I}^{J-1} (\sup(f[x_i,x_{i+1}]) - \inf(f[x_i,x_{i+1}])\,(x_{i+1}-x_i),$$

whereas

$$U(f,P') - L(f,\mathcal{P}') = \sum_{i=0}^{n-1} (\sup(f[x_i,x_{i+1}]) - \inf(f[x_i,x_{i+1}])\,(x_{i+1}-x_i).$$

The only difference between the former sum and the latter is that in the former sum we are summing from I to $J-1$ and in the latter we are summing from 0 to $n-1$, so the latter sum is the former sum together with some additional terms. But every term in either sum is non-negative, because the supremum of f on any subinterval is at least as large as its infimum on that subinterval. Thus:

$$U(f,\mathfrak{p}) - L(f,\mathfrak{p}) \leq U(f,\mathcal{P}') - L(f,\mathcal{P}') < \epsilon.$$

By Theorem 3.6, $f_{[c,d}$ is Darboux integrable.
b) Let $\mathcal{P}$ be a partition of $[a,b]$, and let $\mathcal{P}' := \mathcal{P} \cup \{c\}$. We also put

$$\mathcal{P}_L := \mathcal{P} \cap [a,c] \text{ and } \mathcal{P}_R := \mathcal{P} \cap [c,b],$$

so $\mathcal{P}_L$ is a partition of $[a,c]$ and $\mathcal{P}_R$ is a partition of $[c,b]$. Similarly to part a), upon writing out the partial sums we find immediately that

$$U(f,\mathcal{P}') = U(f|_{[a,c]},\mathcal{P}_L) + U(f|_{[c,b]},\mathcal{P}_R) \text{ and } L(f,\mathcal{P}') = L(f|_{[a,c]},\mathcal{P}_L) + L(f|_{[c,b]},\mathcal{P}_R).$$

Moreover, since $\mathcal{P}'$ is a refinement of $\mathcal{P}$ we have

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

By part a), $f|_{[a,c]} : [a,c] \to \mathbb{R}$ and $f|_{[c,b]} : [c,b] \to \mathbb{R}$ are Darboux integrable, so

$$\begin{aligned} L(f,\mathcal{P}) \leq L(f,\mathcal{P}') &= L(f|_{[a,c]},\mathcal{P}_L) + L(f|_{[c,b]},\mathcal{P}_R) \\ &\leq \int_a^c f + \int_c^b f \\ &\leq U(f|_{[a,c]},\mathcal{P}_L) + U(f|_{[c,b]},\mathcal{P}_R) = U(f,\mathcal{P}') \leq U(f,\mathcal{P}). \end{aligned}$$

Thus $\int_a^c f + \int_c^b f$ lies between every lower sum and every upper sum. Since f is Darboux integrable, the unique such real number is $\int_a^b f$, and we conclude:

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

We have already shown that every Darboux integrable function is bounded: Proposition 3.3. The last, and most important, thing we have to show is this:

THEOREM 3.8. *Let $f : [a,b] \to \mathbb{R}$ be continuous. Then f is Darboux integrable.*

PROOF. The key is that by Theorem 1.28 we know that f is *uniformly* continuous. So let $\epsilon > 0$; we may choose $\delta > 0$ such that for all $x, y \in [a,b]$, if $|x-y| < \delta$ then $|f(x) - f(y)| < \frac{\epsilon}{b-a}$. Now choose $N \in \mathbb{Z}^+$ such that $\frac{b-a}{N} < \delta$ and let $\mathcal{P}_N$ be the partition that divides $[a,b]$ into N subintervals of equal length. Then

$$U(f,\mathcal{P}_N) - L(f,\mathcal{P}_N) = \left(\frac{b-a}{N}\right) \sum_{i=0}^{N-1} (\sup(f[x_i,x_{i+1}]) - \inf(f[x_i,x_{i+1}])) \,. \tag{9}$$

Because $x_{i+1} - x_i = \frac{b-a}{N} < \delta$, on the subinterval $[x_i, x_{i+1}]$ any two values of f differ from each other by less than $\frac{\epsilon}{b-a}$, so

$$\forall 0 \leq i \leq N-1, \ \sup(f[x_i, x_{i+1}]) - \inf(f[x_i, x_{i+1}]) \leq \frac{\epsilon}{b-a}.$$

If we apply this inequality to each term of (9), we now get $\frac{b-a}{N}$ times a sum of N terms, each one of which is at most $\frac{\epsilon}{b-a}$, so we get

$$U(f, \mathcal{P}_N) - L(f, \mathcal{P}_N) \leq \frac{b-a}{N} \cdot N \cdot \frac{\epsilon}{b-a} \leq \epsilon.$$

So f is Darboux integrable by Theorem 3.6. □

Finally the circle has been completed: we shown that the Darboux integral satisfies all of our axioms (I0) through (I3), so we do have a gadget $\int : \mathcal{R}_D[a,b] \to \mathbb{R}$ to plug into the hypothesis of the Fundamental Theorem of Calculus. Thus the Fundamental Theorem of Calculus becomes an unconditional result, and in particular we have shown that every continuous function has an antiderivative.

We were fortunate enough to know the Uniform Continuity Theorem (Theorem 1.28), so we used it to get a very agreeable proof of Theorem 3.8. In contrast to the situation of showing that a continuous function has an antiderivative – which I do not know how to show without somehow constructing a definite integral – there are alternate approaches to Theorem 3.8 that avoid the use of uniform continuity. See for instance [**HC**, Thm. 8.9] or [**No52**].

We end this section with a "supplement" to the Fundamental Theorem of Calculus that is actually more similar to the approach taken to that result in freshman calculus than the approach we have thus far taken here.

One often thinks of the Fundamental Theorem of Calculus as having two parts, with these two parts together showing that integration and differentiation are essentially inverse operations. The first part of the Fundamental Theorem concerns the situation in which we differentiate an integral. This is addressed by part b) of our Theorem 3.1: if f is continuous, then the derivative of $\int_a^x f$ is $f(x)$. The second part of the Fundamental Theorem concerns the situation in which we integrate a derivative. This is addressed by part c) of our Theorem 3.1, which can be restated as: if f is differentiable *and its derivative is continuous*, then $\int_a^b f' = f(b) - f(a)$.

Stating the result this way shows the presence of a hypothesis that is not clearly necessary: namely, that f' be continuous. We could try to push our luck here in two ways: first, a more natural hypothesis here would be that f' is Darboux integrable. Then $\int_a^b f'$ is defined, we so we can ask: is it equal to $f(b) - f(a)$ even when f' is not continuous? The answer is **yes**:

THEOREM 3.9 (Supplement to the Fundamental Theorem of Calculus). *Let $f : [a,b] \to \mathbb{R}$ be differentiable. If $f' : [a,b] \to \mathbb{R}$ is Darboux integrable, then:*

$$\int_a^b f' - f(b) - f(a).$$

PROOF. Let $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ be a partition of $[a,b]$. By the Mean Value Theorem there is $t_i \in (x_i, x_{i+1})$ such that

$$f(x_{i+1}) - f(x_i) = f'(t_i)(x_{i+1} - x_i).$$

Then we have for all $0 \leq i \leq n-1$ that

$$\inf(f'([x_i, x_{i+1}])(x_{i+1} - x_i) \leq f'(t_i)(x_{i+1} - x_i) = f(x_{i+1} - f(x_i)$$
$$= f'(t_i)(x_{i+1} - x_i) \leq \sup(f'[x_i, x_{i+1}])(x_{i+1} - x_i).$$

Summing these inequalities from $i = 0$ to $n - 1$ gives

$$L(f', \mathcal{P}) \leq \sum_{i=0}^{n-1} (f(x_{i+1}) - f(x_i)) = f(b) - f(a) \leq U(f', \mathcal{P}).$$

Thus $f(b) - f(a)$ lies between every lower sum of of f' and every upper sum of f'. Since we are assuming that f' is Darboux integrable, the unique such real number is $\int_a^b$, so we must have $\int_a^b f' = f(b) - f(a)$. □

The second question is whether the Darboux integrability of f' is actually necessary: perhaps whenever f is differentiable, f' is Darboux integrable? The answer is **no**, as is most easily seen by observing that $f' : [a, b] \to \mathbb{R}$ need not be bounded. Part d) of Exercise 3.14 gives examples of unbounded derivatives, while part e) shows a class of functions f for which f' is discontinuous but Darboux integrable — showing that Theorem 3.9 really is an extension of the Fundamental Theorem of Calculus. This class includes our old friend from Example 2.2.

The sufficiently intellectually curious always have further questions. One might say unboundedness is a rather "cheap" reason for a function $f : [a, b] \to \mathbb{R}$ to fail to be Darboux integrable. (Later on we will investigate exactly which bounded functions are Darboux integrable.) Must f' be Darboux integrable if it is bounded? The answer is **no**: see [**Go16**] for an example that is chosen for its relative simplicity.

2.3. Linearity of the Darboux Integral. Before proceeding further, it will be helpful to introduce some further notation regarding the quantity $U(f, \mathcal{P}) - L(f, \mathcal{P})$, which appears in condition (iii) in Darboux's Integrability Criterion (Theorem 3.6) and therefore shows up often in our arguments. If $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ then

$$U(f, \mathcal{P}) - L(f, \mathcal{P}) = \sum_{i=0}^{n-1} \left(\sup(f[x_i, x_{i+1}]) - \inf(f[x_i, x_{i+1}]\right) (x_{i+1} - x_i).$$

For a function $f : I \to \mathbb{R}$ defined on an interval I, let us define the **oscillation of** f **on** I to be

$$\omega(f, I) := \sup(f(I)) - \inf(f(I)) \in [-\infty, \infty].$$

This is an extended real number which lies in $\mathbb{R}$ if and only if f is bounded on I, which will almost always be the case for us. Then we have

$$U(f, \mathcal{P}) - L(f, \mathcal{P}) = \sum_{i=0}^{n-1} \omega(f, [x_i, x_{i+1}])(x_{i+1} - x_i).$$

And let us also put

$$\Delta(f, \mathcal{P}) := U(f, \mathcal{P}) - L(f, \mathcal{P}).$$

Thus f is Darboux integrable if and only if for all $\epsilon > 0$ there is a partition $\mathcal{P}$ of $[a, b]$ with $\Delta(f, \mathcal{P}) < \epsilon$, and moreover if $\mathcal{P}'$ is a partition refining $\mathcal{P}$ then

$$\Delta(f, \mathcal{P}') \leq \Delta(f, \mathcal{P}).$$

THEOREM 3.10. *Let $\mathcal{R}_D[a,b]$ be the set of Darboux integrable functions $f : [a,b] \to \mathbb{R}$. Then the Darboux integral*

$$\int : \mathcal{R}_D[a,b] \to \mathbb{R}$$

*is a **linear functional** – that is:*

a) *The set $\mathcal{R}_D[a,b]$ is a subspace of the vector space of all functions $f : [a,b] \to \mathbb{R}$.*
b) *The function $\int : \mathcal{R}_D[a,b] \to \mathbb{R}$ is a linear map: for all $f, g \in \mathcal{R}_D[a,b]$ and all $\alpha, \beta \in \mathbb{R}$, we have*

$$\int_a^b (\alpha f + \beta g) = \alpha \int_a^b f + \beta \int_a^b g.$$

PROOF. Equivalently, and perhaps more plainly, we must prove that if $f, g : [a,b] \to \mathbb{R}$ are Darboux integrable, then:
(i) For all $\alpha \in \mathbb{R}$, αf is also Darboux integrable, and moreover $\int_a^b (\alpha f) = \alpha \int_a^b f$;
(ii) $f + g$ is also Darboux integrable, and moreover $\int_a^b (f+g) = \int_a^b f + \int_a^b g$.
Assertion (i) is mostly a matter of pulling constants through upper and lower sums, so we leave this as Exercise 3.9.

Now let us show assertion (ii). Let $\epsilon > 0$; because f and g are Darboux integrable, there is a partition $\mathcal{P}_1$ of $[a,b]$ such that $\Delta(f, \mathcal{P}_1) < \frac{\epsilon}{2}$ and a partition $\mathcal{P}_2$ of $[a,b]$ such that $\Delta(f, \mathcal{P}_2) < \frac{\epsilon}{2}$. Let $\mathcal{P}_3$ be a common refinement of $\mathcal{P}_1$ and $\mathcal{P}_2$ (e.g. take $\mathcal{P}_3 = \mathcal{P}_1 \cup \mathcal{P}_2$); then $\Delta(f, \mathcal{P}_3)$ and $\Delta(g, \mathcal{P}_3)$ are each less than $\frac{\epsilon}{2}$.

We observe that for an interval I and for functions $f, g : I \to \mathbb{R}$, we have

$$\sup((f+g)(I)) \leq \sup(f(I)) + \sup(g(I)) \text{ and } \inf((f+g)(I)) \geq \inf(f(I)) + \inf(g(I)).$$

You are asked to show this in Exercise 3.15. Using these inequalities we get

$$L(f, \mathcal{P}_3) + L(g, \mathcal{P}_3) \leq L(f+g, \mathcal{P}_3) \leq U(f+g, \mathcal{P}_3) \leq U(f, \mathcal{P}_3) + U(g, \mathcal{P}_3).$$

This shows that

$$\Delta(f+g, \mathcal{P}_3) \leq \Delta(f, \mathcal{P}_3) + \Delta(g, \mathcal{P}_3) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

and thus $f + g$ is Darboux integrable. Moreover, whenever we have a Darboux integrable function $h : [a,b] \to \mathbb{R}$ and a partition $\mathcal{P}$ of $[a,b]$ such that $\Delta(h, \mathcal{P}) \leq \epsilon$, we know $\int_a^b h$ lies in the interval $[L(h, \mathcal{P}), U(h, \mathcal{P})]$ of length $\Delta(h, P) \leq \epsilon$, so we know that $\int_a^b h$ has distance at most ϵ from each of $U(h, \mathcal{P})$ and $L(h, \mathcal{P})$. So:

$$\int_a^b (f+g) \leq U(f+g, \mathcal{P}_3) \leq U(f, \mathcal{P}_3) + U(g, \mathcal{P}_3) \leq \int_a^b f + \int_a^b g + \epsilon$$

and similarly

$$\int_a^b f + \int_a^b g - \epsilon < L(f, \mathcal{P}_3) + L(g, \mathcal{P}_3) \leq L(f+g, \mathcal{P}_3) \leq \int_a^b (f+g).$$

This shows that $\left| \int_a^b (f+g) - (\int_a^b f + \int_a^b g) \right| \leq \epsilon$. Since this holds for all $\epsilon > 0$, we have $\int_a^b (f+g) = \int_a^b f + \int_a^b g$. $\square$

