MATH 4310 LECTURE NOTES

PETER HUMPHRIES

Contents

1. Real numbers

Recommended reading: [Pug15, §1.2].

1.1. **Review of the real numbers.** This is a short review of the real number system and a set of axioms that describes it. This section of the course will be brief; mostly it is intended to remind you of a few things you studied last year.

The axioms for the real numbers come in various different flavours. First there are the algebraic axioms.

We assume that there are real numbers $0 \in \mathbb{R}$ and $1 \in \mathbb{R}$ with $1 \neq 0$ such that for any real numbers a, b, and c,

(A1) $a + b = b + a$
(A2) $(a + b) + c = a + (b + c)$
(A3) $a + 0 = 0 + a = a$
(A4) For every $a \in \mathbb{R}$, there is exactly one real number, denoted by $-a$, such that $a + (-a) = (-a) + a = 0$

The four axioms for addition say that the real numbers are a *group* under the operation of addition.

The axioms for multiplication say the same for the *nonzero* real numbers:

(A5) $a \times b = b \times a$
(A6) $(a \times b) \times c = a \times (b \times c)$
(A7) $a \times 1 = 1 \times a = a$
(A8) If $a \neq 0$, there is exactly one real number, denoted by a^{-1}, such that $a \times a^{-1} = a^{-1} \times a = 1$

The last algebraic axiom is the distributive law, which links the two operations of addition and multiplication:

(A9) If a, b, and c are real numbers, then $a \times (b + c) = (a \times b) + (a \times c)$

Axioms (A1)—(A9) form the definition of a *field*.

Next, there are the order axioms. We assume that there is a subset P of the set of real numbers, called the set of *positive numbers*, such that:

(A10) For any real number a, exactly one of the following holds: $a = 0$ or $a \in P$ or $-a \in P$
(A11) If $a \in P$ and $b \in P$, then $a + b \in P$ and $ab \in P$

A number a is called *negative* if $-a$ is positive.

We now *define* $a < b$ to mean $b - a \in P$. Something satisfying the axioms (A1)—(A11) is an *ordered field*.

Finally, there is the completeness axiom. This will play an important role in what is to come. Recall that if $A \subseteq \mathbb{R}$ is a set of real numbers, we say that A is *bounded above* if there exists a real number $b \in \mathbb{R}$ such that $a \leq b$ for all $a \in A$. The number b is then called an upper bound for A. A *least upper bound* or *supremum* for a set A is an upper bound that is less than or equal to every upper bound of A. We define a lower bound and a *greatest lower bound* or *infimum* analogously. The completeness axiom is phrased in terms of these:

(A12) Suppose that S is a nonempty set of real numbers that is bounded above. Then S has a supremum.

These twelve axioms determine the real numbers *uniquely*. This is the contents of the following theorem, which we state but do not prove.

Theorem 1.1. *Up to isomorphism, there is precisely one ordered field that obeys the completeness axiom.*

Here by up to isomorphism, we mean the following. Suppose that we have two different sets $\mathbb{R}_a$ and $\mathbb{R}_b$ that satisfy axioms (A1)–(A12). These sets may consist of different elements, so we write 0_a for the zero element in $\mathbb{R}_a$ and 0_b for the zero element in $\mathbb{R}_b$. Similarly, we write $1_a, +_a, \times_a, <_a$ and $1_b, +_b, \times_b, <_b$. Then an *isomorphism* of ordered fields between $\mathbb{R}_a$ and $\mathbb{R}_b$ is a function $f : \mathbb{R}_a \to \mathbb{R}_b$ that satisfies the following:

- $f(0_a) = 0_b$,
- $f(1_a) = 1_b$,
- $f(x +_a y) = f(x) +_b f(y)$ for all $x, y \in \mathbb{R}_a$,
- $f(x \times_a y) = f(x) \times_b f(y)$ for all $x, y \in \mathbb{R}_a$,
- if $x <_a y$, then $f(x) <_b f(y)$,
- for each $x_b \in \mathbb{R}_b$, there exists $x_a \in \mathbb{R}_a$ such that $f(x_a) = f(x_b)$,
- if $f(x) = f(y)$, then $x = y$.

Thus Theorem 1.1 states that if two ordered fields $\mathbb{R}_a$ and $\mathbb{R}_b$ satisfy axioms (A1)—(A12), then there exists an isomorphism between $\mathbb{R}_a$ and $\mathbb{R}_b$.

1.2. **Natural numbers, integers, rational numbers, and complex numbers.** Since $1 \in \mathbb{R}$ and $\mathbb{R}$ is closed under addition, $\mathbb{R}$ contains the *natural numbers* (or *positive integers*)

$$\mathbb{N} \coloneqq \{1, 2, 3, \dots, \},$$

which can be described as the subset of $\mathbb{R}$ containing 1 and satisfying the condition that if $n \in \mathbb{N}$, then also $n + 1 \in \mathbb{N}$. From Axiom (A4), $\mathbb{R}$ contains additive inverses (i.e. real numbers multiplied by -1) and so $\mathbb{R}$ contains the *integers*

$$\mathbb{Z} \coloneqq \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}.$$

From Axiom (A8), $\mathbb{R}$ contains multiplicative inverses and so $\mathbb{R}$ contains the *rational numbers*

$$\mathbb{Q} \coloneqq \left\{ \frac{a}{b} : a \in \mathbb{Z},\ b \in \mathbb{Z} \setminus \{0\} \right\}.$$

In fact, the rational numbers $\mathbb{Q}$ satisfy Axioms (A1)—(A11). However, they do not satisfy Axiom (A12), the completeness axiom. Finally, the *complex numbers*

$$\mathbb{C} \coloneqq \{x + iy : x, y \in \mathbb{R},\ i^2 = -1\}$$

contains the real numbers as a subset. The complex numbers are a field that satisfies the completeness axiom; however, they fail the order axioms (A10) and (A11), since there is no ordering on the complex numbers that divides the nonzero complex numbers into "positive" and "negative" numbers such that "positive" numbers are closed under addition and multiplication.

1.3. **Completeness axiom and limits of sequences.** Properties of the real numbers involving limits almost always involve the completeness axiom. Let me just remind you of one such result.

Theorem 1.2. *Suppose that $(a_n) = a_1, a_2, a_3, \dots$ is a bounded sequence of real numbers that is nondecreasing, so that $a_n \geq a_m$ whenever $n \geq m$. Then the sequence converges to a real number a.*

Proof. Let a be the least upper bound of the subset $A \coloneqq \{a_1, a_2, a_3, \dots\}$ of $\mathbb{R}$. I claim that the sequence (a_n) converges to a. To prove this, we need to show that for every $\varepsilon > 0$ there exists a positive integer N such that $|a_n - a| < \varepsilon$ for all $n \geq N$.

Since a is an upper bound for A, we have for free that $a_n \leq a$ for every $n \in \mathbb{N}$. It remains to show only that $a_n \geq a - \varepsilon$ for all $n \geq N$.

To prove the other one, we use the fact that a is the *least* upper bound. Thus $a - \varepsilon$ is not an upper bound for A. In other words, there is an element $a_N \in A$ such that

$a_N > a - \varepsilon$. Since the sequence (a_n) is nondecreasing, this implies that $a_n > a - \varepsilon$ for all $n \geq N$, and the proof is complete. □

Let me make the observation that the completeness axiom is crucial here. The set of rational numbers $\mathbb{Q}$ is a great example of an ordered field that does not satisfy the completeness axiom, and indeed the result above fails in the setting of rational numbers.

Example 1.3. Consider the sequence (a_n) defined by

$$a_n = 1 + \sum_{k=1}^{n} \frac{1}{k!} = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots + \frac{1}{n!}.$$

This is the sequence of partial sums at $x = 1$ for the Taylor expansion of e^x about $x = 0$. The series has positive terms, so the sequence is nondecreasing. The sequence is also bounded via the ratio test. The limit (as a sequence of real numbers, guaranteed to exist by the theorem above) is e, which is certainly irrational. Therefore there can be no limit of the sequence in $\mathbb{Q}$.

1.4. **The rationals and the reals.** While the rationals fail to satisfy the completeness axiom, they are nonetheless very close to the real numbers in the sense that every real number is arbitrarily close to a rational number.

Theorem 1.4.

(1) *If $x, y \in \mathbb{R}$ with $x > 0$, then there exists a positive integer $n \in \mathbb{N}$ such that $nx > y$.*
(2) *If $x, y \in \mathbb{R}$ with $x < y$, then there exists a rational number $p \in \mathbb{Q}$ such that $x < p < y$.*

Part (1) is the *archimedean property* of $\mathbb{R}$; it has equivalent reformulations, such as there is no real number $x \in \mathbb{R}$ for which $x \geq n$ for every natural number $n \in \mathbb{N}$, or equivalently for every $\varepsilon > 0$, there exists a natural number $n \in \mathbb{N}$ such that $0 < \frac{1}{n} < \varepsilon$.

Part (2) states that the rational numbers $\mathbb{Q}$ are *dense* in the real numbers $\mathbb{R}$: there is a rational number between any two real numbers. So one can think of the rationals as having holes where sequences of rational numbers approach an irrational number, and the real numbers as filling in those wholes via the completeness axiom.

Proof.
(1) Let $A \coloneqq \{nx : n \in \mathbb{N}\}$. Suppose in order to obtain a contradiction that (1) is false. Then y is an upper bound of A. By the completeness axiom, A must then have a *least* upper bound in $\mathbb{R}$, which we call a. Since $x > 0$, we must have that $a - x < a$, and so $a - x$ is not an upper bound of A. Hence there exists some positive integer m such that $a - x < mx$. But then $a < (m + 1)x$, which is in A; however, this contradicts the fact that a is an upper bound of A.
(2) Since $x < y$, we have $y - x > 0$, and so by (1) (with x replaced by $y - x$ and y replaced by 1), there exists a positive integer $n \in \mathbb{N}$ such that $n(y - x) > 1$. We utilise (1) twice more (once with x replaced by 1 and y replaced by nx; once with x replaced by 1 and y replaced by $-nx$) to deduce the existence of positive integers $m_1, m_2 \in \mathbb{N}$ such that $m_1 > nx$ and $m_2 > -nx$. Then $-m_2 < nx < m_1$, and so there must exist an integer $m \in \mathbb{Z}$ (with $-m_2 \leq m \leq m_1$) satisfying $m - 1 \leq nx < m$. By combining these inequalities, we obtain $nx < m \leq 1 + nx < ny$. Since $n > 0$, it follows that $x < \frac{m}{n} < y$. This proves (2) with $p = \frac{m}{n}$. □

2. Countable and uncountable sets

Recommended reading: [Pug15, §1.4–1.5].

2.1. **Cardinality.** We want to develop a way of comparing the sizes of sets. In the case of finite sets this is obvious: two sets are the same size if you can pair them off. We will follow Cantor in extending this idea also to infinite sets. First, we need to recall what it means for a function to be injective, surjective, or bijective.

Definition 2.1. Let A and B be sets and $f : A \to B$ be a function.

- The function f is said to be an *injection* or *injective* or *one-to-one* if $f(x) = f(y)$ implies that $x = y$.
- The function f is said to be a *surjection* or *surjective* or *onto* if for all $b \in B$, there exists some $a \in A$ such that $f(a) = b$.
- The function f is said to be a *bijection* or *bijective* if it is both injective and surjective.

Next, we want to use injective functions to compare the size of sets.

Definition 2.2. Let A and B be sets. We say that A has cardinality no larger than that of B, and write $A \lesssim B$, if there exists an injection from A to B. If there exists a bijection from A to B, then we say that A and B have equal cardinality and write $A \sim B$.

Two sets having equal cardinality defines an equivalence relation.

Definition 2.3. An *equivalence relation* $\sim$ on a set S is a relation satisfying

- $x \sim x$ for all $x \in S$ (reflexivity);
- If $x \sim y$ and $y \sim z$, then $x \sim z$ (transitivity);
- If $x \sim y$ then $y \sim x$ (symmetry).

An *equivalence class* of S with respect to the equivalence relation $\sim$ is a set of the form

$$[x] \coloneqq \{y \in S : y \sim x\}$$

for some $x \in S$, in which case x is said to be a *representative* of the equivalence class $[x]$.

Lemma 2.4. *The relation $A \sim B$ when there exists a bijection from A to B defines an equivalence relation on the set of all sets.*

Proof. Reflexivity is clear by taking $f : A \to A$ to be the identity map. Transitivity is clear, since the composition of two bijective functions is also bijective. Finally, if $f : A \to B$ is a bijection, then the inverse map $f^{-1} : B \to A$ is well-defined and is a bijection, which implies symmetry. □

There is an alternate description of two sets having equal cardinalities: there exist injective maps from A to B and from B to A, so that $A \lesssim B$ and $B \lesssim A$.

Theorem 2.5 (Cantor–Schröder–Bernstein theorem)**.** *Let A and B be sets such that there exists injections from A to B and from B to A. Then there exists a bijection from A to B.*

Proof. Let $f : A \to B$ and $g : B \to A$ be injections. We set $A_0 \coloneqq A$, $B_0 \coloneqq B$, and define inductively $A_j \coloneqq g(B_{j-1})$ and $B_j \coloneqq f(A_{j-1})$. Since f and g are injections, we have that for every $j \in \mathbb{N}$,

$$A \coloneqq A_0 \sim B_1 \sim A_2 \sim \cdots \sim B_{2j-1} \sim A_{2j}, \qquad B \coloneqq B_0 \sim A_1 \sim B_2 \sim \cdots \sim A_{2j-1} \sim B_{2j}.$$

Moreover, $A \coloneqq A_0 \supseteq A_1 \supseteq \cdots \supseteq A_j$ and $B \coloneqq B_0 \supseteq B_1 \supseteq \cdots \supseteq B_j$ for every $j \in \mathbb{N}$. If we have an *equality* in any of these inclusions, so that either $A_{j-1} = A_j$ or $B_{j-1} = B_j$ for some $j \in \mathbb{N}$, then we deduce that $A \sim B$. It remains to deal with the case where these inclusions are all strict. In this case, we define the nonempty sets $A_j^* \coloneqq A_{j-1} \setminus A_j$ and $B_j^* \coloneqq B_{j-1} \setminus B_j$ for each $j \in \mathbb{N}$. Again by the fact that f and g are injections, we have that

$$A_1^* \sim B_2^* \sim \cdots \sim B_{2j}^* \sim A_{2j+1}^*, \qquad B_1^* \sim A_2^* \sim \cdots \sim A_{2j}^* \sim B_{2j+1}^*.$$

Moreover, since the sets $\{A_j^*\}$ are pairwise disjoint, as are $\{B_j^*\}$, we must have that $A_{2j-1}^* \cup A_{2j}^* \sim B_{2j-1}^* \cup B_{2j}^*$ for each $j \in \mathbb{N}$. Setting $A^* \coloneqq \bigcup_{j=1}^{\infty} A_j^*$ and $B^* \coloneqq \bigcup_{j=1}^{\infty} B_j^*$, we see that $A^* \sim B^*$. Finally, we note that A^* is a subset of A and its complement in A is $\bigcap_{j=0}^{\infty} A_j$, and similarly the complement of B^* in B is $\bigcap_{j=0}^{\infty} B_j$. We have that $f(\bigcap_{j=0}^{\infty} A_j) = \bigcap_{j=0}^{\infty} B_j$ and $g(\bigcap_{j=0}^{\infty} B_j) = \bigcap_{j=0}^{\infty} A_j$, so that $\bigcap_{j=0}^{\infty} A_j \sim \bigcap_{j=0}^{\infty} B_j$. Since $A = A^* \cup \bigcap_{j=0}^{\infty} A_j$ and $B = B^* \cup \bigcap_{j=0}^{\infty} B_j$, we conclude that $A \sim B$. □

Closely related to this is the following.

Lemma 2.6. *Let A and B be sets such that there exists an injection from A to B. Then there exists a surjection from B to A.*

Proof. Suppose that $f : A \to B$ is an injection, so that $f(x) = f(y)$ implies that $x = y$. We define $g : B \to A$ by fixing $x_0 \in X$ and letting

$$g(z) \coloneqq \begin{cases} x & \text{if there exists some } x \in A \text{ such that } f(x) = z, \\ x_0 & \text{otherwise.} \end{cases}$$

This is well-defined, since the fact that f is injective implies that given $z \in B$, there exists at most one $x \in A$ such that $f(x) = z$. Moreover, this is surjective as f is injective. □

2.2. **Countable sets.** An equivalence class of sets with regards to cardinality is called a *cardinal*. We are interested in comparing the sizes of sets, namely comparing *cardinalities*. A basic example is a *finite* set.

Definition 2.7. A set A is said to be *finite* if it is the empty set with no elements or if there exists a positive integer $n \in \mathbb{N}$ such that there is a bijection from A to the set $\{1, 2, 3, \ldots, n\}$, in which case we say that A has cardinality n.

Lemma 2.8. *The natural numbers $\mathbb{N}$ are not finite.*

Proof. For each positive integer $n \in \mathbb{N}$, there is no injective map $f : \mathbb{N} \to \{1, 2, 3, \ldots, n\}$, since at least two of $f(1), f(2), f(3), \ldots, f(n), f(n+1)$ must be equal. □

Remark 2.9. This proof uses the *pigeonhole principle*: if we have n elements distributed across m sets and $n > m$, then at least one set has more than one element in it.

Definition 2.10. A set A is said to be *countably infinite* (or *denumerable*) if its cardinality is the same as that of the natural numbers $\mathbb{N}$. A set A is said to be *countable* if it is either finite or countably infinite. A set A is said to be *uncountable* (or *uncountably infinite*) if it is not countable.

Note that if A is countable, then there is a surjection from $\mathbb{N}$ to A and an injection from A to $\mathbb{N}$.

Example 2.11. The even positive integers $2\mathbb{N}$ are countably infinite, since there is a bijective map from $\mathbb{N}$ to $2\mathbb{N}$ given by $n \mapsto 2n$ for each $n \in \mathbb{N}$.

Example 2.12. The integers $\mathbb{Z}$ are countably infinite, since there is a bijective map from $\mathbb{N}$ to $\mathbb{Z}$ given by $1 \mapsto 0$, $2n \mapsto n$, and $2n + 1 \mapsto -n$ for each $n \geq 2$.

Example 2.13. The product $\mathbb{N} \times \mathbb{N} = \{(n, m) : n, m \in \mathbb{N}\}$ is countably infinite, since there is a bijective map from $\mathbb{N}$ to $\mathbb{N} \times \mathbb{N}$ given by

$$1 \mapsto (1,1),\ 2 \mapsto (2,1),\ 3 \mapsto (1,2),\ 4 \mapsto (3,1),\ 5 \mapsto (2,2),\ 6 \mapsto (1,3),\ 7 \mapsto (4,1), \ldots.$$

Lemma 2.14. *The countable union of countable sets is countable.*

Proof. Suppose that we have countable sets $A_1, A_2, \ldots$. Then there exist surjections $a_n : \mathbb{N} \to A_n$, so that $a_n(m)$ is an element of A_n for each positive integer $m \in \mathbb{N}$. Then we have a surjection from $\mathbb{N} \times \mathbb{N}$ to the countable union of the countable sets A_n, $\bigcup_{n=1}^{\infty} A_n$, given by $(n, m) \mapsto a_n(m)$. The result now follows since $\mathbb{N} \times \mathbb{N}$ is countably infinite. □

Example 2.15. The rationals $\mathbb{Q}$ are countably infinite. Indeed, it is the countable union of the sets $\mathbb{Q} \cap (n, n+1]$ for each integer $n \in \mathbb{Z}$: $\mathbb{Q} = \bigcup_{n \in \mathbb{Z}} \mathbb{Q} \cap (n, n+1]$. Finally,

$$\mathbb{Q} \cap (n, n+1] = \{n+1\} \cup \bigcup_{\ell=2}^{\infty} \left\{ n + \frac{m}{\ell} : m \in \{1, \ldots, \ell - 1\} \right\},$$

which is the union of countably many finite sets, and hence is countable.

2.3. **Uncountable sets.** Now we move on to power sets.

Definition 2.16. Given a set A, the *power set* of A, denoted by 2^A, is the set of all subsets of A:

$$2^A \coloneqq \{S : S \subseteq A\}.$$

Example 2.17. The power set of $A = \{1, 2\}$ is

$$2^A = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\},$$

where $\emptyset$ denotes the empty set.

Theorem 2.18 (Cantor's theorem)**.** *Given a set A (finite or infinite), the cardinality of 2^A is strictly larger than A.*

Proof. We must show that there is an injective map from A to 2^A but that there is no bijection. An injective map is given by $x \mapsto \{x\}$ for each $x \in A$. Now suppose in order to obtain a contradiction that a bijection $f : A \to 2^A$ exists. Consider the set

$$S \coloneqq \{x \in A : x \notin f(x)\}.$$

This is well-defined since $f(x)$ is an element of 2^A and so is a subset of A. Since S is a subset of A, it must be an element of 2^A, and since f is a bijection, there must exist some $y \in A$ such that $f(y) = S$. Either $y \in S$ or $y \notin S$. If the former is true, then by definition $y \notin f(y) = S$, which is a contradiction. If the latter is true, then $y \notin f(y) = S$, which is again a contradiction. □

Corollary 2.19. *There is no largest cardinal.*

Corollary 2.20. *The power set of a countable set is uncountable.*

In particular, the power set $2^{\mathbb{N}}$ of the natural numbers $\mathbb{N}$ is *not* countable! Remarkably, this extends to the real numbers.

Theorem 2.21. *The real numbers $\mathbb{R}$ are uncountable.*

Proof. We show that there is an injection from $2^{\mathbb{N}}$ to $\mathbb{R}$. We define a map $f : 2^{\mathbb{N}} \to \mathbb{R}$ by the formula

$$f(S) \coloneqq \sum_{n \in S} 10^{-n}$$

for each subset S of $\mathbb{N}$. Since $\sum_{n=1}^{\infty} 10^{-n}$ is an absolutely convergent series, the series $\sum_{n \in S} 10^{-n}$ is also absolutely convergent, so $f(S)$ is well-defined. To show that f is not injective, we show that if $S_1, S_2 \subseteq \mathbb{N}$ are not equal, then $f(S_1) \neq f(S_2)$. Since these subsets are distinct, the union of complements $(S_1 \setminus S_2) \cup (S_2 \setminus S_1)$ is a nonempty subset of $\mathbb{N}$, which necessarily contains a minimal positive integer n_0. We suppose first that n_0

lies in $S_1 \setminus S_2$, so that $n_0 \in S_1$ but $n_0 \notin S_2$. By the definition of n_0, this means that if $n < n_0$, then either n is both S_1 and S_2 or it is in neither. So we have that

$$\begin{aligned} f(S_1) - f(S_2) &= \sum_{n \in S_1} 10^{-n} - \sum_{n \in S_2} 10^{-n} \\ &= \left(\sum_{\substack{n \in S_1 \\ n < n_0}} 10^{-n} + 10^{-n_0} + \sum_{\substack{n \in S_1 \\ n > n_0}} 10^{-n} \right) - \left(\sum_{\substack{n \in S_2 \\ n < n_0}} 10^{-n} + \sum_{\substack{n \in S_2 \\ n > n_0}} 10^{-n} \right) \\ &= 10^{-n_0} + \sum_{\substack{n \in S_1 \\ n > n_0}} 10^{-n} - \sum_{\substack{n \in S_2 \\ n > n_0}} 10^{-n} \\ &\geq 10^{-n_0} - \sum_{n > n_0} 10^{-n} \\ &= 10^{-n_0} - \frac{1}{9} 10^{-n_0} \\ &> 0. \end{aligned}$$

Similarly, if n_0 lies in $S_2 \setminus S_1$, then $f(S_2) - f(S_1) > 0$. In either case, $f(S_1) \neq f(S_2)$. □

With more work, one can show that there is an injection from $\mathbb{R}$ to $2^{\mathbb{N}}$, so that these sets have the same cardinality. Now that we know that the real numbers are uncountable, we can show that some subsets of the reals are also uncountable.

Example 2.22. The set $(0,1) = \{x \in \mathbb{R} : 0 < x < 1\}$ has the same cardinality as $\mathbb{R}$, since there is a bijection $f : \mathbb{R} \to (0,1)$ given by

$$f(t) = \frac{t}{2\sqrt{1+t^2}} + \frac{1}{2}.$$

3. Metrics and metric spaces

Recommended reading: [Pug15, §2.1], [Tao16, §1.1].

3.1. **Distance functions and metric spaces.** Given two real numbers $x, y \in \mathbb{R}$, the difference $|x - y|$ measures the distance between x and y. We can think of $|x - y|$ as being a function, the *distance function*, from $\mathbb{R} \times \mathbb{R}$ to $[0, \infty)$: it takes two real numbers as inputs and outputs a nonnegative real number that is the distance between the two inputs.

Distance functions can exist on sets other than the real numbers.

Definition 3.1. Let X be a set. A *distance function* on X is a function $d : X \times X \to \mathbb{R}$ satisfying the following three properties:

(1) For any $x, y \in X$, $d(x,y) \geq 0$ and $d(x,y) = 0$ if and only if $x = y$ (positivity);
(2) For any $x, y \in X$, $d(x,y) = d(y,x)$ (symmetry);
(3) For any $x, y, z \in X$, $d(x,z) \leq d(x,y) + d(y,z)$.

Example 3.2. Let $X = \mathbb{R}$. Then $d(x,y) \coloneqq |x - y|$ is a distance function on $\mathbb{R}$.

Example 3.3. Let X be *any* nonempty set. Then

$$d(x,y) \coloneqq \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{otherwise} \end{cases}$$

is a distance function on X.

Now that we have introduced the notion of a distance function, we can define a metric space, which is the key object of this course.

Definition 3.4. A metric space (X, d) is a space X of objects (called *points*) together with a *distance function* $d : X \times X \to [0, \infty)$. We call d the *metric* on X.

When the metric d is clear from context, we write X instead of (X, d).

Example 3.5. The real numbers are a metric space with $X = \mathbb{R}$ and metric $d(x, y) \coloneqq |x - y|$, which we call the *standard metric* on $\mathbb{R}$.

Example 3.6. Let X be a nonempty set, and let

$$d(x, y) \coloneqq \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{otherwise.} \end{cases}$$

Then (X, d) is a metric space. The metric d is called the *discrete metric* on X.

Example 3.7. Let $n \in \mathbb{N}$ be a positive integer and let

$$X = \mathbb{R}^n \coloneqq \{(x_1, \ldots, x_n) : x_1, \ldots, x_n \in \mathbb{R}\}$$

be n-dimensional Euclidean space. This is a metric space with the *Euclidean metric* (or *ℓ^2-metric*)

$$d_2((x_1, \ldots, x_n), (y_1, \ldots, y_n)) \coloneqq \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}.$$

This metric is simply the length of the straight line between the two points $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, which is calculated via Pythagoras' theorem.

Example 3.8. Let $X = \mathbb{R}^n$, but now take the *taxi-cab metric* (or *ℓ^1-metric*)

$$d_1((x_1, \ldots, x_n), (y_1, \ldots, y_n)) \coloneqq |x_1 - y_1| + \cdots + |x_n - y_n|.$$

For $n = 2$, this measures the distance that a taxi would travel between two locations in a city with a grid road layout: only travelling east-west or north-south but never diagonally.

Example 3.9. Let $X = \mathbb{R}^n$ with d the *sup-norm metric* (or *ℓ^∞-metric*)

$$d_\infty((x_1, \ldots, x_n), (y_1, \ldots, y_n)) \coloneqq \sup\{|x_1 - y_1|, \ldots, |x_n - y_n|\}.$$

Example 3.10. Let $X = C([0, 1]) \coloneqq \{f : f \text{ is a continuous function from } [0, 1] \text{ to } \mathbb{R}\}$; in this case, a point in X is a continuous function $f : [0, 1] \to \mathbb{R}$. One choice of metric is the *L^1-metric*

$$d_1(f, g) \coloneqq \int_0^1 |f(t) - g(t)| \, dt.$$

Another is the *sup-norm metric* (or *L^∞-metric*)

$$d_\infty(f, g) \coloneqq \sup_{t \in [0,1]} |f(t) - g(t)|.$$

Metric spaces whose points are functions, such as $C([0, 1])$, are of great prominence later in the course.

3.2. **Generalising concepts from $\mathbb{R}$ to metric spaces.** Much of what this course deals with is understanding how to generalise analytic concepts from the setting of $\mathbb{R}$ (or more generally $\mathbb{R}^n$) to the setting of metric spaces. We define below some notions for $\mathbb{R}$ that require some effort to define for metric spaces.

Definition 3.11. A bounded interval in $\mathbb{R}$ is *open* if it does not contain its endpoints, namely an interval of the form

$$(a, b) \coloneqq \{x \in \mathbb{R} : a < x < b\}.$$

A set of points $A \subseteq \mathbb{R}$ is *open* if for every point $a \in A$, there exists some $\delta > 0$ such that the bounded open interval $(a - \delta, a + \delta)$ centred at a is wholly contained in A: $(a - \delta, a + \delta) \subseteq A$.

Definition 3.12. A point $a \in \mathbb{R}$ is a *limit point* of a set $A \subseteq \mathbb{R}$ if every open interval $(a - \delta, a + \delta)$ centred at a intersects A at a point other than a. A set A is *closed* if it contains all of its limit points.

Definition 3.13. Let (a_n) be a sequence of real numbers. Then the sequence (a_n) *converges* to a real number $a \in \mathbb{R}$, which we denote by $\lim_{n\to\infty} a_n = a$, if for all $\varepsilon > 0$, there exists a positive integer $N = N(\varepsilon) \in \mathbb{N}$ such that for all $n \geq N$, $|a_n - a| < \varepsilon$.

Definition 3.14. A set $A \subseteq \mathbb{R}$ is *compact* if every sequence of points in A has a convergent subsequence with limit in A.

Definition 3.15. Given a set $A \subseteq \mathbb{R}$, a function $f : A \to \mathbb{R}$ is *continuous at a point* $a \in A$ if for all $\varepsilon > 0$, there exists some $\delta = \delta(\varepsilon) > 0$ such that if $|x - a| < \delta$ and $x \in A$, then $|f(x) - f(a)| < \varepsilon$. A function $f : A \to \mathbb{R}$ is *continuous* if it is continuous at *every* point $a \in A$.

Question 3.16. *How does one define each of these concepts in the setting of metric spaces?*

4. Open and closed sets

Recommended reading: [Pug15, §2.3], [Tao16, §1.2].

4.1. **Open and closed sets.** We will begin by focusing on open and closed sets. These are the building blocks of understanding the *topology* of a metric space. Our first step is to generalise the notion of an open interval in $\mathbb{R}$ to the setting of metric spaces; we then use this to define open sets.

Definition 4.1. An *open metric ball* (or simply *open ball* or even just *ball*) in a metric space (X, d) is a set of the form

$$B_r(x) \coloneqq \{y \in X : d(x, y) < r\}$$

for some $x \in X$ and $r > 0$. We call $B_r(x)$ the open ball centred at x of radius r.

Sometimes the notation $B(x, r)$ (or, confusingly, even $B(r, x)$) is used instead of $B_r(x)$.

Example 4.2. For $(X, d) = (\mathbb{R}, d)$, the real numbers with the standard metric, the open ball $B_r(x)$ is simply the open interval $(x - r, x + r)$.

Example 4.3. For (X, d) a metric space with d the discrete metric, we have that

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

Note in particular that the set of points in an open ball depends on the metric: if $X = \mathbb{R}$, then the open ball $B_r(x)$ with respect to the standard metric is never the same as the open ball $B_r(x)$ with respect to the discrete metric.

We use open balls to define what it means for a subset of a metric space to be open, just as we used bounded open intervals to define what it means for a subset of $\mathbb{R}$ to be open.

Definition 4.4. A subset E of a metric space X is said to be *open* if for all $x \in E$, there exists some $r > 0$ such that $B_r(x) \subseteq E$.

That is, E is open if each point in E is surrounded by an open ball that is also in E.

Lemma 4.5. *Open balls are open.*

Proof. Let y be a point in $B_r(x)$, so that $d(x, y) < r$. Let $R \coloneqq r - d(x, y) > 0$. We claim that $B_R(y) \subseteq B_r(x)$. Indeed, if $z \in B_R(y)$, then by the triangle inequality,

$$d(x, z) \leq d(x, y) + d(y, z) < d(x, y) + R = r,$$

which implies that $z \in B_r(x)$. □

The definition of a closed set then builds upon the definition of an open set.

Definition 4.6. A subset E of a metric space X is said to be *closed* if it is the complement of an open set, so that $\{x \in X : x \notin E\}$ is open.

We write E^c or $X \setminus E$ for the complement of E. Note that this definition implies that a set E is open if and only if its complement E^c is closed.

Sets are not doors! A subset $E \subseteq X$ can be open, closed, both, or neither.

Example 4.7. Let (X, d) be $(\mathbb{R}, d)$ with d the standard metric.

- The set $E = (0, 1)$ is open but not closed.
- The set $E = [0, 1]$ is closed but not open.
- The set $E = (0, 1]$ is neither open nor closed.
- Each of the sets $E = \mathbb{R}$ and $E = \emptyset$ is both open and closed.

It is also important to bear in mind that a set being open or closed is dependent not just on the space X but also on the metric d.

Example 4.8. Let $(X, d) = (\mathbb{R}, d)$ with d the standard metric. Then $\{0\}$ is closed but not open. On the other hand, if $(X, d) = (\mathbb{R}, d)$ with d the discrete metric, then $\{0\}$ is both open and closed.

There are various operations one can apply to open and closed sets.

Lemma 4.9.