As one simple application of Theorem 3.10, we can reduce the study of $\int_a^b f$ to the case in which f is non-negative and thus officially not worry about signed areas. Indeed, if $f : [a,b] \to \mathbb{R}$ is Darboux integrable, then it is bounded, so there is a $C \geq 0$ such that $f + C \geq 0$ on $[a,b]$. So $\int_a^b (f+C)$ really does represent the area of the region S_{f+C}, and we can recover $\int_a^b f$ from this as

$$\int_a^b f = \int_a^b (f+C) - \int_a^b C = \int_a^b (f+C) - C(b-a).$$

2.4. Exercises.

EXERCISE 3.4. *Let X be a subset of $\mathbb{R}^N$, and let $Y_1, \ldots, Y_n$ be finitely many subsets of X such that $X = \bigcup_{i=1}^n Y_i$. Let $f : X \to \mathbb{R}^M$.*

a) *Show: f is bounded if and only if, for each $1 \leq i \leq n$, $f|_{Y_i} : Y_i \to \mathbb{R}^M$ is bounded.*
b) *Suppose $M = 1$. Show: f is bounded above if and only if, for each $1 \leq i \leq n$, $f|_{Y_i} : Y_i \to \mathbb{R}^M$ is bounded above. Then show the same with "bounded above" replaced everywhere by "bounded below."*

EXERCISE 3.5. *Let $f : [a,b] \to \mathbb{R}$.*

a) *Suppose that f is bounded above by $M \in \mathbb{R}$: we have $f(x) \leq M$ for all $x \in [a,b]$. Show: for every partition $\mathcal{P}$ of $[a,b]$, we have*

$$\overline{\int_a^b} f \leq U(f,\mathcal{P}) \leq M(b-a).$$

b) *Suppose that f is bounded below by $m \in \mathbb{R}$: we have $f(x) \geq m$ for all $x \in [a,b]$. Show: for every partition $\mathcal{P}$ of $[a,b]$, we have*

$$m(b-a) \leq L(f,\mathcal{P}) \leq \underline{\int_a^b} f.$$

EXERCISE 3.6. *Let $f : [a,b] \to \mathbb{R}$ be defined by $f(x) = C$ for all $x \in [a,b]$.*

a) *Show: for every partition $\mathcal{P}$ of $[a,b]$ we have $U(f,\mathcal{P}) = L(f,\mathcal{P}) = C(b-a)$.*
b) *Deduce: f is Darboux integrable and $\int_a^b C = C(b-a)$. Thus Axiom (I1) holds for the Darboux integral.*

EXERCISE 3.7. *Let $f, g : [a,b] \to \mathbb{R}$ be two Darboux integrable functions with $f \leq g$: that is, for all $x \in [a,b]$, we have $f(x) \leq g(x)$.*

a) *Show: for every partition $\mathcal{P}$ of $[a,b]$ we have $U(f,\mathcal{P}) \leq U(g,\mathcal{P})$ and $L(f,\mathcal{P}) \leq L(g,\mathcal{P})$.*
b) *Deduce: $\int_a^b f \leq \int_a^b g$. Thus Axiom (I2) holds for the Darboux integral.*

EXERCISE 3.8. *Let $f : [a,b] \to \mathbb{R}$ and let $c \in (a,b)$. Suppose that each of $f|_{[a,c]} : [a,c] \to R$ and $f|_{[c,b]} : [c,b] \to \mathbb{R}$ are Darboux integrable. Show: $f : [a,b] \to \mathbb{R}$ is Darboux integrable and $\int_a^b f = \int_a^c f + \int_c^b f$.*
(This is similar to Proposition 3.7 and — hint — can be proved in much the same way. Once we establish this, the proof of Proposition 3.7b) applies verbatim to give $\int_a^b f = \int_a^c f + \int_c^b f$. You can just say so: no need to repeat the argument.)

EXERCISE 3.9. *Suppose $f : [a,b] \to \mathbb{R}$ is Darboux integrable. Show: for all $\alpha \in \mathbb{R}$, the function $\alpha f : [a,b] \to \mathbb{R}$ is also Darboux integrable, and moreover*

$$\int_a^b (\alpha f) = \alpha \int_a^b f.$$

EXERCISE 3.10. *Let $f : [0,1] \to \mathbb{R}$ by $f(x) = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}$.*

a) *Show: $\underline{\int}_0^1 f = 0$ and $\overline{\int}_0^1 f = 1$.*
b) *Deduce: f is bounded function that is not Darboux integrable.*

EXERCISE 3.11. *Let $T : [0,1] \to \mathbb{R}$ be **Thomae's function** of Exercise 2.3, restricted to the unit interval. Recall from that Exercise that f is continuous at every rational number and discontinuous at every irrational number.*[1] *Show: T is Darboux integrable and $\int_0^1 T = 0$.*

EXERCISE 3.12. *Let $f : [a,b] \to \mathbb{R}$ be bounded. Suppose that for all $c \in (a,b]$, the restricted function $f|_{[c,b]} : [c,b] \to \mathbb{R}$ is Darboux integrable. Show that f is Darboux integrable and $\lim_{c \to a^+} \int_c^b f = \int_a^b$.*
(Suggestion: use the fact that if $|f| \leq M$, then on any subinterval $[c,d]$, every upper sum of f is at most $M(d-c)$ and every lower sum of f is at least $-M(d-c)$, and note that these quantities approach 0 with the length of $[c,d]$.

EXERCISE 3.13. *Suppose $f : [a,b] \to \mathbb{R}$ is bounded and has finitely many discontinuities. Show that f is Darboux integrable. (You may, or may not, wish to use Exercise 3.12.)*

EXERCISE 3.14. *Let a, b be positive real numbers, and define $f : \mathbb{R} \to \mathbb{R}$ by*

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

a) *Show: for all values of a and b, $f_{a,b}$ is continuous.*
b) *Show: $f_{a,b}$ is differentiable if and only if $a > 1$.*
(Here and hereafter, the only issues are at $x = 0$; $f_{a,b}$ is certainly infinitely differentiable on $\mathbb{R} \setminus \{0\}$.)
c) *Show: $f'_{a,b}$ is continuous if and only if $a > b + 1$.*
d) *Show: if $a \in (1, b+1)$, then $f'_{a,b}$ is unbounded on any open interval containing 0. Deduce: if $c < 0 < d$, $f'_{a,b}$ is not Darboux integrable on $[c,d]$.*
e) *Show: if $1 < a = b + 1$, then $f'_{a,b}$ exists, is discontinuous precisely at 0, and is bounded on any closed, bounded interval. Using Exercise 3.12, deduce that if $c < 0 < d$, then $f'_{a,b}|_{[c,d]} : [c,d] \to \mathbb{R}$ is Darboux integrable.*

EXERCISE 3.15. *Let $X \subset \mathbb{R}^N$ and let $f, g : X \to \mathbb{R}$.*

a) *Show:*

$$\sup((f+g)(X) \leq \sup(f(X)) + \sup(g(X)).$$

[1] Strictly speaking, one should look again at the endpoints 0 and 1 of the interval: there are some functions $f : \mathbb{R} \to \mathbb{R}$ that are discontinuous at 0 and 1 such that after restricting to $[0,1]$ become continuous at one or both endpoints. But in fact Thomae's function does not have even one-sided limits at any rational point.

b) *Show:*
$$\inf((f+g)(X) \geq \inf(f(X)) + \inf(g(X)).$$

(Comment: in part a), each of the terms is either a real number or ∞. In part b), each of the terms is either a real number or $-\infty$. Standard conventions on the arithmetic of extended real numbers apply, e.g. $\infty + \infty = \infty$ and for all $x \in \mathbb{R}$, $x + \infty = \infty$.)

EXERCISE 3.16 (Mean Value Theorem for Integrals). *Let $f : [a,b] \to \mathbb{R}$ be continuous. Show: there is $c \in [a,b]$ such that*
$$\int_a^b f = f(c) \cdot (b-a).$$

(Hint: let $m = \min f([a,b])$ and $M = \max f([a,b])$. Show that $\frac{\int_a^b f}{b-a} \in [m, M]$. Thus $\frac{\int_a^b f}{b-a}$ is intermediate *between two* values *of f....)*

EXERCISE 3.17 (Integrability of Monotone Functions).

a) *Let $f : [a,b] \to \mathbb{R}$ be an increasing function, and let $\mathcal{P}_n$ be the partition that divides $[a,b]$ into n equally spaced subintervals. Show:*
$$\Delta(f,\mathcal{P}) = U(f,\mathcal{P}) - L(f,\mathcal{P}) = (f(b) - f(a)) \cdot \left(\frac{b-a}{n}\right).$$
Use this to show that f is Darboux integrable.

b) *Show that if $f : [a,b] \to \mathbb{R}$ is decreasing, then it is Darboux integrable.*

c) *Show: if f is monotone, then $\lim_{n\to\infty} \frac{b-a}{n} \sum_{i=0}^{n-1} f(a + i(\frac{b-a}{n})) = \int_a^b f$.*

EXERCISE 3.18 (Monotone Functions Can Be Pretty Discontinuous). *Let $\{x_n\}_{n=1}^{\infty}$ be an injective sequence of real numbers: i.e., for all $m \neq n$ we have $x_m \neq x_n$. For $x \in \mathbb{R}$, let $S_x := \{n \in \mathbb{Z}^+ \mid x_n \leq x\}$. We define a function $f : \mathbb{R} \to \mathbb{R}$ as follows: for $x \in \mathbb{R}$,*
$$f(x) := \sum_{n \in S_x} 2^{-n}.$$

In other words, $f(x)$ is the sum of an infinite series whose nth term is 2^{-n} if $x_n \leq x$ and is 0 if $x_n > x$.

a) *Show that for all $x \in \mathbb{R}$, the infinite series defining $f(x)$ converges and we have $0 \leq f(x) \leq 1$.*
(Suggestion: compare to the geometric series $\sum_{n=1}^{\infty} 2^{-n} = 1$.)

b) *Show: f is increasing.*

c) *Show: for $n \in \mathbb{Z}^+$, $\lim_{x \to x_n^+} f(x) - \lim_{x\to x_n^-} f(x) = 2^{-n}$. Thus f is discontinuous at x_n.*

d) *Show: if $x \in \mathbb{R} \setminus \{x_n \mid n \in \mathbb{Z}^+\}$, then f is continuous at x.*

e) *Deduce: there is an increasing $f : [0,1] \to \mathbb{R}$ that is continuous at every irrational point of $[0,1]$ and discontinuous at every rational point of $[0,1]$.*

3. Riemann's Riemann Integral

In this section we touch upon Riemann's construction of the Riemann integral, which was earlier than Darboux's. Riemann's construction is a bit more technically elaborate than Darboux's – hence our decision to start with, and mostly focus on, Darboux's – but it has its merits, and it is to our advantage to at least be familiar with both.

In order to motivate Riemann's construction, imagine you have a function $f : [a, b] \to \mathbb{R}$ that you know is Darboux integrable: to fix ideas, let us suppose that it is continuous. Can we actually compute $\int_a^b f$?

Perhaps your first idea is to find an antiderivative F of f and use the Fundamental Theorem of Calculus: $\int_a^b f = F(b) - F(a)$. If you think this is how most integrals are actually computed, then you have been misled. Despite the time spent in freshman calculus on integration (i.e., antidifferentiation!) techniques, for any function more complicated than a rational function or a polynomial expression in trigonometric functions, it is quite rare to be able to write down an antiderivative "as an elementary function" of the sort you studied in precalculus. If your function is given by a power series expansion such that $[a, b]$ lies inside the open interval of convergence of the series, then you are in business: as you learned in Math 3100, power series can be integrated term by term, and moreover it is easy to estimate the value of a power series at a non-boundary point of convergence: you can cut off after finitely many terms and use geometric series to get an upper bound for the error. But such functions are a lot more than continuous: they are infinitely differentiable (and in fact, most infinitely differentiable functions are still not given by convergent power series expansions).

Returning to the Darboux integral, we point out that things work out very nicely if $f : [a, b] \to \mathbb{R}$ is monotone, as is explored in Exercise 3.17. To fix ideas, let us suppose that f is increasing. First of all, in this case, on any subinterval $[x_i, x_{i+1}]$ the supremum is just $f(x_{i+1})$, the value at the right endpoint, while the infimum is just $f(x_i)$, the value at the left endpoint. So we can actually compute $U(f, \mathcal{P})$ and $L(f, \mathcal{P})$ for any partition $\mathcal{P}$. Moreover, if you just take the partition $\mathcal{P}_n$ that subdivides $[a, b]$ into n equally spaced subintevals, then in the expression for $\Delta(f, \mathcal{P}_n) = U(f, \mathcal{P}_n) - L(f, \mathcal{P}_n)$ almost everything cancels out, and you are left with $(f(b) - f(a)) \cdot (\frac{b-a}{n})$. Evidently this approaches 0 as n approaches ∞, which already shows that f is Darboux integrable. But moreover, it follows easily from this that

$$\lim_{n\to\infty} U(f, \mathcal{P}_n) = \lim_{n\to\infty} L(f, \mathcal{P}_n) = \int_a^n f,$$

so choosing for instance the lower sum, we get concretely that

$$\int_a^n f = \lim_{n\to\infty} \left(\frac{b-a}{n}\right) \sum_{i=0}^{n-1} f(a + i(\frac{b-a}{n})).$$

Even if this limit is too hard to evaluate exactly (which it usually is), we can still compute, for any n, a lower bound $L(f, \mathcal{P}_n)$ for the integral and an upper bound $U(f, \mathcal{P}_n)$ for the integral, and as n approaches ∞, since each sequence approaches $\int_a^b f$, the gap between them $\Delta(f, \mathcal{P}_n)$ approaches 0. Therefore we can compute $\int_a^b f$ degree to accuracy ϵ by choosing a large n and computing $U(f, \mathcal{P}_n)$, $L(f, \mathcal{P}_n)$ and $\Delta(f, \mathcal{P}_n)$: if $\Delta(f, P_n) \leq \epsilon$; great. If not, try again with a larger value of n.

This works so well that we might try to bootstrap it to other functions, e..g. by breaking up $f : [a, b] \to \mathbb{R}$ into finitely many subintervals such that it is monotone on each one. Unfortunately not every function is piecewise monotone, and for those which are we may have to do quite a lot of work in order to successfully break it up in this way. Or we might try to write $f = g - h$ where each of g and h is monotone, taking advantage of the fact that $\int_a^b f = \int_a^b g - \int_a^b h$. In theory, a large class of functions can be written as the difference of two increasing functions – in particular every function with a continuous derivative can be expressed this way – but in practice finding the g and h is usually not easy.

Why are we avoiding trying to compute the Darboux integral of a non-monotone function $f : [a,b] \to \mathbb{R}$? Because if f is not monotone, then for any partition $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$, then order to compute $U(f, \mathcal{P}_N)$ or $L(f, \mathcal{P}_n)$ we have to solve n optimization problems: we have to maximize (resp. minimize) f on each subinterval $[x_i, x_{i+1}]$. That doesn't sound fun. But we have a more basic issue: which partitions $\mathcal{P}$ should we be using? Darboux integrability means that for each $\epsilon > 0$ there is *some* partition $\mathcal{P}_\epsilon$ of $[a,b]$ for which $\Delta(f, \mathcal{P}_\epsilon) < \epsilon$. It doesn't tell us how to find $\mathcal{P}_\epsilon$. Geometric intuition (recall we have been assuming that f is continuous) suggests we should as in the monotone case be able to use the uniform partitions $\mathcal{P}_n$ for sufficiently large n, or in other words that we should again have

$$\lim_{n\to\infty} \Delta(f, \mathcal{P}_n) = 0, \text{ hence} \lim_{n\to\infty} U(f, \mathcal{P}_n) = \lim_{n\to\infty} L(f, \mathcal{P}_n) = \int_a^b f.$$

We still have the darned upper and lower sums, but....actually, it is clear that the left endpoint sum $\frac{b-a}{n} \sum_{i=0}^{n-1} f(a + i(\frac{b-a}{n}))$ lies in between $L(f, \mathcal{P}_n)$ and $U(f, \mathcal{P}_n)$, so by the Squeeze Theorem for sequences it would indeed then follow that

$$\lim_{n\to\infty} \frac{b-a}{n} \sum_{i=0}^{n-1} f(a + i(\frac{b-a}{n})) = \int_a^b f.$$

Riemann's work shows that all of these things are true and more. There are two key ideas that distinguish Riemann's integral from Darboux's. First, instead of upper and lower sums we work with sums obtained by taking the height of the rectangle to be *any* point in the subinterval $[x_i, x_{i+1}]$. The second idea is that his notion of convergence is *a priori* more demanding than Darboux's in a way that works against you if you are trying to show that a given function is integrable but works for you if you know that it is.

We begin with a function $f : [a,b] \to \mathbb{R}$ and a partition $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ of $[a,b]$, but now we introduce one more piece of data, a **tagging** of $\mathcal{P}$. A tagging is a function $\tau : \{0, 1, \ldots, n\} \to [a,b]$ such that for all i, the point $\tau(i)$ lies in the ith subinterval $[x_i, x_{i+1}]$ determined by the partition $\mathcal{P}$. Instead of using functional notation we may just put $x_i^* = \tau(i)$, and then a tagging is a finite sequence $\{x_0^* \leq x_1^* \leq \ldots \leq x_{n-1}^* \leq x_n^*\}$. Notice that this sequence need not be *quite* injective: we could have $x_i^* = x_{i+1}^*$; this holds if and only if both are equal to x_{i+1}, which is both the right endpoint of $[x_i, x_{i+1}]$ and the left endpoint of $[x_{i+1}, x_{i+2}]$. The pair $(\mathcal{P}, \tau)$ is called a **tagged partition**.

To any tagged partition $(\mathcal{P}, \tau)$ and, of course, a function $f : [a, b] \to \mathbb{R}$ we associate a **Riemann sum**

$$R(f, \mathcal{P}, \tau) \coloneqq \sum_{i=0}^{n-1} f(x_i^*)(x_{i+1} - x_i).$$

It is easy to compare with the upper and lower sums: of course we have

$$\sup(f([x_i, x_{i+1}]) \geq f(x_i^*) \text{ and } \inf(f[x_i, x_{i+1}]) \leq f(x_i^*),$$

so

$$L(f, \mathcal{P}) \leq R(f, \mathcal{P}, \tau) \leq U(f, \mathcal{P}).$$

If f assumes its maximum and minimum value on each subinterval $[x_i, x_{i+1}]$ – so for instance if f is continuous – then $L(f, \mathcal{P})$ and $U(f, \mathcal{P})$ are themselves Riemann sums. In general this is not quite true because the suprema and infima need not be attained, but almost: we will have

$$U(f, \mathcal{P}) = \sup_\tau R(f, \mathcal{P}, \tau) \text{ and } L(f, \mathcal{P}) = \inf_\tau R(f, \mathcal{P}, \tau).$$

Thus for each partition $\mathcal{P}$, the upper sum is the least upper bound of all possible Riemann sums for $\mathcal{P}$ and the lower sum is the greatest lower bound of all possible Riemann sums for $\mathcal{P}$. This is quite clear if f is bounded; it is still true if f is unbounded, but it requires a bit more work:

Proposition 3.11. *Let $f : [a, b] \to \mathbb{R}$ be a function and $\mathcal{P}$ a partition of $[a, b]$.*

a) *If f is unbounded above, then as we range over all possible taggings τ of $[a, b]$, we have*

$$\sup_\tau R(f, \mathcal{P}, \tau) = \infty.$$

b) *If f is unbounded below, then as we range over all possible taggings τ of $[a, b]$, we have*

$$\inf -\tau R(f, \mathcal{P}, \tau) = -\infty.$$

We leave the proof of Proposition 3.11 as Exercise 3.19.

So far this is all pretty similar to Darboux's treatment. The second main idea is that the sense in which the Riemann sums are required to converge to $\int_a^b f$ is quite stringent. To give it, we need just one more definition: for a partition $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ of $[a, b]$, its **mesh** is

$$|\mathcal{P}| \coloneqq \max_{0 \leq i \leq n-1} x_{i+1} - x_i;$$

that is, the mesh of $\mathcal{P}$ is the largest length of a subinterval $[x_i, x_{i+1}]$. For instance, in the uniform partition $\mathcal{P}_n$ all subintervals have length $\frac{b-a}{n}$, so its mesh is $|\mathcal{P}_n| = \frac{b-a}{n}$.

A function $f : [a, b] \to \mathbb{R}$ is **Riemann integrable** if there is $S \in \mathbb{R}$ such that: for all $\epsilon > 0$, there is $\delta > 0$ such that for every partition $\mathcal{P}$ of $[a, b]$ with mesh $|\mathcal{P}| \leq \delta$ and every tagging τ of $\mathcal{P}$, we have

$$|R(f, \mathcal{P}, \tau) - S| \leq \epsilon.$$

We then put $\int_a^b f \coloneqq S$.

Let us check that Riemann integrability implies Darboux integrability: let $\epsilon > 0$.

Then there is $\delta > 0$ such that for any partition $\mathcal{P}$ of mesh less than δ we have $|R(f, \mathcal{P}, \tau) - S| \leq \frac{\epsilon}{2}$, which of course means that $R(f, \mathcal{P}, \tau) \in [S - \frac{\epsilon}{2}, S + \frac{\epsilon}{2}]$. Because the upper sum $U(f, \mathcal{P})$ is the supremum of the $R(f, \mathcal{P}, \tau)$'s as we range over τ and $R(f, \mathcal{P}, \tau) \leq S + \frac{\epsilon}{2}$ for all τ, we get $U(f, \mathcal{P}) \leq S + \frac{\epsilon}{2}$; similarly, we get $L(f, \mathcal{P}) \geq S - \frac{\epsilon}{2}$, so $\Delta(f, \mathcal{P}) \leq \epsilon$, and thus f is Darboux integrable and moreover S is the Darboux integral $\int_a^b f$. (In particular, there is at most one $S \in \mathbb{R}$ satisfying the conditions in the definition of Riemann integrability.)

It is **much less obvious** whether every Darboux integrable function is Riemann integrable. But happily it is true:

Theorem 3.12. a) *For a function $f : [a, b] \to \mathbb{R}$, the following are equivalent:*

(i) *The function f is Darboux integrable.*
(ii) *The function f is Riemann integrable.*
(iii) *For every sequence $\{(\mathcal{P}_n, \tau_n)\}_{n=1}^{\infty}$ of tagged partitions of $[a, b]$ with $|\mathcal{P}_n| \to 0$, the sequence $\{R(f, \mathcal{P}_n, \tau_n)\}_{n=1}^{\infty}$ of Riemann sums is convergent.*

b) *If the equivalent conditions of part a) hold, then for any sequence $\{(\mathcal{P}_n, \tau_n)\}$ of tagged partitions of $[a, b]$ with $|\mathcal{P}_n| \to 0$, we have*

$$\lim_{n\to\infty} R(f, \mathcal{P}_n, \tau_n) = \int_a^b f.$$

We are not going to prove Theorem 3.12 in our course, but you can find the proof in [**HC**, §8.4]. So that you don't feel short-changed, let me mention that most undergraduate analysis texts do not prove this theorem; many of them just develop Darboux's integral and forget to make the connection with Riemann sums.

Let us sum up the state of affairs: because of Theorem 3.12, the Darboux integral and the Riemann integral, although they were defined differently, turn out to be completely equivalent: a function is integrable in sense if and only if it is in the other sense, and if so they return the same real number $\int_a^b f$. So we no longer need to distinguish between them: henceforth we will *only* speak of Riemann integrable functions and the Riemann integral. This is what is most commonly done, even by people who have much less right to conflate the two than we do.

We end this section with one more result that helps to make the Riemann integral more computable.

Theorem 3.13. *Let $f : [a, b] \to \mathbb{R}$ be differentiable with bounded derivative: let $M > 0$ be such that $|f'| \leq M$. For $n \in \mathbb{Z}^+$, let*

$$L_n(f) = \sum_{i=0}^{n-1} f(a + i(\frac{b-a}{n}))(\frac{b-a}{n})$$

*be the **left endpoint Riemann sum** of f. Then*

$$\left| \int_a^b f - L_n(f) \right| \leq \left(\frac{(b-a)^2 M}{2} \right) \frac{1}{n}.$$

Proof. Step 1: We establish the result for $n = 1$. For $x \in [a, b]$, we apply the Mean Value Theorem to f on the interval $[a, x]$: there is $c \in (a, x)$ with

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

Since $|f'(c)| \leq M$, we get

$$-M(x - a) + f(a) \leq f(x) \leq M(x - a) + f(a)$$

and thus

$$\int_a^b (-M(x - a) + f(a)) \leq \int_a^b f \leq \int_a^b (M(x - a) + f(a)).$$

Of course we can evaluate the first and last integrals with the Fundamental Theorem of Calculus, and we get

$$\frac{-M}{2}(b - a)^2 + (b - a)f(a) \leq \int_a^b f \leq \frac{M}{2}(b - a)^2 + (b - a)f(a),$$

which is equivalent to

$$\left| \int_a^b f - L_1(f) \right| \leq \frac{M}{2}(b - a)^2.$$

Step 2: Let $n \in \mathbb{Z}^+$. For $0 \leq i \leq n - 1$, put $x_i^* = a + i\frac{b-a}{n}$. Then:

$$\left| \int_a^b f - L_n(f) \right| = \left| \sum_{i=0}^{n-1} \left(\int_{x_i^*}^{x_{i+1}^*} f - f(x_i^*) \left(\frac{b - a}{n} \right) \right) \right|$$

$$\leq \sum_{i=0}^{n-1} \left| \int_{x_i^*}^{x_{i+1}^*} f - f(x_i^*) \left(\frac{b - a}{n} \right) \right|.$$

Step 1 applies to each term in the last sum to give

$$\left| \int_a^b f - L_n(f) \right| \leq \sum_{n=0}^{n-1} \frac{M}{2} \left(\frac{b - a}{n} \right)^2 = \left(\frac{(b - a)^2 M}{2} \right) \frac{1}{n}. \qquad \square$$

Whereas Theorem 3.12 guarantees us that for any Riemann integral $f : [a, b] \to \mathbb{R}$, we can compute $\int_a^b f$ as the limit $\lim_{n\to\infty} L_n(f)$ of the left endpoint Riemann sums, Theorem 3.13 gives us, for functions with a bounded derivative, a precise **error estimate**: it tells us how large n needs to be in order to for $L_n(f)$ to compute $\int_a^b f$ to any prescribed accuracy, where the bound depends on the size of the derivative. To get a bound on f' essentially amounts to solving one optimization problem, which is great progress over the arbitrarily many optimization problems we had to solve to compute a single upper or lower sum. More basically, this result is telling us that the faster f is changing from point to point in the local sense, the more sample points we will need in order to get a handle on $\int_a^b f$: this makes a lot of sense. On the other hand, of course if we don't know anything about f other than that it is, say, differentiable, then we don't know how many sample points we will need to use to usefully approximate $\int_a^b f$ because for any sample points we choose, for all we know f could be oscillating wildly in between them.

Moreover Theorem 3.13 is the first of an infinite sequence of theorems: the rough form of the kth theorem in the sequence is that if we assume that the kth derivative $f^{(k)}$ of f exists and is bounded, then using the values of f at the points of the partition $\mathcal{P}_n$ of $[a,b]$ into n equally spaced subintervals, one can write down a certain finite sum $S_{k,n}(f)$ that is a certain weighted average of several different Riemann sums, and has the property that

$$\left|S_{k,n}(f) - \int_a^b f\right| \leq C\frac{1}{n^k}.$$

Here C is a certain explicit expression depending only on $(b-a)$, the number n of sample points and an upper bound M for $|f^{(k)}|$. Thus, the more smoothness we assume on f, the more rapidly we can compute $\int_a^b f$. The $k=2$ case is the Trapezoidal Rule [**HC**, Thm. 9.5], while the $k=3$ case is Simpson's rule [**HC**, Thm. 9.8]. The branch of mathematics in which you will learn how this works for all $k \in \mathbb{Z}^+$ and many other similar results is **numerical analysis**.

3.1. Exercises.

Exercise 3.19. *Prove Proposition 3.11.*

Exercise 3.20. *In this exercise we show that* $\lim_{n\to\infty} \sum_{k=1}^n \frac{n}{k^2+n^2} = \frac{\pi}{4}$.

a) *Let* $\mathcal{P}_n$ *be the partition of* $[0,1]$ *into* n *equally spaced subintervals. Let* $f:[0,1]\to\mathbb{R}$ *by* $f(x) = \frac{1}{x^2+1}$. *Show:* $\sum_{k=1}^n \frac{n}{k^2+n^2} = R(f,\mathcal{P}_n,\tau_n)$, *where* τ_n *is the right endpoint tagging: for all* $0 \leq i \leq n-1$, $x_i^* = \frac{i}{n}$.
b) *Use Theorem 3.12 and the Fundamental Theorem of Calculus to evaluate* $\lim_{n\to\infty} R(f,\mathcal{P}_n,\tau_n)$.

4. The Class of Riemann Integrable Functions

4.1. More Riemann Integrable Functions. We had a big fish to catch: the existence of an antiderivative for any continuous function. So we built a big net — the Darboux integral — and with that big net we caught our fish. (Then we discussed the construction of a second net — the Riemann integral — that looked rather different from our first net, but we found that in the end the second net catches exactly the same fish as the first. So we stopped distinguishing between the two nets.) It is now time to ask: what other fish have we caught? That is, what can we say about the class $\mathcal{R}[a,b]$ of Riemann integrable $f:[a,b]\to\mathbb{R}$? Again, we know that this class contains all continuous functions, and we also know that every function in the class is bounded.

The previous exercises Exercise 3.12 and 3.13 give some instances of functions that are discontinuous but Riemann integrable. Let us concentrate on the latter: according to 3.13, if $f:[a,b]\to\mathbb{R}$ is bounded and has only finitely many discontinuities, then f is Riemann integrable. Let us sketch a proof: let $M>0$ be such that $|f|\leq M$, fix $\delta>0$, and choose a partition $\mathcal{P}$ of $[a,b]$ that contains, for each point c of disconinuity of f — to fix ideas, let us assume that the discontinuities occur at interior points of $[a,b]$ – there are consecutive elements $x_i, x_{i+1} \in \mathcal{P}$ with $x_{i+1}-x_i<\delta$. If we remove the open intervals (x_i,x_{i+1}) from $[a,b]$, we get a finite union of closed subintervals — suppose that there are N of them — such that f is continuous on each one, hence Riemann integrable. This means that for any $\epsilon>0$ we can refine $\mathcal{P}$ to a partition $\mathcal{P}_\epsilon$ such that on the Nth subinterval, the difference between the upper sum and the lower sum is at most $\frac{\epsilon}{2N}$, so therefore the sum of the differences of the lower sums is at most $\frac{\epsilon}{2}$. Finally, on each subinterval $[x_i, x_{i+1}]$ we have

$$\Delta(f|_{[x_i,x_{i+1}]}, \mathcal{P}_\epsilon) < 2M\delta.$$

This is because since $|f| \leq M$, its oscillation – i.e., its supremum minus its infimum – is at most $2M$, so we multiply this by the length of the subinterval. Thus overall for this partition $\mathcal{P}_\epsilon$ we find that

$$\Delta(F, \mathcal{P}_\epsilon) < \frac{\epsilon}{2} + 2NM\delta.$$

Since M and N are fixed, we can choose δ sufficiently small so that $2NM\delta < \frac{\epsilon}{2}$, and we win: f is integrable by Darboux's Criterion.