(1) *Let $\{E_\alpha\}_{\alpha \in I}$ be a collection of open sets, where the indexing set I may be uncountable. Then the union $\bigcup_{\alpha \in I} E_\alpha$ is open.*
(2) *Let $\{F_\alpha\}_{\alpha \in I}$ be a collection of closed sets, where the indexing set I may be uncountable. Then the intersection $\bigcap_{\alpha \in I} F_\alpha$ is closed.*
(3) *Let $\{E_1, \ldots, E_n\}$ be a finite collection of open sets. Then the intersection $\bigcap_{m=1}^n E_m$ is open.*
(4) *Let $\{F_1, \ldots, F_n\}$ be a finite collection of closed sets. Then the union $\bigcup_{m=1}^n F_m$ is closed.*

To prove (2) and (4), we must first recall De Morgan's laws for complements of unions or intersects of sets: the complement of the union $\bigcup_{\alpha \in I} X_\alpha$ of subsets $X_\alpha \subseteq X$ is

$$\left(\bigcup_{\alpha \in I} X_\alpha \right)^c = \bigcap_{\alpha \in I} X_\alpha^c;$$

similarly, the complement of the intersection $\bigcap_{\alpha \in I} X_\alpha$ is

$$\left(\bigcap_{\alpha \in I} X_\alpha \right)^c = \bigcup_{\alpha \in I} X_\alpha^c.$$

Proof.
(1) Let $E \coloneqq \bigcup_{\alpha\in I} E_\alpha$. If $x \in E$, then $x \in E_\alpha$ for some $\alpha \in I$. Since E_α is open, there exists some $r > 0$ such that $B_r(x) \subseteq E_\alpha$; as $E_\alpha \subseteq E$, we deduce that $B_r(x) \subseteq E$.
(2) By De Morgan's laws, the complement of $\bigcap_{\alpha\in I} F_\alpha$ is $\bigcup_{\alpha\in I} F_\alpha^c$, which is the union of open sets and hence open.
(3) Let $E \coloneqq \bigcap_{m=1}^n E_m$. If $x \in E$, then $x \in E_m$, so there exists $r_m > 0$ such that $B_{r_m}(x) \subseteq E_m$. Let $r \coloneqq \min\{r_1, \ldots, r_n\}$. Then $B_r(x) \subseteq B_{r_m}(x) \subseteq E_m$. Since this is true for each m, it follows that $B_r(x) \subseteq E$.
(4) By De Morgan's laws, the complement of $\bigcup_{m=1}^n F_m$ is $\bigcap_{m=1}^n F_m^c$, which is the intersection of a finite collection of open sets and hence open. □

It is important to note that the intersection of a collection of open sets need not be open if the indexing set I is not finite; similarly, the union of a collection of closed sets need not be closed if the indexing set I is not finite. The proofs above fail in these settings since the minimum $\min_{\alpha\in I} r_\alpha$ of an *infinite* set of radii need not exist, and the infimum of such a set of radii exists but may be zero.

Example 4.10. Let $(X, d) = (\mathbb{R}, d)$ with d the standard metric. For each $n \in \mathbb{N}$, let $E_n \coloneqq (-\frac{1}{n}, 1 + \frac{1}{n})$, which is open. Then

$$\bigcap_{n=1}^{\infty} E_n = [0, 1],$$

which is not open. Similarly, for each $n \geq 2$, let $F_n \coloneqq [\frac{1}{n}, 1 - \frac{1}{n}]$, which is closed. Then

$$\bigcup_{n=2}^{\infty} F_n = (0, 1),$$

which is not closed.

4.2. **Interior, exterior, boundary, and closure.** We can use open balls to classify points of a subset E of a metric space X.

Definition 4.11. Let $E \subseteq X$ be a subset of a metric space X and let $x \in X$ be a point in X. We say that x is

- an *interior point of* E if there exists some $r > 0$ such that the open ball $B_r(x)$ is wholly contained in E, so that the intersection $B_r(x) \cap E$ of $B_r(x)$ and E is simply $B_r(x)$;
- an *exterior point of* E if there exists some $r > 0$ such that the open ball $B_r(x)$ is wholly contained in the complement $X \setminus E \coloneqq \{y \in X : y \notin E\}$, so that the intersection $B_r(x) \cap E$ is the empty set $\emptyset$;
- a *boundary point of* E if it is neither an interior point of E nor an exterior point of E, so that for every $r > 0$, the intersection $B_r(x) \cap E$ is nonempty but is not equal to $B_r(x)$.

Note that interior points of E lie in E while exterior points of E do not lie in E. Boundary points of E may or may not lie in E.

Definition 4.12. Let $E \subseteq X$ be a subset of a metric space X.

- The *interior of* E, denoted by $\mathrm{Int}(E)$ or E°, is the set of all interior points of E.
- The *exterior of* E, denoted by $\mathrm{Ext}(E)$, is the set of all exterior points of E.
- The *boundary of* E, denoted by ∂E, is the set of all boundary points of E.

Each of these three subsets of X is mutually disjoint, and their union is all of X: $\mathrm{Int}(E) \cup \mathrm{Ext}(E) \cup \partial E = X$. Necessarily, we have that $\mathrm{Int}(E) \subseteq E$ and $\mathrm{Ext}(E) \subseteq E^c$; moreover, we have that $\mathrm{Int}(E^c) = \mathrm{Ext}(E)$ and $\mathrm{Ext}(E^c) = \mathrm{Int}(E)$, and these are all open

sets. On the other hand, the boundary ∂E may intersect nontrivially with E or it may not.

Example 4.13. For $(X, d) = (\mathbb{R}, d)$, the real numbers with the standard metric, the interior of $E = (0, 1]$ is $\text{Int}(E) = (0, 1)$, the exterior is $\text{Ext}(E) = (-\infty, 0) \cup (1, \infty)$, and the boundary is $\partial E = \{0\} \cup \{1\}$. On the other hand, the interior of $E = \mathbb{Q}$ is $\emptyset$, as is the exterior, and the boundary is $\mathbb{R}$.

Finally, we can construct a closed set starting with any set E.

Definition 4.14. The *closure* of a subset E of a metric space X is $\overline{E} \coloneqq \text{Int}(E) \cup \partial E$.

This is a closed set, since it is the complement of the open set $\text{Ext}(E) = \text{Int}(E^c)$, and it contains E, since $E \subseteq \text{Int}(E) \cup \partial E$. As we now show, $\overline{E}$ is the *smallest* closed set containing E.

Lemma 4.15. *Let $E \subseteq X$ be a subset of a metric space X. The interior of E is the largest open set contained in E and is also the union of all open subsets of E, while the closure of E is the smallest closed set containing E and is also the intersection of all closed sets containing E.*

Proof. If $A \subseteq E$ is open, then $A \subseteq \text{Int}(E)$, since for each $x \in A$, there exists $r > 0$ such that $B_r(x) \subseteq A \subseteq E$, and so x is an interior point of E. Thus there is no open subset of E containing $\text{Int}(E)$ except for $\text{Int}(E)$ itself. The union of all such open sets $A \subseteq E$ is also open and is contained in $\text{Int}(E)$, since each set is contained in $\text{Int}(E)$; on the other hand, $\text{Int}(E)$ is such a set itself, and so this union contains $\text{Int}(E)$, which means that this union is *equal* to $\text{Int}(E)$.

Similarly, if $B \supseteq E$ is closed, then $B \supseteq \overline{E}$, since the complement of B is open and contained in E^c, which means it is contained in $\text{Int}(E^c) = \text{Ext}(E)$, and so B contains $\text{Ext}(E)^c = \text{Int}(E) \cup \partial E = \overline{E}$. Thus there is no closed set containing E and contained in $\overline{E}$ except for $\overline{E}$ itself. The intersection of all such closed sets is also closed and contains $\overline{E}$, since each set contains $\overline{E}$; on the other hand, $\overline{E}$ is such a set itself, so this intersection is contained in $\overline{E}$, which means that this intersection is *equal* to $\overline{E}$. □

Finally, we can use the interior, exterior, and boundary of E to clarify whether it is open or closed.

Lemma 4.16. *Let $E \subseteq X$ be a subset of a metric space X. Then E is open if and only if $E = \text{Int}(E)$, while E is closed if and only if $E = \overline{E}$.*

Note that $\text{Int}(E)$ is always contained in E, whereas ∂E need not be; thus a set E is open if and only if it contains none of its boundary ∂E, whereas E is closed if and only if it contains all of its boundary ∂E.

Proof. The first claim is clear from the definition of a set being open. For the second, $\overline{E} = \text{Int}(E) \cup \partial E$ is closed since its complement is $\text{Ext}(E) = \text{Int}(E^c)$, which is open. Conversely, if E is closed, then its complement E^c is open, and hence is equal to $\text{Int}(E^c)$, which is precisely $\text{Ext}(E)$, and so $E = \text{Int}(E) \cup \partial E$. □

Closures may behave differently in some metric spaces than they do for $X = \mathbb{R}$ and d the standard metric.

Lemma 4.17. *Let (X, d) be a metric space, and let $B_r(x) = \{y \in X : d(x, y) < r\}$ be an open ball in X. Then*

(1) $\text{Int}(B_r(x)) = B_r(x)$;
(2) $\text{Ext}(B_r(x)) \supseteq \{y \in X : d(x, y) > r\}$;

(3) $\partial B_r(x) \subseteq \{y \in X : d(x,y) = r\}$;
(4) $\overline{B_r(x)} \subseteq \{y \in X : d(x,y) \leq r\}$.
Equality need not hold for (2), (3), or (4).

Equality *does* hold for (2), (3), and (4) when $X = \mathbb{R}$ and d is the standard metric.

Proof. For counterexamples to equality, we take $X = \{0,1\}$ with $d : X \times X \to [0,\infty)$ the discrete metric, so that $d(0,0) = d(1,1) = 0$ and $d(0,1) = d(1,0) = 1$, and let $x = 0$ and $r = 1$. Then $\text{Int}(B_1(0)) = B_1(0) = \{0\}$. Moreover, $\text{Ext}(B_1(0)) = \{1\}$ whereas $\{y \in X : d(0,y) > 1\} = \emptyset$, $\partial B_1(0) = \emptyset$ whereas $\{y \in X : d(0,y) = 1\} = \{1\}$, and $\overline{B_1(0)} = \{0\}$ whereas $\{y \in X : d(0,y) \leq 1\} = \{0,1\}$.

Next, we note that (1) holds since every point in $B_r(x)$ is an interior point. The containment in (2) holds since if $y \in X$ is such that $R := d(x,y) > r$, then $B_{R-r}(y) \subseteq B_r(x)^c$ by the triangle inequality, and so $y \in \text{Ext}(B_r(x))$. The inclusion in (3) holds from (1) and (2), since $\partial B_r(x) = X \setminus (\text{Int}(B_r(x)) \cup \text{Ext}(B_r(x)))$. Finally, the inclusion in (4) holds from (1), (2), and (3) since $\overline{B_r(x)} = \text{Int}(B_r(x)) \cup \partial B_r(x)$. □

Another way to view closed sets is in terms of limits points or adherent points.

Definition 4.18. Let $E \subseteq X$ be a subset of a metric space X with metric d and let $x \in X$ be a point in X. We say that x is

- a *limit point* of E if for every $r > 0$, $B_r(x) \cap E \setminus \{x\} \neq \emptyset$, so that $B_r(x)$ intersects E at a point other than x;
- an *isolated point* of E if there exists some $r > 0$ such that $B_r(x) \cap E = \{x\}$, so that $B_r(x)$ only intersects E at x;
- an *adherent point* of E if it is either a limit point or an isolated point, so that for every $r > 0$, $B_r(x) \cap E \neq \emptyset$.

Lemma 4.19. *Let $E \subseteq X$ be a subset of a metric space X. Then the closure $\overline{E}$ is the set of all adherent points of E. In particular, E is closed if and only if it contains all of its adherent points.*

Proof. A point $x \in X$ is an adherent point of E if and only if it is not an exterior point of E. Since the complement of $\text{Ext}(E)$ is $\text{Int}(E) \cup \partial E = \overline{E}$, the result follows. □

One can therefore think of taking the closure $\overline{E}$ of a set E as including all of the missing limit points of E.

5. Metric subspaces and relative topology

Recommended reading: [Pug15, §2.3], [Tao16, §1.3].

Open and closed sets in a metric space X depend on the metric d: a set $E \subseteq X$ may be open with respect to one metric d_1 on X but not be open with respect to another metric d_2 on X. Moreover, as we now discuss, the *ambient space* X also determines whether a set is open or closed. This can be seen when creating a new metric space from an old one.

Proposition 5.1. *Let (X,d) be a metric space and let $Y \subseteq X$ be a nonempty subset of X. Define $d|_{Y\times Y} : Y \times Y \to [0,\infty)$ by $d|_{Y\times Y}(x,y) := d(x,y)$ for each $x, y \in Y$. Then $(Y, d|_{Y\times Y})$ is a metric space.*

Proof. This follows immediately from the fact that (X,d) is a metric space. □

We call $(Y, d|_{Y\times Y})$ a *metric subspace* of (X,d).

Example 5.2. Let $X = \mathbb{R}^2$ with $d_2((x_1, x_2), (y_1, y_2)) := \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ the Euclidean metric (or ℓ^2-metric) on X. Inside the plane $\mathbb{R}^2$ is the horizontal axis $Y = \{(x_1, 0) \in \mathbb{R}^2\}$. The metric $d|_{Y\times Y}$ is $d|_{Y\times Y}((x_1, 0), (y_1, 0)) = \sqrt{(x_1 - y_1)^2} = |x_1 - y_1|$.

Note that this restricted metric coincides with the standard metric on $\mathbb{R}$ if we identify $Y = \{(x_1, 0) \in \mathbb{R}^2\}$ with $\mathbb{R}$ by mapping $(x_1, 0) \in Y$ to $x_1 \in \mathbb{R}$.

Question 5.3. *How does one define open and closed sets in metric subspaces?*

This question is answered by the concept of the *relative topology* of a metric subspace.

Definition 5.4. Let (X, d) be a metric space and let $(Y, d|_{Y\times Y})$ be a metric subspace. A subset $E \subseteq Y$ of Y (and hence also a subset of X) is said to be *relatively open with respect to* Y if it is open in the metric subspace $(Y, d|_{Y\times Y})$, while E is said to be *relatively closed with respect to* Y if it is closed in the metric subspace $(Y, d|_{Y\times Y})$.

Proposition 5.5. *Let (X, d) be a metric space and let $(Y, d|_{Y\times Y})$ be a metric subspace. Given $x \in Y$, the open ball $B_r^Y(x)$ in $(Y, d|_{Y\times Y})$ is given by $B_r^Y(x) = B_r(x) \cap Y$.*

Proof. We have that $y \in B_r^Y(x)$ if and only if $y \in Y$ and $d|_{Y\times Y}(x, y) < r$. Since $d|_{Y\times Y}(x, y) = d(x, y)$, this means that $y \in B_r^Y(x)$ if and only if $y \in Y$ and $y \in B_r(x)$. □

Now we can describe a simple criterion for sets in metric subspaces to be open or closed.

Proposition 5.6. *Let (X, d) be a metric space and let $(Y, d|_{Y\times Y})$ be a metric subspace. A subset $A \subseteq Y$ is open in Y if and only if there exists an open set $E \subseteq X$ such that $A = E \cap Y$. Similarly, a subset $A \subseteq Y$ is closed in Y if and only if there exists a closed set $F \subseteq X$ such that $A = F \cap Y$.*

Proof. If E is open in X, then for all $x \in E$, there exists some $r > 0$ such that $B_r(x) \subseteq E$. In particular, for $x \in E \cap Y$, we have that $B_r^Y(x) = B_r(x) \cap Y \subseteq E \cap Y$, and so $E \cap Y$ is open in Y.

Conversely, if A is open in Y, then for each $x \in A$, there exists $r = r_x > 0$ such that $B_{r_x}^Y(x) \subseteq Y$. Define $E := \bigcup_{x\in A} B_{r_x}(x)$. This is a union of open sets, hence open, and it satisfies $E \cap Y = \bigcup_{x\in A} B_{r_x}(x) \cap Y = \bigcup_{x\in A} B_{r_x}^Y(x)$. Clearly this is contained in A, since $B_{r_x}^Y(x) \subseteq A$ for each $x \in A$; on the other hand, this contains A, since $A = \bigcup_{x\in A}\{x\}$ and $x \in B_{r_x}^Y(x)$. Thus this is equal to A.

The result for closed sets now follows by taking complements. □

Some caution is required in working with metric subspaces: it is not necessarily true that an open subset of a metric subspace $(Y, d|_{Y\times Y})$ is open in the larger space (X, d); nor is a closed subset of Y necessarily closed in X.

Example 5.7. Let $X = \mathbb{R}$ and d be the standard metric. Let $Y = (0, 2]$. Then $(0, 1] = Y \cap [0, 1]$ is closed in Y but not closed in X, while $(1, 2] = Y \cap (1, 3)$ is open in Y but not open in X. Finally, $(0, 2] = Y \cap \mathbb{R}$ is both open and closed in Y but neither open nor closed in X.

Later, we will see another way to create new metric spaces from old by constructing the *product* of two metric spaces. An example to keep in mind is $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$.

6. Sequences and convergence

Recommended reading: [Pug15, §2.1, 2.3], [Tao16, §1.4].

6.1. **Sequences and subsequences.** We now move on to the notion of a sequence in a metric space.

Definition 6.1. A *sequence* (x_n) in a metric space (X, d) is a function from $\mathbb{N}$ to X.

So for each $n \in \mathbb{N}$, the n-th element of a sequence (x_n) is the element x_n of X. When $X = \mathbb{R}$, this is just a sequence of real numbers. We also recall the definition of a subsequence.

Definition 6.2. A *subsequence* of a sequence (x_n) in a metric space (X, d) is a sequence of the form (α_n) in X for which there exists a strictly increasing function $j : \mathbb{N} \to \mathbb{N}$ such that $\alpha_n = x_{j(n)}$ for each $n \in \mathbb{N}$.

Informally, a subsequence $(x_{j(n)})$ of a sequence (x_n) arises by omitting some of the terms in the original sequence.

Example 6.3. Let $X = \mathbb{R}$. The sequence $1, 0, 1, 0, 1, 0, \ldots$ contains both $1, 1, 1, 1, \ldots$ and $0, 0, 0, 0, \ldots$ as subsequences.

6.2. **Convergent sequences.** We recall that a sequence (x_n) of real numbers converges to a limit $x \in \mathbb{R}$ if for all $\varepsilon > 0$, there exists some $N = N(\varepsilon) \in \mathbb{N}$ such that $|x_n - x| < \varepsilon$ whenever $n \geq N$. If we view $\mathbb{R}$ as a metric space with the standard metric d, then this means that (x_n) converges to x if and only if the sequence of nonnegative real numbers $d(x_n, x)$ converges to the real number 0. This motivates the definition of convergence in an *arbitrary* metric space.

Definition 6.4. Let (X, d) be a metric space and let (x_n) be a sequence in X. We say that (x_n) *converges* to a point $x \in X$, which we call the *limit* of (x_n), and write $\lim_{n\to\infty} x_n = x$ or $x_n \to x$, if for every $\varepsilon > 0$, there exists a positive integer $N = N(\varepsilon) \in \mathbb{N}$ such that $d(x_n, x) < \varepsilon$ for all $n \geq N(\varepsilon)$.

Equivalently, the sequence (x_n) converges to a limit x in the metric space (X, d) if and only if the sequence of nonnegative real numbers $(d(x_n, x))$ converges to the real number 0 in the metric space $\mathbb{R}$ with the standard metric.

Many of the properties of convergent sequences in metric spaces are the same as those of convergent sequences of real numbers. For example, the limit of a convergent sequence is unique, and any subsequence of a convergent sequence is also convergent with the same limit. On the other hand, a sequence can have a convergent subsequence, yet the sequence itself need not be convergent.

Example 6.5. Let $X = \mathbb{R}$. The sequence $1, 0, 1, 0, 1, 0, \ldots$ in $\mathbb{R}$ does not converge to a limit, yet the two subsequences $1, 1, 1, 1, \ldots$ and $0, 0, 0, 0, \ldots$ both converge, and their limits are distinct.

Definition 6.6. Let (x_n) be a sequence in a metric space (X, d), and let $x \in X$. We say that x is a *limit point* of (x_n) if for every $\varepsilon > 0$ and every $N \geq 1$, there exists some $n \geq N$ such that $d(x_n, x) < \varepsilon$. That is, we do not require that $d(x_n, x) < \varepsilon$ *for all* $n \geq N$, but rather that *there exists* $n \geq N$ for which $d(x_n, x) < \varepsilon$.

Informally, this means that there exist infinitely many elements in the sequence (x_n) that are close to x. Of course, there might well exist infinitely many elements in this sequence that are not close to x!

Proposition 6.7. *A point $x \in X$ is a limit point of a sequence (x_n) in a metric space (X, d) if and only if there exists a subsequence $(x_{j(n)})$ of (x_n) that converges to x.*

Proof. Suppose first that x is a limit point of (x_n). Since x is a limit point, for every $\varepsilon > 0$ and every $N \geq 1$, there exists some $n \geq N$ such that $d(x_n, x) < \varepsilon$. We take $\varepsilon = \frac{1}{2}$ and $N = 1$, so that there exists some $j(1) \in \mathbb{N}$ be such that $d(x_{j(1)}, x) < \frac{1}{2}$. We then take $\varepsilon = \frac{1}{4}$ and $N = j(1)+1$, so that there exists some $j(2) \geq j(1)+1$ such that $d(x_{j(2)}, x) < \frac{1}{4}$. Repeating this process, we obtain a subsequence for which $d(x_{j(n)}, x) < 2^{-n}$, and hence converges to x since for all $\varepsilon > 0$, there exists $M = M(\varepsilon) \in \mathbb{N}$ such that $d(x_{j(n)}, x) < \varepsilon$ whenever $n \geq M(\varepsilon)$.

Conversely, if there exists a subsequence $(x_{j(n)})$ of (x_n) that converges to x, then for all $\varepsilon > 0$, there exists $M = M(\varepsilon) \in \mathbb{N}$ such that $d(x_{j(n)}, x) < \varepsilon$ whenever $n \geq M(\varepsilon)$. In particular, for all $\varepsilon > 0$ and $N \in \mathbb{N}$, there exists $j(n) \geq N$ such that $d(x_{j(n)}, x) < \varepsilon$, and so x is a limit point of (x_n). □

There is a close connection between the closure of a set in a metric space and limit points of sequences.

Proposition 6.8. *Let (X, d) be a metric space and let $E \subseteq X$ be a subset of X. Then $x \in \overline{E}$ if and only if there exists a sequence (x_n) in E such that x_n converges to x.*

Proof. Suppose that there exists a sequence (x_n) in E that converges to $x \in X$. Then for every $\varepsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$ such that $d(x_n, x) < \varepsilon$ whenever $n \geq N$. In particular, $x_{N(\varepsilon)} \in B_\varepsilon(x) \cap E$, and so $B_r(x) \cap E$ is nonempty for each $r > 0$. Thus x is an adherent point of E, and so $x \in \overline{E}$.

Conversely, suppose that $x \in \overline{E}$. Then x is an adherent point of E, and so for all $r > 0$, we have that $B_r(x) \cap E \neq \emptyset$. Choose a sequence (x_n) in E such that $x_n \in B_{2^{-n}}(x) \cap E$ for each $n \in \mathbb{N}$. Then (x_n) converges to x since if $2^{-n} < \varepsilon$, then $x_n \in B_\varepsilon(x)$, so that $d(x_n, x) < \varepsilon$. □

This result can be interpreted in the following two ways.

Corollary 6.9. *Let (X, d) be a metric space and let $E \subseteq X$ be a subset of X. A point $x \in X$ is an adherent point of E if and only if there exists a sequence (x_n) in E such that x_n converges to x. In particular, E is closed if and only if for each sequence (x_n) in E that is convergent in X, the limit x of (x_n) is in E.*

6.3. **Cauchy sequences.** An important notion closely related to convergence of sequences is a Cauchy sequence.

Definition 6.10. A sequence (x_n) in a metric space (X, d) is called a *Cauchy sequence* if for every $\varepsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$ such that $d(x_n, x_m) < \varepsilon$ whenever $n \geq N$ and $m \geq N$.

This looks a lot like the definition of convergence to a limit, except that instead of saying that the terms of the sequence get closer to the limit value, we are saying that the terms get closer to each other.

Theorem 6.11. *A convergent sequence is a Cauchy sequence.*

Proof. Suppose (x_n) is a convergent sequence with limit x. By definition, for every $\varepsilon > 0$ there exists $N(\varepsilon) \in N$ such that $d(x_n, a) < \varepsilon$ whenever $n \geq N(\varepsilon)$. But then if n and m are greater than $N(\frac{\varepsilon}{2})$, we see that by the triangle inequality,

$$d(x_n, x_m) \leq d(x_n, x) + d(x, x_m) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$ □

We are more interested in the converse: is it true that every Cauchy sequence is convergent? For $X = \mathbb{R}$, this is indeed the case, which is a consequence of the following two facts.

Proposition 6.12. *A Cauchy sequence with a convergent subsequence is convergent.*

Proof. Let (x_n) be a Cauchy sequence in a metric space (X, d), and suppose that the subsequence $(x_{j(n)})$ converges to $x \in X$. Then for every $\varepsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$ such that $d(x_{j(n)}, x) < \varepsilon$ whenever $n \geq N$. Moreover, for every $\varepsilon > 0$, there exists $M = M(\varepsilon) \in \mathbb{N}$ such that $d(x_n, x_m) < \varepsilon$ whenever $n \geq M$ and $m \geq M$. Choose $m \geq N(\frac{\varepsilon}{2})$ for which $j(m) \geq M(\frac{\varepsilon}{2})$. Then for any $n \geq M(\frac{\varepsilon}{2})$, we have by the triangle inequality that

$$d(x_n, x) \leq d(x_n, x_{j(m)}) + d(x_{j(m)}, x) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \qquad \square$$

Theorem 6.13 (Bolzano–Weierstrass theorem)**.** *Let (x_n) be a bounded sequence of real numbers. Then (x_n) has a convergent subsequence.*

The standard proof is via a bisection argument; we do not prove this here.

Corollary 6.14. *Every Cauchy sequence of real numbers is convergent.*

Proof. By the previous two results, it suffices to show that a Cauchy sequence (x_n) of real numbers is bounded. By taking $\varepsilon = 1$ and $m = N$ in the definition of a Cauchy sequence, there exists $N \in \mathbb{N}$ such that $|x_n - x_N| < 1$ whenever $n \geq N$. So if $n \geq N$, then $|x_n| \leq |x_N| + 1$ by the triangle inequality. Thus for all $n \in \mathbb{N}$,

$$|x_n| \leq \max\{|x_1|, \ldots, |x_{N-1}|, |x_N| + 1\},$$

and hence (x_n) is bounded. $\square$

7. Completeness

Recommended reading: [Pug15, §2.3, 2.10, 4.5], [Tao16, §1.4, 6.6].

7.1. **Complete metric spaces.** We have seen that every Cauchy sequence of real numbers is convergent. On the other hand, this is not the case for rational numbers: the sequence

$$x_n = 1 + \sum_{k=1}^{n} \frac{1}{k!} = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots + \frac{1}{n!}$$

consists of rational numbers and is Cauchy but does not have a rational limit, since it converges in $\mathbb{R}$ to the irrational number $e \approx 2.718281828\ldots$.

Definition 7.1. A metric space (X, d) is said to be *complete* if every Cauchy sequence in X is convergent in X. A subset $Y \subseteq X$ of a metric space (X, d) is *complete* if the metric subspace $(Y, d|_{Y \times Y})$ is complete.

Example 7.2.

- The real numbers $\mathbb{R}$ with the standard metric d are complete.
- The rational numbers $\mathbb{Q}$ with the standard metric are not complete.
- Any set X with the discrete metric is complete, since every Cauchy sequence (x_n) with respect to the discrete metric is eventually constant for all sufficiently large n by taking $\varepsilon = 1$ in the definition of a Cauchy sequence and recalling that $d(x, y) < 1$ if and only if $x = y$.

Later we will give more complicated examples of complete metric spaces that fall into a more general framework, namely *Banach spaces* and *Hilbert spaces*.

We can think of a complete metric space as having no "holes": whenever a sequence ought to be convergent to an element in X, it is convergent in X.

Completeness is closely related to a set being closed. We can think of a complete metric space as being *intrinsically closed*: if it is a metric subspace of another metric space, then it is a closed subset of this metric space, *regardless* of what this metric space is.

Proposition 7.3. *Let (X, d) be a complete metric space, and let $(Y, d|_{Y\times Y})$ be a metric subspace. Then $(Y, d|_{Y\times Y})$ is complete if Y is closed in X.*

Proof. Suppose that Y is closed and that (x_n) is a Cauchy sequence in Y. Then (x_n) is also a Cauchy sequence in X, and since X is complete, it is convergent to a limit $x \in X$. For any $\varepsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$ such that $d(x_n, x) < \varepsilon$ whenever $n \geq N$. This means that $x_n \in B_\varepsilon(x)$, and hence $B_\varepsilon(x) \cap Y \neq \emptyset$. Thus x is not an exterior point of Y, which means that $x \in \overline{Y} = Y$. □

The converse to this is less restrictive: we do not even need X to be complete.

Proposition 7.4. *Let (X, d) be a metric space, and let $(Y, d|_{Y\times Y})$ be a metric subspace. If $(Y, d|_{Y\times Y})$ is complete, then Y is closed in X.*

Proof. Suppose that Y is complete and let $x \in \overline{Y}$. Then x is an adherent point of Y, and so for every $r > 0$, we have that $B_r(x) \cap Y \neq \emptyset$. Choose a sequence (x_n) in Y by taking $x_n \in B_{2^{-n}}(x) \cap Y$. Then $d(x_n, x) < 2^{-n}$, and so (x_n) converges to $x \in X$, hence is Cauchy in X, hence is Cauchy in Y, and hence is convergent to $y \in Y$. By the uniqueness of limits, we have that $y = x$, and hence $x \in Y$. Thus Y is equal to its closure, and so is closed. □

7.2. **Products of metric spaces.** We showed earlier how to create new metric spaces from old by restricting to *metric subspaces*. Now we move in the opposite direction by taking two metric spaces and creating a new metric space in terms of their *product*.

Definition 7.5. Let (X, d) and (Y, ρ) be metric spaces. The *product* of X and Y is

$$X \times Y \coloneqq \{(x, y) : x \in X,\ y \in Y\}.$$

The *product metric* is

$$(d_{X\times Y}((x_1, y_1), (x_2, y_2)) \coloneqq d(x_1, x_2) + \rho(y_1, y_2).$$

Remark 7.6. There are several other natural metrics that one can define on $X \times Y$.

Theorem 7.7. *Let (X, d) and (Y, ρ) be metric spaces. The product $(X \times Y, d_{X\times Y})$ is a metric space.*

Proof. Positivity, symmetry, and the triangle inequality all hold from the fact that d and ρ are metrics. □

Example 7.8. The product of two copies of $\mathbb{R}$ with the standard metric is $\mathbb{R}^2$ with the ℓ^1-metric (or taxi-cab metric)

$$d((x_1, y_1), (x_2, y_2)) \coloneqq |x_1 - x_2| + |y_1 - y_2|.$$

Similarly, the product of n copies of $\mathbb{R}$ with the standard metric is $\mathbb{R}^n$ with the ℓ^1-metric.

A sequence converging or being Cauchy in a product metric space $X \times Y$ is closely related to a sequence converging or being Cauchy in X and in Y.

Proposition 7.9. *Let $((x_n, y_n))$ be a sequence in $X \times Y$. Then*

(1) *$((x_n, y_n))$ is convergent to (x, y) in $X \times Y$ if and only if both (x_n) converges to x in X and (y_n) converges to y in Y;*
(2) *$((x_n, y_n))$ is Cauchy in $X \times Y$ if and only if both (x_n) is Cauchy in X and (y_n) is Cauchy in Y.*

Proof.
(1) Suppose that $((x_n, y_n))$ converges to (x, y). Then for all $\varepsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$ such that $d(x_n, x) + \rho(y_n, y) < \varepsilon$ for all $n \geq N$. In particular, $d(x_n, x) < \varepsilon$ and $\rho(y_n, y) < \varepsilon$, and so (x_n) converges to x and (y_n) converges to y.

Conversely, if (x_n) converges to x and (y_n) converges to y, then for all $\varepsilon > 0$, there exist $N = N(\frac{\varepsilon}{2}) \in \mathbb{N}$ and $M = M(\frac{\varepsilon}{2})$ such that $d(x_n, x) < \frac{\varepsilon}{2}$ and $\rho(y_m, y) < \frac{\varepsilon}{2}$ whenever $n \geq N$ and $m \geq M$. Thus $d(x_n, x) + \rho(y_n, y) < \varepsilon$ whenever $n \geq \max\{N, M\}$, and so $((x_n, y_n))$ converges to (x, y).
(2) This follows by the same method of proof as for (1). □

Corollary 7.10. *The product metric space $(X \times Y, d_{X\times Y})$ is complete if and only if both (X, d) and (Y, ρ) are complete.*

In particular, a straightforward induction argument shows that for all $n \geq 1$, $\mathbb{R}^n$ with the ℓ^1-metric is a complete metric space due to the fact that $\mathbb{R}$ is complete. Thus product metric spaces not only give us a method for creating new metric spaces from old, but also new complete metric spaces from old complete metric spaces.

Proof. Suppose that X and Y are complete. Let (x_n, y_n) be Cauchy in $X \times Y$. Then (x_n) is Cauchy and (y_n) is Cauchy in Y, hence convergent to $x \in X$ and $y \in Y$ respectively, and so (x_n, y_n) converges to (x, y). Thus $X \times Y$ is complete.

Conversely, suppose that $X \times Y$ is complete, and let (x_n) be Cauchy in X. Then for fixed $y \in Y$, $((x_n, y))$ is Cauchy in $X \times Y$, hence convergent, and so (x_n) is convergent in X. Thus X is complete. The same argument implies that Y is complete. □

7.3. **Uniformly equivalent metrics.** It sometimes happens that two apparently different metrics are for most purposes equivalent. The following definition makes precise one particular kind of equivalence.

Definition 7.11. Let X be a set that carries two distance functions d and ρ. These two distance functions are said to be *uniformly equivalent* if there exist positive constants $C_1, C_2 > 0$ such that

$$C_1\rho(x, y) \leq d(x, y) \leq C_2\rho(x, y)$$

for all $x, y \in X$.

This notion of equivalence is quite strong. In particular, we can prove the following result.

Proposition 7.12. *Suppose that d and ρ are uniformly equivalent metrics on a set X. Then (X, d) is complete if and only if (X, ρ) is complete.*

Proof. Suppose that (X, d) is complete. Let (x_n) be a Cauchy sequence with respect to ρ. Then for all $\varepsilon > 0$, there exists $N = N(\frac{\varepsilon}{C_2}) \in \mathbb{N}$ such $\rho(x_n, x_m) < \frac{\varepsilon}{C_2}$ whenever $m, n \geq N$. This implies that $d(x_n, x_m) < \varepsilon$, and so (x_n) is Cauchy with respect to d. Since (X, d) is complete, (x_n) converges to x in (X, d). In particular, for each $\varepsilon > 0$, there exists $M = M(C_1\varepsilon) \in \mathbb{N}$ such that $d(x_n, x) < C_1\varepsilon$ whenever $n \geq M$. Thus $\rho(x_n, x) < \varepsilon$, and hence (x_n) converges to x in (X, ρ), and so (X, ρ) is complete.