So now we are interested in bounded function $f : [a, b] \to \mathbb{R}$ with *infinitely many* discontinuities. At first glance, such a function looks unlikely to be Riemann integrable, at least to me: by Exercise 1.43, the set of discontinuities of f must have an accumulation point in $[a, b]$, and that seems like it could screw things up – arguments like the one we made for finitely many discontinuities are not going to succeed. (Anyway, in our argument above the *number* N of discontinuities appeared in our bound; if there are infinitely many discontinuities, we certainly cannot do this.) However, again some previous exercises show that we've caught profoundly more fish than we thought: Exercise 3.17 shows that every monotone function $f : [a, b] \to \mathbb{R}$ is Riemann integrable. That is not so surprising, but Exercise 3.18 is: there is a strictly increasing function $f : [a, b] \to \mathbb{R}$ that is discontinuous at every rational point of $[a, b]$! Thus a bounded function can be Riemann integrable even when its set of discontinuities is *dense* in $[a, b]$.

The following result further exhibits the largeness of the class of Riemann integrable functions.

THEOREM 3.14. *Let $f : [a, b] \to [c, d]$ be Riemann integrable, and let $g : [c, d] \to \mathbb{R}$ be continuous. Then the composite function $g \circ f : [a, b] \to \mathbb{R}$ is Riemann integrable.*

We are going to omit the proof of this result because of time constraints and because it is a bit technical: see [**HC**, Thm. 8.17]. It becomes easier in an important special case. For a subset X of $\mathbb{R}^N$, a function $f : X \to \mathbb{R}^M$ is **Lipschitz** if there is a constant $C \in (0, \infty)$ such that

$$\forall x_1, x_2 \in X,\ ||f(x_1) - f(x_2)|| \leq C||x_1 - x_2||.$$

Any C that works here is called a **Lipschitz constant** for f. You should think of Lipschitz as a kind of "super-continuity": indeed such functions are uniformly continuous with $\delta = \frac{\epsilon}{C}$. The following result — the second part of which appeared on the midterm! — showed that this property, although very strong, certainly comes up in nature.

PROPOSITION 3.15. *Let I be an interval, and let $f : I \to \mathbb{R}$ be a differentiable function.*

a) *If f' is bounded, then f is Lipschitz.*
b) *If $I = [a, b]$ and f' is continuous, then f is Lipschitz.*

PROOF. a) Let $M > 0$ such that $|f'| \leq M$. Let $x_1 < x_2$ be elements of I. By the Mean Value Theorem there is $c \in (x_1, x_2)$ such that

$$|f(x_1) - f(x_2)| = |f'(c)||x_1 - x_2| \leq M|x_1 - x_2|.$$

Thus M is a Lipschitz contant for f.
b) If $f' : [a,b] \to \mathbb{R}$ is continuous, then by the Extreme Value Theorem, f' is bounded, so part a) applies to show that f is Lipschitz. $\square$

It turns out to be *much* easier to show Theorem 3.14 if we assume that g is not only continuous but Lipschitz: this is Exercise 3.22. Here is a nice consequence of this:

THEOREM 3.16. *If $f, g : [a,b] \to \mathbb{R}$ are both Riemann integrable, then so is $f \cdot g$.*

PROOF. If $h : [a,b] \to \mathbb{R}$ is any Riemann integrable function, then it follows from Theorem 3.14 that h^2 is also Riemann integrable. In fact, since h is bounded – say $|h| \leq M$ – by Proposition 3.15 we have that $x^2 : [-M, M] \to \mathbb{R}$ is Lipschitz, so this lies in the part of Theorem 3.14 that we (the student who solves the right exercises and I) have proved. Now here is a dirty trick;

$$fg = \frac{(f+g)^2 - f^2 - g^2}{2},$$

which shows that fg is Riemann integrable, since we know that linear combinations of Riemann integrable functions are Riemann integrable and that squares of Riemann integrable functions are Riemann integrable. $\square$

4.2. The Riemann-Lebesgue Criterion. In fact there is a precise characterization of which bounded functions $f : [a,b] \to \mathbb{R}$ are Riemann integrable. This is usually called **Lebesgue's Criterion**, after the leading mathematician Henri Lebesgue who constructed a superior version of the integral to Riemann's. (Lebesgue is also the founder of the subject of measure theory referred to above. Most of Math 8100 concerns measure theory and the Lebesgue integral.) However my former colleague Roy Smith showed me exactly where this criterion occurs in a work of Riemann, so I will call it the Riemann-Lebesgue criterion.

We actually need a tiny piece of measure theory even to state this criterion, namely we need the notion of a subset X of $\mathbb{R}$ having **measure zero**. For this, let $\{[a_n, b_n]\}_{n=1}^{\infty}$ be a sequence of closed bounded intervals. We say that this sequence **covers** X if

$$X \subseteq \bigcup_{n=1}^{\infty} [a_n, b_n],$$

or in words, if every element of X lies in at least one of the subintervals $[a_n, b_n]$. To this sequence of intervals we attach a **total length**

$$\mathbb{L}(\{[a_n, b_n]\} := \sum_{n=1}^{\infty} (b_n - a_n) \in [0, \infty].$$

In other words, we really do just add up the lengths of the subintervals. This is an infinite series with non-negative terms, so it either converges or diverges to ∞. The idea is that, if the total length is finite, it should give an upper bound on the "length" of X. This intriguing idea is the beginning of measure theory, but we only need this one thing: we say that X has **measure zero** if for all $\epsilon > 0$, there is a covering $\{[a_n, b_n]\}$ of X of total length at most ϵ.

This concept is addressed in the exercises. It is pretty clear that every finite subset of $\mathbb{R}$ has measure zero: if we are allowed to use degenerate closed intervals $[a, a]$ this is truly obvious, but actually in the definition of measure zero it doesn't matter whether we are allowed this or not, so it might be more educational to always use intervals of positive length. More generally, we say that a subset $X \subseteq \mathbb{R}$ is **countable** if there is a surjective sequence in X, i.e., a surjective function $x_\bullet : \mathbb{Z}^+ \to X$. This includes finite subsets, certainly. Moreover, each of $\mathbb{Z}^+$, $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$ are countable (these are developed in the exercises). Then any countable subset X has measure zero – again, this is clear if we can use intervals $[a, a]$ but is not much harder to show even if we can't. There are also uncountable subsets of measure zero: one very famous one, the **Cantor set**, is developed in the exercises.

And here is the result:

THEOREM 3.17 (Riemann-Lebesgue Criterion). *For a function $f : [a, b] \to \mathbb{R}$, the following are equivalent:*

(i) *f is Riemann integrable.*
(ii) *f is bounded, and its set of discontinuities has measure zero.*

We will not give a proof of Theorem 3.17. Most proofs use somewhat more advanced material, but this is not necessary: see [**HC**, §8.5] for a proof that you have all the prerequisites to read.

Nevertheless we can stop to appreciate Theorem 3.17: it tells us *exactly* what fish we've caught with our integral! Moreover, if you know this result than many of our other results on Riemann integrability follow immediately. It is easy to see from the definition of measure zero that a finite union of sets, each of measure zero, also has measure zero. (It is not much harder to see that moreover if $\{X_n\}_{n=1}^\infty$ is an infinite sequence of sets of measure zero then $\bigcup_{n=1}^\infty X_n$ also has measure zero...but we don't need this here.)

So: let $f, g : [a, b] \to \mathbb{R}$ be Riemann integrable, so each is bounded and is discontinuous only a set of measure zero. Then:

• For $\alpha \in \mathbb{R}$, αf is bounded and has the same discontinuities as f, so is Riemann integrable.
• $f + g$ is bounded (if $|f| \leq M_1$ and $|g| \leq M_2$ then $|f + g| \leq M_1 + M_2$). If the set of discontinuties of f is X_f and the set of discontinuities of g is X_g, then the set of discontinuities of $f + g$ is contained in $X_f \cup X_g$, so has measure zero. So $f + g$ is Riemann integrable.
• Almost the identical argument works to show that $f \cdot g$ is Riemann integrable (only modification: if $|f| \leq M_1$ and $|g| \leq M_2$, then $|fg| \leq M_1 M_2$).
• If f is monotone, then it is bounded – $f([a, b])$ lies in the interval in between $f(a)$ and $f(b)$ – and it can be shown that the set of discontinuities of f is *countable*. So f is Riemann integrable.
• Suppose $f : [a, b] \to [c, d]$ and $g : [c, d] \to \mathbb{R}$ is continuous. Then g is bounded, hence so is $g \circ f$. Moreover, since g is continuous, the set of discontinuities of $g \circ f$ is contained in the set of discontinuities of f, and a subset of a set of measure zero has measure zero. So $g \circ f$ is Riemann integrable.

4.3. Exercises.

Exercise 3.21. *Let $f : I \to [c,d]$ be a bounded function, and let $g : [c,d] \to \mathbb{R}$ be a Lipschitz function with Lipschitz constant C. Show:*

$$\omega(g \circ f, I) \leq C\omega(f, I).$$

Exercise 3.22. *Let $f : [a,b] \to [c,d]$ be Darboux integrable, and let $g : [a,b] \to \mathbb{R}$ be Lipschitz with Lipschitz constant C. Show that $g \circ f : [a,b] \to \mathbb{R}$ is Riemann integrable as follows: let $\epsilon > 0$, and choose a partition $\mathcal{P}_\epsilon$ for which $\Delta(f, \mathcal{P}_\epsilon) < \frac{\epsilon}{C}$. Use Exercise 3.21 to show that $\Delta(g \circ f, \mathcal{P}_\epsilon) < \epsilon$.*

Exercise 3.23. *Let $f : [a,b] \to \mathbb{R}$ be Riemann integrable.*

a) *Show that $|f| : [a,b] \to \mathbb{R}$ is Riemann integrable.*
(Suggestions: the absolute value function is Lipschitz, so you can apply Exercise 3.22. Or you can show that for any subinterval I of $[a,b]$ we have $\omega(|f|, I) \leq \omega(f, I)$.)

b) *Show the **Integral Triangle Inequality**:*

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

A nonempty set X is **countable** if there is a surjective function $f : \mathbb{Z}^+ \to X$. By definition, the empty set is also countable.

Exercise 3.24. *Let X be a set.*

a) *Show: if X is finite, then X is countable.*
b) *Show: if X is infinite and countable, then there is a bijection $f : \mathbb{Z}^+ \to X$.*

Thus every countable set is in bijection with exactly one of the following: (i) the empty set; (ii) the set $\{1, \ldots, n\}$ for some $n \in \mathbb{Z}^+$; or (iii) $\mathbb{Z}^+$.

Exercise 3.25. a) *Show: every subset of a countable set is countable.*

b) *Let $\iota : Y \to X$ be an injective function. Show: if X is countable, then so is Y.*

Exercise 3.26. *Show: if $n \in \mathbb{Z}^+$ and $X_1, \ldots, X_n$ are countable sets, then their union $\bigcup_{n=1}^{\infty}$ is countable.*

Exercise 3.27. a) *Let X and Y be sets. Show: if X is countable and there is a surjection $f : X \to Y$, then Y is countable.*

b) *Show: $\mathbb{Z}^+ \times \mathbb{Z}^+$ is countable.*
c) *Show: the set $\mathbb{Q}$ of rational numbers is countable.*
*(Suggestion: since $\mathbb{Q} = \mathbb{Q}^{>0} \cup \{0\} \cup \mathbb{Q}^{<0}$ and multiplication by -1 gives a bijection from $\mathbb{Q}^{>0}$ to $\mathbb{Q}^{<0}$, by Exercise **??** it is enough to show that $\mathbb{Q}^{>0}$ is countable. Do this by finding a surjective function $f : \mathbb{Z}^+ \times \mathbb{Z}^+ \to \mathbb{Q}$.)*

Exercise 3.28.

a) *Let $X \subseteq \mathbb{R}$ be a countable subset. Show: X has measure zero.*
b) *Deduce from the Riemann-Lebesgue Criterion that $\mathbb{R}$ is uncountable.*
(Hint: if $\mathbb{R}$ were countable, then every bounded function would be Riemann integrable.)

The previous exercise gives a proof of the uncountability of $\mathbb{R}$ that is striking, but is also overkill: the Riemann-Lebesgue Criterion is a difficult result whose proof we have omitted. The next exercise outlines a classic proof (due to G. Cantor) of the uncountability of $\mathbb{R}$.

Exercise 3.29. *In this exercise we refer to decimal expansions of real numbers. Some real numbers have a unique decimal expansion, but others have (exactly) two different decimal expansions: a real number admitting a decimal expansion ending with all* 0*'s also has a decimal expansion ending with all* 9*'s. For the sake of definiteness, when we refer to "the decimal expansion" of* $x \in \mathbb{R}$ *we will exclude an expansion ending with all* 9*'s.*

Let $f : \mathbb{Z}^+ \to \mathbb{R}$ *be any function, and put* $x_n := f(n)$*. Build a real number* $x = 0.d_1 d_2 \cdots d_n \cdots \in [0,1]$ *as follows: for all* $n \in \mathbb{Z}^+$*, the* n*th decimal digit* d_n *of* x *is different from the* n*th decimal digit of* x_n *and also different from* 0 *and* 9*. (This still leaves us at least* 7 *choices.) Show: for no* $n \in \mathbb{Z}^+$ *do we have* $x = x_n$*, and deduce that* f *is not surjective.*

Exercise 3.30. *Let* I *be an interval, and let* $f : I \to \mathbb{R}$ *be a monotone function. Let* X *be the set of* $c \in I$ *such that* f *is discontinuous at* c*. Show:* X *is countable. (Suggestion: we may assume* f *is increasing. An increasing function* f *can only be discontinuous at* c *if* $\lim_{x \to c^-} f(x) < \lim_{x \to c^+} f(x)$*. If so, there is a rational number lying strictly in between the left hand limit and the right hand limit. Use this to build an injective function* $\iota : X \to \mathbb{Q}$ *and then apply Exercise 3.25b).)*

Compare Problem 3.18 with Exercise 3.30. Things are getting subtle: monotone functions can have infinitely many discontinuities on a bounded interval, but still their set of discontinuities is "small" in a strong sense.