The same argument shows that if (X, ρ) is complete, then so is (X, d). □

Example 7.13. Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ be elements of $\mathbb{R}^n$. For each $p \geq 1$, the ℓ^p-metrics

$$d_p(x, y) \coloneqq (|x_1 - y_1|^p + \cdots + |x_n - y_n|^p)^{\frac{1}{p}}$$

on $\mathbb{R}^n$ are all uniformly equivalent, and they are all uniformly equivalent to the ℓ^∞-metric (or sup-norm metric)

$$d_\infty(x,y) \coloneqq \max_{1\le m\le n} |x_m - y_m|.$$

To see this, note that

$$d_p(x,y) = (|x_1-y_1|^p + \cdots + |x_n-y_n|^p)^{\frac{1}{p}} \le \left(n \max_{1\le m\le n} |x_m-y_m|^p\right)^{\frac{1}{p}} = n^{1/p} d_\infty(x,y),$$

whereas

$$d_\infty(x,y) = \left(\max_{1\le m\le n} |x_m-y_m|^p\right)^{\frac{1}{p}} \le (|x_1-y_1|^p + \cdots + |x_n-y_n|^p)^{\frac{1}{p}} = d_p(x,y).$$

Thus each metric d_p is uniformly equivalent to d_∞, and consequently each is uniformly equivalent to any other. In particular, the metric spaces $(\mathbb{R}^n, d_p)$ and $(\mathbb{R}^n, d_\infty)$ are all complete, since the ℓ^1-metric d_1 is a product of $\mathbb{R}$ with itself n times, and all the others are uniformly equivalent to this. Taking $p=2$ shows that $\mathbb{R}^n$ with the Euclidean metric d_2 is a complete metric space.

7.4. **The contraction mapping theorem.** The fact that complete metric spaces have no "holes" leads to some useful consequences. One such consequence is the contraction mapping theorem, which is also known as the Banach fixed point theorem. This is an extremely useful theorem, which is used to prove existence of all kinds of things: solutions of algebraic equations, solutions of differential equations (ordinary and partial), optimal strategies in game theory and in economics, optimal control in engineering, and so on.

Definition 7.14. Let (X,d) be a metric space, and let $f : X \to X$ be a function from X to itself. A point $z \in X$ is said to be a *fixed point* of f if $f(z) = z$. The function f is said to be a *contraction* if there exists some $\lambda \in [0,1)$ such that for every $x, y \in X$,

$$d(f(x), f(y)) \le \lambda d(x,y).$$

Note that λ must be *independent* of x and y. Thus a contraction maps two points closer together. The notion of a fixed point is quite natural: in dynamical systems (such as physical systems, but also more general situations such as population modelling or economic modelling), the space X can represent some kind of "state space" describing the configuration of the system at a given time. The map f represents how the system changes from one "cycle" to the next — a year, perhaps, in population modelling, or some fixed interval of measurement in other systems. A fixed point then represents a steady state of the system.

Theorem 7.15 (Contraction mapping theorem)**.** *Let (X,d) be a complete metric space and let $f : X \to X$ be a contraction. Then f has a* unique fixed *point $z \in X$. Furthermore, for any $x \in X$, the sequence $(f^n(x))$ converges to z as n tends to infinity, and*

$$d(f^n(x), z) \le \lambda^n d(x,z).$$

Here $f^n(x)$ means $f(\cdots(f(x)))$; that is, $f^n : X \to X$ is defined by $f^n \coloneqq f \circ \cdots \circ f$.

The theorem says several things: First, there is a fixed point. Secondly, there cannot be more than one fixed point. And thirdly, the fixed point can be approximated as closely as desired by repeatedly applying the contraction to an arbitrary initial guess.

Proof. Take an arbitrary point $x_1 \in X$ and define a sequence (x_n) in X by $x_{n+1} = f(x_n) = f^n(x_1)$. We claim that

$$d(x_{n+1}, x_n) \le \lambda^{n-1} d(x_2, x_1).$$

We prove this by induction. The base case $n = 1$ is immediate. For the induction step, we suppose that $d(x_n, x_{n-1}) \leq \lambda^{n-2} d(x_2, x_1)$. Then by the fact that f is a contraction and the induction assumption,

$$d(x_{n+1}, x_n) = d(f(x_n), f(x_{n-1})) \leq \lambda d(x_n, x_{n-1}) \leq \lambda^{n-1} d(x_2, x_1).$$

Using this, we show that (x_n) is a Cauchy sequence. If $n, m \in \mathbb{N}$ with $n > m$, then by the triangle inequality,

$$\begin{aligned} d(x_n, x_m) &\leq d(x_n, x_{n-1}) + d(x_{n-1}, x_{n-2}) + \cdots + d(x_{m+1}, x_m) \\ &\leq (\lambda^{n-2} + \lambda^{n-3} + \cdots + \lambda^{m-1}) d(x_2, x_1) \\ &\leq \frac{\lambda^{m-1}}{1-\lambda} d(x_1, x_2). \end{aligned}$$

If $d(x_1, x_2) = 0$, then $d(x_n, x_m) = 0$. Otherwise, the fact that $0 \leq \lambda < 1$ means that λ^N tends to zero as N tends to infinity, and so for all $\varepsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$ such that $\lambda^{N-1} < \frac{\varepsilon(1-\lambda)}{d(x_1, x_2)}$, which implies that $d(x_n, x_m) < \varepsilon$ whenever $m, n \geq N$.

Since (x_n) is a Cauchy sequence and X is complete, this sequence has a limit $z \in X$. We show that this limit z is a fixed point of f. Indeed, for any $n \in \mathbb{N}$, we have by the triangle inequality and the fact that f is a contraction that

$$\begin{aligned} d(z, f(z)) &\leq d(z, x_n) + d(x_n, f(x_n)) + d(f(x_n), f(z)) \\ &\leq (1 + \lambda) d(z, x_n) + d(x_n, x_{n+1}). \end{aligned}$$

Fix $\varepsilon > 0$. We choose n sufficiently large such that $d(z, x_n) < \frac{\varepsilon}{3}$ (since (x_n) converges to z) and $d(x_n, x_{n+1}) < \frac{\varepsilon}{3}$ (since (x_n) is Cauchy). Since $\lambda < 1$, it follows that $d(z, f(z)) < \varepsilon$. Since $\varepsilon > 0$ was arbitrary, we deduce that $d(z, f(z)) = 0$, and hence that $z = f(z)$.

Next, we show that this fixed point $z \in X$ is unique. If $y \in X$ is another fixed point, then

$$d(z, y) = d(f(z), f(y)) \leq \lambda d(z, y)$$

as f is a contraction. Since $\lambda < 1$, this is only possible if $d(z, y) = 0$, so that $z = y$.

Finally, we prove that $d(f^n(x), z) \leq \lambda^n d(x, z)$ for any $x \in X$ by induction. The base case $n = 0$ is clear. Now suppose that $d(f^{n-1}(x), z) \leq \lambda^{n-1} d(x, z)$. Then

$$d(f^n(x), z) = d(f^n(x), f(z)) \leq \lambda d(f^{n-1}(x), z) \leq \lambda^n d(x, z). \qquad \square$$

Example 7.16. Consider the map $f(x) = \frac{x}{2} + \frac{1}{x}$, which maps $[1, \infty)$ to itself. A fixed point of f must satisfy $z = \frac{z}{2} + \frac{1}{z}$ or equivalently $z^2 = 2$, and so the only fixed point in $[1, \infty)$ is $z = \sqrt{2}$. We have that

$$f(y) - f(x) = \left(\frac{1}{2} - \frac{1}{xy} \right) (y - x),$$

and so f is a contraction with $\lambda \leq 1/2$ provided that $xy \geq 1$, which is ensured by taking $x, y \in [1, \infty)$. Taking any $x_1 \in [1, \infty)$ and repeatedly applying f will give a sequence $(x_n) \coloneqq (f^{n-1}(x_1))$ that converges to $\sqrt{2}$, with the error decreasing by more than half with each iteration.

7.5. **Completions of metric spaces.** We saw previously that a closed subset of a complete metric space is complete as a metric subspace. The metric subspace $\mathbb{Q}$ of the metric space $(\mathbb{R}, |x - y|)$ is not closed and is not complete. On the other hand, its closure is $\mathbb{R}$, which is complete.

Definition 7.17. A subset E of a metric space (X, d) is *dense* in X if $\overline{E} = X$.

Equivalently, every point in X is the limit of a Cauchy sequence in E. With this in mind, we can think of $\mathbb{R}$ as an *enlarging* of $\mathbb{Q}$ that results in a complete metric space in which $\mathbb{Q}$ is dense. A priori, there is no reason to assume that such an enlarging is *unique*. To enforce uniqueness, we need to introduce the notion of an isometry.

Definition 7.18. A map $\imath : X \to Y$ from one metric space (X, d) to another metric space (Y, ρ) is an *isometry* if

$$\rho(\imath(x), \imath(y)) = d(x, y)$$

for every $x, y \in X$.

Isometries preserve distances. The image $\imath(X) \subseteq Y$ of an isometry is effectively the *same* metric space as X, since all metric space properties are the same if the distance function is the same. Note that an isometry is necessarily injective, but may not be surjective.

Definition 7.19. Let (X, d) be a metric space. A *completion* of (X, d) is a complete metric space $(\overline{X}, \overline{d})$ together with an isometry $\imath : X \to \overline{X}$ such that $\imath(X)$ is dense in $\overline{X}$.

The idea is that $(\overline{X}, \overline{d})$ contains an isometric copy of (X, d), and that no smaller complete subspace of $(\overline{X}, \overline{d})$ does. Indeed, given any complete space containing an isometric copy of (X, d), we could take the closure of the image set, and this would be complete and have a dense isometric copy of (X, d) (and so would be a completion of X in the above sense).

Example 7.20. A completion of a metric subspace $(Y, d|_{Y\times Y})$ of a complete metric space (X, d) is its closure $(\overline{Y}, d|_{\overline{Y}\times\overline{Y}})$.

Theorem 7.21. *Every metric space (X, d) has a completion $(\overline{X}, \overline{d})$ under some isometry $\imath : X \to \overline{X}$. Furthermore, the completion is unique, in the sense that if (Y, ρ) is any other completion with isometry $\jmath : X \to Y$, then there exists an isometric bijection ψ from $(\overline{X}, \overline{d})$ to (Y, ρ) such that $\psi \circ \imath = \jmath$.*

Proof. Let $\mathcal{S}$ denote the set of all Cauchy sequences in X, and define an equivalence relation $\sim$ on $\mathcal{S}$ by $(x_n) \sim (y_n)$ if and only if $\lim_{n\to\infty} d(x_n, y_n) = 0$. We write

$$[(x_n)] \coloneqq \{(y_n) \in \mathcal{S} : (y_n) \sim (x_n)\}$$

for the equivalence class of the Cauchy sequence (x_n). We then let

$$\overline{X} \coloneqq \{[(x_n)] : (x_n) \in \mathcal{S}\}$$

be the set of equivalence classes $[(x_n)]$ of Cauchy sequences (x_n) in X. This is a metric space with metric

$$\overline{d}([(x_n)], [(y_n)]) \coloneqq \lim_{n\to\infty} d(x_n, y_n).$$

Positivity, symmetry, and the triangle inequality then follow from the fact that d is a metric on X. However, we additionally need to show that this limit exists and that this metric is well-defined, in the sense that the limit is independent of the choice of representatives of the equivalence classes $[(x_n)], [(y_n)]$.

To see that this limit exists, we note that as $(x_n), (y_n)$ are Cauchy sequences, there exists some $N = N(\frac{\varepsilon}{2}) \in \mathbb{N}$ and $M = M(\frac{\varepsilon}{2}) \in \mathbb{N}$ such that

$$d(x_n, x_m) < \frac{\varepsilon}{2} \quad \text{for } n, m \geq N, \qquad d(y_n, y_m) < \frac{\varepsilon}{2} \quad \text{for } n, m \geq M.$$

So by the triangle inequality, for $n, m \geq \max\{N, M\}$, we have that

$$d(x_n, y_n) - d(x_m, y_m) \leq d(x_n, x_m) + d(y_n, y_m) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

and similarly

$$d(x_m, y_m) - d(x_n, y_n) \leq d(x_m, x_n) + d(y_m, y_n) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Thus $|d(x_m, y_m) - d(x_n, y_n)| < \varepsilon$ for all $n, m \geq \max\{N, M\}$, so that $(d(x_n, y_n))$ is a Cauchy sequence in $\mathbb{R}$, and hence is convergent.

To see that $\overline{d}$ is well-defined, so that this limit is independent of the choice of representatives (x_n) and (y_n) of the equivalence classes $[(x_n)]$ and $[(y_n)]$, we let $(x_n), (x'_n) \in [(x_n)]$ and $(y_n), (y'_n) \in [(y_n)]$ be pairs of representatives of the equivalence classes $[(x_n)]$ and $[(y_n)]$. By the triangle inequality, we have that

$$d(x'_n, y'_n) - d(x_n, y_n) \leq d(x'_n, x_n) + d(y_n, y'_n)$$

for all $n \in \mathbb{N}$, and similarly

$$d(x_n, y_n) - d(x'_n, y'_n) \leq d(x_n, x'_n) + d(y'_n, y_n).$$

Since $\lim_{n\to\infty} d(x_n, x'_n) = \lim_{n\to\infty} d(y_n, y'_n) = 0$, we deduce that $\lim_{n\to\infty} d(x_n, y_n) = \lim_{n\to\infty} d(x'_n, y'_n)$; thus this limit is indeed independent of the choice of representatives.

An isometry from X to $\overline{X}$ is given by $\imath(x) \coloneqq [(x_n)]$ with (x_n) the constant sequence given by $x_n = x$ for all $n \in \mathbb{N}$. That this is an isometry is clear, since

$$\overline{d}(\imath(x), \imath(y)) = \lim_{n\to\infty} d(x, y) = d(x, y).$$

We claim that $\imath(X)$ is dense in $\overline{X}$. Indeed, if $[(y_n)] \in \overline{X}$, then there exists a sequence $(\imath(x_m))$ that converges to $[(y_n)]$ simply by taking $x_m = y_m$, so that

$$\overline{d}(\imath(x_m), [(y_n)]) = \lim_{n\to\infty} d(y_m, y_n).$$

Since (y_n) is a Cauchy sequence, there exists $N = N(\varepsilon) \in \mathbb{N}$ such that $d(y_m, y_n) < \varepsilon$ whenever $n, m \geq N$. Thus if $m \geq N$, $\overline{d}(\imath(x_m), [(y_n)]) < \varepsilon$, as desired.

We next show that $(\overline{X}, \overline{d})$ is complete. Let $(\overline{x_m})$ be a Cauchy sequence in $\overline{X}$, so that for each $m \in \mathbb{N}$, $\overline{x_m}$ is an equivalence class of sequences $[(x_{m,k})]$. Since the sequence $(\overline{x_m})$ is Cauchy, for all $\varepsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$ such that

$$\overline{d}(\overline{x_m}, \overline{x_n}) \coloneqq \lim_{k\to\infty} d(x_{m,k}, x_{n,k}) < \varepsilon$$

whenever $n, m \geq N$. For each $n \in \mathbb{N}$, the sequence $(x_{n,k})$ in X is a Cauchy sequence in X, so that for all $\varepsilon > 0$, there exists $M = M(n, \varepsilon) \in \mathbb{N}$ such that $d(x_{n,k}, x_{n,\ell}) < \varepsilon$ whenever $k, \ell \geq M$. We then define $y_{m,k} \coloneqq x_{m,M(m,1/k)}$, which is a Cauchy sequence in X satisfying $\lim_{k\to\infty} d(y_{m,k}, x_{m,k}) = 0$, so that $[(y_{m,k})] = [(x_{m,k})] = \overline{x_m}$. We claim that $(y_{m,m})$ is a Cauchy sequence in X, so that $\overline{x} \coloneqq [(y_{m,m})] \in \overline{X}$. Indeed, by the triangle inequality, we have that

$$d(y_{m,m}, y_{n,n}) \leq d(y_{m,m}, y_{m,k}) + d(y_{m,k}, y_{n,k}) + d(y_{n,k}, y_{n,n})$$

for any $k, m, n \in \mathbb{N}$. For $m, n \geq \max\{\frac{3}{\varepsilon}, N(\frac{\varepsilon}{3})\}$, we see that by taking the limit as k tends to infinity,

$$d(y_{m,m}, y_{n,n}) < \frac{1}{m} + \frac{\varepsilon}{3} + \frac{1}{n} \leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

Finally, we claim that $\lim_{m\to\infty} \overline{d}(\overline{x_m}, \overline{x}) = 0$. This follows from the fact that $\overline{d}(\overline{x_m}, \overline{x}) = \lim_{k\to\infty} d(y_{m,k}, y_{m,m})$, which is less than ε provided that $m \geq N(\varepsilon)$.

The final step is to show that the completion is unique. Suppose that (Y, ρ) is also a completion of X with $\jmath : X \to Y$ an isometry. The key idea is to construct a map $\psi : \overline{X} \to Y$ given by $\psi([(x_n)]) \coloneqq \lim_{n\to\infty} \jmath(x_n)$ and then confirm that this is an isometric bijection satisfying $\psi \circ \imath = \jmath$.

To see that ψ is well-defined, we note that the limit exists for any representative (x_n) of the equivalence class $[(x_n)]$ since (x_n) is Cauchy in X, so that $(\jmath(x_n))$ is Cauchy in

Y as $\imath$ is an isometry, and hence convergent as Y is complete. Moreover, this limit is independent of choice of representative, since if $(y_n) \sim (x_n)$ and $\lim_{n\to\infty} \jmath(x_n) = x$, then the fact that j is an isometry implies that

$$\rho(x, \jmath(y_n)) \leq \rho(x, \jmath(x_n)) + \rho(\jmath(x_n), \jmath(y_n)) = \rho(x, \jmath(x_n)) + d(x_n, y_n),$$

by the triangle inequality, and this converges to 0 as n tends to infinity.

Next, we show that ψ is an isometry: we have that

$$\begin{aligned}\rho(\psi([(x_n)]), \psi([(y_n)])) &= \rho\left(\lim_{n\to\infty} \jmath(x_n), \lim_{n\to\infty} \jmath(y_n)\right) \\ &= \lim_{n\to\infty} \rho(\jmath(x_n), \jmath(y_n)) \\ &= \lim_{n\to\infty} d(x_n, y_n) \\ &= \overline{d}([(x_n)], [(y_n)]).\end{aligned}$$

The validity of the second equality may be justified as follows. Letting $x = \lim_{n\to\infty} \jmath(x_n)$ and $y = \lim_{n\to\infty} \jmath(y_n)$, we have that

$$\rho(x, y) - \rho(\jmath(x_n), \jmath(y_n)) \leq \rho(x, \jmath(x_n)) + \rho(\jmath(y_n), y)$$

and similarly

$$\rho(\jmath(x_n), \jmath(y_n)) - \rho(x, y) \leq \rho(x, \jmath(x_n)) + \rho(\jmath(y_n), y)$$

by the triangle inequality. Since the right-hand side tends to 0 as n tends to infinity, the left-hand side must also tend to 0.

Since ψ is an isometry, it is injective. Moreover, ψ is surjective: for any $y \in Y$, y must lie in the closure of $\jmath(X)$, so that there exists some sequence (x_n) in X such that $\lim_{n\to\infty} \jmath(x_n) = y$. But then the sequence (x_n) is Cauchy since $\jmath$ is an isometry, $(\jmath(x_n))$ is Cauchy, and $\psi([(x_n)]) = y$.

Finally, for each $x \in X$, we have that $\psi \circ \imath(x) = \psi([(x_n)])$, where $x_n = x$ for all $n \in \mathbb{N}$. Thus

$$\psi \circ \imath(x) = \lim_{n\to\infty} \jmath(x_n) = \lim_{n\to\infty} \jmath(x) = \jmath(x),$$

which shows that $\psi \circ \imath = \jmath$. □

8. Normed spaces

Recommended reading: [Pug15, §1.3].

We take a brief detour to introduce certain types of metric spaces that are of central importance in analysis. These spaces are such that X is a *vector space* and the metric on X is particularly well-behaved.

8.1. Normed spaces and inner product spaces.

Definition 8.1. A *normed space* $(V, \|\cdot\|)$ is a real vector space V together with a norm, which is a function $\|\cdot\| : V \to [0, \infty)$ satisfying the following three properties:

(i) $\|v\| \geq 0$ for all $v \in V$ with equality if and only if $v = 0$ (positivity);
(ii) For all $\alpha \in \mathbb{R}$ and $v \in V$, we have that $\|\alpha v\| = |\alpha|\|v\|$ (homogeneity);
(iii) For all $v, w \in V$, we have that $\|v + w\| \leq \|v\| + \|w\|$ (subadditivity).

Theorem 8.2. *Let $(V, \|\cdot\|)$ be a normed space. Define $d : V \times V \to [0, \infty)$ by $d(v, w) := \|v - w\|$. Then (V, d) is a metric space.*

Proof. Positivity of this function follows from positivity of the norm. Symmetry follows from homogeneity upon taking $\alpha = -1$ and replacing v with $x - y$. The triange inequality follows by subadditivity by replacing v with $x - y$ and w with $y - z$. □

A special case of a normed space is an inner product space.

Definition 8.3. An *inner product space* $(V, \langle\cdot,\cdot\rangle)$ is a real vector space V together with an inner product, which is a function $\langle\cdot,\cdot\rangle : V \to \mathbb{R}$ satisfying the following three properties:

(i) $\langle v, v\rangle \geq 0$ for all $v \in V$ with equality if and only if $v = 0$ (positive definiteness);
(ii) For all $v, w \in V$, we have that $\langle v, w\rangle = \langle w, v\rangle$ (symmetry);
(iii) For all $\alpha \in \mathbb{R}$ and $u, v, w \in V$, we have that $\langle v + u, w\rangle = \langle v, w\rangle + \langle u, w\rangle$ and $\langle \alpha v, w\rangle = \alpha\langle v, w\rangle$ (linearity).

Proposition 8.4. *Let $(V, \langle\cdot,\cdot\rangle)$ be an inner product space. Then V is a normed space with norm $\|v\| \coloneqq \sqrt{\langle v, v\rangle}$, and hence a metric space with metric $\sqrt{\langle v - w, v - w\rangle}$.*

Proof. Positivity and homogeneity are immediate. Subadditivity follows in the same way as the proof of the triangle inequality for $\sqrt{\langle x - y, x - y\rangle}$ via the Cauchy–Schwarz inequality. □

Remark 8.5. The notion of a normed space and of an inner product space can be generalised to *complex* vector spaces with a little extra work. For normed spaces, there is no difference, save that homogeneity allows for $\alpha \in \mathbb{C}$, so that $|\alpha|$ denotes the modulus of the complex number α.

For inner product spaces, the chief difference is that inner products on complex vector spaces may be *complex*-valued rather than just real-valued; moreover, they do not satisfy symmetry, but rather *conjugate symmetry*: for all $v, w \in V$, we have that $\langle v, w\rangle = \overline{\langle w, v\rangle}$. Note that combining this with linearity, we have that $\langle \alpha v, w\rangle = \alpha\langle v, w\rangle$ but $\langle v, \beta w\rangle = \overline{\beta}\langle v, w\rangle$ for $v, w \in V$ and $\alpha, \beta \in \mathbb{C}$.

Thus every inner product space is a normed space, and every normed space is a metric space. We will now go through some examples of inner product spaces and normed spaces (albeit without proofs that these are indeed inner product spaces or normed spaces).

Example 8.6. Let $X = \mathbb{R}^n$. Then X is an inner product space with inner product

$$\langle x, y\rangle \coloneqq x_1y_1 + x_2y_2 + \cdots + x_ny_n,$$

and hence a normed space with norm

$$\|x\| \coloneqq \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2},$$

as well as a metric space with (Euclidean) metric

$$d_2(x, y) \coloneqq \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}.$$

Example 8.7. Let $\ell^2(\mathbb{N})$ denote the space of all functions $a : \mathbb{N} \to \mathbb{R}$ that satisfy

$$\sum_{n=1}^{\infty} |a(n)|^2 < \infty.$$

(Alternatively, we can view $a \in \ell^2(\mathbb{N})$ as a real-valued sequence $(a_n) \subset \mathbb{R}$.) Then $\ell^2(\mathbb{N})$ is an inner product space with inner product

$$\langle a, b\rangle \coloneqq \sum_{n=1}^{\infty} a(n)b(n),$$

and hence a normed space with norm

$$\|a\|_{\ell^2} \coloneqq \sqrt{\sum_{n=1}^{\infty} |a(n)|^2}$$

and a metric space with metric

$$d_2(a,b) \coloneqq \sqrt{\sum_{n=1}^{\infty} |a(n) - b(n)|^2}.$$

Remark 8.8. To prove that this is an inner product space, one must first confirm that this is a vector space with addition given by $(a+b)(n) \coloneqq a(n) + b(n)$, scalar multiplication given by $(\alpha a)(n) \coloneqq \alpha a(n)$, and the zero vector in $\ell^2(\mathbb{N})$ is the function $a : \mathbb{N} \to \mathbb{R}$ given by $a(n) \coloneqq 0$ for all $n \in \mathbb{N}$. One must then show that this inner product satisfies positive definiteness, symmetry, and linearity.

Example 8.9. Let $1 \leq p < \infty$, and let $\ell^p(\mathbb{N})$ denote the space of all functions $a : \mathbb{N} \to \mathbb{R}$ that satisfy

$$\sum_{n=1}^{\infty} |a(n)|^p < \infty.$$

(Alternatively, we can view $a \in \ell^p(\mathbb{N})$ as a real-valued sequence $(a_n) \subset \mathbb{R}$.) Then $\ell^p(\mathbb{N})$ is a normed space with norm

$$\|a\|_{\ell^p} \coloneqq \left(\sum_{n=1}^{\infty} |a(n)|^p \right)^{\frac{1}{p}}.$$

Similarly, let $\ell^\infty(\mathbb{N})$ denote the space of all functions $a : \mathbb{N} \to \mathbb{R}$ that satisfy

$$\sup_{n \in \mathbb{N}} |a(n)| < \infty.$$

Then $\ell^\infty(\mathbb{N})$ is a normed space with norm

$$\|a\|_{\ell^\infty} \coloneqq \sup_{n \in \mathbb{N}} |a(n)|.$$

Remark 8.10. The spaces $\ell^p(\mathbb{N})$ are vector spaces with addition, scalar multiplication, and the zero vector exactly as for $\ell^2(\mathbb{N})$. Proving that $\| \cdot \|_{\ell^p}$ is a norm on $\ell^p(\mathbb{N})$ takes a little work; in order to prove subadditivity, one needs to first prove a generalisation of Hölder's inequality in this setting. If $p \neq 2$, the normed space $(\ell^p(\mathbb{N}), \| \cdot \|_{\ell^p})$ can be proven to *not* be an inner product space: there is no inner product $\langle \cdot, \cdot \rangle$ on $\ell^p(\mathbb{N})$ that satisfies $\sqrt{\langle a, a \rangle} = \|a\|_p$.

These spaces are known as *sequence spaces*. The norm $\| \cdot \|_{\ell^p}$ is called the *ℓ^p-norm*.

Example 8.11. Let $C([0,1])$ denote the space of all continuous functions $f : [0,1] \to \mathbb{R}$. Then $C([0,1])$ is an inner product space with inner product

$$\langle f, g \rangle \coloneqq \int_0^1 f(t) g(t)\, dt,$$

and hence a normed space with norm

$$\|f\|_{L^2} \coloneqq \sqrt{\int_0^1 f(t)^2\, dt}.$$

Remark 8.12. Again, to prove that this is an inner product space, one must first confirm that $C([0,1])$ is a vector space with addition given by $(f+g)(x) \coloneqq f(x) + g(x)$, scalar multiplication given by $(\alpha f)(x) \coloneqq \alpha f(x)$, and the zero vector in $C([0,1])$ is the function $f : [0,1] \to \mathbb{R}$ given by $f(x) \coloneqq 0$ for all $x \in [0,1]$. One must then show that this inner product satisfies positive definiteness, symmetry, and linearity.

Example 8.13. Let $1 \leq p < \infty$, and let $C([0,1])$ denote the space of all continuous functions $f : [0,1] \to \mathbb{R}$. Then $C([0,1])$ is a normed space with norm

$$\|f\|_{L^p} \coloneqq \left(\int_0^1 |f(t)|^p \, dt \right)^{\frac{1}{p}}.$$

Similarly, $C([0,1])$ is a normed space with norm

$$\|f\|_{L^\infty} \coloneqq \sup_{t \in [0,1]} |f(t)|.$$

Remark 8.14. Proving that $\| \cdot \|_{L^p}$ is a norm on $C([0,1])$ again takes a little work, since a generalisation of Hölder's inequality in this setting is needed in order to prove subadditivity. If $p \neq 2$, the normed space $(C([0,1]), \| \cdot \|_{L^p})$ can be proven to *not* be an inner product space: there is no inner product $\langle \cdot, \cdot \rangle$ on $C([0,1])$ that satisfies $\sqrt{\langle f, f \rangle} = \|f\|_p$.

These spaces are known as *function spaces*. The norm $\| \cdot \|_{L^p}$ is called the *L^p-norm*.

8.2. **Banach spaces and Hilbert spaces.** Normed spaces or inner product spaces that are *complete* as metric spaces are a special type of metric space.

Definition 8.15. A *Banach space* is a complete normed vector space.

Definition 8.16. A *Hilbert space* is a complete inner product space.

Thus a Hilbert space is a Banach space, but not every Banach space is a Hilbert space. Hilbert spaces arise naturally in the field of quantum mechanics. The study of Hilbert and Banach spaces leads to the field of *functional analysis*. Many fields of mathematics, and nearly all subfields of analysis, deal with Hilbert or Banach spaces in some way or another.

Theorem 8.17. *For each $1 \leq p \leq \infty$, the normed space $(\ell^p(\mathbb{N}), \| \cdot \|_{\ell^p})$ is a Banach space. In particular, the inner product space $(\ell^2(\mathbb{N}), \langle \cdot, \cdot \rangle)$ is a Hilbert space.*

Thus these normed spaces have no "holes".

Proof. We prove this for $p = 1$; the main ideas of the proof work for each $1 \leq p \leq \infty$ with some minor changes. We write a for an element of $\ell^1(\mathbb{N})$, viewing a as a function from $\mathbb{N}$ to $\mathbb{R}$, so that $a(n) \in \mathbb{R}$ for each $n \in \mathbb{N}$. Now let (a_n) be a Cauchy sequence in $\ell^1(\mathbb{N})$, so that for each $\varepsilon >$, there exists $N = N(\varepsilon) \in \mathbb{N}$ such that

$$\|a_n - a_m\|_{\ell^1} \coloneqq \sum_{k=1}^{\infty} |a_n(k) - a_m(k)| < \varepsilon$$

whenever $n, m \geq N$. In particular, if $n, m \geq N$, then for each fixed $k \in \mathbb{N}$, we have that $|a_n(k) - a_m(k)| < \varepsilon$. So for each fixed $k \in \mathbb{N}$, the sequence $(a_n(k))$ of real numbers $a_n(k)$ is a Cauchy sequence in $\mathbb{R}$, and hence convergent in $\mathbb{R}$ since $\mathbb{R}$ is complete, so that $a_n(k)$ converges to some real number $a(k)$ as n tends to infinity.

In this way, we construct a sequence $a : \mathbb{N} \to \mathbb{R}$. We claim that $a \in \ell^1(\mathbb{N})$. To see this, we first fix $K \in \mathbb{N}$. For any $n \in \mathbb{N}$, we have by the triangle inequality that

$$\sum_{k=1}^{K} |a(k)| \leq \sum_{k=1}^{K} |a_n(k)| + \sum_{k=1}^{K} |a(k) - a_n(k)| \leq \sum_{k=1}^{\infty} |a_n(k)| + \sum_{k=1}^{K} |a(k) - a_n(k)|.$$

The first term is simply $\|a_n\|_{\ell^1}$. Since (a_n) is a Cauchy sequence in $\ell^1(\mathbb{N})$, it is bounded in $\ell^1(\mathbb{N})$, so that there exists some constant $C > 0$ independent of $n \in \mathbb{N}$ such that

$\|a_n\|_{\ell^1} \leq C$. It follows that

$$\sum_{k=1}^{K} |a(k)| \leq C + \sum_{k=1}^{K} |a(k) - a_n(k)|$$

for all $n \in \mathbb{N}$. Since the sequence $(a_n(k))$ in $\mathbb{R}$ converges to $a(k)$ for each fixed $k \in \mathbb{N}$, for all $\varepsilon > 0$, there exists $N_k = N_k(\frac{\varepsilon}{K})$ such that $|a(k) - a_n(k)| < \frac{\varepsilon}{K}$ whenever $n \geq N_k$. Thus for $n \geq \max\{N_1, \ldots, N_K\}$, we have that $\sum_{k=1}^{K} |a(k) - a_n(k)| < \varepsilon$, so that $\sum_{k=1}^{K} |a(k)| < C + \varepsilon$. Since $\varepsilon > 0$ was arbitrary, we have that $\sum_{k=1}^{K} |a(k)| \leq C$. Moreover, since $K \in \mathbb{N}$ was also arbitrary, we have that $\|a\|_{\ell^1} \coloneqq \sum_{k=1}^{\infty} |a(k)| \leq C < \infty$, and hence $a \in \ell^1(\mathbb{N})$.

It remains to show that the sequence (a_n) in $\ell^1(\mathbb{N})$ converges to a. To see this, we again first fix $K \in \mathbb{N}$. Since (a_n) is Cauchy, for all $\varepsilon > 0$, there exists $N = N(\frac{\varepsilon}{2})$ such that $\|a_n - a_m\|_{\ell^1} < \frac{\varepsilon}{2}$ whenever $n, m \geq N$. It follows that for $n, m \geq N$, we have by the triangle inequality that

$$\begin{aligned}
\sum_{k=1}^{K} |a_n(k) - a(k)| &\leq \sum_{k=1}^{K} |a_m(k) - a(k)| + \sum_{k=1}^{K} |a_n(k) - a_m(k)| \\
&\leq \sum_{k=1}^{K} |a_m(k) - a(k)| + \sum_{k=1}^{\infty} |a_n(k) - a_m(k)| \\
&< \sum_{k=1}^{K} |a_m(k) - a(k)| + \frac{\varepsilon}{2}.
\end{aligned}$$

Since the sequence $(a_m(k))$ in $\mathbb{R}$ converges to $a(k)$ for each fixed $k \in \mathbb{N}$, for all $\varepsilon > 0$, there exists $N_k = N_k(\frac{\varepsilon}{2K})$ such that $|a_m(k) - a(k)| < \frac{\varepsilon}{2K}$ whenever $m \geq N_k$. It follows that $\sum_{k=1}^{K} |a_m(k) - a(k)| < \frac{\varepsilon}{2}$ for $m \geq \max\{N, N_1, \ldots, N_K\}$, and so $\sum_{k=1}^{K} |a_n(k) - a(k)| < \varepsilon$ for $n \geq N$. Since $K \in \mathbb{N}$ was arbitrary, we have that $\|a_n - a\|_{\ell^1} \coloneqq \sum_{k=1}^{\infty} |a_n(k) - a(k)| < \varepsilon$. Thus (a_n) converges to a, and so $\ell^1(\mathbb{N})$ is complete. □

For $C([0,1])$, the story is different: this is *not* complete with respect to the L^p-norm *except* when $p = \infty$.

Theorem 8.18. *For each $1 \leq p < \infty$, the normed space $(C([0,1]), L^p)$ is* not *a Banach space. In particular, the inner product space $(C([0,1]), L^2)$ is* not *a Hilbert space. On the other hand, $(C([0,1]), L^\infty)$* is *a Banach space.*

Thus the normed spaces $(C([0,1]), L^p)$ with $1 \leq p < \infty$ have "holes", where a Cauchy sequence of continuous functions does not have a limit in $C([0,1])$. When $p = \infty$, these "holes" disappear. The completions of $(C([0,1]), L^p)$ are Banach spaces denoted by $L^p([0,1])$; these spaces have a natural description as the spaces of *Lebesgue-integrable functions* for which the L^p-norm is finite. In particular, these spaces arise in the theory of Lebesgue integration and more generally in measure theory.

Proof. We prove only that $(C([0,1]), L^\infty)$ is complete. Let (f_n) be a Cauchy sequence in $(C([0,1]), L^\infty)$, so that for every $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $\|f_n - f_m\|_{L^\infty} \coloneqq \sup_{x \in [0,1]} |f_n(x) - f_m(x)| < \varepsilon$ for all $x \in [0,1]$ whenever $n, m \geq N$. Then for each *fixed* $x \in [0,1]$, the sequence $(f_n(x))$ of real numbers is a Cauchy sequence in $\mathbb{R}$, and hence convergent to some real number $f(x)$ since $\mathbb{R}$ is complete. We claim that for $\sup_{x \in [0,1]} |f_n(x) - f(x)| < \varepsilon$ if $n \geq N(\varepsilon)$. Indeed, for any $x \in [0,1]$, we have that

$$|f_n(x) - f(x)| \leq |f_n(x) - f_m(x)| + |f_m(x) - f(x)$$

for any $m \in \mathbb{N}$ by the triangle inequality. If $m \geq N(\varepsilon)$, the first term is less than ε. Moreover, since $f_m(x)$ converges to $f(x)$, for all $\varepsilon' > 0$, there exists $M = M(\varepsilon', x) \in \mathbb{N}$ such that $|f_m(x) - f(x)| < \varepsilon'$ for all $m \geq M$. Thus $|f_n(x) - f(x)| < \varepsilon + \varepsilon'$, and since $\varepsilon' > 0$ was arbitrary, we have that $|f_n(x) - f(x)| < \varepsilon$. Since this is true for all $x \in [0,1]$ independently of ε, we deduce that $\sup_{x\in[0,1]} |f_n(x) - f(x)| < \varepsilon$ if $n \geq N(\varepsilon)$.

It remains to show that the function $f : [0,1] \to \mathbb{R}$ given by $f(x) \coloneqq \lim_{n\to\infty} f_n(x)$ is *continuous*, since all we know so far is that for each $x \in [0,1]$, $f(x)$ is the limit of the convergent sequence of real numbers $(f_n(x))$. We must show that for each $x \in [0,1]$, f is continuous at x, so that for all $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon, x)$ such that $|f(x) - f(y)| < \varepsilon$ whenever $|x - y| < \delta$. For any $y \in [0,1]$, we have that for $n \geq N(\frac{\varepsilon}{3})$,

$$\begin{aligned} |f(x) - f(y)| &\leq |f(x) - f_n(x)| + |f_n(x) - f_n(y)| + |f_n(y) - f(y)| \\ &\leq 2 \sup_{t\in[0,1]} |f(t) - f_n(t)| + |f_n(x) - f_n(y)| \\ &\leq \frac{2\varepsilon}{3} + |f_n(x) - f_n(y)|. \end{aligned}$$

Since f_n is continuous at x, there exists $\delta = \delta(\frac{\varepsilon}{3}, n, x) > 0$ such that $|f_n(x) - f_n(y)| < \frac{\varepsilon}{3}$ whenever $|x-y| < \delta$. Thus $|f(x)-f(y)| < \varepsilon$ whenever $|x-y| < \delta$, and so f is continuous at x. □

Remark 8.19. The method of proof works in more general settings. In particular, let Ω be a closed subset of $\mathbb{R}^n$, which need not be bounded (such as $\mathbb{R}^n$ itself), and let $C_b(\Omega)$ denote the vector space consisting of functions $f : \Omega \to \mathbb{R}$ that are bounded, so that $\sup_{t\in\Omega} |f(t)| < \infty$. Let $\|\cdot|_{L^\infty}$ denote the norm $\|f\|_{L^\infty} \coloneqq \sup_{t\in\Omega} |f(t)|$. Then $(C_b(\Omega), \|\cdot\|_{L^\infty})$ is a Banach space. If we take $n = 1$ and $\Omega = \mathbb{N}$, then this is simply $(\ell^\infty(\mathbb{N}), \|\cdot|_{\ell^\infty})$.

9. Compactness

Recommended reading: [Pug15, §2.4], [Tao16, §1.5].

9.1. **Sequential compactness.** Recall that the Bolzano–Weierstrass theorem states that every bounded sequence (x_n) in $\mathbb{R}$ (with the standard metric) has a convergent subsequence. We used this to show that every Cauchy sequence in $\mathbb{R}$ has a convergent subsequence, so that $\mathbb{R}$ is complete. The Bolzano–Weierstrass theorem does not necessarily hold for arbitrary metric spaces, but it does motivate the following definition.

Definition 9.1. A metric space (X, d) is *sequentially compact* (or just *compact*) if every sequence (x_n) in X has a convergent subsequence. A subset $Y \subseteq X$ of a metric space (X, d) is *compact* if the metric subspace $(Y, d|_{Y\times Y})$ is compact.

A consequence of the Bolzano–Weierstrass theorem is that every closed bounded subset E of $\mathbb{R}$ is compact as a metric subspace of $\mathbb{R}$. The converse is also true, which thereby gives necessary and sufficient conditions for a subset of $\mathbb{R}$ to be compact.

Theorem 9.2 (Heine–Borel theorem)**.** *Let E be a subset of $\mathbb{R}$. Then E is compact if and only if it is closed and bounded.*

Example 9.3. The whole space $\mathbb{R}$ is not compact, since the sequence (x_n) with $x_n = n$ has no convergent subsequence in $\mathbb{R}$. The open bounded set $(0,1)$ is not compact since the sequence (x_n) with $x_n = 2^{-n}$ has no subsequence that converges to an element of $(0,1)$.

Question 9.4. *How does one generalise this to necessary and sufficient conditions for a subset of an arbitrary metric space to be compact?*

It is not hard to see that being closed and bounded are *necessary* conditions for sequential compactness. To prove this, we must first define what it means for a metric space to be bounded.

Definition 9.5. A metric space (X, d) is *bounded* if there exists some $r > 0$ and $x \in X$ such that $X = B_r(x) = \{y \in X : d(x, y) < r\}$. A subset $Y \subseteq X$ of a metric space (X, d) is *bounded* if the metric subspace $(Y, d|_{Y\times Y})$ is bounded.

Lemma 9.6. *Let (X, d) be a metric space and let E be a compact subset of X. Then E is closed and bounded.*

Proof. We prove the contrapositive. Suppose that E is not bounded, so that for each $x \in E$, there does not exist some $r > 0$ such that $E \subseteq B_r(x)$. Thus if we fix $x \in E$, then for each $n \in \mathbb{N}$, there exists some $x_n \notin B_n(x)$. The sequence (x_n) is such that every subsequence is unbounded, and hence does not converge; consequently, E is not compact.

Similarly, suppose that E is not closed, so that there exists some limit point $x \in X$ of E that is not in E itself. Let (x_n) be a sequence in E that converges to x. Then every subsequence converges to x, which is not in E, and so no subsequence converges to a point in E itself. Thus again E is not compact. □

Completeness is also a necessary condition. For $\mathbb{R}$, this is immediate, since $\mathbb{R}$ is complete and so any subset of $\mathbb{R}$ is closed if and only if it is complete as a metric subspace of $\mathbb{R}$. In particular, the Heine–Borel theorem can be rephrased as stating that a subset E of $\mathbb{R}$ is compact if and only if it is complete and bounded.

Lemma 9.7. *Let (X, d) be a compact metric space. Then X is complete.*

Proof. We prove the contrapositive. Suppose that X is not complete, so that there exists a Cauchy sequence (x_n) in X that is not convergent in X. Since a Cauchy sequence with a convergent subsequence is itself convergent, this means that (x_n) must not contain a convergent subsequence, and hence X is not compact. □

However, being closed and bounded are not *sufficient* conditions for sequential compactness; nor is completeness sufficient.

Example 9.8. Let $\ell^\infty(\mathbb{N})$ be the Banach space of bounded functions $a : \mathbb{N} \to \mathbb{R}$ with the ℓ^∞-norm $\|a\|_{\ell^\infty} \coloneqq \sup_{n\in\mathbb{N}} |a(n)|$. Let $E \coloneqq \{a \in \ell^\infty(\mathbb{N}) : \|a\|_{\ell^\infty} \leq 1\}$ be the closed unit ball centred at the origin. This is closed and bounded; it is also complete since it is a closed subset of the complete normed space $\ell^\infty(\mathbb{N})$. However, it is not compact. To see this, we consider the sequence (a_n) of functions $a_n \in \ell^\infty(\mathbb{N})$ for which $a_n : \mathbb{N} \to \mathbb{R}$ is given by

$$a_n(k) \coloneqq \begin{cases} 1 & \text{if } k = n, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\|a_n - a_m\|_{\ell^\infty} = \sup_{k\in\mathbb{N}} |a_n(k) - a_m(k)| = \begin{cases} 0 & \text{if } m = n, \\ 1 & \text{otherwise.} \end{cases}$$

Thus no subsequence of (a_n) is Cauchy, as any two distinct elements in this sequence are always distance 1 apart, and so there cannot exist a convergent subsequence.

Example 9.9. Let X be the Banach space $C([0, 1])$ of continuous functions from $[0, 1]$ to $\mathbb{R}$ with the metric inherited from the L^∞-norm $\|f\|_{L^\infty} \coloneqq \sup_{x\in[0,1]} |f(x)|$. Let $E \coloneqq \{f \in C([0, 1]) : \|f\|_{L^\infty} \leq 1\}$ be the closed unit ball centred at the origin. This is closed and bounded; it is also complete since it is a closed subset of the complete metric space

$C([0,1])$. However, it is not compact. To see this, we consider the sequence (f_n) of functions $f_n \in C([0,1])$ for which $f_n : [0,1] \to \mathbb{R}$ is given by

$$f_n(x) \coloneqq \begin{cases} 0 & \text{if } 0 \leq x \leq 2^{-n}, \\ 2^{n+1}(x - 2^{-n}) & \text{if } 2^{-n} \leq x \leq \frac{3}{2}2^{-n}, \\ 2^{n+1}(2^{1-n} - x) & \text{if } \frac{3}{2}2^{-n} \leq x \leq 2^{1-n}, \\ 0 & \text{if } 2^{1-n} \leq x \leq 1. \end{cases}$$

Then f_n is continuous on $[0,1]$ and is such that

$$\|f_n - f_m\|_{L^\infty} = \sup_{x \in [0,1]} |f_n(x) - f_m(x)| = \begin{cases} 0 & \text{if } m = n, \\ 1 & \text{otherwise.} \end{cases}$$

Thus (f_n) has no convergent subsequences, since any two distinct elements in this sequence are always distance 1 apart.

Spaces can fail to be compact for two main reasons:

(1) the space is "too big at infinity", so that any sequence that "runs off to infinity" cannot have a convergent subsequence (take $X = \mathbb{R}$ as an example);
(2) the space is "too big locally", so that closed bounded sets like closed unit balls have "too many points" and there are "too many directions" in which one can travel from a given point (take $X = \ell^\infty(\mathbb{N})$ as an example).

The second issue is not easily overcome. For the first issue, we know from the Heine–Borel theorem that even though $X = \mathbb{R}$ is not compact, every closed bounded subset of it is. This motivates the following generalisation of compactness.

Definition 9.10. A metric space (X, d) is *locally compact* if for every $x \in X$ and for every $r > 0$, the closure $\overline{B_r(x)}$ of the open ball $B_r(x)$ is compact. A subset $Y \subseteq X$ of a metric space (X, d) is *locally compact* if the metric subspace $(Y, d|_{Y \times Y})$ is locally compact.

One should think of a space being locally compact as allowing for the possibility that X "runs off to infinity" but nonetheless requiring that X not be "too big locally"; in particular, while $\mathbb{R}$ *is* locally compact, both $\ell^\infty(\mathbb{N})$ and $C([0,1])$ are *not* locally compact. Local compactness has some nice consequences.

Proposition 9.11. *Let $E \subseteq X$ be a locally compact subset of a metric space (X, d). Then for each $y \in X$, there exists some $z \in E$ such that $d(z, y) = \inf_{x \in E} d(x, y)$.*

We do not prove this here, since it requires some additional machinery. This result shows that the local compactness of E implies that any point of X has a closest point in E. In particular, a closed subset of $\mathbb{R}^n$ always contains a closest point to any given point in $\mathbb{R}^n$. This may *fail* for spaces that are not locally compact: there need not exist some $z \in E$ for which the infimum $\inf_{x \in E} d(x, y)$ is attained by a minimiser $d(z, y)$.

While local compactness is a good substitute for compactness in many situations, it is still weaker than compactness. Nonetheless, it is still sufficiently strong to ensure completeness.

Lemma 9.12. *Let (X, d) be a locally compact metric space. Then X is complete.*

Proof. Let (x_n) be a Cauchy sequence in X. Then (x_n) is bounded, so that there exists some $x \in X$ and $r > 0$ such that $(x_n) \subseteq \overline{B_r(x)}$. Since X is locally compact, $\overline{B_r(x)}$ is compact, and hence complete, and so there exists some $y \in \overline{B_r(x)}$ such that (x_n) converges to y in $\overline{B_r(x)}$. But then (x_n) converges to y in X, and so X is complete. □

On the other hand, if X itself is compact, then closed subsets are also compact.

Proposition 9.13. *Let (X, d) be a compact metric space and let $Y \subseteq X$ be a subset of X. Then Y is compact if and only if Y is closed in X.*

Proof. Suppose that Y is compact. Let (x_n) be a sequence in Y that converges to $x \in \overline{Y} \subseteq X$; consequently, every subsequence of (x_n) converges to x. Since Y is compact, (x_n) has a convergent subsequence that converges to some $y \in Y$; by the uniqueness of limit points, we deduce that $x = y$, and so $x \in Y$. Thus $Y = \overline{Y}$, and hence Y is closed.

Conversely, suppose that Y is closed. Let (x_n) be a sequence in $Y \subseteq X$. Since X is compact, there exists a subsequence $(x_{j(n)})$ of (x_n) that converges to $x \in X$. Necessarily, we must have that $x \in \overline{Y}$; since Y is closed, we have that $Y = \overline{Y}$, so that $x \in Y$, and hence (x_n) has a convergent subsequence in Y, which implies that Y is compact. □

This tells us that metric subspaces of compact metric spaces are compact if and only if they are closed. Product metric spaces behave even more nicely with regards to compactness.

Proposition 9.14. *Let (X, d) and (Y, ρ) be metric spaces. Then $(X \times Y, d_{X\times Y})$ is compact if and only if both X and Y are compact.*

Proof. Suppose first that X and Y are compact. Let $((x_n, y_n))$ be a sequence in $X \times Y$. Then (x_n) has a convergent subsequence $(x_{j(n)})$ and $(y_{j(n)})$ has a convergent subsequence $(y_{k(j(n))})$, and so $((x_{k(j(n))}, y_{k(j(n))})$ is convergent. Thus $X \times Y$ is compact.

Conversely, suppose that $X \times Y$ is compact and let (x_n) be a sequence in X. Fix $y \in Y$. Then the sequence $((x_n, y))$ in $X \times Y$ has a convergent subsequence $((x_{j(n)}, y))$, and so the sequence (x_n) in X has a convergent subsequence $(x_{j(n)})$. Thus X is compact; the same argument also shows that Y is compact. □

Uniformly equivalent metric spaces also behave nicely with regards to compactness.

Proposition 9.15. *Suppose that d and ρ are uniformly equivalent metrics on a set X. Then (X, d) is compact if and only if (X, ρ) is compact.*

Proof. Suppose that (X, d) is compact and let (x_n) be a sequence in X. Then (x_n) has a convergent subsequence $(x_{j(n)})$ with respect to d, so that there exists some $x \in X$ such that for all $\varepsilon > 0$, there exists $N = N(C_1\varepsilon) \in \mathbb{N}$ such that $d(x_{j(n)}, x) < C_1\varepsilon$ whenever $n \geq N(C_1\varepsilon)$. As d and ρ are uniformly equivalent,

$$\rho(x_{j(n)}, x) \leq \frac{1}{C_1} d(x_{j(n)}, x) < \varepsilon,$$

and so $(x_{j(n)})$ converges to x with respect to ρ. Thus (X, ρ) is compact.

The same argument shows that if (X, ρ) is compact, then so is (X, d). □

9.2. **Total boundedness.** We have shown that being closed and bounded is *insufficient* for a set in a metric space to be compact. To find necessary and sufficient conditions, we introduce the following notion.

Definition 9.16. A metric space (X, d) is *totally bounded* if for every $r > 0$, there exists some $N = N(r) \in \mathbb{N}$ and a corresponding finite collection of points $\{x_1, \ldots, x_N\} \subseteq X$ such that $X = \bigcup_{n=1}^{N} B_r(x_n)$. A subset $Y \subseteq X$ of a metric space (X, d) is *totally bounded* if the metric subspace $(Y, d|_{Y\times Y})$ is totally bounded.

Total boundedness simply means that given any radius $r > 0$, we can completely cover X by a finite collection of open balls of radius r. This is *stronger* than boundedness.

Lemma 9.17. *A totally bounded metric space is bounded.*

Proof. Suppose that (X, d) is totally bounded, so that there exists some $N = N(1) \in \mathbb{N}$ and a collection of points $\{x_1, \dots, x_N\} \subseteq X$ such that $X = \bigcup_{n=1}^{N} B_1(x_n)$. Then for any $x \in X$, there exists $m \in \{1, \dots, N\}$ such that $x \in B_1(x_m)$, in which case

$$d(x, x_1) \leq d(x, x_m) + d(x_m, x_1) < 1 + d(x_m, d_1) \leq 1 + \max_{1 \leq m \leq N} d(x_m, d_1).$$

Thus if we denote the right-hand side above by R, then every point $x \in X$ lies in $B_R(x_1)$, and so X is bounded. □

We showed previously that a totally bounded subset $Y \subseteq X$ is bounded. The converse *is* true when $X = \mathbb{R}^n$ with d the Euclidean metric but is *not* true in general.

Example 9.18. Let $X \coloneqq \{a \in \ell^\infty(\mathbb{N}) : \|a\|_{\ell^\infty} \leq 1\}$ be the closed unit ball centred at the origin in $\ell^\infty(\mathbb{N})$. This is bounded, since it is contained inside the open ball of radius 2. On the other hand, it is not totally bounded. To see this, for each $n \in \mathbb{N}$, let $a_n \in X$ be given by

$$a_n(k) \coloneqq \begin{cases} 1 & \text{if } k = n, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$d_\infty(a_n, a_m) \coloneqq \sup_{k \in \mathbb{N}} |a_n(k) - a_m(k)| = \begin{cases} 0 & \text{if } n = m, \\ 1 & \text{otherwise.} \end{cases}$$

So if $a \in X$, then there exists at most a single positive integer n for which $a_n \in B_{1/2}(a)$, since if $a_n, a_m \in B_{1/2}(a)$, then $d_\infty(a_n, a_m) \leq d_\infty(a_n, a) + d_\infty(a, a_m) < 1$, and hence $n = m$. Thus infinitely many balls of radius $1/2$ are needed to cover X.

As we now shall show, total boundedness is a necessary condition for compactness.

Proposition 9.19. *Let (X, d) be a compact metric space. Then X is totally bounded.*

Proof. We prove the contrapositive. If X is not totally bounded, then there exists some $r > 0$ such that $X \setminus \bigcup_{n=1}^{N} B_r(y_n) \neq \emptyset$ for every finite collection of points $\{y_1, \dots, y_N\} \subseteq X$. We now construct a sequence (x_n) in X by choosing $x_1 \in X$, then taking $x_2 \in X \setminus B_r(x_1)$, and inductively $x_n \in X \setminus \bigcup_{m=1}^{n-1} B_r(x_m)$; such an element $x_n \in X$ exists for all $n \in \mathbb{N}$ since X is not totally bounded. The sequence (x_n) satisfies $d(x_n, x_m) \geq r$ for every $m, n \in \mathbb{N}$ with $m \neq n$, since if $m > n$ then $x_m \notin B_r(x_n)$. Thus (x_n) has no convergent subsequence, and so X is not compact. □

Finally, we arrive at necessary and sufficient conditions for a metric space to be compact.

Theorem 9.20. *A metric space (X, d) is compact if and only if it is complete and totally bounded.*

This is the natural generalisation of the Heine–Borel theorem to the setting of metric spaces; in particular, it is a useful way to check whether a space is compact. Since a subset of a complete metric space is complete if and only if it is closed, we have the following corollary.

Corollary 9.21. *A subset $Y \subseteq X$ of a complete metric space (X, d) is compact if and only if it is closed and totally bounded.*

Proof of Theorem 9.20. We have already shown that a compact metric space (X, d) is complete and totally bounded. Conversely, let (X, d) be complete and totally bounded, and let (x_n) be a sequence in X. We shall show that (x_n) has a Cauchy subsequence via the total boundedness of X; since X is complete, this subsequence is therefore convergent, and so X is compact.

Since X is totally bounded, we may take $r = 2^{-k}$ for each $k \in \mathbb{N}$, so that for each $k \in \mathbb{N}$, there exists a finite collection of points $\{y_{k,1}, \dots, y_{k,N(k)}\} \subseteq X$ such that $X = \bigcup_{\ell=1}^{N(k)} B_{2^{-k}}(y_{k,\ell})$. We now construct a subsequence of (x_n) as follows. First, there exists $\ell_1 \in \{1, \dots, N(1)\}$ such that infinitely many terms of the sequence (x_n) lie in $B_{2^{-1}}(y_{1,\ell_1})$, since there are infinitely many terms in this sequence and only finitely many balls $B_{2^{-1}}(y_{1,\ell})$ that cover X. We choose

$$j(1) \coloneqq \min\{k \in \mathbb{N} : x_k \in B_{2^{-1}}(y_{1,\ell_1})\},$$

so that $x_{j(1)} \in B_{2^{-1}}(y_{1,\ell_1})$. Similarly, there exists $\ell_2 \in \{1, \dots, N(2)\}$ such that infinitely many elements of the sequence (x_n) that lie in $B_{2^{-1}}(y_{1,\ell_1})$ also lie in $B_{2^{-2}}(y_{2,\ell_2})$, since again there are infinitely many terms in this sequence and only finitely many balls that cover X, and in particular cover $B_{2^{-1}}(y_{1,\ell_1})$. We choose

$$j(2) \coloneqq \min\{k > j(1) : x_k \in B_{2^{-1}}(y_{1,\ell_1}) \cap B_{2^{-2}}(y_{2,\ell_2})\},$$

so that $x_{j(2)} \in B_{2^{-1}}(y_{1,\ell_1}) \cap B_{2^{-2}}(y_{2,\ell_2})$.

In this way, we inductively deduce the existence of $\ell_n \in \{1, \dots, N(n)\}$ such that infinitely many elements of (x_n) lie in $\bigcap_{m=1}^{n} B_{2^{-m}}(y_{m,\ell_m})$, and we choose

$$j(n) \coloneqq \min\{k > j(n-1) : x_k \in \bigcap_{m=1}^{n} B_{2^{-m}}(y_{m,\ell_m})\}.$$

This subsequence is such that if $m > n$, then $x_{j(m)}, x_{j(n)} \in B_{2^{-n}}(y_{n,\ell_n})$, and so

$$d(x_{j(m)}, x_{j(n)}) \leq d(x_{j(m)}, y_{n,\ell_n}) + d(y_{n,\ell_n}, x_{j(n)}) < 2^{-n} + 2^{-n} = 2^{1-n}.$$

Thus this subsequence is Cauchy, since for all $\varepsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$ given by the least integer for which $2^{1-n} \leq \varepsilon$ such that $d(x_{j(m)}, x_{j(n)}) < \varepsilon$ whenever $m > n \geq N(\varepsilon)$. □

9.3. **Topological compactness.** Up until now, we have been working with the notion of *sequential compactness*, which concerns the existence of convergent subsequences. There is an alternative definition of compactness that does not involve subsequences at all. Instead, this definition is *topological*: it uses only notions involving open sets and does not invoke properties of the metric d.

In practice, this topological definition of compactness is more useful, since it remains valid in the more general setting of *topological spaces*, which is a generalisation of metric spaces to spaces that do not necessarily have a metric but nonetheless have a complete classification of open and closed sets. On the other hand, this topological definition of compactness may be somewhat less intuitive than sequential compactness.

We first require the following definitions.

Definition 9.22. Let (X, d) be a metric space. An *open cover* of X is a collection $\mathcal{F}$ of open subsets $E \subseteq X$ such that $\bigcup_{E \in \mathcal{F}} E = X$. A *subcover* of an open cover $\mathcal{F}$ is a subset of $\mathcal{F}$ that is also an open cover of X.

Each element of an open cover $\mathcal{F}$ is an open set $E \subseteq X$. An open cover need not be a *finite* collection of open sets; indeed, it need not even be *countable*! Similarly, a subcover of an open cover need not be countable, let alone finite.

Definition 9.23. A metric space (X, d) is *topologically compact* if every open cover of X has a finite subcover. A subset $Y \subseteq X$ of a metric space (X, d) is *topologically compact* if the metric subspace $(Y, d|_{Y \times Y})$ is topologically compact.

Thus topological compactness means that if we cover X by some collection of open sets, we can always pick a finite subcollection of these open sets that still covers X. A priori, this notion seems entirely distinct from the notion of sequential compactness. Nonetheless, we shall prove the following.

Theorem 9.24. *A metric space (X, d) is sequentially compact if and only if it is topologically compact.*

These notions of compactness for metric spaces are therefore *equivalent*. This is a useful result, since it is often more straightforward to prove that a metric space is topological compact compared to proving that it is sequentially compact.

The proof of Theorem 9.24 proceeds in several parts. That sequential compactness implies topological compactness can be proven via the following lemma, which is known as the *Lebesgue covering lemma*.

Lemma 9.25. *Let (X, d) be a sequentially compact metric space, and let $\mathcal{F}$ be an open cover of X. There exists some $r = r(\mathcal{F}) > 0$ such that for every $x \in X$, there exists some $E \in \mathcal{F}$ for which $B_r(x) \subseteq E$.*

Proof. Suppose in order to obtain a contradiction that X is sequentially compact and that $\mathcal{F}$ is an open cover of X but that for all $r > 0$, there exists some $x \in X$ such that $B_r(x) \cap E^c \neq \emptyset$ for every open set $E \in \mathcal{F}$.

We take $r = 2^{-n}$ for every $n \in \mathbb{N}$, so that there exists some $x_n \in X$ for which $B_{2^{-n}}(x_n) \cap E^c \neq \emptyset$ for each $E \in \mathcal{F}$. Since X is sequentially compact, the sequence (x_n) has a subsequence $(x_{j(n)})$ that converges to some $x \in X$, so that for all $\varepsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$ such that $d(x_{j(n)}, x) < \varepsilon$ whenever $n \geq N(\varepsilon)$. Since $\mathcal{F}$ is an open cover of X, there exists some $E \in \mathcal{F}$ such that $x \in E$, and since E is open, there exists some $r > 0$ such that $B_r(x) \subseteq E$.

On the other hand, if $n \in \mathbb{N}$ is such that $n \geq \max\{N(\frac{r}{2}), 1 - \frac{\log r}{\log 2}\}$, then for each $y \in B_{2^{-j(n)}}(x_{j(n)})$, we have that

$$d(y, x) \leq d(y, x_{j(n)}) + d(x_{j(n)}, x) < 2^{-j(n)} + \frac{r}{2} \leq 2^{-n} + \frac{r}{2} \leq r,$$

and so $y \in B_r(x)$. It follows that $B_{2^{-j(n)}}(x_{j(n)}) \subseteq B_r(x) \subseteq E$, which contradicts the assumption that $B_{2^{-n}}(x_n) \cap E^c \neq \emptyset$ for each $E \in \mathcal{F}$. □

Corollary 9.26. *A sequentially compact metric space is topologically compact.*

Proof. Let (X, d) be sequentially compact and let $\mathcal{F}$ be an open cover of X. By the Lebesgue covering lemma, there exists some $r > 0$ such that for every $x \in X$, there exists some $E \in \mathcal{F}$ for which $B_r(x) \subseteq E$. The metric space X is totally bounded since it is sequentially compact, so that there exists a finite collection $\{x_1, \ldots, x_N\}$ of points in X that are such that $\bigcup_{n=1}^{N} B_r(x_n) = X$. Then for each $n \in \{1, \ldots, N\}$, there exists some $E_n \in \mathcal{F}$ for which $B_r(x_n) \subseteq E_n$ by the Lebesgue covering lemma, and hence

$$\bigcup_{n=1}^{N} E_n \supseteq \bigcup_{n=1}^{N} B_r(x_n) = X.$$

Thus $\{E_1, \ldots, E_N\}$ is a finite subcover of X, and so X is topologically compact. □

To prove that topological compactness implies sequential compactness, we use the fact that sequential compactness is equivalent to completeness and totally boundedness. We first show that topological compactness implies completeness.

Proposition 9.27. *A topologically compact metric space (X, d) is complete.*

Proof. We prove the contrapositive. Suppose that X is not complete. Let $(\overline{X}, \overline{d})$ be the completion of (X, d) via an isometry $\imath : X \to \overline{X}$. Since X is not complete, this isometry is not surjective, so that there exists some $x \in \overline{X} \setminus \imath(X)$. Fix such an $x \in \overline{X} \setminus \imath(X)$, and for each $n \in \mathbb{N}$, define $E_n \coloneqq \{y \in X : \overline{d}(\imath(y), x) > 2^{-n}\}$.

We first prove that E_n is open for each $n \in \mathbb{N}$. Let $y \in E_n$, so that $R \coloneqq \overline{d}(\imath(y), x) - 2^{-n} > 0$. If $z \in B_R(y)$, then by the triangle inequality and the fact that $\imath$ is an isometry,

$$\begin{aligned}\overline{d}(\imath(z), x) &\geq \overline{d}(\imath(y), x) - \overline{d}(\imath(y), \imath(z)) \\ &= 2^{-n} + R - d(y, z) \\ &> 2^{-n} + R - R \\ &= 2^{-n},\end{aligned}$$

and so $z \in E_n$. Thus $B_R(y) \subseteq E_n$, which implies that E_n is open.

Next, we observe that

$$\bigcup_{n=1}^{\infty} E_n = \{y \in X : \overline{d}(\imath(y), x) > 0\} = \{y \in X : \imath(y) \neq x\} = X$$

since $x \notin \imath(X)$, and so the collection $\mathcal{F} \coloneqq \{E_n : n \in \mathbb{N}\}$ is an open cover of X.

Finally, we show that any finite subset $\{E_{n_1}, \dots, E_{n_N}\}$ of $\mathcal{F}$ cannot cover X. Indeed, for $M \coloneqq \max\{n_1, \dots, n_N\}$, we have that

$$\bigcup_{m=1}^{N} E_{n_m} = E_M = \left\{y \in X : \overline{d}(\imath(y), x) > 2^{-M}\right\}$$

since $E_n \supseteq E_m$ whenever $n \geq m$. On the other hand, $\overline{X}$ being the completion of X means that $\imath(X)$ is dense in $\overline{X}$, and hence for each $\varepsilon > 0$ and each $x \in \overline{X}$, there exists $y \in X$ such that $\overline{d}(\imath(y), x) < \varepsilon$; taking $\varepsilon = 2^{-M}$, we thereby see that there exists some $y \in X$ for which $y \notin E_M = \bigcup_{m=1}^{N} E_{n_m}$. Thus no finite subset $\{E_{n_1}, \dots, E_{n_N}\}$ of $\mathcal{F}$ can cover X, and hence X is not topologically compact. □

To complete the proof of Theorem 9.24, we must show that topological compactness implies total boundedness.

Proposition 9.28. *A topologically compact metric space (X, d) is totally bounded.*

Proof. Fix $r > 0$, and let $\mathcal{F} \coloneqq \{B_r(x) : x \in X\}$. This is an open cover of X, and so there exists a finite subcover, which is to say a finite collection of points $x_1, \dots, x_N \in X$ such that $X = \bigcup_{n=1}^{N} B_r(x_n)$. Since such a collection of points exists for each $r > 0$, we deduce that X is totally bounded. □

Since completeness and total boundedness implies sequential compactness, we deduce the following.

Corollary 9.29. *A topologically compact metric space is sequentially compact.*

This completes the proof of Theorem 9.24, and so we are now free to use the notion of compactness freely without first specifying topological compactness versus sequential compactness.

10. Continuity

Recommended reading: [Pug15, §2.2], [Tao16, §2.1–2.4].