CHAPTER 4

Convex Functions

1. The Basic Definition

Let $X \subseteq \mathbb{R}^N$, and let $f : X \to \mathbb{R}$ b a function. We define the **epigraph** of f,

$$\mathrm{Epi}(f) := \{(x_1, \ldots, x_N, y) \in X \times \mathbb{R} \mid y \geq f(x_1, \ldots, x_N)\}.$$

Compare the epigraph to the *graph* of f:

$$\mathrm{Graph}(f) := \{(x_1, \ldots, x_N, y) \in X \times \mathbb{R} \mid y = f(x_1, \ldots, x_N)\}.$$

The only difference in the two deifnitions is the $\geq$ appearing in the definition of the epigraph: thus the epigraph is indeed the set of points that lie "on top of" the graph $\mathrm{Graph}(f)$ in the sense that for all $P = (x_1, \ldots, x_N) \in X$, the point (P, y) lies on the epigraph if and only if the final coordinate y is equal or larger to the final coordinate of the corresponding point $(P, f(P))$ of $\mathrm{Graph}(f)$. In particular:

$$\mathrm{Graph}(f) \subseteq \mathrm{Epi}(f).$$

Proposition 4.1. *Let $X \subseteq \mathbb{R}^N$ be a convex set. For a function $f : X \to \mathbb{R}$, the following are equivalent:*

(i) *The epigraph $\mathrm{Epi}(f)$ of f is a convex subset of $\mathbb{R}^{N+1}$.*
(ii) *The function f satisfies the **Secant Graph Inequality**: for all $P, Q \in X$ and all $0 \leq \lambda \leq 1$, we have*

$$f((1-\lambda)P + \lambda Q) \leq (1-\lambda)f(P) + \lambda f(Q).$$

Proof. (i) $\implies$ (ii): Suppose that $\mathrm{Epi}(f)$ is a convex subset of $\mathbb{R}^{N+1}$, and let $P, Q \in X$ and $\lambda \in [0,1]$. Then the points

$$\widetilde{P} := (P, f(P)) \text{ and } \widetilde{Q} := (Q, f(Q))$$

lie on the graph of f, hence also on the epigraph $\mathrm{Epi}(f)$. Since $\mathrm{Epi}(f)$ is convex, we have $(1-\lambda)\tilde{P} + \lambda\tilde{Q}$ also lies in the epigraph, which means that its final coordinate — which is $(1-\lambda)f(P) + \lambda f(Q)$ — is greater than or equal to the value of f on its first N coordinates — which is $f((1-\lambda)P + \lambda Q)$. This is precisely condition (ii).
(ii) $\implies$ (i): Assume the Secant Graph Inequality, and let $\widetilde{P} = (P, y_P)$ and $\widetilde{Q} = (Q, y_Q)$ be two points of $\mathrm{Epi}(f)$: thus we have $y_P \geq f(P)$ and $y_Q \geq f(Q)$. Let $\lambda \in [0,1]$. By the Secant Graph Inequality, we have:

$$(1-\lambda)y_P + \lambda y_Q \geq (1-\lambda)f(P) + \lambda f(Q) \geq f((1-\lambda)P + \lambda Q).$$

This inequality indeed shows that $(1-\lambda)\widetilde{P} + \lambda\widetilde{Q}$ lies in $\mathrm{Epi}(f)$. $\square$

We call a function satisfying the equivalent properties of Proposition 4.1 **convex**.

2. Secant Inequalities

PROPOSITION 4.2. *Let $I \subseteq \mathbb{R}$ be an interval, and let $f : I \to \mathbb{R}$ be a function. The following are equivalent:*

(i) *f is convex.*

(ii) *(**Three Secant Inequality**) For all $a < x < b \in I$,*

$$\frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a} \leq \frac{f(b) - f(x)}{b - x}.$$

(iii) *(**Two Secant Inequality**) For all $a < x < b \in I$,*

$$\frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a}.$$

(iv) *(**Generalized Two Secant Inequality**) For all $a < b \leq c < d \in I$,*

$$\frac{f(b) - f(a)}{b - a} \leq \frac{f(d) - f(c)}{d - c}.$$

PROOF. (i) $\iff$ (iii): Clearing denominators in the Two Secant Inequality, we get: for all $a < x < b \in I$,

$$f(x) \leq f(a) + \left(\frac{f(b) - f(a)}{b - a}\right)(x - a). \tag{10}$$

The function on the right hand side is the secant line between $(a, f(a))$ and $(b, f(b))$, so (10) is precisely the Secant Graph Inequality, which was one of our two definitions of convex functions.

(ii) $\implies$ (iii): This is immediate.

(i) $\implies$ (ii): Let $a < x < b \in I$. As we saw, the Secant Graph Inequality implies the Two Secant Inequality:

$$\frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a},$$

which is half of the Three Secant Inequality. To get the other half: the secant line between $(a, f(a))$ and $(b, f(b))$ can also be written as

$$f(b) + \frac{f(a) - f(b)}{b - a}(b - x),$$

so the Secant Graph Inequality gives

$$f(x) \leq f(b) + \frac{f(a) - f(b)}{b - a}(b - x),$$

which is equivalent to

$$\frac{f(b) - f(a)}{b - a} \leq \frac{f(b) - f(x)}{b - x}.$$

(iv) $\implies$ (iii): Taking $b = c$ in the Generalized Two Secant Inequality we get the Two Secant Inequality (with different names for the three points).

(iii) $\implies$ (iv): We may assume that $b < c$, otherwise as above the two inequalities are the same. If we apply the Two Secant inequality to $a < b < c$ and then to $b < c < d$, we get

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

and

$$\frac{f(c) - f(b)}{c - b} \leq \frac{f(d) - f(c)}{d - c},$$

and thus

$$\frac{f(b)-f(a)}{b-a} \leq \frac{f(d)-f(c)}{d-c}. \quad \square$$

$\square$

3. Continuity

It is not *quite* true that convex functions must be continuous:

EXAMPLE 3.1. *Let $f : (0,1) \to \mathbb{R}$ by $f(x) = x^2$. Then f is continuous. In freshman calculus, to determine the concavity of f we would take its second derivative; since $f''(x) = 2 > 0$, we would say that f is "concave up," which is what we are now calling convex. Soon enough we will establish that this is actually true: a twice differentiable function $f : I \to \mathbb{R}$ is convex if and only if $f'' \geq 0$. At the moment we are interested in different questions: how can we extend f to be defined at 0 and 1 so as to be (i) continuous and (ii) convex?*

Since 0 and 1 are accumulation points of $(0,1)$, there is at most one way to extend f continuously to 0 and at most one way to extend f continuously to 1: namely, we must take $f(0) = 0$ and $f(1) = 1$. But what about convexity? Assuming, as we are, that

$$\operatorname{Epi}(f) = \{(x,y) \in \mathbb{R} \mid 0 < x < 1,\ y \geq x^2\}$$

is convex, we find that if we put $f(0) = A$ and $f(1) = B$, then the epigraph of the extension is convex if and only if $A \geq 0$ and $B \geq 1$. Thus it is possible to extend f to $[0,1]$ so as to be convex and not continuous.

However, we should not give up so easily. It turns out that a convex function $f : I \to \mathbb{R}$ must be continuous except possibly on the endpoints of the interval, and actually even more is true.

THEOREM 4.3. *Let $I \subseteq \mathbb{R}$ be an interval, and let $f : I \to \mathbb{R}$. Then the restriction of f to the interior I° of I is continuous. Moreover, for any subinterval $[a,b]$ of I°, we have that $f|_{[a,b]} : [a,b] \to \mathbb{R}$ is Lipschitz: there is $C \geq 0$ such that for all $x, y \in [a,b]$ we have $|f(x) - f(y)| \leq C|x-y|$.*

PROOF. Let $[a,b] \subset I^\circ$; we may then choose $u, v, w, z \in I^\circ$ with $u < v < a$ and $b < w < z$. Now let $x \leq y \in [a,b]$. Applying the Generalized Two Secant Inequality twice, we get

$$\frac{f(v)-f(u)}{v-u} \leq \frac{f(y)-f(x)}{y-x} \leq \frac{f(z)-f(w)}{z-w}.$$

It follows that

$$\frac{|f(x)-f(y)|}{|x-y|} \leq \max\left(\left|\frac{f(v)-f(u)}{v-u}\right|, \left|\frac{f(z)-f(w)}{z-w}\right| \right),$$

so

$$C := \max\left(\left|\frac{f(v)-f(u)}{v-u}\right|, \left|\frac{f(z)-f(w)}{z-w}\right| \right)$$

is a Lipschitz constant for f on $[a,b]$.

Lipschitz functions are continuous, so f is continuous on every closed subinterval of I°. For any $c \in I^\circ$ there is $\delta > 0$ such that $[c-\delta, c+\delta] \subset I^\circ$, and this shows that f is continuous at c. $\square$

4. Derivatives

The following result tells us that what we learned about convex functions in freshman calculus was actually correct.

THEOREM 4.4. *Let $I \subseteq \mathbb{R}$ be an interval, and let $f : I \to \mathbb{R}$ be a function.*

a) *Suppose that f is differentiable. Then the following are equivalent:*
   (i) *f is convex.*
   (ii) *f' is increasing.*

b) *Suppose that f is twice differentiable. Then the following are equivalent.*
   (i) *f is convex.*
   (ii) *$f'' \geq 0$ on I.*

PROOF. a) (i) $\implies$ (ii): Suppose f is convex. Let $a < x \leq b \in I$, and define

$$s(x) := \frac{f(x) - f(a)}{x - a},$$

while for $x \in [a, b)$, we define

$$S(x) := \frac{f(b) - f(x)}{b - x}.$$

Since f is convex, we may apply the Three Secant Inequality for $a < x < b$:

$$s(x) \leq s(b) = S(a) \leq S(x).$$

Taking limits, we get:

$$f'(a) = \lim_{x \to a^+} s(x) \leq s(b) = S(a) \leq \lim_{x \to b^-} S(x) = f'(b).$$

(ii) $\implies$ (i): Let $a < b \in I$, and consider $s : (a, b] \to \mathbb{R}$ as in part a). Since f is differentiable, so is s, and

$$s'(x) = \frac{(x - a)f'(x) - (f(x) - f(a))}{(x - a)^2}.$$

By the Mean Value Theorem, there is $y \in (a, x)$ such that $\frac{f(x)-f(a)}{x-a} = f'(y)$. Since f' is increasing, we have

$$\frac{f(x) - f(a)}{x - a} = f'(y) \leq f'(x),$$

or equivalently:

$$(x - a)f'(x) - (f(x) - f(a)) \geq 0.$$

Thus $s'(x) \geq 0$ for all $x \in (a, b]$, so indeed s is increasing. In particular we have for all $a < x \leq b$ that

$$\frac{f(x) - f(a)}{x - a} \leq s(x) \leq s(b) = \frac{f(b) - f(a)}{b - a}.$$

This is the Two Secant inequality, so f is convex.
b) We know that a differentiable function defined on an interval is increasing if and only if its derivative is non-negative at all points on that interval. Applying this observation to f', we get that: f is convex if and only if f' is increasing if and only if $f'' \geq 0$. $\square$

COROLLARY 4.5. *Suppose that a function $f : I \to \mathbb{R}$ is both convex and differentiable. Then f' is continuous.*

PROOF. By Theorem 4.4a), f' is increasing. By Darboux's Theorem (Theorem 2.8), f' is a Darboux function, so by Corollary 2.25, f' is continuous. □

Here is a simple but imporant extremal property of differentiable convex functions.

COROLLARY 4.6. *Let I be an interval, and let $f : I \to \mathbb{R}$ be a differentiable convex function. Suppose that there is $c \in I$ such that $f'(c) = 0$. Then:*

a) *The function f attains a global minimum at c.*
b) *If f is not constant on any nontrivial subinterval of I, then f attains a* strict *global minimum at c: that is, for all $x \neq c$, we have $f(x) > f(c)$.*

PROOF. a) By Theorem 4.4, the function f' is increasing. Seeking a contradiction, we suppose that if f does not attain a minimum at c.
Case 1: Suppose there is some $d > c$ such that $f(d) < f(c)$. By the Mean Value Theorem, there is $z \in (c, d)$ such that $f'(z) < 0$. But since f' is increasing and $c < z$, this implies that $f'(c) \le f'(z) < 0$, so $f'(c) < 0$, a contradiction.
Case 2: Suppose there is some $b < c$ such that $f(b) < f(c)$. Now Mean Value Theorem implies there is $z \in (b, c)$ such that $f'(z) > 0$. But since f' is increasing and $z < c$, this implies that $f'(c) \ge f'(z) > 0$, a contradiction.
b) If there is some other point $e \neq c$ such that f attains a minimum at e, then by Corollary 2.7 we must have $f'(e) = 0$. Since f' is increasing, this means that f' is identically zero on the interval between c and e, so f is constant on the interval between c and e. □

5. Supporting Lines and Differentiability

Let I be an interval, let $f : I \to \mathbb{R}$ be a function, and let $c \in I$. A linear function[1] $\ell : I \to \mathbb{R}$ is a **supporting line for f at c** if

$$\ell(c) = f(c) \text{ and } \forall x \in I,\ f(x) \ge \ell(x).$$

Thus a supporting line is a line that is always equal or below the graph of the function and is equal at c.

EXAMPLE 5.1.

a) *Consider the function $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^2$. The line $\ell = 0$ is a supporting line for f at 0. It is hard not to notice that ℓ is none other than the tangent line to f at 0. It is moreover easy to see that this is the* only *supporting line at 0: suppose $\ell(x) = mx$ with $m \neq 0$. Then setting $f(x) = \ell(x)$ gives $x^2 = mx$, with solutions $x = 0$ and $x = m$. This is not yet a contradiction: a supporting line at c is allowed to "touch" the graph of f at points other than c. But now plug in $x = \frac{m}{2}$: we have*

$$f(\frac{m}{2}) = \frac{m^2}{4} < \frac{m^2}{2} = \ell(x).$$

In Exercise 4.2a) you are asked to show that for all $c \in \mathbb{R}$, the unique supporting line at c is the tangent line.

If instead we had $f(x) = -x^2$, then f could not have a supporting line

[1]Analysts and algebraists, sadly, don't quite agree on what a linear function from $\mathbb{R}$ to $\mathbb{R}$ is. An algebraist would say that it is a linear transformation, hence of the form αx for some $\alpha \in \mathbb{R}$. An analyst would say that $f(x) = mx + b$ for any $m, b \in \mathbb{R}$; equivalently, these are the functions with constant derivative. I am an algebraist but this is an analysis course, so *here*, by a linear function I mean $f(x) = mx + b$.

at any point: for sufficiently large $|x|$, $-x^2$ grows faster in absolute value than any linear function $\ell(x)$, so we will certainly have $-x^2 < \ell(x)$ for all sufficiently large $|x|$. Again, we can't help but notice: x^2 is convex – its second derivative is $2 > 0$ – while $-x^2$ is not – its second derivative is $-2 < 0$. This suggests that the existence of supporting lines has something to do with convexity.

b) *Consider the function $g : \mathbb{R} \to \mathbb{R}$ by $g(x) = |x|$. At $c = 0$ the line $y = x$ is clearly a supporting line. But also the line $y = -x$ is also a supporting line! In Exercise 4.2b) you are asked to show that the supporting lines for g at $c = 0$ are precisely those of the form $\ell(x) = mx$ with $m \in [-1, 1]$. You are also asked to show that if $c > 0$ the unique supporting line to g at c is $y = x$, while if $c < 0$ the unique supporting line to g at c is $y = -x$.*

Again the function g is convex: its epigraph is the intersection of two closed halfplanes, hence is a convex set. This reinforces our feeling that convexity is related to the existence *of supporting lines and further suggests that the differentiability of a convex function is related to the* uniqueness *of supporting lines.*

Theorem 4.7. *Let I be an open interval. For a function $f : I \to \mathbb{R}$, the following are equivalent:*

(i) *f is convex.*
(ii) *f admits a supporting line at each $c \in I$.*

Proof. (i) $\implies$ (ii): Neither of the two properties (i) or (ii) is disturbed by translating the coordinate axes, so we may assume that $c = 0$ and $f(0) = 0$. Suppose that f is convex.

Let $\alpha \in I \setminus \{0\}$. For all $\lambda_1, \lambda_2 > 0$ such that $\lambda_1\alpha, -\lambda_2\alpha \in I$, the Secant Graph Inequality gives

$$0 = (\lambda_1 + \lambda_2) f\left(\frac{\lambda_1}{\lambda_1 + \lambda_2}(-\lambda_2\alpha) + \frac{\lambda_2}{\lambda_1 + \lambda_2}(\lambda_1\alpha)\right) \le \lambda_1 f(-\lambda_2\alpha) + \lambda_2 f(\lambda_1\alpha),$$

or

$$\frac{-f(-\lambda_2\alpha)}{\lambda_2} \le \frac{f(\lambda_1\alpha)}{\lambda_1}.$$

It follows that

$$-\infty < \sup_{\lambda_2} \frac{-f(-\lambda_2\alpha)}{\lambda_2} \le \inf_{\lambda_1} \frac{f(\lambda_1\alpha)}{\lambda_1} < \infty,$$

so there is $m \in \mathbb{R}$ such that

$$\frac{-f(-\lambda_2\alpha)}{\lambda_2} \le m \le \frac{f(\lambda_1\alpha)}{\lambda_1}.$$

Equivalently, for all $t \in \mathbb{R}$ such that $t\alpha \in I$ we have $f(t\alpha) \ge mt$, which shows that $\ell(x) = \frac{m}{\alpha}x$ is a supporting line for f at $c = 0$.

(ii) $\implies$ (i): Suppose that f has a supporting line at each point of I, and choose one: for all $c \in I$, let $\ell_c : I \to \mathbb{R}$ be a supporting line for f at c. Then for all $x \in I$, we have

$$\forall c \in I,\ f(x) \ge \ell_c(x) \text{ and } f(c) = \ell_c(c),$$

so

$$f(x) = \sup_{c \in I} \ell_c(x).$$

Since linear functions are convex, it follows that f is the supremum of a family of convex functions, hence f is convex by Exercise 4.7. □

The proof of (i) $\implies$ (ii) in Theorem 4.7 is perhaps a bit too slick for its own good, but if one looks carefully one can detect a whiff of one-sided derivatives of f. The proof of (ii) $\implies$ (i) is more transparent: it gives a geometric sense in which f is put together from its family of supporting lines.

The next result is the deepest result on convex functions for which we will give a complete proof. We know that a convex function defined on an open interval must be continuous but need *not* be differentiable: e.g. $|x|$ has a corner point at 0. In fact a general convex function is differentiable except only at certain corner points, of which there are cannot be too many.

THEOREM 4.8. *Let I be an open interval, and let $f : I \to \mathbb{R}$ be convex.*

a) *For all $c \in I$, the left-hand derivative $f'_-(c)$ and the right-hand derivative $f'_+(c)$ each exist. Moreover, for all $x_1 < x_2 \in I$ we have*

$$f'_-(x_1) \leq f'_+(x_1) \leq f'_-(x_2) \leq f'_+(x_2). \tag{11}$$

In particular f'_- and f'_+ are increasing.

b) *The function f'_- is left-continuous on I: for all $c \in I$, $\lim_{x\to c^-} f(x) = f(c)$. Similarly, the function f'_+ is right-continuous on I: for all $c \in I$, $\lim_{x\to c^+} f(x) = f(c)$.*

c) *For $c \in I$, the following are equivalent:*
   (i) *f is differentiable at c.*
   (ii) *f'_- is continuous at c.*
   (iii) *f'_+ is continuous at c.*

d) *The set of points of I at which f fails to be differentiable is countable.*

e) *Let $c \in I$. A line ℓ passing through $(c, f(c))$ is a supporting line for f at c if and only if its slope m satisfies*

$$f'_-(c) \leq m \leq f'_+(c).$$

PROOF. a) Fix $c \in I$, and define

$$\varphi_c : I \setminus \{c\} \to \mathbb{R},\ \varphi(x) := \frac{f(x) - f(c)}{x - c}.$$

For notational simplicity, put

$$I_- = (-\infty, c) \cap I \text{ and } I_+ = (c, \infty) \cap I.$$

By the Three Secant Inequality, φ_c is increasing on I_- and is also increasing on I_+ and moreover every value φ_c takes on I_- is less than or equal to every value it takes on I_+. It follows that

$$f'_-(c) = \lim_{x\to c^-} \varphi(x) = \sup \varphi(I_-) \leq \inf \varphi(I_+) = \lim_{x\to c^+} \varphi(x) = f'_+(c).$$

Let $x_1 < x_2$ in I, and choose $z \in (x_1, x_2)$. Using part a) and the Three Secant Inequality, we get

$$f'_-(x_1) \leq f'_+(x_1) \leq \varphi_{x_1}(z) = \frac{f(z) - f(x_1)}{z - x_1}$$
$$\leq \frac{f(z) - f(x_2)}{z - x_2} = \varphi_{x_2}(z) \leq f'_-(x_2) \leq f'_+(x_2).$$

b) This is proved in Exercises 4.9 and 4.10.
c) Let $c \in I$. From (11) we deduce:

$$\lim_{x \to c^-} f'_+(x) \leq f'_-(c) \leq f'_+(c) \leq \lim_{x \to c^+} f'_-(x).$$

So: if f'_- is continuous at c, then $f'_-(c) \leq f'_+(c) \leq f'_-(c)$, so the left and right hand derivatives are equal at c and thus f is differentiable at c. Similarly, if f'_+ is continuous at c, then $f'_+(c) \leq f'_-(c) \leq f'_+(c)$; again f is differentiable at c.

It follows from (11) that

$$\lim_{x \to c^-} f'_-(x) = \lim_{x \to c^-} f'_+(x) \text{ and } \lim_{x \to c^+} f'_-(x) = \lim_{x \to c^+} f'_+(x).$$

If f'_- is *not* continuous at c, then since it is increasing and left-continuous, while f'_+ is increasing and right-continuous, we must have

$$f'_-(c) < \lim_{x \to c^+} f'_-(x) \leq \lim_{x \to c^+} f'_+(x) = f'_+(c).$$

Thus $f'_(c) < f'_+(c)$ and f is not differentiable at c. A similar argument shows that if f'_+ is not continuous at c then f is not differentiable at c.
d) By part c), we know that f fails to be differentiable precisely where the increasing function f'_- fails to be continuous; this latter set is countable by Exercise 3.30.
e) Let $c, x \in I$. If $x < c$, then $f'_-(c) \geq \varphi(x) = \frac{f(x)-f(c)}{x-c}$. Clearing denominators and using that $x - c < 0$, we get

$$\forall x < c,\ f(x) \geq f(c) + f'_-(c)(x-c). \tag{12}$$

Clearly this also holds, with equality, when $x = c$. If $x > c$, then $f'_+(c) \leq \varphi(x) = \frac{f(x)-f(c)}{x-c}$. Now $x - c > 0$, so clearing denominators gives

$$\forall x > c,\ f(x) \geq f(c) + f'_+(c)(x-c). \tag{13}$$

So if $f'_-(c) \leq m \leq f'_+(c)$, then combining (12) and (13) we get:

$$\forall x \in I,\ f(x) \geq f(c) + m(x-c),$$

so $\ell_{m,c}(x) := m(x-c) + f(c)$ is a supporting line for f at c. The fact that in order for $\ell_{m,c}(x)$ to be supporting line we must have $f'_-(c) \leq m \leq f'_+(c)$ follows easily from the definitions of the one-sided derivatives. We leave this as Exercise 4.11. $\square$

We want to mention two final results that we will *not* prove here. They are safely at the graduate level. First:

THEOREM 4.9 (Lebesgue). *Let I be an interval, and let $f : I \to \mathbb{R}$ be monotone. Then the set X of points at which f is not differentiable has measure zero.*

This is a remarkable theorem. We know that monotone functions are continuous except on a *countable* set of points: this is Exercise 3.30. We interpret this as saying that the set of discontinuities of a monotone function is "small" in a very strong sense. The set of points at which a monotone function fails to be differentiable must of course contain its set of discontinuities, but Lebesgue's Theorem says that this set is still "small," however in a different and weaker sense.

There are certainly sets of measure zero that are uncountable. A classic one is the **Cantor set**, which is obtained from $[0, 1]$ by repeatedly removing the open middle third of each line segment remaining. This is more of a hand-wave than a definition: we do not wish to enter into serious discussion of the Cantor set in our course; rather we will refer the reader to [**GT**, §2.15]. It turns out that there is also a **Cantor function** $f : [0, 1] \to \mathbb{R}$. This amazing function is continuous, increasing and fails to be differentiable precisely on points of the Cantor set. In particular, the set of points of nondifferentiability *is* uncountable.

Lebesgue's Theorem suggests investigating the twice differentiability of a convex function $f : I \to \mathbb{R}$. Indeed, if f is convex and differentiable, then by Theorem 4.4b) we know that f' is increasing, so by Lebesgue's Theorem $f'' = (f')'$ exists except on a set of measure zero.

Our definition of f'' is $(f')'$; with this definition, in order to ask whether f is twice differentiable at c it is not enough to know that $f'(c)$ exists; we need to know that f' exists in some interval $(c-\delta, c+\delta)$. As we have seen, for a convex function f this need not be the case. However, there are various ways to extend this definition, after which it turns out that a convex function is twice differentiable except on a set of measure zero. Such results go under the name **Alexandroff's Theorem**.

6. Jensen's Inequality

Theorem 4.10 (Jensen's Inequality). *] Let $f : I \to \mathbb{R}$ be continuous and convex. For $x_1, \ldots, x_n \in I$ and $\lambda_1, \ldots, \lambda_n \in [0, 1]$ with $\lambda_1 + \ldots + \lambda_n = 1$, we have*

$$f(\lambda_1 x_1 + \ldots + \lambda_n x_n) \le \lambda_1 f(x_1) + \ldots + \lambda_n f(x_n).$$

Proof. We go by induction on n, the case $n = 1$ being trivial. So suppose the result holds for $n \in \mathbb{Z}^+$, and let $x_1, \ldots, x_n, x_{n+1} \in I$ and $\lambda_1, \ldots, \lambda_n, \lambda_{n+1} \in [0, 1]$ with $\lambda_1 + \ldots + \lambda_{n+1} = 1$. If $\lambda_{n+1} = 0$ we are reduced to the case of n variables, which holds by induction. If $\lambda_{n+1} = 1$ then $\lambda_1 = \ldots = \lambda_n = 0$ and equality holds trivially. So the nontrivial case is $\lambda_{n+1} \in (0, 1)$. Now we write

$$\lambda_1 x_1 + \ldots + \lambda_{n+1} x_{n+1} = (1 - \lambda_{n+1}) \left(\frac{\lambda_1}{1 - \lambda_{n+1}} x_1 + \ldots + \frac{\lambda_n}{1 - \lambda_{n+1}} x_n \right) + \lambda_{n+1} x_{n+1},$$

and use the Secant Graph Inequality to get

$$f(\lambda_1 x_1 + \ldots + \lambda_{n+1} x_{n+1}) = f((1 - \lambda_{n+1}) \left(\frac{\lambda_1}{1 - \lambda_{n+1}} x_1 + \ldots + \frac{\lambda_n}{1 - \lambda_{n+1}} x_n \right) + \lambda_{n+1} x_{n+1})$$

$$\le (1 - \lambda_{n+1}) f \left(\frac{\lambda_1}{1 - \lambda_{n+1}} x_1 + \ldots + \frac{\lambda_n}{1 - \lambda_{n+1}} x_n \right) + \lambda_{n+1} f(x_{n+1}).$$

Since $\frac{\lambda_1}{1-\lambda_{n+1}}, \ldots, \frac{\lambda_n}{1-\lambda_{n+1}}$ are non-negative numbers that sum to 1, by induction the above expression is less than or equal to

$$(1 - \lambda_{n+1}) \left(\frac{\lambda_1}{1 - \lambda_{n+1}} f(x_1) + \ldots + \frac{\lambda_n}{1 - \lambda_{n+1}} f(x_n) \right) + \lambda_{n+1} f(x_{n+1})$$

$$= \lambda_1 f(x_1) + \ldots + \lambda_n f(x_n) + \lambda_{n+1} f(x_{n+1}). \qquad \square$$

Although there is nothing to the proof of Theorem 4.10 other than induction applied to the Secant Graph Inequality, nevertheless when we apply it to various specific convex functions f we get a remarkably rich supply of nontrivial inequalities. Indeed, mathematics is rife with "named inequalities," most of which have no apparent relation to each other and must be proved by various (sometimes *ad hoc*) techniques. Above I said "most": the exception to this is Jensen's Inequality, which provides a common source for so many individual inequalities, as we will now see.

Here is a quick tour of several named inequalities. A special case of the first is the Arithmetic Geometric Mean Inequality, which pops up now and then in undergraduate level mathematics and can be proved by induction albeit in a rather elaborate way [**AM**, Thm. 7.1]. The ones in the middle are probably unfamiiliar at the undergraduate level but are relevant to graduate level real analysis. Our tour ends with a visit to our oldest friend.