10.1. **Continuity, limits, and sequences.** The definitions of continuity and of limits are very similar to those for real functions: they encapsulate the idea that the images of nearby points are nearby.

Definition 10.1. Let (X, d) and (Y, ρ) be metric spaces. A function $f : X \to Y$ has *limit $z \in Y$ as y tends to x*, which we denote by $\lim_{y\to x} f(y) = z$, if for every $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon, x) > 0$ such $\rho(f(y), z) < \varepsilon$ for all $y \in X$ with $0 < d(y, x) < \delta$.

A function $f : X \to Y$ is *continuous at $x \in X$* if $\lim_{y\to x} f(y) = f(x)$, so that for every $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon, x) > 0$ such $\rho(f(y), p) < \varepsilon$ for all $y \in X$ with $d(y, x) < \delta$.

A function $f : X \to Y$ is *continuous* if it is continuous at x for every $x \in X$.

Note that these definitions can be described in terms of balls. We have that $\lim_{y\to x} f(y) = z$ if for every $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon, x) > 0$ such $f(y) \in B_\varepsilon(z)$ for all $y \in B_\delta(x)\setminus\{x\}$. A function $f : X \to Y$ is continuous at $x \in X$ if for every $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon, x) > 0$ such $f(y) \in B_\varepsilon(z)$ for all $y \in B_\delta(x)$.

These definitions seem straightforward, but some caution should be exercised: things are not always as well-behaved as in the case of real functions of one variable.

Example 10.2. Consider the case where ρ is the discrete metric on Y. Then the only functions that are continuous at $x \in X$ are those that are constant on some ball centred at x, since there must exist some $\delta = \delta(1, x) > 0$ such that $y \in B_\delta(x)$ implies that $f(y) \in B_1(f(x))$, which means that $f(y) = f(x)$.

Example 10.3. Constant functions are of course continuous for any X and Y.

Example 10.4. Consider the case where d is the discrete metric on X. Then *every* function f is continuous, since for any $\varepsilon > 0$, we can choose $\delta = \delta(\varepsilon, x) = 1$: for any $y \in B_1(x)$, we have $y = x$, and hence $f(y) = f(x) \in B_\varepsilon(f(x))$.

Bizarrely, we also have that $\lim_{y\to x} f(y) = z$ for *every* $z \in Y$: for any $\varepsilon > 0$, we have $B_1(x) \setminus \{x\} = \emptyset$, so it is vacuously true that $y \in B_1(x) \setminus \{x\}$ implies that $f(y) \in B_\varepsilon(z)$ for any $z \in Y$.

More generally, even if d is not the discrete metric on X, suppose that $x \in X$ is an isolated point, so that there exists some $r > 0$ such that $B_r(x) = \{x\}$. Then it remains true that every function f is continuous at x and that $\lim_{y\to x} f(y) = z$ for every $z \in Y$.

This example shows that it only really makes sense to consider limits of functions at points that are not isolated — equivalently, at points that are limit points of X — so that $B_r(x) \setminus \{x\} \neq \emptyset$ for all $r > 0$.

Proposition 10.5. *Let x be a limit point of X, and let $f : X \to Y$ be a function. Then the limit of f at x is unique.*

Proof. Suppose that $\lim_{y\to x} f(y) = z_1$ and $\lim_{y\to x} f(y) = z_2$. Then for all $\varepsilon > 0$, there exists $\delta_1 = \delta(\varepsilon, z_1) > 0$ such that $f(y) \in B_{\frac{\varepsilon}{2}}(z_1)$ whenever $y \in B_{\delta_1}(x)\setminus\{x\}$ and there exists $\delta_2 = \delta(\varepsilon, z_2) > 0$ such that $f(y) \in B_{\frac{\varepsilon}{2}}(z_2)$ whenever $y \in B_{\delta_2}(x) \setminus \{x\}$. Since x is a limit point, upon setting $\delta \coloneqq \min\{\delta_1, \delta\} > 0$, we see that there exists some $y \in B_\delta(x) \setminus \{x\}$, and so $f(y) \in B_{\frac{\varepsilon}{2}}(z_1) \cap B_{\frac{\varepsilon}{2}}(z_2)$. By the triangle inequality,

$$\rho(z_1, z_2) \leq \rho(z_1, f(y)) + \rho(f(y), z_2) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, we deduce that $\rho(z_1, z_2) = 0$, so that $z_1 = z_2$. □

It is useful to have an alternative characterisation of continuity in terms of sequences.

Proposition 10.6. *Let $f : X \to Y$ be a function. Then f is continuous at $x \in X$ if and only if whenever (x_n) is a sequence in X that converges to x, the sequence $(f(x_n))$ in Y is convergent to $f(x)$.*

Proof. If f is continuous at x, then for every $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon, x) > 0$ such that $y \in B_\delta(x)$ implies that $f(y) \in B_\varepsilon(f(x))$. If (x_n) is a sequence that converges to $x \in X$, then for every $\varepsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$ such that $n \geq N$ implies $d(x_n, x) < \varepsilon$. So if $n \geq N(\delta(\varepsilon))$, we have that $d(x_n, x) < \delta$, so that $\rho(f(x_n), f(x)) < \varepsilon$, and hence $f(x_n)$ converges to $f(x)$.

For the converse, we prove the contrapositive. Suppose that f is not continuous at x, so that there exists some $\varepsilon > 0$ such that for all $\delta > 0$, there exists $y \in B_\delta(x)$ such that $\rho(f(y), f(x)) \geq \varepsilon$. We take $\delta = 2^{-n}$ for each $n \in \mathbb{N}$, so that there exists $x_n \in B_{2^{-n}}(x)$ such that $\rho(f(x_n), f(x)) \geq \varepsilon$. Then the sequence (x_n) in X converges to $x \in X$ but the sequence $(f(x_n))$ in Y does not converge to $f(x) \in Y$. □

10.2. **Continuity and open and closed sets.** Now we want to look at another important way of thinking about continuous functions.

Proposition 10.7. *A function $f : X \to Y$ is continuous if and only if whenever $U \subseteq Y$ is an open subset of Y, $f^{-1}(U) \coloneqq \{x \in X : f(x) \in U\} \subseteq X$ is an open subset of X. Equivalently, f is continuous if and only if whenever $V \subseteq Y$ is closed in Y, $f^{-1}(V) \subseteq X$ is closed in X.*

Proof. For the former, suppose that f is continuous, and let $U \subseteq Y$ be open. Let $x \in f^{-1}(U) \subseteq X$, so that $f(x) \in U$. Since U is open, there exists some $r > 0$ such that $B_r(f(x)) \subseteq U$. Since f is continuous at x, there exists some $\delta > 0$ such that $f(y) \in B_r(f(x))$ whenever $y \in B_\delta(x)$. Thus $f(y) \in U$, and so $B_\delta(x) \subseteq f^{-1}(U)$, which means that $f^{-1}(U)$ is open.

Conversely, suppose that $f^{-1}(U)$ is open in X whenever U is open in Y. Then for any $x \in X$ and for all $\varepsilon > 0$, we may take $U = B_\varepsilon(f(x))$, so that $f^{-1}(U)$ is open and contains x. Since $f^{-1}(U)$ is open, there exists some $\delta > 0$ such that $B_\delta(x) \subseteq f^{-1}(U)$, and so if $y \in B_\delta(x)$ then $f(y) \in U = B_\varepsilon(f(x))$, which is precisely the definition of continuity.

The latter follows from the former since V being closed is equivalent to $Y \setminus V =: U$ being closed, and $f^{-1}(V)$ being closed is equivalent to $X \setminus f^{-1}(V) = f^{-1}(U)$ being open. □

The nice thing about this way of thinking about continuity is that there is no explicit mention of the distance function, only of the open sets. Once we know which sets are open, we know which functions are continuous. In particular, any two metrics on the same space which have the same open sets also have the same continuous functions.

Example 10.8. Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) \coloneqq \begin{cases} 0 & \text{if } x \leq 0, \\ 1 & \text{if } x > 0. \end{cases}$$

Then if $U = (-1, 1)$, we have that $f^{-1}(U) = (-\infty, 0]$; since U is open but $f^{-1}(U)$ is not, we deduce that f is not continuous.

Note that if $f : X \to Y$ is continuous, then given an open set $E \subseteq X$, it need not be the case that $f(E) \subseteq Y$ is open! Thus it is important that the definition of continuity involving open sets involves the *preimage* $f^{-1}(U)$ rather than the *image* $f(E)$.

Example 10.9. The function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = \sin x$ is continuous, but if we take $E = (0, 2\pi)$, then $f(E) = [-1, 1]$: E is open but $f(E)$ is not.

10.3. **Continuity and compactness.** The following is a fundamental result relating compactness and continuity.

Proposition 10.10. *Let $f : X \to Y$ be continuous, and suppose that (X, d) is compact. Then $f(X) \coloneqq \{y \in Y : y = f(x) \text{ for some } x \in X\} \subseteq Y$ is a compact metric subspace of Y.*

Proof. Let (y_n) be a sequence in $f(X)$, so that there exists a sequence (x_n) in X with $f(x_n) = y_n$. Since X is compact, there exists a convergent subsequence $(x_{j(n)})$ that converges to some $x \in X$, and since f is continuous, $(f(x_{j(n)})) = (y_{j(n)})$ converges to $f(x) \in f(X)$. Therefore (y_n) has a convergent subsequence, and so $f(X)$ is compact. □

This is a very simple observation, but it has some very useful consequences. One such consequence is a generalisation to compact metric spaces the result that continuous functions achieve their infimum and supremum on closed bounded intervals.

Proposition 10.11. *Let $f : X \to \mathbb{R}$ be continuous, and suppose that (X, d) is compact. Then there exists $x_+, x_- \in X$ such that $f(x_+) = \sup_{x \in X} f(x)$ and $f(x_-) = \inf_{x \in X} f(x)$. In particular, f is bounded.*

This is sometimes known as the *maximum principle*.

Proof. Since X is compact and f is continuous, $f(X) \subset \mathbb{R}$ is compact. Since $\sup\{y : y \in f(X)\} = \sup\{f(x) : x \in X\}$ is a limit point of the compact set $f(X) \subset \mathbb{R}$, which is closed and bounded, we must have that $\sup\{f(x) : x \in X\} \in f(X)$, so that there exists some $x_+ \in X$ for which $f(x_+)$ is equal to this supremum. The same argument implies the existence of $x_- \in X$ for which $f(x_-) = \inf_{x \in X} f(x)$. □

10.4. **Continuity and restrictions, compositions, product spaces, and uniformly equivalent metrics.** Continuity is a property that is preserved under many operations. A first such operation is restriction to metric subspaces.

Proposition 10.12. *Let $f : X \to Y$ be continuous at $x \in X$. Then if $E \subseteq X$ is a nonempty subset of X containing x, the restriction $f|_E : E \to Y$ of f from X to E is continuous at $x \in E$. In particular, the restriction of a continuous function to a nonempty set is continuous.*

Proof. Since $f : X \to Y$ is continuous at x, for all $\varepsilon > 0$, there exists $\delta > 0$ such that if $y \in B_\delta(x)$ then $f(y) \in B_\varepsilon(f(y))$. In particular, if $y \in B_\delta^E(x) = \{y \in E : d(x, y) < \delta\} \subseteq \{y \in X : d(x, y) < \delta\} = B_\delta(x)$, then $f(y) \in B_\varepsilon(f(y))$, and so $f|_E : E \to Y$ is continuous at x. □

Continuity is also preserved under composition.

Proposition 10.13. *Let X, Y, Z be metric spaces, and let $f : X \to Y$ be continuous at $x \in X$ and $g : Y \to Z$ be continuous at $f(x) \in Y$. Then $g \circ f : X \to Z$ is continuous at $x \in X$. In particular, compositions of continuous functions are continuous.*

Proof. Since g is continuous at $f(x) \in Y$, for every $\varepsilon > 0$, there exists some $\delta_2 = \delta_2(\varepsilon, f(x)) > 0$ such that $g(z) \in B_\varepsilon(g(f(x)))$ whenever $z \in B_{\delta_2}(f(x))$. Then since f is continuous at $y \in X$, by taking $\varepsilon = \delta_2$, we see that there exists some $\delta_1 = \delta_1(\delta_2, x) > 0$ such that $f(y) \in B_{\delta_2}(f(x))$ whenever $y \in B_{\delta_1}(x)$. Taking $z = f(y)$, we deduce that $g \circ f$ is continuous at x, since for all $\varepsilon > 0$, there exists $\delta_1 > 0$ such that $g(f(y)) \in B_\varepsilon(g(f(x)))$ whenever $y \in B_{\delta_1}(x)$. □

Note that the fact that compositions of continuous functions are continuous has a short proof via the definition of continuity in terms of open sets. Given $U \subseteq Z$, we have that

$$(g \circ f)^{-1}(U) = \{x \in X : g(f(x)) \in U\} = \{x \in X : f(x) \in g^{-1}(U)\} = f^{-1}(g^{-1}(U)).$$

If U is open in Z, then g being continuous implies that $g^{-1}(U)$ is open in Y, and f being continuous implies that $f^{-1}(g^{-1}(U)) = (g \circ f)^{-1}(U)$ is open in X, which means that $g \circ f$ is continuous.

Similarly, continuity is preserved for product metrics. Recall our definition of the metric space structure on the product of two metric spaces: if (X, d) and (Y, ρ) are metric spaces, then $X \times Y$ is a metric space with the distance function $d_{X\times Y}((x_1, y_1), (x_2, y_2)) \coloneqq d(x_1, x_2) + \rho(y_1, y_2)$.

Proposition 10.14. *The projections $\pi_1 : X \times Y \to X$ and $\pi_2 : X \times Y \to Y$ given by $\pi_1(x, y) = x$ and $\pi_2(x, y) = y$ for all $(x, y) \in X \times Y$ are continuous.*

Proof. We take $\delta = \varepsilon$ to see that if $d_{X\times Y}((x_1, y_1), (x_2, y_2)) < \delta = \varepsilon$, then

$$d(\pi_1(x_1, y_1), \pi_1(x_2, y_2)) = d(x_1, x_2) \leq d(x_1, x_2) + \rho(y_1, y_2) = d_{X\times Y}((x_1, y_1), (x_2, y_2)) = \varepsilon.$$

□

We can use this to determine continuity of functions from a metric space to a product metric space.

Corollary 10.15. *A function $f : X \to Y \times Z$ is continuous at $x \in X$ if and only if both $f_1 : X \to Y$ and $f_2 : X \to Z$ are continuous at $x \in X$, where $f(x) = (f_1(x), f_2(x))$.*

Proof. If f is continuous at $x \in X$, then $f_1 = \pi_1 \circ f$ and $f_2 = \pi_2 \circ f$ are compositions of continuous functions at x, and hence continuous at x.

Conversely, suppose that f_1 and f_2 are both continuous at x. Then for any sequence (x_n) in X converging to $x \in X$, the sequence $(f_1(x_n))$ in Y converges to $f_1(x) \in Y$ and the sequence $(f_2(x_n))$ in Z converges to $(f_2(x)) \in Z$. This ensures that the sequence $((f_1(x_n), f_2(x_n))) = (f(x_n))$ in $Y \times Z$ converges to $(f_1(x), f_2(x)) = f(x) \in Y \times Z$, and hence f is continuous at x. □

A useful observation is that the distance function itself is a continuous function.

Proposition 10.16. *Let $d : X \times X \to \mathbb{R}$ be a metric on a space X. Then d is continuous.*

Proof. We must show that for all $(x_1, x_2) \in X \times X$ and for all $\varepsilon > 0$, there exists some $\delta = \delta(\varepsilon, x_1, x_2) > 0$ such that $|d(x_1, x_2) - d(y_1, y_2)| < \varepsilon$ whenever $d_{X\times X}((x_1, x_2), (y_1, y_2)) < \delta$. Taking $\delta = \varepsilon$, we see that by the triangle inequality,

$$\begin{aligned} d(x_1, x_2) - d(y_1, y_2) &\leq d(x_1, y_1) + d(y_1, y_2) + d(y_2, x_2) - d(y_1, y_2) \\ &= d(x_1, y_1) + d(x_2, y_2) \\ &= d_{X\times X}((x_1, x_2), (y_1, y_2)), \\ d(y_1, y_2) - d(x_1, x_2) &\leq d(y_1, x_1) + d(x_1, x_2) + d(x_2, y_2) - d(x_1, x_2) \\ &= d(x_1, y_1) + d(x_2, y_2) \\ &= d_{X\times X}((x_1, x_2), (y_1, y_2)). \end{aligned}$$

Thus $|d(x_1, x_2) - d(y_1, y_2)| < \varepsilon$. □

We move on to uniformly equivalent metrics. For this, we have the following observation.

Proposition 10.17. *Suppose that d_1 and d_2 are uniformly equivalent metrics on a space X and ρ_1 and ρ_2 are uniformly equivalent metrics on a space Y. Then a function $f : X \to Y$ is continuous at $x \in X$ with respect to the metric d_1 on X and ρ_1 on Y if and only if it is also continuous at x with respect to d_2 and ρ_2.*

Proof. Uniform equivalence means that there exists $C_1, C_2 > 0$ such that

$$C_1 d_1(x_1, x_2) \leq d_2(x_1, x_2) \leq C_2 d_1(x_1, x_2)$$

for all $x_1, x_2 \in X$ and $D_1, D_2 > 0$ such that

$$D_1 \rho_1(y_1, y_2) \leq \rho_2(y_1, y_2) \leq D_2 \rho_1(y_1, y_2)$$

for all $y_1, y_2 \in Y$. Now if f is continuous at x with respect to d_1 and ρ_1, then for all $\varepsilon > 0$, there exists $\delta_1 = \delta_1(\frac{\varepsilon}{D_2}, x)$ such that $\rho_1(f(x_1), f(x_2)) < \frac{\varepsilon}{D_2}$ whenever $d_1(x_1, x_2) < \delta_1$. Taking $\delta_2 = C_1\delta_1$, we see that $\rho_2(f(x_1), f(x_2)) < \varepsilon$ whenever $d_2(x_1, x_2) < \delta_2$, and so f is continuous at x with respect to d_2 and ρ_2.

The same argument shows if f is continuous at x with respect to d_2 and ρ_2, then it is also continuous at x with respect to d_1 and ρ_1. □

Note also that if instead of uniform equivalence, we assume only one of the inequalities holds in each case, so that $d_1 \leq \frac{1}{C_1}d_2$ and $\rho_2 \leq D_2\rho_1$, then the proof above goes through, so that a function continuous at x with respect to d_1 and ρ_1 is necessarily continuous at x with respect to d_2 and ρ_2.

A consequence of this result is that functions to or from $\mathbb{R}^n$ are continuous with respect to the taxi-cab metric (or ℓ^1-metric) if and only if they are continuous with respect to the Euclidean metric (or ℓ^2-metric).

There are some other nice consequences in the particular case when $Y = \mathbb{R}$.

Proposition 10.18. *Let $f, g : X \to \mathbb{R}$ be continuous functions. Then the sum $f + g$ and the product fg are also continuous.*

Proof. To show that $f + g : X \to \mathbb{R}$ is continuous, it suffices to show that $h_1 : \mathbb{R}^2 \to \mathbb{R}$ given by $h_1(x, y) \coloneqq x + y$ is continuous, since $f + g$ is the composition of the map $(f, g) : X \times Y \to \mathbb{R}^2$ and $h_1 : \mathbb{R}^2 \to \mathbb{R}$. Similarly, in order to show that $fg : X \to \mathbb{R}$ is continuous, we must show that $h_2 : \mathbb{R}^2 \to \mathbb{R}$ given by $h_2(x, y) \coloneqq xy$ is continuous, since fg is the composition of $(f, g) : X \times Y \to \mathbb{R}^2$ and $h_2 : \mathbb{R}^2 \to \mathbb{R}$.

For the former, we have that if $(x_1, x_2), (y_1, y_2) \in \mathbb{R}^2$, then by taking $\delta = \varepsilon$, we see that if $d_1((x_1, x_2), (y_1, y_2)) < \delta$, then

$$\begin{aligned} d(h_1(x_1, x_2), h_1(y_1, y_2)) &= |x_1 + x_2 - y_1 - y_2| \\ &\leq |x_1 - y_1| + |x_2 - y_2| \\ &= d_1((x_1, x_2), (y_1, y_2)) \\ &< \varepsilon. \end{aligned}$$

For the latter, we take $\delta = \min\{1, \frac{\varepsilon}{1/2+|x_1|+|x_2|}\}$ to see that if $d_1((x_1, x_2), (y_1, y_2)) < \delta$, then

$$\begin{aligned} d(h_2(x_1, x_2), h_2(y_1, y_2)) &= |x_1x_2 - y_1y_2| \\ &= |x_1(x_2 - y_2) - x_2(y_1 - x_1) - (y_1 - x_1)(y_2 - x_2)| \\ &\leq |x_1||x_2 - y_2| + |x_2||y_1 - x_1| + |y_1 - x_1||y_2 - x_2| \\ &\leq (|x_1| + |x_2|)(|y_1 - x_1| + |y_2 - x_2|) + \frac{1}{2}(|y_1 - x_1| + |y_2 - x_2|)^2 \\ &= (|x_1| + |x_2|)d_1((x_1, x_2), (y_1, y_2)) + \frac{1}{2}d_1((x_1, x_2), (y_1, y_2))^2 \\ &< \varepsilon. \end{aligned}$$

□

10.5. **Uniform continuity.** Recall that a function $f : X \to Y$ is continuous if for all $x \in X$ and $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon, x) > 0$ such that $f(y) \in B_\varepsilon(f(x))$ whenever $y \in B_\delta(x)$.

Definition 10.19. A function $f : X \to Y$ is *uniformly continuous* if for all $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that for all $x \in X$, we have that $f(y) \in B_\varepsilon(f(x))$ whenever $y \in B_\delta(x)$.

This looks a lot like the definition of continuity, except that δ depends only on ε and not on x. Uniform continuity implies continuity, but the converse does not hold. It is important to realise that uniform continuity is not the same as just continuity at every point: it is *stronger* than continuity. If $X = \mathbb{R}$ and f is *differentiable*, one can think of uniform continuity as being implied by f' being *bounded* on all of $\mathbb{R}$.

Example 10.20. The function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$ is continuous at every point, but is not uniformly continuous, since $|f(y) - f(x)| = |y - x||x + y|$ always approaches infinity as $x \to \pm\infty$ whenever $|y - x|$ is fixed, and so is not bounded by ε for any fixed $|y - x|$.

Nonetheless, when the metric space X is compact, continuity and uniform continuity are equivalent.

Proposition 10.21. *Let (X, d) and (Y, ρ) be metric spaces, and suppose that X is compact. Then a function $f : X \to Y$ is continuous if and only if it is uniformly continuous.*

Proof. If f is uniformly continuous, then it is also continuous.

Conversely, suppose that f is continuous, and fix $\varepsilon > 0$. For each $x \in X$, f is continuous at x, and so there exists some $\delta = \delta(\frac{\varepsilon}{2}, x) > 0$ such that $y \in B_{\delta(\frac{\varepsilon}{2},x)}(x)$ implies that $f(y) \in B_{\frac{\varepsilon}{2}}(f(x))$. In particular, this remains true if $y \in B_{\frac{1}{2}\delta(\frac{\varepsilon}{2},x)}(x)$.

Now consider the collection of balls $\{B_{\frac{1}{2}\delta(\frac{\varepsilon}{2},x)}(x) : x \in X\}$. This is an open cover of X as $x \in B_{\frac{1}{2}\delta(\frac{\varepsilon}{2},x)}(x)$ for each $x \in X$, and since X is compact, there exists a finite subcover, which is to say a finite collection of points $\{x_1, \dots, x_N\} \subseteq X$ such that $\bigcup_{n=1}^{N} B_{\frac{1}{2}\delta(\frac{\varepsilon}{2},x_n)}(x_n) = X$.

Let $\delta \coloneqq \min_{1\le n\le N} \frac{1}{2}\delta(\frac{\varepsilon}{2}, x_n) > 0$. For each $x \in X$, there exists some $n \in \{1, \dots, N\}$ such that $x \in B_{\frac{1}{2}\delta(\frac{\varepsilon}{2},x_n)}(x_n)$, so that $f(x) \in B_{\frac{\varepsilon}{2}}(f(x_n))$. If $y \in B_\delta(x)$, so that $d(y, x) < \delta$, then

$$d(y, x_n) \le d(y, x) + d(x, x_n) < \delta + \frac{1}{2}\delta\left(\frac{\varepsilon}{2}, x_n\right) < \delta\left(\frac{\varepsilon}{2}, x_n\right),$$

so that $y \in B_{\delta(\frac{\varepsilon}{2},x_n)}(x_n)$ and hence $f(y) \in B_{\frac{\varepsilon}{2}}(f(x_n))$. But then

$$\rho(f(y), f(x)) \le \rho(f(y), f(x_n)) + \rho(f(x_n), f(x)) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Thus for all $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) = \min_{1\le n\le N} \frac{1}{2}\delta(\frac{\varepsilon}{2}, x_n) > 0$ such that for all $x \in X$, $y \in B_\varepsilon(f(x))$ whenever $y \in B_\delta(x)$. It follows that f is uniformly continuous. □

11. Connectedness

Recommended reading: [Pug15, §2.5], [Tao16, §2.4].

11.1. **Connected and disconnected sets.** Before we discuss further properties of continuity, especially with regards to sequences of continuous functions, we move on to one final concept for subsets of metric spaces.

Definition 11.1. A metric space (X, d) is *disconnected* if there exist disjoint nonempty open sets $E_1, E_2 \subseteq X$ such that $E_1 \cup E_2 = X$. The space X is *connected* if it is nonempty and not disconnected. A subset $Y \subseteq X$ of a metric space X is *disconnected* if the metric subspace $(Y, d|_{Y\times Y})$ is disconnected and is *connected* if $(Y, d|_{Y\times Y})$ is connected.

Thus a connected space X cannot be separated into two disjoint open subsets.

Example 11.2. Let $Y \coloneqq [-2, -1] \cup [1, 2]$ be the metric subspace of $\mathbb{R}$ with d the standard metric. Then Y is disconnected since both $[-2, -1]$ and $[1, 2]$ are open sets in Y, as $[-2, -1] = (-\frac{5}{2}, -\frac{1}{2}) \cap Y$ and $[1, 2] = (\frac{1}{2}, \frac{5}{2}) \cap Y$. Similarly, $(-1, 0) \cup (0, 1)$ is disconnected, since both $(-1, 0)$ and $(0, 1)$ are open sets in Y.

Lemma 11.3. *A metric space (X, d) is disconnected if and only if X contains a nonempty proper subset E that is both open and closed.*

Proof. If X is disconnected, then there exist disjoint nonempty open sets $E_1, E_2 \subseteq X$ such that $E_1 \cup E_2 = X$. Then E_1 is a nonempty proper subset of X that is both open and closed, since E_1 is the complement of the open set E_2.

Conversely, if X contains a nonempty proper subset E that is both open and closed, then E^c is nonempty, disjoint from E, is open as it is the complement of the closed set E, and satisfies $E \cup E^c = X$. □

The classification of connected subsets of $\mathbb{R}$ is quite straightforward.

Proposition 11.4. *A subset E of $\mathbb{R}$ is connected if and only if whenever $x, y \in E$ with $x < y$, then $[x, y] \subseteq E$.*

Thus connected subsets of $\mathbb{R}$ that contain two points necessarily contain every point in the interval between those two points.

Proof. Suppose that there exists $x, y \in E$ with $x < y$ such that there exists some point $z \in [x, y]$ for which $z \notin E$. We claim that E is not connected. Indeed, we may write E as the union of $E \cap (-\infty, z)$ and $E \cap (z, \infty)$. These are clearly disjoint, and they are nonempty since the former contains x and the latter contains y. Finally, they are open in E since the intervals $(-\infty, z)$ and (z, ∞) are open in $\mathbb{R}$. Thus E is disconnected.

Conversely, suppose in order to obtain a contradiction that whenever $x, y \in E$ with $x < y$, then $[x, y] \subseteq E$, but that E is disconnected, so that there exists disjoint nonempty open sets $E_1, E_2 \subseteq E$ such that $E = E_1 \cup E_2$. Choose $x \in E_1$ and $y \in E_2$, and assume without loss of generality that $x < y$ (for otherwise relabel x as y and y as x).

Let $F \coloneqq [x, y]$; by assumption, $F \subseteq E$. The set $F \cap E_1$ is bounded and nonempty, and hence it has a supremum $z \coloneqq \sup F \cap E_1$. Since $F \subseteq E$, either $z \in E_1$ or $z \in E_2$. If the former holds, then $z \neq y$, but E_1 is open in $E \supseteq F$, which means that there exists some ball $B_r(z)$ such that $B_r(z) \cap F \subseteq E_1$, which contradicts the fact that z is the supremum of $F \cap E_1$. Similarly, if the latter holds, then $z \neq x$, but E_2 is open in $E \supseteq F$, which means that there exists some ball $B_r(z)$ such that $B_r(z) \cap F \subseteq E_2$, which contradicts the fact that z is the supremum of $F \cap E_1$. □

11.2. **Connectedness and continuity.** Just like compactness, connectedness is preserved by continuous functions.

Proposition 11.5. *Let (X, d) and (Y, ρ) be metric spaces, and let $f : X \to Y$ be a continuous function. Let $E \subseteq X$ be connected. Then $f(E) \coloneqq \{y \in Y : y = f(x) \text{ for some } x \in E\}$ is connected.*

Proof. Let U be a subset of $f(E)$ that is open and closed in $f(E)$. Since f is continuous, $f^{-1}(U)$ is both open and closed in E. Since E is connected, $f^{-1}(U)$ must be either E itself or $\emptyset$, in which case U must be either $f(E)$ or $\emptyset$. Thus $f(E)$ contains no nonempty proper subsets that are both open and closed, and hence $f(E)$ is connected. □

Corollary 11.6 (Intermediate value theorem)**.** *Let (X, d) be a metric space and let $f : X \to \mathbb{R}$ be continuous. Let $E \subseteq X$ be connected, and suppose that $x, y \in E$ are such that $f(x) \leq f(y)$. Then for all $c \in [f(x), f(y)]$, there exists some $z \in E$ for which $f(z) = c$.*

Proof. Since E is connected, its image $f(E)$ is connected, and hence if $f(x), f(y) \in f(E)$ with $f(x) \leq f(y)$, then $c \in f(E)$ for all $c \in [f(x), f(y)]$. □

12. Uniform convergence

Recommended reading: [Pug15, §4.1], [Tao16, §3.1–3.6].

12.1. **Spaces of continuous functions.** We now move on to the important notion of convergence of *sequences of functions* between metric spaces. We have already seen some special cases of this, since sequences in normed spaces such as $(C([0,1]), \|\cdot\|_{L^p})$ or $(\ell^p(\mathbb{N}), \|\cdot\|_{\ell^p})$ are sequences of functions. In many situations, it is of use to consider metric spaces whose elements are functions themselves, especially when it comes to continuous functions.

Definition 12.1. Let (X,d) and (Y,ρ) be metric spaces. The space of continuous functions from X to Y is denoted by $C(X,Y)$. The space of bounded continuous functions from X to Y is denoted by $C_b(X,Y)$, so that if $f \in C_b(X,Y)$, there exists some $y \in Y$ and $r > 0$ for which $f(X) \subseteq B_r(y)$, so that $f(x) \in B_r(y)$ for all $x \in X$.

Note that if Y is compact, then Y is bounded, so that $C_b(X,Y) = C(X,Y)$, as every continuous function $f : X \to Y$ is bounded. More interestingly, if instead X is compact, then $C_b(X,Y) = C(X,Y)$ as well, since the image of a compact set under a continuous function is compact, and hence bounded.

Lemma 12.2. *The space $C_b(X,Y)$ is a metric space with respect to the supremum metric*

$$d_\infty(f,g) \coloneqq \sup_{x\in X} \rho(f(x),g(x)).$$

Note that if f and g are *unbounded*, then $d_\infty(f,g)$ need not be finite!

Proof. Positivity and symmetry are clear, while the triangle inequality holds since

$$\begin{aligned} d_\infty(f,h) &= \sup_{x\in X} \rho(f(x),h(x)) \\ &\le \sup_{x\in X} (\rho(f(x),g(x)) + \rho(g(x),h(x))) \\ &\le \sup_{x\in X} \rho(f(x),g(x)) + \sup_{x\in X} \rho(g(x),h(x)) \\ &= d_\infty(f,g) + d_\infty(g,h). \end{aligned}$$

□

A special case of importance is when $Y = \mathbb{R}$.

Definition 12.3. Let $C(X) \coloneqq C(X,\mathbb{R})$ denote the space of continuous functions from X to $\mathbb{R}$, and let $C_b(X)$ denote the metric subspace of $C(X)$ consisting of continuous functions $f : X \to \mathbb{R}$ that are bounded.

Since the scalar multiple of a continuous function $f : X \to \mathbb{R}$ is continuous and the sum of two continuous functions is continuous, $C(X)$ is a vector space. In particular, $C_b(X)$ is not just a metric space but a normed space with respect to the sup-norm

$$\|f\|_{L^\infty} \coloneqq \sup_{x\in X} |f(x)|.$$

Proposition 12.4. *Let (X,d) and (Y,ρ) be metric spaces, and suppose that Y is complete. Then $C_b(X,Y)$ is a complete metric space.*

We previously proved this when $X = [0,1]$ (so that boundedness was automatic) and $Y = \mathbb{R}$, so that $C_b(X,Y) = C([0,1])$. The same proof goes through for $C_b(X,Y)$ with minor notational changes, so we do not reproduce this here.

12.2. **Pointwise convergence.** There are several different concepts for convergence of sequences of functions. Perhaps the most straightforward type of convergence is *pointwise convergence*.

Definition 12.5. A sequence of functions (f_n) from a metric space (X, d) to another metric space (Y, ρ) is said to *converge pointwise* to a function $f : X \to Y$ if $\lim_{n\to\infty} f_n(x) = f(x)$ for each $x \in X$. The function f is said to be the *pointwise limit* of f.

Thus the sequence (f_n) converges pointwise to f if for each $x \in X$ and for all $\varepsilon > 0$, there exists $N = N(\varepsilon, x) \in \mathbb{N}$ such that $\rho(f_n(x), f(x)) < \varepsilon$ whenever $n \geq N$. Note that this definition makes no use of the metric on X!

Example 12.6. The sequence of functions (f_n) from $\mathbb{R}$ to $\mathbb{R}$ given by $f_n(x) \coloneqq \frac{x}{n}$ converges pointwise to the zero function $f(x) \coloneqq 0$.

By the uniqueness of limits, a sequence of functions (f_n) may converge pointwise to at most one function f, and need not converge pointwise at all.

When $Y = \mathbb{R}$, so that each f_n is real-valued, we may consider a *series* of functions $\sum_{n=1}^{\infty} f_n(x)$. We say that this series is pointwise convergent if the sequence of partial sums (g_n) given by $g_n(x) \coloneqq \sum_{m=1}^{n} f_n(x)$ is pointwise convergent.

Pointwise convergence is a very *weak* concept: it *fails* to preserve many nice properties.

Example 12.7. The sequence of functions (f_n) from $[0, 1]$ to $\mathbb{R}$ given by $f_n(x) \coloneqq x^n$ converges pointwise to the function

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

Each function f_n is continuous — in fact, each is infinitely differentiable — but f is not, and so continuity and differentiability are not preserved; in particular, $\lim_{n\to\infty} \frac{d}{dx} f_n(x)$ need not be equal to $\frac{d}{dx} \lim_{n\to\infty} f_n(x)$. Furthermore, we have that $\lim_{x\nearrow 1} f_n(x) = 1$ for every $n \in \mathbb{N}$, yet $\lim_{x\nearrow 1} f(x) = 0$, so limits are not preserved; in particular, $\lim_{n\to\infty} \lim_{x\to a} f_n(x) = \lim_{n\to\infty} f_n(a) = f(a)$ need not be equal to $\lim_{x\to a} \lim_{n\to\infty} f_n(x) = \lim_{x\to a} f(x)$.

Example 12.8. The sequence of functions (f_n) from $\mathbb{R}$ to $\mathbb{R}$ given by

$$f_n(x) \coloneqq \begin{cases} -n & \text{if } x < -n, \\ x & \text{if } -n \leq x \leq n, \\ n & \text{if } x > n \end{cases}$$

converges pointwise to the function $f(x) \coloneqq x$. Each f_n is bounded, but f is not, so boundedness is not preserved.

Example 12.9. The sequence of functions (f_n) from $[0, 1]$ to $\mathbb{R}$ given by

$$f_n(x) \coloneqq \begin{cases} 0 & \text{if } 0 \leq x \leq 2^{-n}, \\ 2^{2(n+1)}(x - 2^{-n}) & \text{if } 2^{-n} \leq x \leq \frac{3}{2}2^{-n}, \\ 2^{2(n+1)}(2^{1-n} - x) & \text{if } \frac{3}{2}2^{-n} \leq x \leq 2^{1-n}, \\ 0 & \text{if } 2^{1-n} \leq x \leq 1. \end{cases}$$

converges pointwise to $f(x) \coloneqq 0$. Each of these is continuous, as is f. However, $\int_0^1 f_n(x)\,dx = 1$ for all $n \in \mathbb{N}$, whereas $\int_0^1 f(x)\,dx = 0$, and so $\lim_{n\to\infty} \int_0^1 f_n(x)\,dx \neq \int_0^1 \lim_{n\to\infty} f_n(x)\,dx$. So limits of integrals are not preserved.

The weakness of pointwise convergence is that we have no control over the dependence on x with respect to the *rate* of convergence, or more precisely the dependence of $N = N(\varepsilon, x)$ on x. This can be seen for $f_n(x) = x^n$: for $0 \leq x < 1$, this converges pointwise to 0, but the convergence is much slower near 1 than near 0.

12.3. **Uniform convergence.** A strengthening of pointwise convergence removes this issue of the dependence on x with respect to the rate of convergence.

Definition 12.10. A sequence of functions (f_n) from a metric space (X, d) to another metric space (Y, ρ) is said to *converge uniformly* to a function $f : X \to Y$ if for all $\varepsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$ such that for all $x \in X$, $\rho(f_n(x), f(x)) < \varepsilon$ whenever $n \geq N$. The function f is said to be the *uniform limit* of (f_n).

The crucial difference between the definitions of pointwise convergence and uniform convergence is that in the former, $N = N(\varepsilon, x)$ is allowed to depend on x, whereas in the latter, $N = N(\varepsilon)$ is *independent* of x. This is the same type of distinction between continuity and uniform continuity.

It is immediately clear that if (f_n) converges uniformly to f, then it also converges pointwise to the same function f. The converse, however, is not true.

Example 12.11. The sequence of functions (f_n) from $[0, 1]$ to $\mathbb{R}$ given by $f_n(x) \coloneqq \frac{x}{n}$ converges uniformly to the zero function $f(x) \coloneqq 0$. To see this, fix $\varepsilon > 0$. Then letting N be the least integer larger than $\frac{1}{\varepsilon}$, we see that for all $x \in [0, 1]$, if $n \geq N$ then

$$|f_n(x) - f(x)| = \left|\frac{x}{n} - 0\right| = \frac{x}{n} \leq \frac{1}{n} \leq \frac{1}{N} < \varepsilon.$$

Example 12.12. The sequence of functions (f_n) from $[0, 1]$ to $\mathbb{R}$ given by $f_n(x) \coloneqq x^n$ does *not* converge uniformly to any function $f : [0, 1] \to \mathbb{R}$.

Again, when $Y = \mathbb{R}$, so that each f_n is real-valued, we may consider the *series* of functions $\sum_{n=1}^{\infty} f_n$. We say that this series is uniformly convergent if the sequence of partial sums (g_n) given by $g_n \coloneqq \sum_{m=1}^{n} f_n$ is uniformly convergent. A useful consequence of uniform convergence is a condition for when a series of functions from X to $\mathbb{R}$ converges uniformly.

Theorem 12.13 (Weierstrass M-test)**.** *Let (X, d) be a metric space, and let (f_n) be a sequence of functions from X to $\mathbb{R}$. Suppose that there exists a sequence (M_n) of nonnegative real numbers for which $\sup_{x \in X} |f_n(x)| \leq M_n$. Then if $\sum_{n=1}^{\infty} M_n$ converges, the series $\sum_{n=1}^{\infty} f_n$ converges uniformly to a function $f : X \to \mathbb{R}$.*

Proof. If $\sum_{n=1}^{\infty} M_n$ converges, then for each $\varepsilon > 0$, there exists some $N = N(\varepsilon) \in \mathbb{N}$ such that whenever $m > n \geq N$, we have that $\sum_{\ell=n+1}^{m} M_\ell < \varepsilon$. In particular, for all $x \in X$, we have that

$$|g_m(x) - g_n(x)| = \left|\sum_{\ell=n+1}^{m} f_\ell(x)\right| \leq \sum_{\ell=n+1}^{m} |f_\ell(x)| \leq \sum_{\ell=n+1}^{m} M_\ell < \varepsilon.$$

It follows that for each $x \in X$, the sequence $(g_n(x))$ in $\mathbb{R}$ is Cauchy; since $\mathbb{R}$ is complete, this converges to some $f(x) \in \mathbb{R}$. Thus (g_n) converges pointwise to f.

To show that this convergence is *uniform*, we note that $|g_n(x) - g_m(x)| < \frac{\varepsilon}{2}$ for any $m > n \geq N = N(\frac{\varepsilon}{2})$. Moreover, since $(g_m(x))$ converges to $f(x)$, there exists $M = M(\frac{\varepsilon}{2}, x) \in \mathbb{N}$ such that $|g_m(x) - f(x)| < \frac{\varepsilon}{2}$ whenever $m \geq M$. So for any $n \geq N(\frac{\varepsilon}{2})$ and $m \geq \max\{n + 1, N(\frac{\varepsilon}{2})\}$,

$$|g_n(x) - f(x)| \leq |g_n(x) - g_m(x)| + |g_m(x) - f(x)| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Since $N = N(\frac{\varepsilon}{2})$ is independent of x, we deduce that convergence is uniform. □

Unlike pointwise convergence, many nice properties are preserved by uniform convergence. An important such example is continuity.

Proposition 12.14. *Let (f_n) be a sequence of functions from (X, d) to (Y, ρ) that converges uniformly to $f : X \to Y$. If each f_n is continuous at $x \in X$, then f is continuous at x. In particular, the uniform limit f of a sequence of continuous functions f_n is continuous.*

In general, the converse is *not* true; there exist sequences of continuous functions that converge pointwise to a continuous function but do not converge uniformly to that function.

Proof. Suppose that (f_n) converges uniformly and that each f_n is continuous at x. Then for all $\varepsilon > 0$, there exists $N = N(\frac{\varepsilon}{3}) \in \mathbb{N}$ such that for all $y \in X$, $\rho(f_n(y), f(y)) < \frac{\varepsilon}{3}$ whenever $n \geq N$, while for all $\varepsilon > 0$, there exists $\delta = \delta(\frac{\varepsilon}{3}, x, n)$ such that $\rho(f_n(x), f_n(y)) < \frac{\varepsilon}{3}$ whenever $d(x, y) < \delta$. Taking $n = N = N(\frac{\varepsilon}{3})$, it follows that if $d(x, y) < \delta(\frac{\varepsilon}{3}, x, N(\frac{\varepsilon}{3}))$

$$\begin{aligned}\rho(f(x), f(y)) &\leq \rho(f(x), f_N(x)) + \rho(f_N(x), f_N(y)) + \rho(f_N(y), f(y)) \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \\ &= \varepsilon,\end{aligned}$$

and so f is continuous at x. □

Similarly, uniform convergence preserves boundedness.

Proposition 12.15. *Let (f_n) be a sequence of functions from (X, d) to (Y, ρ) that converges uniformly to $f : X \to Y$. If each f_n is bounded, so that $f_n(X)$ is a bounded subset of Y, then f is also bounded.*

Proof. Since (f_n) converges uniformly to f, for all $\varepsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$ such that for all $x \in X$, $\rho(f_n(x), f(x)) < \varepsilon$ whenever $n \geq N$. Since f_n is bounded, there exists some $y = y(n) \in Y$ and $R = R(n) > 0$ such that $\rho(f_n(x), y) < R$ for all $x \in X$. Taking $n = N = N(\varepsilon)$, it follows that

$$\rho(f(x), y) \leq \rho(f(x), f_N(x)) + \rho(f_N(x), y) < \varepsilon + R.$$

Taking $r \coloneqq \varepsilon + R$, we deduce that $f(x) \in B_r(y)$ for all $x \in X$, so that f is bounded. □

These two results can be reinterpreted in terms of convergence of sequences in the metric space $(C_b(X, Y), d_\infty)$.

Corollary 12.16. *A sequence (f_n) in the metric space $(C_b(X, Y), d_\infty)$ is convergent if and only if this sequence of continuous bounded functions (f_n) from X to Y is uniformly convergent.*

Proof. The sequence (f_n) in $C_b(X, Y)$ converges to some $f \in C_b(X, Y)$ if for all $\varepsilon > 0$, there exists some $N = N(\varepsilon) \in \mathbb{N}$ such that $\sup_{x \in X} \rho(f_n(x), f(x)) < \varepsilon$ whenever $n \geq N$, which implies that (f_n) converges uniformly to f.

Conversely, if (f_n) is uniformly convergent and each f_n is continuous and bounded, then its uniform limit f is also continuous and bounded. Uniform convergence states that for all $\varepsilon > 0$, there exists some $N = N(\varepsilon) \in \mathbb{N}$ such that for all $x \in X$, $\rho(f_n(x), f(x)) < \varepsilon$ whenever $n \geq N$, which is the same as convergence in $C_b(X, Y)$ with respect to the supremum metric. □

Another property preserved by uniform convergence is integration.

Proposition 12.17. *Let (f_n) be a sequence of continuous functions from $[0,1]$ to $\mathbb{R}$ that is uniformly convergent. Then $\lim_{n\to\infty} \int_0^1 f_n(x)\,dx = \int_0^1 \lim_{n\to\infty} f_n(x)\,dx$.*

Proof. Since each f_n is continuous and the sequence (f_n) is uniformly convergent, there exists a continuous function $f : [0,1] \to \mathbb{R}$ such that for all $\varepsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$ such that for all $x \in X$, we have that $|f_n(x) - f(x)| < \varepsilon$ whenever $n \geq N$. It follows that

$$\begin{aligned}\left|\int_0^1 f_n(x)\,dx - \int_0^1 f(x)\,dx\right| &= \left|\int_0^1 (f_n(x) - f(x))\,dx\right| \\ &\leq \int_0^1 |f_n(x) - f(x)|\,dx \\ &< \int_0^1 \varepsilon\,dx \\ &= \varepsilon,\end{aligned}$$

and so the sequence of real numbers $(\int_0^1 f_n(x)\,dx)$ converges to the real number $\int_0^1 f(x)\,dx$. □

Remark 12.18. Buried in this proof is the key fact that $|\int_0^1 f(x)\,dx| \leq \int_0^1 |f(x)|\,dx$ whenever f is continuous. This result remains true without the assumption of continuity provided one assumes that each function f_n is Riemann-integrable. Moreover, much *weaker* conditions can be imposed on the sequence (f_n) and the method of convergence to f once one instead works with *Lebesgue-integrable* functions.

Unfortunately, differentiation is *not* preserved by uniform convergence: even if a sequence of functions (f_n) from $[0,1]$ to $\mathbb{R}$ converges uniformly to f and each function f_n is differentiable, it need not be the case that f is differentiable or that (f_n') converges uniformly to f'.

Example 12.19. The sequence (f_n) from $[0, 2\pi]$ to $\mathbb{R}$ given by $f_n(x) \coloneqq n^{-1/2} \sin nx$ converges uniformly to $f(x) = 0$, but $f_n'(x) = n^{1/2} \cos nx$ does not converge pointwise to $f'(x) = 0$, let alone uniformly to f'.

While pointwise convergence alone is weaker than uniform convergence, one can impose additional conditions on a sequence (f_n) and on the metric space X in order to strengthen pointwise convergence to uniform convergence.

Theorem 12.20 (Dini's theorem)**.** *Let (X, d) be a compact metric space and let (f_n) be a nondecreasing sequence in $C(X)$, so that for all $x \in X$, we have that $f_n(x) \geq f_m(x)$ whenever $n \geq m$. Suppose that (f_n) converges pointwise to some function $f \in C(X)$. Then (f_n) converges to f in $C(X)$ with respect to the sup-norm metric, so that (f_n) converges uniformly to f.*

Proof. Fix $\varepsilon > 0$, and define

$$E_n \coloneqq \{x \in X : f_n(x) > f(x) - \varepsilon\}.$$

Then $E_n = (f - f_n)^{-1}((-\infty, \varepsilon))$ is the preimage of the open set $(-\infty, \varepsilon) \subseteq \mathbb{R}$ by the continuous function $f - f_n$, and hence is an open subset of X. Since $f_n(x)$ is nondecreasing, we have that $E_n \supseteq E_m$ whenever $n \geq m$. Moreover, for each $x \in X$, there exists some $n \in \mathbb{N}$ for which $x \in E_n$ since $f_n(x)$ converges to $f(x)$.

It follows that $\{E_n : n \in \mathbb{N}\}$ is an open cover of X. Since X is compact, this open cover has a finite subcover $\{E_{n_1}, \ldots, E_{n_N}\}$; as $E_n \supseteq E_m$ whenever $n \geq m$, this finite subcover is simply of the form E_M with $M \coloneqq \max\{n_1, \ldots, n_N\}$. It follows that $X = E_M$, and hence

if $n \geq M$ then $f_n(x) > f(x) - \varepsilon$ for all $x \in X$. Since (f_n) is nondecreasing, we must also have that $f_n(x) \leq f(x)$, and so we deduce that (f_n) converges uniformly to f. □

Compactness is an essential assumption in Dini's theorem.

Example 12.21. Let $X = (0,1)$ and let $f_n(x) = -\frac{1}{nx+1}$. This is a nondecreasing sequence of continuous functions that converges pointwise to $f(x) \coloneqq 0$, but it does not converge uniformly to f.

13. The Arzelà–Ascoli theorem

Recommended reading: [Pug15, §4.3, 4.8].

13.1. **Equicontinuity.** We now move on to the notion of *equicontinuity*, which is introduced in order to answer the following question.

Question 13.1. *Which subspaces of $C(X,Y)$ are compact? That is, given a subspace of $C(X,Y)$, when can we ensure that every sequence in this subspace has a convergent subsequence?*

As stated, this question is ill-posed, since in general $C(X,Y)$ is not even a metric space! In particular, we need to impose some conditions (such as restricting to *bounded* functions, which can be ensured by taking X to be *compact*) in order for the supremum metric $d_\infty(f,g) \coloneqq \sup_{x\in X} \rho(f(x),g(x))$ to always be finite. Nonetheless, for the time being, we will consider subsets of $C(X,Y)$ without any impositions.

Definition 13.2. Let (X,d) and (Y,ρ) be metric spaces. A subspace $\mathcal{F} \subseteq C(X,Y)$ is said to be *equicontinuous at $x \in X$* if for every $\varepsilon > 0$, there exists some $\delta = \delta(\varepsilon, x) > 0$ such that for every $f \in \mathcal{F}$, we have that $f(y) \in B_\varepsilon(f(x))$ whenever $y \in B_\delta(x)$. Additionally, $\mathcal{F}$ is said to be *equicontinuous* if it is equicontinuous at each $x \in X$.

This looks very similar to the definition of continuity, but the crux is that the number $\delta = \delta(\varepsilon, x)$ is *independent* of $f \in \mathcal{F}$, and hence can be chosen to be the same for *every* function f in the subspace $\mathcal{F}$.

Recall that a function $f : X \to Y$ is *uniformly continuous* if for all $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that for all $x \in X$, we have that $f(y) \in B_\varepsilon(f(x))$ whenever $y \in B_\delta(x)$; in particular, δ may depend only on ε and not on x. With this in mind, we define the following strengthening of equicontinuity.

Definition 13.3. Let (X,d) and (Y,ρ) be metric spaces. A subspace $\mathcal{F} \subseteq C(X,Y)$ is said to be *uniformly equicontinuous* if for every $\varepsilon > 0$, there exists some $\delta = \delta(\varepsilon) > 0$ such that for every $f \in \mathcal{F}$ and every $x \in X$, we have that $f(y) \in B_\varepsilon(f(x))$ whenever $y \in B_\delta(x)$.

Uniform equicontinuity is a very strong condition! It not only means that each function $f \in \mathcal{F} \subseteq C(X,Y)$ is uniformly continuous, but that the choice of δ in the definition of uniform continuity is *independent* of f. An example to keep in mind is when X and Y are both subsets of $\mathbb{R}$ and each $f \in \mathcal{F}$ is differentiable; $f \in \mathcal{F}$ being uniformly continuous is implied by $\sup_{x\in X} |f'(x)|$ being finite, while $\mathcal{F}$ being uniformly equicontinuous is implied by $\sup_{f\in\mathcal{F}} \sup_{x\in X} |f'(x)|$ being finite.

There is a simple way to strengthen equicontinuity to uniform equicontinuity. Just as continuous functions on a compact space are uniformly continuous, equicontinuous families of functions on a compact metric space are uniformly equicontinuous.

Theorem 13.4. *Let (X,d) be a compact metric space and let $\mathcal{F}$ be a subspace of $C(X,Y)$. Then $\mathcal{F}$ is equicontinuous if and only if it is uniformly equicontinuous.*

The proof is very similar to the proof of the fact that continuous functions on compact metric spaces are uniformly continuous. Indeed, taking $\mathcal{F}$ to be a set consisting of one element, this is precisely the same proof.

Proof. Clearly uniform equicontinuity implies equicontinuity.

For the converse, first fix $\varepsilon > 0$. Since $\mathcal{F}$ is equicontinuous, for each $x \in X$, there exists some $\delta = \delta(\frac{\varepsilon}{2}, x) > 0$ such that for all $f \in \mathcal{F}$, we have that $f(y) \in B_{\frac{\varepsilon}{2}}(f(x))$ whenever $y \in B_{\delta(\frac{\varepsilon}{2},x)}(x)$. In particular, this remains true if $y \in B_{\frac{1}{2}\delta(\frac{\varepsilon}{2},x)}(x)$.

Now consider the collection of balls $\{B_{\frac{1}{2}\delta(\frac{\varepsilon}{2},x)}(x) : x \in X\}$. This is an open cover of X as $x \in B_{\frac{\delta_x}{2}}(x)$ for each $x \in X$, and since X is compact, there exists a finite subcover, which is to say a finite collection of points $\{x_1, \dots, x_N\} \subseteq X$ for which $\bigcup_{n=1}^{X} B_{\frac{1}{2}\delta(\frac{\varepsilon}{2},x_n)}(x_n) = X$.

Let $\delta \coloneqq \min_{1\le n\le N} \frac{1}{2}\delta(\frac{\varepsilon}{2}, x_n) > 0$. For each $x \in X$, there exists some $n \in \{1, \dots, N\}$ such that $x \in B_{\frac{1}{2}\delta(\frac{\varepsilon}{2},x_n)}(x_n)$, so that $f(x) \in B_{\frac{\varepsilon}{2}}(f(x_n))$ for all $f \in \mathcal{F}$. If $y \in B_\delta(x)$, so that $d(y, x) < \delta$, then

$$d(y, x_n) \le d(y, x) + d(x, x_n) < \delta + \frac{1}{2}\delta\left(\frac{\varepsilon}{2}, x_n\right) < \delta\left(\frac{\varepsilon}{2}, x_n\right),$$

so that $y \in B_{\delta(\frac{\varepsilon}{2},x_n)}(x_n)$ and hence $f(y) \in B_{\frac{\varepsilon}{2}}(f(x_n))$. But then

$$\rho(f(y), f(x)) \le \rho(f(y), f(x_n)) + \rho(f(x_n), f(x)) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Thus for all $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) = \min_{1\le n\le N} \frac{1}{2}\delta(\frac{\varepsilon}{2}, x_n) > 0$ such that for every $f \in \mathcal{F}$ and $x \in X$, we have that $y \in B_\varepsilon(f(x))$ whenever $y \in B_\delta(x)$. It follows that $\mathcal{F}$ is uniformly equicontinuous. □

13.2. **The Arzelà–Ascoli theorem.** The definition of equicontinuity was introduced in order to facilitate the formulation of the following theorem.

Theorem 13.5 (Arzelà–Ascoli theorem)**.** *Let (X, d) and (Y, ρ) be metric spaces, and suppose that X is compact, so that $C(X, Y) = C_b(X, Y)$ is a metric space with respect to the supremum metric, and that Y is locally compact. Then a subspace $\mathcal{F}$ of $C(X, Y)$ is compact if and only if it is closed, bounded, and equicontinuous.*

This is a fundamental theorem in analysis: it gives necessary and sufficient conditions for a sequence of continuous functions to have a convergent subsequence with respect to the supremum metric. It is perhaps most natural to consider the Arzelà–Ascoli theorem when $X = [0, 1]$ and $Y = \mathbb{R}$ (and this is the standard formulation of this).

Corollary 13.6. *Every bounded equicontinuous sequence of functions from $[0, 1]$ to $\mathbb{R}$ has a uniformly convergent subsequence.*

Like the proof of the equivalence of sequential and topological compactness, the proof of the Arzelà–Ascoli theorem proceeds in several parts. We must show that a closed, bounded, and equicontinuous subspace of $C(X, Y)$ is compact, or equivalently is both complete and totally bounded. We first prove the completeness of $\mathcal{F}$.

Proposition 13.7. *Let (X, d) and (Y, ρ) be metric spaces, and suppose that X is compact and Y is locally compact. Then a closed subspace $\mathcal{F}$ of $C(X, Y)$ is complete.*

Proof. Since X is compact, $C(X, Y) = C_b(X, Y)$, and since Y is locally compact, and hence complete, $C_b(X, Y)$ is complete. Thus $C(X, Y)$ is complete. It remains to note that a closed subset of a complete metric space is complete. □

It remains to prove the total boundedness of $\mathcal{F}$.

Proposition 13.8. *Let (X, d) and (Y, ρ) be metric spaces, and suppose that X is compact and Y is locally compact. Then a bounded equicontinuous subspace $\mathcal{F}$ of $C(X, Y)$ is totally bounded.*

Proof. Fix $r > 0$. We must show that there exists a finite collection of functions $\{f_1, \dots, f_L\} \subseteq \mathcal{F}$ such that for all $f \in \mathcal{F}$, there exists $\ell \in \{1, \dots, L\}$ for which $d_\infty(f, f_\ell) < r$.

Since X is compact and $\mathcal{F}$ is equicontinuous, it is also uniformly equicontinuous, so that there exists some $\delta = \delta(\frac{r}{4}) > 0$ such that for all $f \in \mathcal{F}$ and $x \in X$, we have that $f(y) \in B_{\frac{r}{4}}(f(x))$ whenever $y \in B_\delta(x)$. Since X is compact, it is totally bounded; choosing the radius to be δ, we deduce that there exists a finite collection of points $\{x_1, \dots, x_N\} \subseteq X$ for which $\bigcup_{n=1}^N B_\delta(x_n) = X$. Finally, since $\mathcal{F}$ is bounded, there exists some $y \in Y$ and $R > 0$ such that $\mathcal{F} \subseteq C(X, \overline{B_R(y)})$. Since Y is locally compact, $\overline{B_R(y)}$ is compact, and hence totally bounded; choosing the radius to be $\frac{r}{4}$, we deduce the existence of a finite collection of points $\{y_1, \dots, y_M\}$ such that $\bigcup_{m=1}^M B_{\frac{r}{4}}(y_m) \supseteq \overline{B_R(y)}$.

With this in mind, we now construct a finite collection of functions in $\mathcal{F}$ as follows. Let

$$Z \coloneqq \{\sigma : \{1, \dots, N\} \to \{1, \dots, M\}\}$$

denote the set of functions from $\{1, \dots, N\}$ to $\{1, \dots, M\}$. This is a finite set (with M^N elements). Let

$$Z_{\mathcal{F}} \coloneqq \{\sigma \in Z : \text{there exists } f \in \mathcal{F} \text{ such that } f(x_n) \in B_{\frac{r}{4}}(y_{\sigma(n)}) \text{ for all } 1 \leq n \leq N\}.$$

This is a subset of Z; it is nonempty, since given $f \in \mathcal{F}$, there must exist some $\sigma \in Z$ for which $f(x_n) \in B_{\frac{r}{4}}(y_{\sigma(n)})$ for all $n \in \{1, \dots, N\}$. For each $\sigma \in Z_{\mathcal{F}}$, we can choose some $f_\sigma \in \mathcal{F}$ such that $f_\sigma(x_n) \in B_{\frac{r}{4}}(y_{\sigma(n)})$ for each $n \in \{1, \dots, N\}$. Then $\{f_\sigma : \sigma \in Z_{\mathcal{F}}\}$ is a finite set of functions in $\mathcal{F}$.

We claim that $\bigcup_{\sigma \in Z_{\mathcal{F}}} B_r(f_\sigma) \supseteq \mathcal{F}$, from which total boundedness follows. To see this, we note that given $f \in \mathcal{F}$, as $\overline{B_R(y)} \subseteq \bigcup_{m=1}^M B_{\frac{r}{4}}(y_m)$, for each $n \in \{1, \dots, N\}$, we may choose a corresponding $\sigma(n) \in \{1, \dots, M\}$ such that $f(x_n) \in B_{\frac{r}{4}}(y_{\sigma(n)})$. This defines some $\sigma \in Z_{\mathcal{F}}$. We claim that $f \in B_r(f_\sigma)$. Indeed, given $x \in X$, as $X = \bigcup_{n=1}^N B_\delta(x_n)$, there exists some $n \in \{1, \dots, N\}$ for which $x \in B_\delta(x_n)$, and so

$$\begin{aligned}
\rho(f(x), f_\sigma(x)) &\leq \rho(f(x), f(x_n)) + \rho(f(x_n), y_{\sigma(n)}) + \rho(y_{\sigma(n)}, f_\sigma(x_n)) + \rho(f_\sigma(x_n), f_\sigma(x)) \\
&< \frac{r}{4} + \frac{r}{4} + \frac{r}{4} + \frac{r}{4} \\
&= r.
\end{aligned}$$

Here the first and last terms are bounded by $\frac{r}{4}$ due to uniform equicontinuity, since f and f_σ are in $\mathcal{F}$ and $x \in B_\delta(x_n)$. The second term is bounded by $\frac{r}{4}$ from the definition of σ in terms of f, and the third term is bounded by $\frac{r}{4}$ due to the choice of f_σ in terms of σ. $\square$

Thus we have proven that if X is compact and Y is locally compact, then a closed, bounded, and equicontinuous subset $\mathcal{F}$ of $C(X, Y)$ is compact.

It remains to prove the converse, namely that compactness implies closedness, boundedness, and equicontinuity.

Proposition 13.9. *Let (X, d) and (Y, ρ) be metric spaces, and suppose that X is compact and Y is locally compact. Then a compact subspace $\mathcal{F}$ of $C(X, Y)$ is closed, bounded, and equicontinuous.*

Once more, the proof is very similar to that of compactness implying uniform continuity.

Proof. Since $\mathcal{F}$ is compact, it is closed and bounded, so it remains to prove equicontinuity. For this, we only need the fact that $\mathcal{F}$ is totally bounded, which follows from compactness. Since $\mathcal{F}$ is totally bounded, for all $r > 0$, there exists a finite collection of functions $\{f_1, \dots, f_N\} \subseteq \mathcal{F}$ such that for all $f \in \mathcal{F}$, there exists some $n \in \{1, \dots, N\}$ for which $d_\infty(f, f_n) < r$. We fix $\varepsilon > 0$ and take $r = \frac{\varepsilon}{3}$. Now since each f_n is uniformly continuous due to the fact that X is compact, there exists some $\delta = \delta(\frac{\varepsilon}{3}, f_n) > 0$ such that for all $x \in X$, we have that $f_n(y) \in B_{\frac{\varepsilon}{3}}(f_n(x))$ whenever $y \in B_{\delta(\frac{\varepsilon}{3}, f_n)}(x)$.

Let $\delta \coloneqq \min_{1 \le n \le N} \delta(\frac{\varepsilon}{3}, f_n) > 0$. Then for all $x \in X$ and $f \in \mathcal{F}$, there exists some $n \in \{1, \dots, N\}$ for which $d_\infty(f, f_n) < \frac{\varepsilon}{3}$, and so whenever $y \in B_\delta(x)$, we have that

$$\begin{aligned} \rho(f(x), f(y)) &\le \rho(f(x), f_n(x)) + \rho(f_n(x), f_n(y)) + \rho(f_n(y), f(y)) \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \\ &= \varepsilon. \end{aligned}$$

Thus we have shown that for every $\varepsilon > 0$, there exists some $\delta = \delta(\varepsilon) > 0$ such that for every $f \in \mathcal{F}$ and $x \in X$, we have that $f(y) \in B_\varepsilon(f(x))$ whenever $y \in B_\delta(x)$. In particular, $\mathcal{F}$ is equicontinuous (and in fact uniformly equicontinuous). □

The conditions in the Arzelà–Ascoli theorem truly are necessary: dropping a single one causes compactness to fail.

Example 13.10.

- Let $\mathcal{F}$ be the subset of $C([0,1])$ consisting of constant functions. This is closed and equicontinuous but not bounded and not compact.
- Let $\mathcal{F}$ be the subset of $C([0,1])$ consisting of functions f satisfying $\sup_{x \in [0,1]} |f(x)| \le 1$, so that $\mathcal{F}$ is the closed unit ball in $C([0,1])$. This is closed and bounded but not equicontinuous and not compact.
- Let $\mathcal{F}$ be the subset of $C([0,1])$ consisting of functions f satisfying both

$$\sup_{x \in [0,1]} |f(x)| < 1, \qquad \sup_{\substack{x,y \in [0,1] \\ x \ne y}} \frac{|f(x) - f(y)|}{|x - y|} \le 1.$$

  This is bounded and equicontinuous but not closed and not compact.

An important application of the Arzelà–Ascoli theorem is to prove the *Peano existence theorem* on the existence of a solution to an ordinary differential equation, which we state here without proof.

Theorem 13.11 (Peano existence theorem)**.** *Let n be a positive integer and let $\Omega \subseteq \mathbb{R} \times \mathbb{R}^n$ be an open set. Let $f : \Omega \to \mathbb{R}$ be continuous. Then for any $(x_0, y_0) \in \Omega$, there exists some $h > 0$ and a corresponding solution $y : (x_0 - h, x_0 + h) \to \mathbb{R}^n$ of the initial value problem*

$$\begin{aligned} y' &= f(x, y), \\ y(x_0) &= y_0. \end{aligned}$$

In general, this ordinary differential equation need not have a "nice" closed-form solution, nor need the solution be unique. Nonetheless, the Peano existence theorem *guarantees* that a solution always exists.

14. The Stone–Weierstrass theorem

Recommended reading: [Pug15, §4.4], [Tao16, §3.8].

We deal with one last topic relating to metric spaces before launching into the theory of the Lebesgue measure and the Lebesgue integral. This last topic is an important result in approximation: the fact that continuous functions (on a closed interval, for example) can be approximated uniformly by polynomials. This result is called the *Weierstrass approximation theorem*.

We take a more general approach, which, in principle, remains valid on arbitrary compact metric spaces, and applies not only to approximation by polynomials, but also in a range of other situations. This theorem is called the *Stone–Weierstrass theorem*. To state the theorem, we first need to introduce some definitions and notation.

14.1. **Subalgebras.** The Stone–Weierstrass theorem applies to approximation by elements of a *subalgebra* of the Banach space $C(X)$ of continuous real-valued functions on a compact metric space X.

Definition 14.1. A *subalgebra* of $C(X)$ is a subset of $C(X)$ that is closed under the algebraic operations of addition, scalar multiplication, and multiplication. That is, a subset $\mathcal{A} \subseteq C(X)$ is a subalgebra if whenever f and g are in $\mathcal{A}$ and c is a real number, then cf, $f+g$, and fg are also in $\mathcal{A}$.

Note that the zero function is necessarily in a nonempty subalgebra, since subalgebras are closed under scalar multiplication, and in particular under multiplication by the scalar $c = 0$.