THEOREM 4.11 (Weighted Arithmetic Geometric Mean Inequality). *Let $x_1, \ldots, x_n \geq 0$, and let $\lambda_1, \ldots, \lambda_n \in [0,1]$ be such that $\lambda_1 + \ldots + \lambda_n = 1$.*

a) *We have:*

$$x_1^{\lambda_1} \cdots x_n^{\lambda_n} \leq \lambda_1 x_1 + \ldots + \lambda_n x_n. \tag{14}$$

b) ***(Arithmetic Geometric Mean (AGM) Inequality)*** *We have:*

$$(x_1 \cdots x_n)^{1/n} \leq \frac{x_1 + \ldots + x_n}{n}.$$

PROOF. a) We are going to apply Jensen's Inequality with $f(x) = e^x$, a function that is convex because for all $x \in \mathbb{R}$, $f''(x) = e^x > 0$. We may assume that $x_1, \ldots, x_n > 0$; otherwise, at least one is zero, so the left hand side of (14) is 0 and the right hand side is non-negative. For $1 \leq i \leq n$, put $y_i := \log x_i$. Then:

$$x_1^{\lambda_1} \cdots x_n^{\lambda_n} = e^{\log(x_1^{\lambda_1} \cdots x_n^{\lambda_n})} = e^{\lambda_1 y_1 + \ldots + \lambda_n y_n}$$
$$\leq \lambda_1 e^{y_1} + \ldots + \lambda_n e^{y_n} = \lambda_1 x_1 + \ldots + \lambda_n x_n.$$

b) This follows from part a) by taking $\lambda_1 = \cdots = \lambda_n = \frac{1}{n}$. $\square$

THEOREM 4.12 (Young's Inequality). *Let $x, y \geq 0$, and let $p, q > 1$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Then:*

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}. \tag{15}$$

PROOF. As for the last result, the inequality holds trivially if either x or y is 0, so we may assume $x, y > 0$. Now we apply Theorem 4.11a) with $n = 2$, $x_1 = x^p$, $x_2 = y^q$, $\lambda_1 = \frac{1}{p}$ and $\lambda_2 = \frac{1}{q}$. We get:

$$xy = (x^p)^{1/p}(y^q)^{1/q} = x_1^{\lambda_1} x_2^{\lambda_2} \leq \lambda_1 x_1 + \lambda_2 x_2 = \frac{x^p}{p} + \frac{y^q}{q}.$$ $\square$

THEOREM 4.13 (Hölder's Inequality). *Let $x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathbb{R}$, and let $p, q > 1$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Then:*

$$|x_1 y_1| + \ldots + |x_n y_n| \leq \left(|x_1|^p + \ldots + |x_n|^p\right)^{1/p} \left(|y_1|^q + \ldots + |y_n|^q\right)^{1/q}. \tag{16}$$

PROOF. The result is clear if either $x_1 = \ldots = x_n = 0$ or $y_1 = \ldots = y_n = 0$, so we assume neither is the case. For $1 \leq i \leq n$, we apply Young's Inequality with

$$x = \frac{|x_i|}{(|x_1|^p + \ldots + |x_n|^p)^{1/p}} \text{ and } y = \frac{|y_i|}{(|y_1|^q + \ldots + |y_n|^q)^{1/q}}$$

and sum the resulting inequalities from $i = 1$ to n. We get:

$$\frac{\sum_{i=1}^n |x_i y_i|}{(|x_1|^p + \ldots + |x_n|^p)^{1/p} (|y_1|^q + \ldots + |y_n|^q)^{1/q}} \leq \frac{1}{p} + \frac{1}{q} = 1.$$ $\square$

Theorem 4.14 (Minkowski's Inequality).
For $x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathbb{R}$ and $p \geq 1$, we have:
(17)
$$\left(|x_1+y_1|^p+\ldots+|x_n+y_n|^p\right)^{1/p} \leq \left(|x_1|^p+\ldots+|x_n|^p\right)^{1/p}+\left(|y_1|^p+\ldots+|y_n|^p\right)^{1/p}.$$

Proof. Once again we may assume that $x_1, \ldots, x_n$ are not all 0 and that $y_1, \ldots, y_n$ are not all 0. The case $p = 1$ is much less interesting than the rest: in that case the inequality reads:

$$|x_1 + y_1| + \ldots + |x_n + y_n| \leq |x_1| + |y_1| + \ldots |x_n| + |y_n|,$$

which is a consequence of the Triangle Inequality in $\mathbb{R}$. Now suppose that $p > 1$. Setting $q := \frac{1}{1-\frac{1}{p}}$ we get: $q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. Now:

$$|x_1 + y_1|^p + \ldots + |x_n + y_n|^p = |x_1 + y_1||x_1 + y_1|^{p-1} + \ldots + |x_n + y_n||x_n + y_n|^{p-1}$$
$$= |x_1||x_1+y_1|^{p-1}+\ldots+|x_n||x_n+y_n|^{p-1}+|y_1||x_1+y_1|^{p-1}+\ldots+|y_n||x_n+y_n|^{p-1}.$$

Applying Hölder's Inequality and using $(p-1)q = p$, we get that the last expression is at most

$$(|x_1|^p + \ldots + |x_n|^p)^{1/p} (|x_1 + y_1|^p + \ldots + |x_n + y_n|^p)^{1/q}$$
$$+ (|y_1|^p + \ldots + |y_n|^p)^{1/p} (|x_1 + y_1|^p + \ldots + |x_n + y_n|^p)^{1/q}$$
$$= \left((|x_1|^p+\ldots|x_n|^p)^{1/p}+(|y_1|^p+\ldots+|y_n|^p)^{1/p}\right)\cdot\left(|x_1+y_1|^p+\ldots+|x_n+y_n|^p\right)^{1/q}.$$

Dividing both sides by the second factor on the right hand side, we get

$$(|x_1 + y_1|^p + \ldots + |x_n + y_n|^p)^{1/p}$$
$$= (|x_1 + y_1|^p + \ldots + |x_n + y_n|^p)^{1-1/q} \leq (|x_1|^p + \ldots |x_n|^p)^{1/p}+(|y_1|^p + \ldots + |y_n|^p)^{1/p}$$
□.

If in Minkowski's Inequality we take $p = 2$ and write $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, then we get

$$||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||.$$

Thus Minkowski's Inequality is a parameterized family of inequalities that specializes to the Triangle Inequality in $\mathbb{R}^n$ when $p = 2$.

7. Exercises

Exercise 4.1. *Let I be an interval, and let $f : I \to \mathbb{R}$. Show: f is convex if and only if satisfies the **Grand Two Secant Inequality**: for all $a, b, c, d \in I$ with $a < b$, $c < d$ and $a < c$, we have $\frac{f(b)-f(a)}{b-a} \leq \frac{f(d)-f(c)}{d-c}$.*

Exercise 4.2.

a) *Let $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^2$. Show: for all $c \in \mathbb{R}$, the unique supporting line to f at c is the tangent line $\ell_c(x) = y - c^2 = 2c(x - c)$.*
b) *Let $g : \mathbb{R} \to \mathbb{R}$ by $g(x) = |x|$. Show:*
   (i) *The supporting lines to g at 0 are precisely $\ell(x) = mx$ for $m \in [-1, 1]$.*
   (ii) *If $c > 0$, the unique supporting line to g at c is $\ell(x) = x$.*
   (iii) *If $c < 0$, the unique supporting line to g at c is $\ell(x) = -x$.*

Exercise 4.3. *Let $f : [a, b] \to \mathbb{R}$ be convex.*

a) *Show that f attains a maximum value at $x = a$ or $x = b$.*
b) *Suppose moreover that f is nonconstant. Show: for all $c \in (a, b)$, we have $f(c) < \max(f(a), f(b))$.*

EXERCISE 4.4. *Let $f : [a, b] \to \mathbb{R}$ be convex.*

a) *Show: if f is* not *monotone, then there is a unique $c \in (a, b)$ such that f is decreasing on $[a, c]$ and increasing on $[c, b]$.*
b) *Deduce: f attains a minimum value.*

EXERCISE 4.5.

a) *Let $f : (a, b) \to \mathbb{R}$ be convex. Show: $\lim_{x\to a^+} f(x) \in (-\infty, \infty]$ and $\lim_{x\to b^-} f(x) \in (-\infty, \infty]$.*
b) *Let $f : [a, b] \to \mathbb{R}$ be convex. Show:*

$$\lim_{x\to a^+} f(x) \leq f(a) \text{ and } \lim_{x\to b^-} f(b).$$

In particular, both of these limits are finite.

EXERCISE 4.6. *Let $f : \mathbb{R} \to \mathbb{R}$ be a bounded convex function. Show: f is constant.*

EXERCISE 4.7. *Let I be an interval in R, and let $\{f_j : I \to \mathbb{R}\}_{j\in J}$ be a family of convex functions. Let $f : I \to \mathbb{R}$ be a function such that for all $x \in I$ we have*

$$f(x) = \sup_{j\in J} f_j(x).$$

Show: f is convex.

EXERCISE 4.8. *Let $f : [a, b] \to \mathbb{R}$ be convex and continuous. Prove the* ***Hermite-Hadamard inequality****:*

$$f\left(\frac{a+b}{2}\right) \leq \frac{\int_a^b f}{b-a} \leq \frac{f(a)+f(b)}{2}.$$

(Hint: let $l(x) = m(x - \frac{a+b}{2})$ be a supporting line to f at $\frac{a+b}{2}$, and let $L(x)$ be the secant line between a and b – i.e., the unique line passing through $(a, f(a))$ and $(b(, f(b))$. We know that

$$\forall x \in [a, b],\ l(x) \leq f(x) \leq L(x).$$

Write out $\int_a^b l \leq \int_a^b f \leq \int_a^b L$ and simplify.)

EXERCISE 4.9. *Let I be an open interval, and let $\{f_n : I \to \mathbb{R}\}$ be a sequence of functions. Suppose that all of the following hold:*

(i) *For all $n \in \mathbb{Z}^+$ and all $x \in I$ we have $f_n(x) \leq f_{n+1}(x)$.*
(ii) *Each f_n is continuous and increasing.*
(iii) *The sequence converges pointwise on I to a function g.*

a) *Show: g is left-continuous: for every point $c \in I$, $\lim_{x\to c^-} f(x) = f(c)$.*
b) *Give an example to show that g need not be right-continuous.*
c) *State an analogue of part a) for which the conclusion is that f is right-continuous at c.*

EXERCISE 4.10. *Let I be a open interval, and let $f : I \to \mathbb{R}$ be a convex function.*

a) *Let* $n \in \mathbb{Z}^+$, *and define* $g_n(x) = \frac{f(x-\frac{1}{n})-f(x)}{\frac{-1}{n}}$. *(The domain of* g_n *is the set of* $x \in I$ *such that* $x - \frac{1}{n} \in I$.*) Show:* g_n *is continuous and increasing.*
b) *Show: for all* n, *we have* $g_n \leq g_{n+1}$.
c) *Show: for all* $c \in I$, *we have* $\lim_{n\to\infty} g_n(c) = f'_-(c)$. *Deduce from Exercise 4.9 that* f'_- *is left-continuous.*
d) *Adapt the above argument to show that* f'_+ *is right-continuous.*

EXERCISE 4.11. *Let* I *be an open interval, let* $f : I \to \mathbb{R}$ *be a convex function, and let* $c \in I$. *Show directly from the definitions of* $f'_-(c)$ *and* $f'_+(c)$ *that if* ℓ *is a supporting line for* f *at* c *with slope* m, *then* $f'_-(c) \leq m \leq f'_+(c)$.

EXERCISE 4.12. *Let* I *be an interval. A function* $f : I \to \mathbb{R}$ *is* ***midpoint-convex*** *if for all* $x_1 < x_2$ *in* I *we have*

$$f\left(\frac{x_1+x_2}{2}\right) \leq \frac{f(x_1)+f(x_2)}{2}.$$

It is immediate that convex functions are midpoint-convex. Show: if f *is midpoint-convex and continuous, then* f *is convex.*[2]

EXERCISE 4.13. *Let* $f : \mathbb{R} \to \mathbb{R}$ *be increasing. Define* $F : \mathbb{R} \to \mathbb{R}$ *by* $F(x) := \int_0^x f$. *(It is a standard convention that when* $a > b$, *by* $\int_a^b f$ *we mean* $-\int_b^a f$.*) Note that by Exercise 3.17, this makes sense even if* f *is not continuous.*

a) *Show:* F *is convex. (Suggestion: by Theorem 3.1a),* F *is continuous. Show that* F *is midpoint-convex and apply Exercise 4.12.)*
b) *Show: for* $c \in \mathbb{R}$, *we have that* F *is differentiable at* c *if and only if* f *is continuous at* c.
c) *Show: for any countable subset* $X \subset \mathbb{R}$, *there is a convex function* $F : \mathbb{R} \to \mathbb{R}$ *that is differentiable at* c *if and only if* $c \in \mathbb{R} \setminus X$. *In particular, there is a convex function that is differentiable at every irrational number and fails to be differentiable at every rational number.*
(Suggestion: combine part b) with Exercise 3.18.)

[2]It turns out that there are functions that are midpoint-convex but not convex, but this nontrivial fact is more a fact of set theory than of analysis.

CHAPTER 5

Metric Spaces

1. A look ahead

In the last two chapters we sketched out some of the terrain of more advanced analysis, in which function theory and set theory interact in more subtle ways. In truth, graduate level real analysis is a very challenging course that relatively few students will take. So I want to end by previewing a different course: undergraduate general topology. This course gives a generalization and abstraction of most of the material from Chapter 1, to the context of metric spaces.

Let X be a set. A **metric function** is a function $d : X \times X \to \mathbb{R}$ such that all of the following hold:

(D1) (Positive Definiteness) For all $x, y \in X$, we have $d(x, y) \geq 0$, with equality if and only if $x = y$.
(D2) (Symmetry) For all $x, y \in X$, we have $d(x, y) = d(y, x)$.
(D3) (Triangle Inequality) For all $x, y, z \in X$, we have $d(x, z) \leq d(x, y) + d(y, z)$.

A **metric space** is a pair (X, d), where X is a set and $d : X \times X \to \mathbb{R}$ is a metric function. In our course, the shining example was to take $X = \mathbb{R}^N$ and d to be the Euclidean distance function: $d(\mathbf{x}, \mathbf{y}) := ||\mathbf{x} - \mathbf{y}||$. Many other examples come from this, since if (X, d) is a metric space and Y is any subset of X, then if $d_Y : Y \times Y \to \mathbb{R}$ is just the metric function restricted to $Y \times Y$, then d_Y is a metric function on Y, so (Y, d_Y) is again a metric space.

In Chapter 1 we studied:

- Convergence of sequences in $\mathbb{R}^N$.
- Continuity of functions $f : X \to \mathbb{R}^M$ where X is a subset of $\mathbb{R}^N$.
- Bounded, open and closed sets in $\mathbb{R}^N$.
- Sequential compactness of subsets of $\mathbb{R}^N$.

These concepts translate essentially verbatim to the context of a general metric space (X, d), and there is a useful general theory that parallels *much* of what we did in Euclidean spaces. However, in several ways, $\mathbb{R}^N$ and various subsets of it (especially, closed and bounded subsets) behave more nicely than an arbitrary metric space. When this occurs, it is important to think deeply about why: usually one can isolate a certain specific feature of $\mathbb{R}^N$ and use it to define classes of metric spaces in which these good things continue to happen.

Let us now give just a flavor of this.

We said that a sequence $\{x_n\}$ in $\mathbb{R}^N$ converges to a point p in $\mathbb{R}^N$ if the real sequence $d(x_n, p)$ converges to 0. This definition makes sense in any metric space, and the geometric intuitition is the same: all sufficiently large terms of the sequence should lie arbitrarily close to the limit. (It does *not* make sense to talk about divergence to ∞ without some extra structure.) In any metric space, a sequence can have at most one limit, and if a sequence converges to p than all of its subsequences converge to p. And again, it is interesting to explore to what extent we can get a divergent sequence to converge by passing to subsequences.

We can define open and closed balls in any metric space (X, d) in exactly the same way, for $x \in X$ and $\epsilon > 0$, we put

$$B^\circ(x, \epsilon) \coloneqq \{y \in X \mid d(x, y) < \epsilon\} \text{ and } B^\bullet(x, \epsilon) \coloneqq \{y \in X \mid d(x, y) \leq \epsilon\}.$$

They may not *look* like balls anymore – we will see an interesting example of this later – but if you think about it, the finer geometry of balls was never really used.[1]

Again we can define a subset U of a metric space (X, d) to be **open** if for every $x \in U$ there is $\epsilon > 0$ such that $B^\circ(x, \epsilon) \subseteq U$. Moreover we can define limit points of a subset Y in the same way: these are the limits of convergent sequences whose terms lie in Y. (We can also define accumulation points.) Then we can say that a subset Y is **closed** if it contains all of its limit points. Again it turns out that $Y \subseteq X$ is closed if and only if its complement $X \setminus Y$ is open...and the proof is really the same. We can also define boundedness: a subset $Y \subseteq X$ is bounded if it lies in some closed ball $B^\bullet(x, R)$. Equivalently, for a nonempty subset Y of a metric space X we can define its **diameter**

$$\operatorname{diam}(Y) \coloneqq \sup\{\ d(y_1, y_2) \mid y_1, y_2 \in Y\} \in [0, \infty]$$

and put $\operatorname{diam} \varnothing = 0$; then a subset Y is bounded if and only if it has finite diameter.