Example 14.2. The following are subalgebras:

- The subset of $C(X)$ consisting of constant functions;
- The subset of $C(X)$ consisting only of the zero function;
- The subset of $C(X)$ consisting of those continuous functions that vanish on a given subset $E \subset X$;
- The subset of $C([0,1])$ consisting of *polynomials*;
- The subset of $C(S^1)$ consisting of *trigonometric polynomials*, where $S^1 \coloneqq \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ denotes the unit circle in $\mathbb{R}^2$, and a trigonometric polynomial is a finite linear combination of functions of the form

$$\Re((x+iy)^n) = \sum_{\substack{m=0 \\ m \text{ even}}}^{n} \binom{n}{m} x^{n-m} (-1)^{\frac{m}{2}} y^m,$$

$$\Im((x+iy)^n) = \sum_{\substack{m=0 \\ m \text{ odd}}}^{n} \binom{n}{m} x^{n-m} (-1)^{\frac{m-1}{2}} y^m$$

  with n a nonnegative integer; these correspond to $\cos(n\theta)$ and $\sin(n\theta)$ respectively via the mapping from S^1 to $[0, 2\pi)$ given by $x = \cos\theta$ and $y = \sin\theta$ with $\theta \in [0, 2\pi)$. Equivalently, we may view S^1 as the unit circle in $\mathbb{C}$, namely $\{z \in \mathbb{C} : |z| = 1\}$, so that $z = e^{i\theta} = \cos\theta + i\sin\theta$, and then trigonometric polynomials are of the form

$$\Re(z^n) = \cos(n\theta),$$
$$\Im(z^n) = \sin(n\theta).$$

In the proof of the Stone–Weierstrass theorem, we will need the following simple observation.

Lemma 14.3. *If $\mathcal{A}$ is a subalgebra of $C(X)$, then so is $\overline{\mathcal{A}}$.*

Here by $\overline{\mathcal{A}}$ we mean the closure of $\mathcal{A}$ in the Banach space $(C(X), \|\cdot\|_{L^\infty})$.

Proof. Suppose that $f \in \overline{\mathcal{A}}$ and that $c \in \mathbb{R}$; we may assume that $c \neq 0$, since clearly $0f = 0 \in \mathcal{A}$. There exists a sequence (f_n) in $\mathcal{A}$ that converges to $f \in \overline{\mathcal{A}}$, so that for all $\varepsilon > 0$, there exists $N = N(\frac{\varepsilon}{|c|}) \in \mathbb{N}$ such that $\|f_n - f\|_{L^\infty} < \frac{\varepsilon}{|c|}$ whenever $n \geq N$. Since $\mathcal{A}$ is a subalgebra, we have that $cf_n \in \mathcal{A}$ for each $n \in \mathbb{N}$, and so (cf_n) converges to cf, which must also be in $\overline{\mathcal{A}}$, since $\|cf_n - cf\|_{L^\infty} = |c|\|f_n - f\|_{L^\infty} < \varepsilon$ whenever $n \geq N$ by the homogeneity of the norm. Thus $\overline{\mathcal{A}}$ is closed under scalar multiplication.

Next, suppose that $f, g \in \overline{\mathcal{A}}$. Then again there exist sequences $(f_n), (g_n)$ in $\mathcal{A}$ that converge to $f, g \in \overline{\mathcal{A}}$ respectively, so that for all $\varepsilon > 0$, there exists $N = N(\frac{\varepsilon}{2}) \in \mathbb{N}$ and $M = M(\frac{\varepsilon}{2}) \in \mathbb{N}$ such that $\|f_n - f\|_{L^\infty} < \frac{\varepsilon}{2}$ whenever $n \geq N$ and $\|g_n - g\|_{L^\infty} < \frac{\varepsilon}{2}$ whenever $n \geq M$. Since $\mathcal{A}$ is a subalgebra, we have that $f_n + g_n \in \mathcal{A}$ for each $n \in \mathbb{N}$, and so $(f_n + g_n)$ converges to $f + g$, which must also be in $\overline{\mathcal{A}}$, since

$$\|(f_n + g_n) - (f + g)\|_{L^\infty} \leq \|f_n - f\|_{L^\infty} + \|g_n - g\|_{L^\infty} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

whenever $n \geq \max\{N, M\}$ by the subadditivity of the norm. Thus $\overline{\mathcal{A}}$ is closed under addition.

Finally, if $f, g \in \overline{\mathcal{A}}$, then it is trivially true that $fg \in \overline{A}$ if either f or g are 0. Otherwise, we again let $(f_n), (g_n)$ be sequences in $\mathcal{A}$ that converge to $f, g \in \overline{\mathcal{A}}$ respectively. Since $\mathcal{A}$ is a subalgebra, we know that $f_n g_n$ is in $\mathcal{A}$ for each $n \in \mathbb{N}$. We claim that $(f_n g_n)$ converges to fg, so that $fg \in \overline{A}$, which implies that $\overline{\mathcal{A}}$ is closed under multiplication. For all $\varepsilon > 0$, there exists $N = N(\frac{\varepsilon}{3\|g\|_{L^\infty}}) \in \mathbb{N}$ and $M = M(\frac{\varepsilon}{3\|f\|_{L^\infty}}) \in \mathbb{N}$ such that $\|f_n - f\|_{L^\infty} < \frac{\varepsilon}{3\|g\|_{L^\infty}}$ for $n \geq N$ and $\|g_n - g\|_{L^\infty} < \frac{\varepsilon}{3\|f\|_{L^\infty}}$ for $n \geq M$, and additionally there exists $L = L(\sqrt{\frac{\varepsilon}{3}}) \in \mathbb{N}$ such that $\|f_n - f\|_{L^\infty} < \sqrt{\frac{\varepsilon}{3}}$ for $n \geq L$ and $K = K(\sqrt{\frac{\varepsilon}{3}})$ such that $\|g_n - g\|_{L^\infty} < \sqrt{\frac{\varepsilon}{3}}$ for $n \geq K$. We deduce via the subadditivity of the norm that for $n \geq \max\{N, M, L, K\}$,

$$\begin{aligned}
\|f_n g_n - fg\|_{L^\infty} &= \|(f_n - f)(g_n - g) + (f_n - f)g + f(g_n - g)\|_{L^\infty} \\
&\leq \|f_n - f\|_{L^\infty}\|g_n - g\|_{L^\infty} + \|f_n - f\|_{L^\infty}\|g\|_{L^\infty} + \|f\|_{L^\infty}\|g_n - g\|_{L^\infty} \\
&< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \\
&= \varepsilon.
\end{aligned}$$

Here we have used the fact that for all $f_1, f_2 \in C(X)$,

$$\|f_1 f_2\|_{L^\infty} = \sup_{x \in X} |f_1(x) f_2(x)| \leq \left(\sup_{x \in X} |f_1(x)|\right)\left(\sup_{x \in X} |f_2(x)|\right) = \|f_1\|_{L^\infty}\|f_2\|_{L^\infty}. \qquad \square$$

It is certainly not true that every continuous function can be uniformly approximated by functions in any given subalgebra; the subalgebra consisting of the zero function is little use, for example. We need to impose some further conditions. One of the conditions we need is the following.

Definition 14.4. A subalgebra $\mathcal{A}$ of $C(X)$ *separates points* if for every pair of distinct points $x, y \in X$ with $x \neq y$, there exists some $f \in \mathcal{A}$ for which $f(x) \neq f(y)$.

Example 14.5.

- The subalgebra of $C([0,1])$ consisting of polynomials separates points, since if $y \neq x$, then $f(z) \coloneqq (z - x)(y - x)$ satisfies $f(x) = 0$ and $f(y) = (y - x)^2 > 0$;
- The subalgebra of $C(S^1)$ consisting of trigonometric polynomials separates points, since if $(x_1, y_1), (x_2, y_2) \in S^1$ are not equal, then either $x_1 \neq x_2$ or $y_1 \neq y_2$, and accordingly the trigonometric polynomial $f(x, y) = x$ or $f(x, y) = y$ separates these points.

It is clear that a subalgebra $\mathcal{A}$ cannot be dense in $C(X)$ if it does not separate points, since if there are points $x, y \in X$ such that $x \neq y$ but that $f(x) = f(y)$ for all $f \in \mathcal{A}$, then the continuous function $g \in C(X)$ given by $g(z) \coloneqq d(z, x)$ cannot be uniformly approximated by elements of $\mathcal{A}$.

A second condition we require is that the subalgebra be *nowhere vanishing*.

Definition 14.6. A subalgebra $\mathcal{A}$ of $C(X)$ is *nowhere vanishing* if for every $x \in X$, there exists some $f \in \mathcal{A}$ for which $f(x) \neq 0$.

Example 14.7. Once more, the subalgebra of polynomials in $C([0, 1])$ is nowhere vanishing: it contain the constant functions, and each nonzero constant function is nonzero at every point. The same goes for the subalgebra of trigonometric polynomials in $C(S^1)$.

Note that it is sufficient (but not necessary) for the subalgebra to contain the constant functions for it to be nowhere vanishing. The nowhere vanishing condition is necessary for the subalgebra to be dense: if $\mathcal{A}$ vanishes at some point $x \in X$, then the constant function $1 \in C(X)$ cannot be uniformly approximated by functions in $\mathcal{A}$.

14.2. **The Stone–Weierstrass theorem.** With these definitions in hand, we may finally state the Stone–Weierstrass theorem.

Theorem 14.8 (Stone–Weierstrass theorem)**.** *Let X be a compact metric space and let $\mathcal{A}$ be a subalgebra of $C(X)$ that separates points and is nowhere vanishing. Then $\mathcal{A}$ is dense in $C(X)$, so that $\overline{\mathcal{A}} = C(X)$.*

Thus if the two conditions we have just given are satisfied, then any function in $C(X)$ can be approximated uniformly by functions in $\mathcal{A}$, since if $f \in \overline{\mathcal{A}} = C(X)$, then for all $\varepsilon > 0$, there exists some $g \in B_\varepsilon(f) \cap \mathcal{A}$, so that

$$\|f - g\|_{L^\infty} = \sup_{x \in X} |f(x) - g(x)| < \varepsilon.$$

Corollary 14.9 (Weierstrass approximation theorem)**.** *For all $f \in C([0, 1])$ and for all $\varepsilon > 0$, there exists a polynomial $g : [0, 1] \to \mathbb{R}$ such that $\sup_{x \in [0,1]} |f(x) - g(x)| < \varepsilon$. In particular, there exists a sequence of polynomials (g_n) that converges uniformly to f.*

Similarly, for all $f \in C(S^1)$ and for all $\varepsilon > 0$, there exists a trigonometric polynomial $g : S^1 \to \mathbb{R}$ such that $\sup_{(x,y) \in S^1} |f(x, y) - g(x, y)| < \varepsilon$. In particular, there exists a sequence of trigonometric polynomials (g_n) that converges uniformly to f.

We prove the Stone–Weierstrass theorem in several steps. Our goal is to show that for any fixed $f \in C(X)$ and any fixed $\varepsilon > 0$, there exists some $g \in \mathcal{A}$ for which $d_\infty(f, g) < \varepsilon$.

For our first step, we need the following minor lemma.

Lemma 14.10. *Let $g : \mathbb{R} \to \mathbb{R}$ be given by $g(x) \coloneqq |x|$. Then g is continuous.*

Proof. This is clearly continuous at any point $x \neq 0$, since $g(x) = -x$ for $x < 0$ and $g(x) = x$ for $x > 0$, so we can just choose $\delta = \min\{\varepsilon, |x|\}$. Finally, it is also continuous at $x = 0$ by choosing $\delta = \varepsilon$, since if $|x - 0| < \delta$, then $|g(x) - g(0)| = |x| < \varepsilon$. □

Now we prove the following.

Lemma 14.11. *Let X be a compact metric space and let $\mathcal{A}$ be a subalgebra of $C(X)$. If $f \in \mathcal{A}$, then $|f| \in \overline{\mathcal{A}}$.*

Proof. We first observe that $|f|$ is continuous, being the composition of the continuous functions f and $|x|$. We shall produce a sequence (f_n) in $\mathcal{A}$ that converges in $C(X)$ to $|f| \in C(X)$, so that $|f|$ is a limit point of $\mathcal{A}$ and hence an element of $\overline{\mathcal{A}}$.

If $\sup_{x\in X}|f(x)| = 0$, then $f(x) = 0$ for all $x \in X$, and we can simply take $f_n(x) = 0$ for all $x \in X$ and $n \in \mathbb{N}$, which is necessarily in $\mathcal{A}$ since $\mathcal{A}$ is closed under scalar multiplication.

If $\sup_{x\in X}|f(x)| =: \|f\|_{L^\infty} > 0$, then we define a sequence (f_n) in $\mathcal{A}$ by setting $f_1(x) = 0$ for all $x \in X$, which is necessarily in $\mathcal{A}$ since $\mathcal{A}$ is closed under scalar multiplication, and defining iteratively

$$f_{n+1}(x) \coloneqq f_n(x) + \frac{1}{2\|f\|_{L^\infty}}\left(f(x)^2 - f_n(x)^2\right).$$

This is in $\mathcal{A}$ since $\mathcal{A}$ is closed under scalar multiplication, multiplication, and addition.

We claim that $0 \le f_n(x) \le |f(x)|$ for all $x \in X$ and $n \in \mathbb{N}$, from which it follows that $(f_n(x))$ is a nondecreasing sequence of nonnegative real numbers bounded above by $|f(x)|$. The base case $n = 1$ is trivially true. For the induction step, we note that

$$f_{n+1}(x) = f_n(x) + \frac{1}{2\|f\|_{L^\infty}}\left(f(x)^2 - f_n(x)^2\right) \ge f_n(x)$$

by the induction hypothesis. To show that $f_{n+1}(x) \le |f(x)|$, we note that

$$f_{n+1}(x) = \frac{1}{2\|f\|_{L^\infty}}f(x)^2 + \|f\|_{L^\infty} g\left(\frac{f_n(x)}{\|f\|_{L^\infty}}\right), \quad \text{where} \quad g(t) \coloneqq t - \frac{1}{2}t^2.$$

Since

$$0 \le \frac{f_n(x)}{\|f\|_{L^\infty}} \le 1$$

by the induction hypothesis and since g is increasing on $[0, 1]$, it follows that

$$g\left(\frac{f_n(x)}{\|f\|_{L^\infty}}\right) \le g\left(\frac{|f(x)|}{\|f\|_{L^\infty}}\right)$$

for all $x \in X$ by the induction hypothesis $f_n(x) \le |f(x)|$. Thus

$$f_{n+1}(x) \le \frac{1}{2\|f\|_{L^\infty}}f(x)^2 + \|f\|_{L^\infty} g\left(\frac{|f(x)|}{\|f\|_{L^\infty}}\right) = |f(x)|.$$

Thus for each $x \in X$, $(f_n(x))$ is a nondecreasing sequence of real numbers bounded above by $|f(x)|$, and hence this sequence converges to some real number $\widetilde{f}(x)$. We claim that $\widetilde{f}(x) = |f(x)|$. Indeed, we have that

$$\widetilde{f}(x) = \lim_{n\to\infty} f_{n+1}(x) = \lim_{n\to\infty}\left(f_n(x) + \frac{1}{2}\left(f(x)^2 - f_n(x)^2\right)\right) = \widetilde{f}(x) + \frac{1}{2}\left(f(x)^2 - \widetilde{f}(x)^2\right).$$

Thus $\widetilde{f}(x)^2 = f(x)^2$, and since $\widetilde{f}(x)$ is nonnegative, due to the fact that each $f_n(x)$ is nonnegative, we must have that $\widetilde{f}(x) = |f(x)|$.

Thus we have shown that (f_n) is a nondecreasing sequence in $\mathcal{A} \subseteq C(X)$ that converges pointwise to $|f| \in C(X)$. It remains to apply Dini's theorem, which tells us that the sequence of functions (f_n) converges uniformly, so that (f_n) converges in $C(X)$ to $|f|$ with respect to the sup-norm metric. □

With this in hand, we may now approximate the *minimum* and *maximum* of two functions.

Lemma 14.12. *Let X be a compact metric space and let $\mathcal{A}$ be a subalgebra of $C(X)$. If f, g are in $\mathcal{A}$, then $\min\{f, g\}$ and $\max\{f, g\}$ are in $\overline{\mathcal{A}}$. In particular, if $f_1, \ldots, f_n$ are all in $\mathcal{A}$, then $\min\{f_1, \ldots, f_n\}$ and $\max\{f_1, \ldots, f_n\}$ are in $\overline{\mathcal{A}}$.*

Proof. We have that

$$\min\{f,g\} = \frac{1}{2}\left(f+g-|f-g|\right),$$
$$\max\{f,g\} = \frac{1}{2}\left(f+g+|f-g|\right),$$

and these are both in $\overline{\mathcal{A}}$ since $\overline{\mathcal{A}}$ is closed under scalar multiplication and addition.

For taking the minimum or maximum of more than two functions, we replace $\mathcal{A}$ with $\overline{\mathcal{A}}$ and proceed by induction, using the key fact that $\overline{\mathcal{A}}$ is closed and hence equal to its own closure, as well as the fact that $\min\{f_1,\ldots,f_n,f_{n+1}\} = \min\{\min\{f_1,\ldots,f_n\},f_{n+1}\}$ and $\max\{f_1,\ldots,f_n,f_{n+1}\} = \max\{\max\{f_1,\ldots,f_n\},f_{n+1}\}$. □

Our next step is to construct a distinguished element of $\mathcal{A}$ dependent on $f \in C(X)$.

Lemma 14.13. *Let X be a compact metric space and let $\mathcal{A}$ be a subalgebra of $C(X)$ that separates points and is nowhere vanishing. Given any $f \in C(X)$ and any two points $x, y \in X$, there exists a function $h_{x,y} \in \mathcal{A}$ satisfying $h_{x,y}(x) = f(x)$ and $h_{x,y}(y) = f(y)$.*

Here we need to use *both* of our assumptions on the subalgebra, since such a construction would be impossible were the subalgebra to vanish at a point or to not separate points.

Proof. If $x = y$, the fact that $\mathcal{A}$ is nowhere vanishing means that there exists a function $\xi \in \mathcal{A}$ such that $\xi(x) \neq 0$. We now define $h_{x,x}(z) \coloneqq \frac{f(x)}{\xi(x)}\xi(z)$, so that $h_{x,x} \in \mathcal{A}$, as $\mathcal{A}$ is closed under scalar multiplication, and $h_{x,x}(x) = f(x)$.

If $x \neq y$, the fact that $\mathcal{A}$ separates points means that there exists a function $g \in \mathcal{A}$ such that $g(x) \neq g(y)$, while the fact that $\mathcal{A}$ is nowhere vanishing means that there exist functions $\xi, \eta \in \mathcal{A}$ such that $\xi(x) \neq 0$ and $\eta(y) \neq 0$. We now define

$$h_{x,y}(z) \coloneqq \frac{f(x)}{(g(x)-g(y))\xi(x)}(g(z)-g(y))\xi(z) + \frac{f(y)}{(g(x)-g(y))\eta(y)}(g(x)-g(z))\eta(z).$$

This is well-defined and is an element of $\mathcal{A}$ since $\mathcal{A}$ is closed under scalar multiplication, multiplication, and addition. Moreover, we have that $h_{x,y}(x) = f(x)$ and $h_{x,y}(y) = f(y)$. □

Remark 14.14. If $\mathcal{A}$ contains the constant functions, we can just choose $\xi = \eta = 1$, so that

$$h_{x,y}(z) \coloneqq \begin{cases} f(x) & \text{if } x = y, \\ \dfrac{f(x)(g(z)-g(y)) + f(y)(g(x)-g(z))}{(g(x)-g(y))} & \text{if } x \neq y. \end{cases}$$

We use this distinguished function $h_{x,y} \in \mathcal{A}$ to approximate $f \in C(X)$ from below.

Lemma 14.15. *Let X be a compact metric space and let $\mathcal{A}$ be a subalgebra of $C(X)$ that separates points and is nowhere vanishing. Given any $f \in C(X)$, any $\varepsilon > 0$, and any $x \in X$, there exists a function $h_x \in \overline{\mathcal{A}}$ for which $h_x(x) = f(x)$ and $h_x(z) < f(z) + \varepsilon$ for all $z \in X$.*

Proof. Fix $f \in C(X)$, $\varepsilon > 0$, and $x \in X$. Since f is continuous (indeed, it is uniformly continuous since X is compact), for each $y \in X$, there exists some $\delta_1 = \delta_1(\frac{\varepsilon}{2}, y) > 0$ such that $f(z) \in B_{\frac{\varepsilon}{2}}(f(y))$ whenever $z \in B_{\delta_1}(y)$. Moreover, given such $x, y \in X$, we take $h_{x,y}$ as in the previous lemma; since $h_{x,y}$ is continuous, there exists some $\delta_2 = \delta_2(\frac{\varepsilon}{2}, x, y) > 0$ such that $h_{x,y}(z) \in B_{\frac{\varepsilon}{2}}(h_{x,y}(y))$ whenever $z \in B_{\delta_2}(y)$. Letting $\delta(\varepsilon, x, y) \coloneqq \min\{\delta_1, \delta_2\}$ and using the fact that $h_{x,y}(y) = f(y)$, we deduce that if $z \in B_{\delta(\varepsilon,x,y)}(y)$, then

$$|h_{x,y}(z) - f(z)| \leq |h_{x,y}(z) - f(y)| + |f(y) - f(z)| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

In particular, if $z \in B_{\delta(\varepsilon,x,y)}(y)$, then $h_{x,y}(z) < f(z) + \varepsilon$.

The collection of open sets $\{B_{\delta(\varepsilon,x,y)}(y) : y \in X\}$ is an open cover of X, and so by the compactness of X, there exists a finite subcover, which is to say a finite collection of points $\{y_1, \dots, y_N\} \subseteq X$ such that $\bigcup_{n=1}^{N} B_{\delta(\varepsilon,x,y_n)}(y_n) = X$.

Now define $h_x \coloneqq \min\{h_{x,y_1}, \dots, h_{x,y_N}\}$. This is a function in $\overline{\mathcal{A}}$ since each h_{x,y_n} is in $\mathcal{A}$. We have that $h_{x,y_n}(x) = f(x)$ for each $n \in \{1, \dots, N\}$, and hence $h_c(x) = x$. Now for each $z \in X$, there exists some $n \in \{1, \dots, N\}$ for which $z \in B_{\delta(\varepsilon,x,y_n)}(y_n)$, and hence

$$h_x(z) = \min\{h_{x,y_1}(z), \dots, h_{x,y_N}(z)\} \leq h_{x,y_n}(z).$$

It remains to note that $h_{x,y_n}(z) < f(z) + \varepsilon$ since $z \in B_{\delta(\varepsilon,x,y_n)}(y_n)$. □

Finally, we also use this distinguished function $h_x \in \overline{\mathcal{A}}$ to simultaneously approximate f from above, which completes the proof of the Stone–Weierstrass theorem.

Proof of the Stone–Weierstrass theorem. We shall show that given any $f \in C(X)$ and any $\varepsilon > 0$, there exists some $h \in \overline{\mathcal{A}}$ for which

$$\|h - f\|_{L^\infty} \coloneqq \sup_{z \in X} |h(z) - f(z)| < \varepsilon.$$

This means that f is a limit point of $\overline{\mathcal{A}}$, and hence in the closure of $\overline{\mathcal{A}}$, but since $\overline{\mathcal{A}}$ is closed, it is equal to its own closure, and so $f \in \overline{\mathcal{A}}$. It follows that $\overline{\mathcal{A}} = C(X)$.

We have shown that given $f \in C(X)$, $\varepsilon > 0$, and $x \in X$, there exists a function $h_x \in \overline{\mathcal{A}}$ satisfying $h_x(x) = f(x)$ and $h_x(z) < f(z) + \varepsilon$ for all $z \in X$. Since f is continuous, there exists some $\delta_1 = \delta_1(\frac{\varepsilon}{2}, x) > 0$ such that $f(z) \in B_{\frac{\varepsilon}{2}}(f(x))$ whenever $z \in B_{\delta_1}(x)$, and since h_x is continuous, there exists some $\delta_2 = \delta_2(\frac{\varepsilon}{2}, x) > 0$ such that $h_x(z) \in B_{\frac{\varepsilon}{2}}(h_x(x))$ whenever $z \in B_{\delta_2}(x)$. Letting $\delta(\varepsilon, x) \coloneqq \min\{\delta_1, \delta_2\}$ and using the fact that $h_x(x) = f(x)$, we deduce that if $z \in B_{\delta(\varepsilon,x)}(x)$, then

$$|h_x(z) - f(z)| \leq |h_x(z) - f(x)| + |f(x) - f(z)| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

In particular, if $z \in B_{\delta(\varepsilon,x)}(x)$, then $h_x(z) > f(z) - \varepsilon$. On the other hand, each h_x satisfies $h_x(z) < f(z) + \varepsilon$ for all $z \in X$.

The collection of open sets $\{B_{\delta(\varepsilon,x)}(x) : x \in X\}$ is an open cover of X, and so by the compactness of X, there exists a finite subcover, which is to say a finite collection of points $\{x_1, \dots, x_M\} \subseteq X$ such that $\bigcup_{m=1}^{M} B_{\delta(\varepsilon,x_m)}(x_m) = X$.

Now define $h \coloneqq \max\{h_{x_1}, \dots, h_{x_M}\}$. This is a function in $\overline{\mathcal{A}}$ since each h_{x_m} is in $\overline{\mathcal{A}}$. By construction, we have that $h(z) < f(z) + \varepsilon$ for every $z \in X$ since $h_{x_m}(z) < f(z) + \varepsilon$ for each $m \in \{1, \dots, M\}$. On the other hand, we also have that $h(z) > f(z) - \varepsilon$, since there exists some $m \in \{1, \dots, M\}$ for which $z \in B_{\delta(\varepsilon,x_m)}(x_m)$, in which case

$$h(z) = \max\{h_{x_1}(z), \dots, h_{x_M}(z)\} \geq h_{x_m}(z) > f(z) - \varepsilon.$$

Thus we have shown that $|h(z) - f(z)| < \varepsilon$ for every $z \in X$. □

15. The Lebesgue measure

Recommended reading: [Pug15, §6.1, 6.2, 6.4], [SS05, §1.1–1.4], [Tao16, §7.1–7.5].

The Riemann integral gives a way of evaluating integrals of certain "nice" functions; functions for which the Riemann integral exists are called *Riemann integrable*. However, there are many functions, such as various discontinuous functions, that are *not* Riemann integrable.

Example 15.1. The function

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

is not Riemann integrable.

Moreover, the Riemann integral is not so well-behaved in some ways: the pointwise limit of Riemann integrable functions need not be Riemann integrable, for example.

For this reason, we study the notion of the *Lebesgue integral*, which is a more general version of integration on $\mathbb{R}^n$. In order to set up the Lebesgue integral rigorously, we first require the notion of the *Lebesgue measure*, which is a method of determining the volume of a set in $\mathbb{R}^n$. As we shall shortly discover, this is no easy task.

Our first goal is to define the concept of a *measurable set* Ω in $\mathbb{R}^n$ and to determine the *Lebesgue measure* $m(\Omega)$ of such a set.

15.1. **Outer measure.** We begin by introducing the notion of an *outer measure*. This is a weaker notion than Lebesgue measure, yet more intuitive. The definition of outer measure builds on the definition of the volume of a box.

Definition 15.2. A *closed box* or *closed rectangle* or simply *box* B in $\mathbb{R}^n$ is a set of the form

$$B = [a_1, b_1] \times \cdots \times [a_n, b_n] = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_j \in [a_j, b_j] \text{ for each } j \in \{1, \ldots, n\}\}$$

for some $a_1, b_1, \ldots, a_n, b_n \in \mathbb{R}$ with $a_j \leq b_j$ for each $j \in \{1, \ldots, n\}$. An *open box* or *open rectangle* is a set of the form

$$S = (a_1, b_1) \times \cdots \times (a_n, b_n) = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_j \in (a_j, b_j) \text{ for each } j \in \{1, \ldots, n\}\}.$$

The *volume* of a closed box $B = [a_1, b_1] \times \cdots \times [a_n, b_n]$ is

$$\operatorname{vol}(B) := \prod_{j=1}^{n} (b_j - a_j) = (b_1 - a_1) \cdots (b_n - a_n).$$

When $n = 1$, by "volume" we mean *length*, while for $n = 2$, we mean *area*.

Lemma 15.3. *Let $B, B_1, \ldots, B_L$ be boxes, and suppose that $B \subseteq \bigcup_{\ell=1}^{L} B_\ell$. Then* $\operatorname{vol}(B) \leq \sum_{\ell=1}^{L} \operatorname{vol}(B_\ell)$.

Proof. We first consider the case where $B = \bigcup_{\ell=1}^{L} B_\ell$ and the boxes $B_1, \ldots, B_L$ are *almost disjoint*: the interiors of any two boxes are disjoint. Consider the grid formed by extending the sides of all these boxes. This yields finitely many boxes $\widetilde{B_1}, \ldots, \widetilde{B_M}$ and a partition $J_1, \ldots, J_M$ of the integers between 1 and M such that the unions

$$B = \bigcup_{m=1}^{M} \widetilde{B_m}, \qquad B_\ell = \bigcup_{m \in J_\ell} \widetilde{B_m}$$

are almost disjoint. Then $\operatorname{vol}(B) = \sum_{m=1}^{M} \operatorname{vol}(\widetilde{B_m})$, since the boxes $\widetilde{B_1}, \ldots, \widetilde{B_M}$ are such that the sums of the side lengths are partitions of the side lengths of B. Similarly, $\operatorname{vol}(B_\ell) = \sum_{m \in J_\ell} \operatorname{vol}(\widetilde{B_m})$ for the same reason, and hence

$$\operatorname{vol}(B) = \sum_{m=1}^{M} \operatorname{vol}(\widetilde{B_m}) = \sum_{\ell=1}^{L} \sum_{m \in J_\ell} \operatorname{vol}(\widetilde{B_m}) = \sum_{\ell=1}^{L} \operatorname{vol}(B_\ell).$$

Finally, for the more general case where $B \subseteq \bigcup_{\ell=1}^{L} B_\ell$ and the boxes $B_1, \ldots, B_L$ need not be almost disjoint, we use the same approach. There are two alterations: a box $\widetilde{B_m}$

may not be interesect B, and additionally a box $\widetilde{B_m}$ may lie in more than one box B_ℓ and so may be counted more than once in the expression $\sum_{\ell=1}^{L} \mathrm{vol}(B_\ell)$. It follows that

$$\mathrm{vol}(B) \leq \sum_{m=1}^{M} \mathrm{vol}(\widetilde{B_m}) \leq \sum_{\ell=1}^{L} \sum_{m \in J_\ell} \mathrm{vol}(\widetilde{B_m}) = \sum_{\ell=1}^{L} \mathrm{vol}(B_\ell). \qquad \square$$

Definition 15.4. Let $\Omega \subseteq \mathbb{R}^n$ be a subset of $\mathbb{R}^n$. We say that a collection of boxes $\{B_j\}_{j \in J}$ *covers* Ω if $\Omega \subseteq \bigcup_{j \in J} B_j$.

Here J is simply some indexing set: it could be finite, countably infinite, or uncountable. For the purpose of measure theory, it is essential that we only consider *countable* (that is, finite or countably infinite) covers; that is, we take J to be a subset of $\mathbb{N}$. From here onwards, we shall always assume that this indeed the case: $J \subseteq \mathbb{N}$. Now we may define the outer measure.

Definition 15.5. Let $\Omega \subseteq \mathbb{R}^n$ be a subset of $\mathbb{R}^n$. The *outer measure* $m^*(\Omega)$ of Ω is the quantity

$$m^*(\Omega) \coloneqq \inf \left\{ \sum_{j \in J} \mathrm{vol}(B_j) : \{B_j\}_{j \in J} \text{ covers } \Omega \text{ and } J \subseteq \mathbb{N} \right\}.$$

Thus the outer measure of a set Ω is the infimum over all countable covers of Ω by boxes of the sums of the volumes of these boxes. The outer measure $m^*(\Omega)$ always exists (if we allow for the possibility that $m^*(\Omega) = \infty$), since every set in $\mathbb{R}^n$ can be covered by a countable collection of boxes. In particular, for all $\varepsilon > 0$, there exists a cover $\{B_j\}_{j \in J}$ of Ω such that

$$m^*(\Omega) \geq \sum_{j \in J} \mathrm{vol}(B_j) - \varepsilon.$$

Since $\mathrm{vol}(B)$ is nonnegative for each box B, $\sum_{j \in J} \mathrm{vol}(B_j)$ is nonnegative, and so $m^*(\Omega) \geq 0$. On the other hand, it could well be the case that $m^*(\Omega) = 0$: if Ω is a single point $\{(x_1, \ldots, x_n)\}$, then this is itself a closed box with all side lengths equal to zero. Moreover, it also could be the case that $m^*(\Omega) = \infty$: if $\Omega = \mathbb{R}^n$, then there is no cover of Ω by boxes whose sums of volumes is finite.