If (X, d_X) and (Y, d_Y) are two metric spaces and $f : X \to Y$ is a function between them, then all of the following definitions go through using the metric functions instead of Euclidean norms: continuous, uniformly continuous, Lipschitz. Just to spell out the first one: we say that $f : X \to Y$ is **continuous** at $c \in X$ if for all $\epsilon > 0$, there is $\delta > 0$ such that for all $x \in X$, if $d(x, c) < \delta$ then $d(f(x), f(c)) < \epsilon$. Once again continuous functions are characterized by preservation of limits of convergent sequences, we have a sequential characterization of uniform continuity, and so forth.

For a subset Y of a metric space (X, d), we say that Y is **sequentially compact** if every sequence $\{x_n\}$ in Y admits a subsequence converging to an element of Y. This is the same definition as before. Again, as before it is easy to prove that a sequentially compact subset must be closed (otherwise take a sequence converging to a limit point of Y that does not lie in Y) and bounded (otherwise we can build a sequence in Y for which any two distinct terms have distance at least one from each other, and then there is no convergent subsequence). But now a surprise occurs: for a subset Y of a general metric space X, being closed and bounded is not sufficient for sequential compactness.

[1] Perhaps the closest we came to this was showing that balls are convex. Convexity does *not* make sense in an arbitrary metric space. One needs the structure of a real vector space for this.

EXAMPLE 1.1. *For a set X, the **discrete metric** d_d on X is*

$$d(x,y) := \begin{cases} 0 & x = y \\ 1 & x \neq y \end{cases}.$$

For any $c \in X$ and any $0 < \epsilon < 1$, we have $B^\circ(c,\epsilon) = \{x\}$. Then (X, d_d) is bounded: if $X \neq \varnothing$, then $\operatorname{diam}(X) = 1$. *Also X is closed as a subset of itself.*

But: if a sequence $\{x_n\}$ in a metric space converges to a point c, then for all $\epsilon > 0$, we must have $x_n \in B^\circ(c,\epsilon)$ for all sufficiently large n. So a sequence in the discrete metric space (X, d_X) converges to c if and only if all sufficiently large terms are equal to c, i.e., is eventually constant. If X is moreover infinite, then there is an injective sequence $x_\bullet : \mathbb{Z}^+ \to X$. Every subsequence of $\{x_n\}$ remains injective and therefore divergent. Therefore X itself is closed, bounded but not sequentially compact. (Exactly the same holds for every *infinite subset Y of X.)*

If one looks back at the proof of the Bolzano-Weierstrass Theorem in $\mathbb{R}^N$, we get referred back to Bolzano-Weierstrass in $\mathbb{R}$ which was proved in a previous course using the completeness properties of $\mathbb{R}$. In a metric space X we do *not* have a notion of ordering of the points, so upper bounds and Dedekind completeness doesn't make sense. However, Cauchy sequences do: a sequence $\{x_n\}$ in a metric space (X, d) is **Cauchy** if for all $\epsilon > 0$ there is $N \in \mathbb{Z}^+$ such that for all $m, n \geq N$ we have $d(x_m, x_n) < \epsilon$. Again it is easy to see that convergent sequences are Cauchy but the converse does *not* generally hold: we say that a metric space is **complete** if every Cauchy sequence in that space is convergent.

For instance, there are Cauchy sequences in $\mathbb{Q}$ (which becomes a metric space by restricting the metric function on $\mathbb{R}$) that converge only to elements of $\mathbb{R}$ – e.g. a sequence of rational approximations to $\sqrt{2}$ – so $\mathbb{Q}$ is not a complete metric space. Actually, a little thought shows that this phenomenon is much more general: if (X, d) is any metric space whatsoever and Y is a subset of X that is *not* closed, then by definition there is a sequence $\{y_n\}$ in Y converging to an element $x \in X \setminus Y$; any subsequence still converges to x and therefore not to any element of Y. So incomplete metric spaces abound.

I want to end by telling you some striking and important theorems in metric topology to convince you that there is more to learn here. The first one introduces an alternate take on sequential compactness that is more prevalent in advanced mathematics. Namely, a subset Y of a metric space (X, d) is **compact** if for every family $\{U_i\}_{i \in I}$ of open subsets of X that covers Y in the sense that $Y \subseteq \bigcup_{i \in I} U_i$, there is a finite subset $J \subseteq I$ such that $\bigcup_{i \in J} U_i$ also covers Y. In brief: "every open cover of Y has a finite subcover." As we saw in our course, sequential definitions seem to be easier to understand and process than set-theoretic definitions, though both are important. Indeed, we could have made good use of compactness in our course but we were able to make do with sequential compactness instead.

One more definition: a subset Y of a metric space X is **totally bounded** if for every $\epsilon > 0$, it admits a finite cover by subsets of diameter at most ϵ. Equivalently, for every $\epsilon > 0$, Y admits a finite cover by closed ϵ-balls. Since sets of finite diameter are bounded and finite unions of bounded sets are bounded, certainly totally bounded implies bounded. The terminology is of course suggesting that totally bounded could be stronger. This is an absolutely key example of a difference between $\mathbb{R}^N$ and a general metric space: in $\mathbb{R}^N$ every bounded set is totally bounded (this was a homework problem). However, an infinite set with the discrete metric is bounded but not totally bounded: the only sets of diameter at most $\frac{1}{2}$ are single points, an an infinite set is not a finite union of singleton subsets!

Now we can state what is probably the most important theorem of metric topology:

THEOREM 5.1. *For a metric space (X, d), the following are equivalent:*

(i) *X is compact: every open cover of X has a finite subcover.*
(ii) *X is sequentially compact: every sequence in X has a convergent subsequence.*
(iii) *X is "accumulation point compact": every infinite subset of X has an accumulation point in X.*
(iv) *X is complete (Cauchy sequences converge) and totally bounded (for all $\epsilon > 0$, X can be covered by finitely many subsets of arbitrarily small diameter).*

PROOF. See e.g. [**GT**, Thm. 2.78]. (Warning: in those notes, where we say "limit point" they say "adherent point" and where we say "accumulation point" they say "limit point.") □

Theorem 5.1 is not really a generalization of the Bolzano-Weiestrass Theorem in $\mathbb{R}^N$, but it helps us to understand Bolzano-Weierstrass more deeply: it shows that the two key facts that go into it are the completeness of $\mathbb{R}$ (from which the completeness of $\mathbb{R}^N$ follows almost immediately) and the fact that bounded subsets in Euclidean space are totally bounded, which can be thought of as a consequence of the Archmedean property in the form that if you divide a real number by 2 enough times, it gets arbitrarily small.

If we have two metric spaces (X, d_X) and (Y, d_Y), a map $f : X \to Y$ is an **isometric embedding** if it preserves distances between points:

$$\forall x_1, x_2 \in X,\ d_Y(f(x_1), f(x_2)) = d_X(x_1, x_2).$$

Such maps are in particular Lipschitz with Lipschitz constant 1, so they are uniformly continuous, and so forth. But really this is much stronger: Lipschitz maps are maps that only stretch distances between points by a bounded factor, while isometric embeddings *preserve* distances. You should think of an isometric embedding $f : X \to Y$ as giving you a "perfect copy" $f(X)$ of X as a subset of Y.

Now here is another big theorem:

THEOREM 5.2. *Let (X, d_X) be a metric space. Then there is a metric space $(\tilde{X}, d_{\tilde{X}})$ and an isometric embedding*

$$\iota : X \to \tilde{X}$$

such that:

(i) *$\tilde{X}$ is a complete metric space, and*
(ii) *The image $\iota(X)$ is dense in $\tilde{X}$: that is, for every $x \in \tilde{X}$ there is a sequence $\{x_n\}$ in X such that $\iota(x_n) \to x$.*

The metric space $\tilde{X}$ is called the **completion** of the metric space X, and it can be thought of as "filling in the missing holes" that prevent Cauchy sequences in X from converging. Moreover, the completion $\tilde{X}$ is essentially *unique*, although I don't have the time to explain exactly what that means here. This is such a profound idea: you have a space in which not every Cauchy sequence converges, which robs you of and essential tool to show convergence of sequences. So you faithfully embed your space inside a larger space (in a parsimonious way: every point you have added is the limit of a sequence in your original space) and in that larger space all Cauchy sequences converge.

One thing that the formalism of metric spaces buys you is the idea to consider two different metric functions *on the same space*. This turns out to be very natural and useful, because indeed there is often more than one sense in which things can get "close together" and you want to compare the two. Let me end by mentioning an example of this: let

$$X = \mathcal{C}[a,b] = \{\text{continuous } f : [a,b] \to \mathbb{R}\}$$

be the set of continuous real-valued functions defined on $[a,b]$. We want to make this into a metric space, i.e., we want to measure the distance between two functions. How might we do this?

One way to do this was given in Math 3100: for $f, g \in \mathcal{C}[a,b]$, put

$$d_\infty(f,g) \coloneqq \max_{x\in[a,b]} |f(x) - g(x)|.$$

(It is easy to see that this is a metric: the triangle inequality follows from the usual triangle equality in $\mathbb{R}$.) Convergence of sequences in the d_∞-metric is precisely uniform convergence. However, there is another metric that is arguably even more natural: for $f, g \in \mathcal{C}[a,b]$, put

$$d_1(f,g) \coloneqq \int_a^b |f - g|.$$

I claim that d_1 is a metric function. This time the Triangle Inequality is not the hardest part: for $f, g, h \in \mathcal{C}[a,b]$ we have

$$d_1(f,h) = \int_a^b |f-h| \leq \int_a^b (|f-g| + |g-h|) = \int_a^b |f-g| + \int_a^b |g-h) = d_1(f,g) + d_1(g,h).$$

Because $|f-g| = |g-f|$, clearly $d_1(f,g) = d_1(g,f)$. Also clearly $f_1(f,g) \geq 0$, because the integral of a non-negative function is non-negative. Also clearly $d_1(f,f) = 0$. However, it takes some work to show that if if $d_1(f,g) = 0$ then $f = g$: this comes down to showing: if $f : [a,b] \to \mathbb{R}$ is continuous and non-negative, then $\int_a^b f = 0$ implies $f = 0$. (This is a good exercise! I recommend you try it.) Note that everthing we've said so far about d_1 holds for all Riemann integrable functions but here we need to assume f is continuous.

We have moreover that

$$d_1(f,g) = \int_a^b |f-g| \leq \int_a^b d_\infty(f,g) = (b-a)d_\infty(f,g),$$

so that the d_1-metric is, up to the constant $(b-a)$, the smaller of the two, and intuitively it measures the distance in a more refined way: whereas $d_\infty(f,g)$ measures the *maximum* distance between $f(x)$ and $g(x)$, $\frac{1}{b-a}d_1(f,g)$ measures the *average* distance between $f(x)$ and $g(x)$. In more advanced analysis *both* of these metrics are important, and they fit into an infinite family of metrics d_p for $p \in [1,\infty]$.

It turns out that $\mathcal{C}[a,b]$ with the d_∞-metric is complete: this is a variant of the Math 3100 fact that a uniform limit of continuous functions remains continuous. On the other hand, $\mathcal{C}[a,b]$ with the d_1-metric is *not* complete.

EXAMPLE 1.2. *For $n \in \mathbb{Z}^+$, let $f_n : [0,2]$ be the function*

$$f_n(x) = \begin{cases} x^n & \text{if } x \in [0,1) \\ 1 & \text{if } x \in [1,2] \end{cases}.$$

This sequence converges pointwise the to the function:

$$f : [0,2] \to \mathbb{R},\ f(x) = \begin{cases} 0 & \text{if } x \in [0,1) \\ 1 & \text{if } x \in [1,2] \end{cases},$$

which is discontinuous at 1*. Since f is bounded with a single discontinuity, it is Riemann integrable, and*

$$d_1(f_n,f) = \int_0^2 |f_n - f| = \int_0^1 |x^n| + \int_1^2 0 = \int_0^1 x^n = \frac{1}{n+1} \to 0.$$

This implies that the sequence $\{f_n\}$ is Cauchy in $\mathcal{C}[a,b]$ with the d_1-metric: indeed, for all $m,n \in \mathbb{Z}^+$, we have

$$d_1(f_m,f_n) \leq d_1(f_m,f) + d_1(f_n,f),$$

so if we choose $N \in \mathbb{Z}^+$ such that for all $n \geq N$ we have $d_1(f_n,f) < \frac{\epsilon}{2}$ then for all $m,n \geq N$ we have $d_1(f_m,f_n) < \epsilon$. If there were a continuous *function g such that $f_n \to g$ in the d_1-metric, then for all $n \in \mathbb{Z}^+$ we have*

$$d_1(f,g) \leq d_1(f,f_n) + d_2(f_n,g),$$

so

$$d_1(f,g) \leq \lim_{n\to\infty} d_1(f,f_n) + d_2(f_n,g) = 0.$$

Since f and g are both continuous on $[1,2]$*, we must have $g(x) = f(x) = 1$ for all $x \in [1,2]$. Similarly, for any $\delta \in (0,1)$, we have*

$$0 = \int_0^2 |f-g| = \int_0^\delta |f-g| + \int_\delta^2 |f-g|;$$

since both terms are non-negative, we conclude $\int_0^\delta |f-g| = 0$. Since f and g are continuous on $[0,\delta]$, we must have $g = f = 0$. Therefore we know that g is continuous on $[0,2]$, is equal to 0 for all $x \in [0,1)$ and is equal to 1 for all $x \in [1,2]$. But there is no such function, so $\{f_n\}$ is Cauchy in $(\mathcal{C}[a,b],d_1)$ but not convergent.

But all is not lost! By Theorem 5.2 one can consider the completion of $\mathcal{C}[a,b]$ with respect to the d_1-metric. This is called the **Lebesgue space** $L^1([a,b])$...and now we are back to real analysis. Indeed, these Lebesgue spaces are discussed in Math 8100 perhaps more than any other topic.

Bibliography

[AM] P.L. Clark, *Introduction to Advanced Mathematics.* `http://alpha.math.uga.edu/~pete/3200_supplemental.pdf`

[Cl22] P.L. Clark, *A Mean Value Inequality for all functions.* `http://alpha.math.uga.edu/~pete/MVT_Dini.pdf`

[Go16] R.A. Gordon, *A bounded derivative that is not Riemann integrable.* Math. Mag. 89 (2016), 364–370.

[GT] P.L. Clark, *General Topology.* `http://alpha.math.uga.edu/~pete/pointset.pdf`

[HC] P.L. Clark, *Honors Calculus* `http://alpha.math.uga.edu/~pete/2400full.pdf`

[No52] M.J. Norris, *Integrability of continuous functions.* Amer. Math. Monthly 59 (1952), 244–245.

[R] W. Rudin, *Principles of mathematical analysis.* Third edition. International Series in Pure and Applied Mathematics. McGraw-Hill, New York-Auckland-Düsseldorf, 1976.

[SS] M.R. Adams and P.L. Clark, *Sequences and series.*