Lemma 15.6. *The outer measure of a closed box is equal to its volume.*

Proof. Since $\{B\}$ is a cover of B, clearly $m^*(B) \leq \mathrm{vol}(B)$. Conversely, let $\{B_j\}_{j \in J}$ be a covering of B by closed boxes. It suffices to prove that $\mathrm{vol}(B) \leq \sum_{j \in J} \mathrm{vol}(B_j)$. For each fixed $\varepsilon > 0$, we choose for each $j \in J$ an open box S_j containing B_j and satisfying $\mathrm{vol}(\overline{S_j}) \leq \mathrm{vol}(B_j) + \frac{\varepsilon}{2^j}$; such an open box exists by lengthening each side a small amount. Since B is compact and $\{S_j\}$ is an open cover of B, there exists a finite subcover $\{S_{j_1}, \ldots, S_{j_M}\}$ of B by open boxes. By taking the closure of these balls, we deduce that $B \subseteq \bigcup_{m=1}^{M} \overline{S_{j_m}}$. Thus

$$\mathrm{vol}(B) \leq \sum_{m=1}^{M} \mathrm{vol}(\overline{S_{j_m}}) \leq \sum_{m=1}^{M} \left(\mathrm{vol}(B_{j_m}) + \frac{\varepsilon}{2^{j_m}} \right) \leq \sum_{j \in J} \mathrm{vol}(B_j) + \varepsilon,$$

at which point we deduce that $\mathrm{vol}(B) \leq \sum_{j \in J} \mathrm{vol}(B_j)$ since $\varepsilon > 0$ was arbitrary, and so $\mathrm{vol}(B) \leq m^*(B)$. $\square$

Lemma 15.7. *The outer measure of an open box is equal to the volume of its closure.*

Proof. If S is an open box, then $m^*(S) \leq \mathrm{vol}(\overline{S})$ since $\overline{S}$ is a closed box containing S. For the reverse inequality, we note that this is immediate if $\mathrm{vol}(\overline{S}) = 0$. Otherwise, we take

a closed box B contained in S of volume $(1-\varepsilon)\operatorname{vol}(\overline{S})$, which exists by shortening each side a small amount. Since any cover of S by boxes is a cover of B, we have that

$$m^*(S) \geq m^*(B) = \operatorname{vol}(B) = (1-\varepsilon)\operatorname{vol}(\overline{S}).$$

Since $\varepsilon > 0$ was arbitrary, we deduce that $m^*(S) \geq \operatorname{vol}(\overline{S})$. □

Proposition 15.8. *The outer measure has the following properties.*

(1) *Empty set: the empty set has outer measure zero.*
(2) *Monotonicity: if $\Omega_1 \subseteq \Omega_2$, then $m^*(\Omega_1) \leq m^*(\Omega_2)$.*
(3) *Countable subadditivity: given $J \subseteq \mathbb{N}$, if $\Omega = \bigcup_{j\in J} \Omega_j$, then $m^*(\Omega) \leq \sum_{j\in J} m^*(\Omega_j)$.*
(4) *Translation invariance: given $\Omega \subseteq \mathbb{R}^n$ and $x \in \mathbb{R}^n$, we have that $m^*(\Omega + x) = m^*(\Omega)$, where $\Omega + x \coloneqq \{y \in \mathbb{R}^n : y - x \in \Omega\}$.*
(5) *Dilation equivariance: given $\Omega \subseteq \mathbb{R}^n$ and $\delta \in (0,\infty)$, we have that $m^*(\delta\Omega) = \delta^n m^*(\Omega)$, where $\delta\Omega \coloneqq \{y \in \mathbb{R}^n : \delta^{-1}y \in \Omega\}$.*

Proof.
(1) Any box is a cover of the empty set; in particular, we could take a box of volume zero (with a side of length zero).
(2) Any cover of Ω_2 by boxes is also a cover of Ω_1.
(3) If $m^*(\Omega_j) = \infty$ for some $j \in J$, then this is trivially true. Otherwise, for each $\varepsilon > 0$ and each $j \in J$, there exists a covering $\{B_{j,k}\}_{k\in K}$ of Ω_j by closed boxes such that

$$\sum_{k\in K} \operatorname{vol}(B_{j,k}) \leq m^*(\Omega_j) + \frac{\varepsilon}{2^j}.$$

Then Ω is covered by $\{B_{j,k}\}_{j\in J,\, k\in K}$ and satisfies

$$m^*(\Omega) \leq \sum_{\substack{j\in J\\ k\in K}} \operatorname{vol}(B_{j,k}) \leq \sum_{j\in J} \left(m^*(\Omega_j) + \frac{\varepsilon}{2^j}\right) \leq \sum_{j\in J} m^*(\Omega_j) + \varepsilon.$$

Since this is true for every $\varepsilon > 0$, we deduce that $m^*(\Omega) \leq \sum_{j\in J} m^*(\Omega_j)$.
(4) This follows from the fact that the volume of a box is unchanged by translation together with the fact that $\{B_j\}_{j\in J}$ is a cover of Ω by boxes if and only if $\{B_j + x\}_{j\in J}$ is a cover of $\Omega + x$ by boxes.
(5) The volume of the dilation $\delta B = [\delta a_1, \delta b_1] \times \cdots \times [\delta a_n, \delta b_n]$ of a box $B = [a_1, b_1] \times \cdots \times [a_n, b_n]$ is

$$\operatorname{vol}(\delta B) \coloneqq \prod_{j=1}^n (\delta b_j - \delta a_j) = \delta^n \prod_{j=1}^n (\delta b_j - \delta a_j) =: \delta^n \operatorname{vol}(B).$$

The result then follows from the fact that $\{B_j\}_{j\in J}$ is a cover of Ω by boxes if and only if $\{\delta B_j\}_{j\in J}$ is a cover of $\delta\Omega$ by boxes. □

Already we are able to show a connection between cardinality and outer measure.

Corollary 15.9. *A countable set has outer measure zero. In particular, $\mathbb{Q}^n$ has outer measure zero.*

Proof. If Ω is countable, then $\Omega = \bigcup_{j\in J} \Omega_j$, where J is countable and each Ω_j consists of a single point. Since a single point has outer measure zero, it follows that

$$0 \leq m^*(\Omega) \leq \sum_{j\in J} m^*(\Omega_j) = \sum_{j\in J} 0 = 0. \qquad \square.$$

As we shall later see, however, there exist uncountable sets that also have outer measure zero, so cardinality alone is not enough to classify sets of outer measure zero.

We have shown countable subadditivity. We would prefer something stronger: finite additivity, namely if $\Omega = \Omega_1 \cup \Omega_1$ with Ω_1 and Ω_2 *disjoint*, then $m^*(\Omega) = m^*(\Omega_1) + m^*(\Omega_2)$. A weaker version of this is the following.

Lemma 15.10 (Finite additivity of the outer Lebesgue measure for separated sets)**.** *If* $\Omega = \Omega_1 \cup \Omega_2$ *and*

$$\delta \coloneqq \inf_{\substack{x_1 \in \Omega_1 \\ x_2 \in \Omega_2}} d_2(x_1, x_2) > 0,$$

then

$$m^*(\Omega) = m^*(\Omega_1) + m^*(\Omega_2).$$

Proof. By countable subadditivity, we have that $m^*(\Omega) \leq m^*(\Omega_1) + m^*(\Omega_2)$. For the reverse inequality, fix $\varepsilon > 0$, and choose a covering $\{B_j\}_{j \in J}$ of Ω for which $\Omega \subseteq \bigcup_{j \in J} B_j$ and $m^*(\Omega) \geq \sum_{j \in J} \operatorname{vol}(B_j) - \varepsilon$. Without loss of generality, we may suppose that each box has diameter less than δ, since otherwise we can divide up these boxes into smaller boxes. Thus each box B_j intersects at most one of Ω_1 and Ω_2. Letting J_1 denote the indices for which $B_j \cap \Omega_1 \neq \emptyset$ and similarly defining J_2 for Ω_2, so that $J_1 \cup J_2 \subseteq J$ and $J_1 \cap J_2 = \emptyset$, we get coverings $\Omega_1 \subseteq \bigcup_{j \in J_1} B_j$ and $\Omega_2 \subseteq \bigcup_{j \in J_2} B_j$. Thus

$$m^*(\Omega_1) + m^*(\Omega_2) \leq \sum_{j \in J_1} \operatorname{vol}(B_j) + \sum_{j \in J_2} \operatorname{vol}(B_j) \leq \sum_{j \in J} \operatorname{vol}(B_j) \leq m^*(\Omega) + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, the desired inequality follows. □

We shall later show that unfortunately finite additivity for the outer measure does *not* hold without this condition on the distance between Ω_1 and Ω_2. Nonetheless, things do not go awry if we are dealing with almost disjoint boxes.

Lemma 15.11. *Given a set* $\Omega \subseteq \mathbb{R}^n$ *that is a countable union of almost disjoint boxes* $\{B_j\}_{j \in J}$*, so that* $\Omega = \bigcup_{j \in J} B_j$*, we have that*

$$m^*(\Omega) = \sum_{j \in J} \operatorname{vol}(B_j).$$

Proof. We have that $m^*(\Omega) \leq \sum_{j \in J} \operatorname{vol}(B_j)$ by countable subadditivity. For the reverse inequality, fix $\varepsilon > 0$. For each $j \in J$, let $\widetilde{B_j}$ denote a box strictly contained in B_j such that $\operatorname{vol}(\widetilde{B_j}) \geq \operatorname{vol}(B_j) - \frac{\varepsilon}{2^j}$, which exists by shortening each side a small amount. Then any finite collection of such boxes $\widetilde{B_j}$ are not just almost disjoint but completely disjoint and are a positive distance apart, which implies by monotonicity and finite additivity that for every $N \in \mathbb{N}$,

$$m^*(\Omega) \geq m^* \left(\bigcup_{\substack{j=1 \\ j \in J}}^{N} \widetilde{B_j} \right) = \sum_{\substack{j=1 \\ j \in J}}^{N} \operatorname{vol}(\widetilde{B_j}) \geq \sum_{\substack{j=1 \\ j \in J}}^{N} \left(\operatorname{vol}(B_j) - \frac{\varepsilon}{2^j} \right) \geq \sum_{\substack{j=1 \\ j \in J}}^{N} \operatorname{vol}(B_j) - \varepsilon.$$

Since $N \in \mathbb{N}$ was arbitrary, we have that $m^*(\Omega) \geq \sum_{j \in J} \operatorname{vol}(B_j) - \varepsilon$, and since $\varepsilon > 0$ was arbitrary, we conclude that $m^*(\Omega) \geq \sum_{j \in J} \operatorname{vol}(B_j)$. □

15.2. **Measurable sets.** Open sets are of particular importance in measure theory.

Lemma 15.12. *Every open subset $E \subseteq \mathbb{R}^n$ of $\mathbb{R}^n$ may be written as a countable union of almost disjoint boxes.*

Proof. We prove something slightly stronger in terms of cubes rather than boxes. We begin by considering the grid in $\mathbb{R}^n$ obtained by taking closed cubes of side length 1 whose vertices have integer coordinates. We take all of these cubes that lie ocmpletely inside E to be part of this countable union. For the cubes that intersect both E and E^c nontrivially, we further bisect these into 2^n additional cubes of side length $1/2$. Again, we take all of these cubes that lie completely inside E to be part of this countable union.

We iterate this process by bisecting cubes of side length $2^{-\ell}$ that intersect both E and E^c into further cubes of side length $2^{-\ell-1}$, and continue this process as ℓ tend to infinity. We end up with a countable union of countable sets, which is countable. By construction, these cubes are almost disjoint. Finally, the union of these cubes is equal to E, since if $x \in E$, there exists a cube of side length $2^{-\ell}$ that contains x and is entirely contained in E, since E is open. □

This gives us a straightforward way of determining the outer measure of an open set E.

Corollary 15.13. *Every open set $E \subseteq \mathbb{R}^n$ can be written as a countable union of almost disjoint boxes $\{B_j\}_{j\in J}$ for which $m^*(E) = \sum_{j\in J} \operatorname{vol}(B_j)$.*

Notably, this is true regardless of how we write E as a countable union of disjoint boxes; there may be more than one way to do so, but the outer measure remains unchanged.

Next, we observe that we can use open sets to determine the outer measure of an *arbitrary* set $\Omega \subseteq \mathbb{R}^n$.

Lemma 15.14. *Given $\Omega \subseteq \mathbb{R}^n$, we have that*

$$m^*(\Omega) = \inf_{\substack{E \supseteq \Omega \\ E \text{ open}}} m^*(E),$$

where the infimum is over all open sets $E \subseteq \mathbb{R}^n$ containing Ω.

Proof. Via monotonicity, we have that $m^*(\Omega) \leq m^*(E)$ for every open set $E \supseteq \Omega$. For the reverse inequality, we must show that for every $\varepsilon > 0$, there exists an open set $E \supseteq \Omega$ for which $m^*(E) \leq m^*(\Omega) + \varepsilon$. We begin by choosing a covering $\{B_j\}_{j\in J}$ of Ω by boxes such that $\Omega \subseteq \bigcup_{j\in J} B_j$ with

$$\sum_{j\in J} \operatorname{vol}(B_j) \leq m^*(\Omega) + \frac{\varepsilon}{2}.$$

Let S_j denote an open box containing B_j and satisfying $m^*(S_j) \leq \operatorname{vol}(B_j) + \frac{\varepsilon}{2^{j+1}}$, which can be done by lengthening each side by a small amount. Then $E = \bigcup_{j\in J} S_j$ is a union of open sets, and hence open, and so by countable subadditivity,

$$m^*(E) \leq \sum_{j\in J} m^*(S_j) \leq \sum_{j\in J} \left(\operatorname{vol}(B_j) + \frac{\varepsilon}{2^{j+1}}\right) \leq \sum_{j\in J} \operatorname{vol}(B_j) + \frac{\varepsilon}{2} \leq m^*(\Omega) + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, the desired inequality follows. □

As we have mentioned previously, finite additivity may fail for the outer measure of two sets in $\mathbb{R}^n$. There are several approaches to overcome the failure of finite additivity. The approach that we choose is to only work with certain classes of sets in $\mathbb{R}^n$: *measurable sets*.

Definition 15.15. A subset $\Omega \subseteq \mathbb{R}^n$ is *Lebesgue measurable* or simply *measurable* if for all $\varepsilon > 0$, there exists an open set $E \supseteq \Omega$ containing Ω for which

$$m^*(E \setminus \Omega) < \varepsilon.$$

The *Lebesgue measure* or simply *measure* $m(\Omega)$ of a measurable set Ω is simply its outer measure, namely

$$m(\Omega) \coloneqq m^*(\Omega).$$

The Lebesgue measure has some straightforward properties.

Proposition 15.16.

(1) *Every open set in $\mathbb{R}^n$ is measurable.*
(2) *If $m^*(\Omega) = 0$, then Ω is measurable.*
(3) *A countable union of measurable sets is measurable.*
(4) *Compact sets are measurable.*
(5) *Closed sets are measurable.*
(6) *The complement of a measurable set is measurable.*
(7) *A countable intersection of measurable sets is measurable.*
(8) *A translation of a measurable set is measurable.*
(9) *A dilation of a measurable set is measurable.*

Proof.
(1) This follows by taking $E = \Omega$ and recalling that the empty set has outer measure zero.
(2) For every $\varepsilon > 0$, there exists an open set $E \supseteq \Omega$ for which $m^*(E \setminus \Omega) < \varepsilon$. Since $E = \Omega \cup (E \setminus \Omega)$, it follows by countable subadditivity that $m^*(E) < m^*(\Omega) + \varepsilon$. So if $m^*(\Omega) = 0$, then as $E \setminus \Omega \subseteq E$, we have by monotonicity that $m^*(E \setminus \Omega) \leq m^*(E) < \varepsilon$.
(3) Suppose that $\Omega = \bigcup_{j \in J} \Omega_j$, where each Ω_j is measurable. Fix $\varepsilon > 0$. Then for each $j \in N$, there exists an open set E_j with $E_j \supseteq \Omega_j$ and $m^*(E_j \setminus \Omega_j) < \frac{\varepsilon}{2^j}$. The countable union $E = \bigcup_{j \in J} E_j$ is open, contains Ω, and satisfies $E \setminus \Omega \subseteq \bigcup_{j \in J}(E_j \setminus \Omega_j)$, and so by monotonicity and countable subadditivity,

$$m^*(E \setminus \Omega) \leq m^* \left(\bigcup_{j \in J} (E_j \setminus \Omega_j) \right) \leq \sum_{j \in J} m^*(E_j \setminus \Omega_j) < \sum_{j \in J} \frac{\varepsilon}{2^j} \leq \varepsilon.$$

(4) Suppose that Ω is compact, so that it is bounded, and hence contained inside a box, and so has finite outer measure. Fix $\varepsilon > 0$. There exists an open set $E \supseteq \Omega$ for which $m^*(\Omega) > m^*(E) - \varepsilon$. Since Ω is closed, $E \setminus \Omega$ is open, and so may be written as a countable union of almost disjoint boxes: $E \setminus \Omega = \bigcup_{j \in J} B_j$. For each integer $N \in \mathbb{N}$, the finite union $K_N \coloneqq \bigcup_{\substack{j=1 \\ j \in J}}^{N} B_j$ is compact, which implies that

$$\inf_{\substack{x_1 \in \Omega \\ x_2 \in K_N}} d_2(x_1, x_2) > 0$$

since Ω and K_N are compact and disjoint. Since $\Omega \cup K_N \subseteq E$, we deduce by monotonicity and finite additivity that

$$m^*(E) \geq m^*(\Omega \cup K_N) = m^*(\Omega) + m^*(K_N) = m^*(\Omega) + \sum_{\substack{j=1 \\ j \in J}}^{N} m^*(B_j).$$

It follows that $\sum_{\substack{j=1\\ j\in J}}^{N} m^*(B_j) < \varepsilon$. Since $N \in \mathbb{N}$ was arbitrary, we thereby have that $\sum_{j\in J} m^*(B_j) < \varepsilon$. Via countable subadditivity, we deduce that

$$m^*(E \setminus \Omega) \leq \sum_{j\in J} m^*(B_j) < \varepsilon,$$

and so Ω is measurable.
(5) If Ω is closed, then Ω is equal to a countable union of compact sets, namely $\Omega = \bigcup_{m=1}^{\infty} \Omega \cap \overline{B_m(0)}$, where $B_m(0)$ denotes the open ball of radius m centred at 0. Since a countable union of measurable sets is measurable, Ω is measurable.
(6) If Ω is measurable, then for each $j \in \mathbb{N}$, there exists an open set $E_j \subseteq \Omega$ for which $m^*(E_j \setminus \Omega) \leq 2^{-j}$. The complement $F_j := E_j^c$ is closed, hence measurable, and so the countable union $F := \bigcup_{j=1}^{\infty} F_j$ is also measurable. We then have that $F \subseteq \Omega^c$ and $\Omega^c \setminus F \subseteq E_N \setminus \Omega$ for each $N \in \mathbb{N}$, and so by monotonicity,

$$m^*(\Omega^c \setminus F) \leq m^*(E_N \setminus \Omega) \leq 2^{-N}.$$

Since $N \in \mathbb{N}$ was arbitrary, it follows that the left-hand side is equal to 0, and so $\Omega^c \setminus F$ is measurable. Thus Ω^c is itself measurable, being the union the measurable sets F and $\Omega^c \setminus F$.
(7) If Ω_j is measurable, then Ω_j^c is measurable, being the complement of a measurable set. Thus $\bigcup_{j\in J} \Omega_j^c$ is measurable, being the countable union of measurable sets. Finally, $\bigcap_{j\in J} \Omega_j = \left(\bigcup_{j\in J} \Omega_j^c\right)^c$ is measurable, being the complement of a measurable set.
(8) If Ω is measurable, then for all $\varepsilon > 0$, there exists some open set $E \supseteq \Omega$ for which $m^*(E \setminus \Omega) < \varepsilon$. Then given $x \in \mathbb{R}^n$, $E + x$ is also open and satisfies $E + x \supseteq \Omega + x$ and $(E + x) \setminus (\Omega + x) = (E \setminus \Omega) + x$, and so by the translation invariance of the outer measure,

$$m^*((E + x) \setminus (\Omega + x)) = m^*((E \setminus \Omega) + x) = m^*(E \setminus \Omega) < \varepsilon.$$

(9) If Ω is measurable, then given $\delta \in (0, \infty)$, for all $\varepsilon > 0$, there exists some open set $E \supseteq \Omega$ for which $m^*(E \setminus \Omega) < \frac{\varepsilon}{\delta^n}$. As δE is also open and satisfies $\delta E \supseteq \delta\Omega$ and $(\delta E) \setminus (\delta\Omega) = \delta(E \setminus \Omega)$, and so by the dilation equivariance of the outer measure,

$$m^*((\delta E) \setminus (\delta\Omega)) = m^*(\delta(E \setminus \Omega)) = \delta^n m^*(E \setminus \Omega) < \varepsilon. \qquad \square$$

Thus we have shown that the family of measurable sets is closed under some familiar operations of set theory:

- countable unions;
- countable intersections;
- complements.

This means that the family of measurable sets is a *σ-algebra*, namely a collection of sets closed under countable unions, countable intersections, and complements. For this reason, the family of measurable sets is called the *σ-algebra of Lebesgue measurable sets*. There are plenty of other σ-algebras; perhaps the most natural is the *Borel σ-algebra*, which is the smallest σ-algebra in $\mathbb{R}^n$ that contains all open sets.

The σ-algebra of measurable sets is closed under under *countable* unions and intersections, not just *finite* unions and intersections. As we shall later see, however, it is not closed under *uncountable* unions and intersections.

We are now able to show that countable additivity holds for the Lebesgue measure.

Theorem 15.17 (Countable additivity of the Lebesgue measure)**.** *Suppose that $\Omega = \bigcup_{j\in J} \Omega_j$, where $\{\Omega_j\}_{j\in J}$ is a countable collection of disjoint measurable sets. Then Ω is*

measurable and

$$m(\Omega) = \sum_{j\in J} m(\Omega_j).$$

Proof. The measurability of Ω follows from the fact that a countable union of measurable sets is measurable. To prove countable additivity, we begin by observing that the inequality $m(\Omega) \leq \sum_{j\in J} m(\Omega_j)$ holds by countable subadditivity.

For the reverse inequality, we first prove this in the special case that each Ω_j is bounded. Fix $\varepsilon > 0$. Since each complement Ω_j^c is measurable, there exists an open set $E_j \supseteq \Omega_j^c$ for which $m^*(E_j \setminus \Omega_j^c) < \frac{\varepsilon}{2^j}$, and hence a closed set $F_j \coloneqq E_j^c \subseteq \Omega_j$ for which $m^*(\Omega_j \setminus F_j) < \frac{\varepsilon}{2^j}$, so that by finite subadditivity,

$$m(\Omega_j) = m^*(\Omega_j) \leq m^*(F_j) + m^*(\Omega_j \setminus F_j) \leq m(F_j) + \frac{\varepsilon}{2^j}.$$

For each $N \in \mathbb{N}$, the collection $\{F_j : j \in J \cap \{1, \dots, N\}\}$ consists of sets that are compact and disjoint, and hence a positive distance apart, so that

$$m\left(\bigcup_{\substack{j=1\\ j\in J}}^{N} F_j\right) = \sum_{\substack{j=1\\ j\in J}}^{N} m(F_j)$$

by finite additivity. Since $\Omega \supseteq \bigcup_{\substack{j=1\\ j\in J}}^{N} F_j$, we have by monotonicity that

$$m(\Omega) \geq \sum_{\substack{j=1\\ j\in J}}^{N} m(F_j) \geq \sum_{\substack{j=1\\ j\in J}}^{N} \left(m(\Omega_j) - \frac{\varepsilon}{2^j}\right) \geq \sum_{\substack{j=1\\ j\in J}}^{N} m(\Omega_j) - \varepsilon.$$

Since $N \in \mathbb{N}$ was arbitrary, we therefore have that $m(\Omega) \geq \sum_{j\in J} m(\Omega_j) - \varepsilon$, and since $\varepsilon > 0$ was arbitrary, we deduce that $m(\Omega) \geq \sum_{j\in J} m(\Omega_j)$.

Now we remove the assumption that each Ω_j is bounded. For $k \in \mathbb{N}$, let $B_k(0)$ denote the ball of radius k centred at 0, so that $B_k(0) \subseteq B_{k+1}(0)$ for all $k \in \mathbb{N}$ and $\bigcup_{k=1}^{\infty} B_k(0) = \mathbb{R}^n$. We let

$$C_k \coloneqq \begin{cases} B_1(0) & \text{for } k = 1, \\ B_k(0) \setminus B_{k-1}(0) & \text{for } k \geq 2. \end{cases}$$

We then define sets $\Omega_{j,k} \coloneqq \Omega_j \cap C_k$, so that $\{\Omega_{j,k}\}_{j\in J,\, k\in\mathbb{N}}$ is a countable collection of disjoint sets satisfying $\bigcup_{\substack{j\in J\\ k\in\mathbb{N}}} \Omega_{j,k} = \Omega$. Moreover, each $\Omega_{j,k}$ is bounded, being contained in $B_k(0)$, and measurable, since C_k is measurable, being the complement of two open sets, and hence $\Omega_{j,k}$ is measurable, being the intersection of two measurable sets. It follows that for each $j \in J$, $\{\Omega_{j,k}\}_{k\in\mathbb{N}}$ is a collection of disjoint measurable bounded sets satisfying $\bigcup_{k=1}^{\infty} \Omega_{j,k} = \Omega_j$. Thus by reducing to the bounded case, we have shown that

$$m(\Omega) = \sum_{\substack{j\in J\\ k\in\mathbb{N}}} m(\Omega_{j,k}) = \sum_{j\in J} m(\Omega_j). \qquad \square$$

A useful consequence concerns the measure of the union or intersection of nested sets.

Lemma 15.18. *Let $\{\Omega_k\}_{k\in\mathbb{N}}$ be a countably infinite collection of measurable sets of $\mathbb{R}^n$.*
(1) *If $\Omega_k \subseteq \Omega_\ell$ whenever $k \leq \ell$, then*

$$m\left(\bigcup_{k=1}^{\infty} \Omega_k\right) = \lim_{K\to\infty} m(\Omega_K).$$

(2) *If $\Omega_k \supseteq \Omega_\ell$ whenever $k \leq \ell$ and $m(\Omega_1)$ is finite, then*

$$m\left(\bigcap_{k=1}^{\infty} \Omega_k\right) = \lim_{K\to\infty} m(\Omega_K).$$

Note that (2) is false if $m(\Omega_1) = \infty$: take, for example, $\Omega_k \coloneqq B_k(0)^c$, so that $m(\Omega_k) = \infty$ for all $k \in \mathbb{N}$ whereas $\bigcap_{k=1}^{\infty} \Omega_k = \emptyset$, so that $m(\bigcap_{k=1}^{\infty} \Omega_k) = 0$.

Proof.
(1) Let

$$A_k \coloneqq \begin{cases} \Omega_1 & \text{if } k = 1, \\ \Omega_k \setminus \Omega_{k-1} & \text{if } k \geq 2. \end{cases}$$

Each A_k is measurable, since complements of measurable sets are measurable, and the collection $\{A_k\}$ is disjoint. Moreover, $\bigcup_{k=1}^{\infty} \Omega_k = \bigcup_{k=1}^{\infty} A_k$, and this is measurable since these are countable unions of measurable sets. By countable additivity, we deduce that

$$m\left(\bigcup_{k=1}^{\infty} \Omega_k\right) = m\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} m(A_k) = \lim_{K\to\infty} \sum_{k=1}^{K} m(A_k) = \lim_{K\to\infty} m\left(\bigcup_{k=1}^{K} A_k\right).$$

It remains to note that $\bigcup_{k=1}^{K} A_k = \Omega_K$ for any $K \in \mathbb{N}$.
(2) For each $k \in \mathbb{N}$, define $A_k \coloneqq \Omega_k \setminus \Omega_{k+1}$, which is measurable since complements of measurable sets are measurable. Moreover, we have that

$$\Omega_1 = \bigcap_{k=1}^{\infty} \Omega_k \cup \bigcup_{k=1}^{\infty} A_k,$$

and this union is disjoint. Finally, all of these sets are measurable, and so by countable additivity, we have that

$$m(\Omega_1) = m\left(\bigcap_{k=1}^{\infty} \Omega_k\right) + \sum_{k=1}^{\infty} m(A_k) = m\left(\bigcap_{k=1}^{\infty} \Omega_k\right) + \sum_{k=1}^{\infty} \left(m(\Omega_k) - m(\Omega_{k+1})\right).$$

We have that

$$\sum_{k=1}^{\infty} \left(m(\Omega_k) - m(\Omega_{k+1})\right) = \lim_{K\to\infty} \sum_{k=1}^{K-1} \left(m(\Omega_k) - m(\Omega_{k+1})\right) = m(\Omega_1) - \lim_{K\to\infty} m(\Omega_K)$$

as this is a telescoping sum. Since $m(\Omega_1)$ is finite, we therefore have that

$$m\left(\bigcap_{k=1}^{\infty} \Omega_k\right) = \lim_{K\to\infty} m(\Omega_K). \qquad \square$$

15.3. **Nonmeasurable sets.** We have shown that measurable sets satisfy countable additivity. We shall now show that there exist sets for which this fails — indeed, sets for which *finite additivity* fails. This implies the existence of *nonmeasurable sets*. We only prove this for $\mathbb{R}^n$ with $n = 1$, though this method can be extended to arbitrary $n \in \mathbb{N}$. The construction of a nonmeasurable set relies on the axiom of choice applied to an explicit equivalence relation among real numbers in the interval $[0, 1]$.

Definition 15.19. We say that two numbers $x, y \in [0, 1]$ are *equivalent* and write $x \sim y$ if $x - y \in \mathbb{Q}$.

This is an equivalence relation, since it satisfies the following three conditions:

- $x \sim x$ for all $x \in S$ (reflexivity);
- If $x \sim y$ and $y \sim z$, then $x \sim z$ (transitivity);

• If $x \sim y$ then $y \sim x$ (symmetry).

Recall that an *equivalence class* $\mathcal{E}_\alpha$ of $[0,1]$ with respect to this equivalence relation $\sim$ is a set of the form

$$\mathcal{E}_\alpha \coloneqq \{x \in [0,1] : x \sim x_\alpha\} = \{x \in [0,1] : x - x_\alpha \in \mathbb{Q}\}$$

for some $x_\alpha \in [0,1]$, in which case x_α is said to be a *representative* of the equivalence class $\mathcal{E}_\alpha$. Necessarily $\mathcal{E}_\alpha$ is a countable set, since it consists of real numbers that are rational translates of each other.

Two equivalence classes $\mathcal{E}_\alpha, \mathcal{E}_\beta$ are either equal or are disjoint, and [0,1] is the disjoint union of all such equivalence classes, so that $[0,1] = \bigcup_\alpha \mathcal{E}_\alpha$ as α varies over an indexing set of equivalence classes. Since $[0,1]$ is uncountable and each $\mathcal{E}_\alpha$ is countable, this indexing set must be uncountable.

Definition 15.20. Let $\mathcal{N} \subset [0,1]$ be a set obtained by choosing exactly one element x_α from each equivalence class $\mathcal{E}_\alpha$ and letting $\mathcal{N}$ be the union of these chosen elements.

Note that we are using the axiom of choice to construct $\mathcal{N}$; indeed, we are using the axiom of choice for uncountable sets, since this indexing set is uncountable.

Theorem 15.21. *The set $\mathcal{N}$ is not measurable. In particular, there exist nonmeasurable sets.*

Proof. Let $\{r_k\}$ be an enumeration of all the rational numbers in $[-1,1]$ and define for each $k \in \mathbb{N}$ the translate $\mathcal{N}_k \coloneqq \mathcal{N} + r_k$. We claim that these sets are disjoint and that

$$[0,1] \subset \bigcup_{k=1}^{\infty} \mathcal{N}_k \subset [-1,2].$$

To prove disjointness, we suppose that $\mathcal{N}_k \cap \mathcal{N}_{k'} \neq \emptyset$, so that there exist rational numbers $r_k, r_{k'} \in [-1,1]$ and real numbers $x_\alpha, x_\beta \in \mathcal{N}$ for which $x_\alpha + r_k = x_\beta + r'_k$, so that $x_\alpha - x_\beta = r_k - r_{k'} \in \mathbb{Q}$. It follows that $x_\alpha \sim x_\beta$; since $\mathcal{N}$ only contains *one* element from each equivalence class, we must have that $x_\alpha = x_\beta$, so that $r_k = r_{k'}$, and hence $k = k'$. Thus $\mathcal{N}_k \cap \mathcal{N}_{k'} = \emptyset$ whenever $k \neq k'$.

Next, we note that the fact that $\bigcup_{k=1}^{\infty} \mathcal{N}_k \subset [-1,2]$ is a simple consequence of the fact that $r_k \in [-1,1]$ and $\mathcal{N} \subset [0,1]$, so that $\mathcal{N}_k \subset [-1,2]$ for each $k \in \mathbb{N}$.

Finally, if $x \in [0,1]$, then $x \sim x_\alpha$ for some x_α, so that $x - x_\alpha = r$ for some $r \in [-1,1]$. But then $r = r_k$ for some $k \in \mathbb{N}$, so that $x \in \mathcal{N}_k$. Thus $[0,1] \subset \bigcup_{k=1}^{\infty} \mathcal{N}_k$.

With this in hand, we now suppose in order to obtain a contradiction that $\mathcal{N}$ is measurable. Then for each $k \in \mathbb{N}$, $\mathcal{N}_k \coloneqq \mathcal{N} + r_k$ is the translate of a measurable set, hence measurable, and so $\bigcup_{k=1}^{\infty} \mathcal{N}_k$ is measurable, being the countable union of measurable sets. By monotonicity, we have that

$$1 = m([0,1]) \leq m\left(\bigcup_{k=1}^{\infty} \mathcal{N}_k\right) \leq m([-1,2]) = 3.$$

Moreover, $\bigcup_{k=1}^{\infty} \mathcal{N}_k$ is a disjoint union of measurable sets, and so by countable additivity and translation invariance, we have that

$$m\left(\bigcup_{k=1}^{\infty} \mathcal{N}_k\right) = \sum_{k=1}^{\infty} m(\mathcal{N}_k) = \sum_{k=1}^{\infty} m(\mathcal{N} + r_k) = \sum_{k=1}^{\infty} m(\mathcal{N}).$$

Thus we have shown that

$$1 \leq \sum_{k=1}^{\infty} m(\mathcal{N}) \leq 3.$$

This yields the desired contradiction, since neither $m(\mathcal{N}) = 0$ nor $m(\mathcal{N}) > 0$ allow for this inequality to hold. □

We used the fact that measurable sets satisfy countable additivity to prove the existence of a nonmeasurable set. Now we show the existence of sets that fail finite additivity with respect to the outer measure.

Corollary 15.22. *The outer measure m^* is not finitely additive. In particular, it is not countably additive.*

Proof. We recall that

$$[0,1] \subset \bigcup_{k=1}^{\infty} \mathcal{N}_k \subset [-1,2].$$

By monotonicity, countable subadditivity, and translation invariance of the *outer measure*, we have that

$$1 = m^*([0,1]) \leq m^*\left(\bigcup_{k=1}^{\infty} \mathcal{N}_k\right) \leq \sum_{k=1}^{\infty} m^*(\mathcal{N}_k) = \sum_{k=1}^{\infty} m^*(\mathcal{N} + r_k) = \sum_{k=1}^{\infty} m^*(\mathcal{N}).$$

It follows that $m^*(\mathcal{N}) > 0$, so that there exists a positive integer K for which $m^*(\mathcal{N}) > \frac{3}{K}$.

By monotonicity,

$$m^*\left(\bigcup_{k=1}^{K} \mathcal{N}_k\right) \leq m^*\left(\bigcup_{k=1}^{\infty} \mathcal{N}_k\right) \leq m^*([-1,2]) = 3,$$

whereas by translation invariance,

$$\sum_{k=1}^{K} m^*(\mathcal{N}_k) = \sum_{k=1}^{K} m^*(\mathcal{N} + r_k) = \sum_{k=1}^{K} m^*(\mathcal{N}) = K m^*(\mathcal{N}) > 3.$$

Thus $m^*\left(\bigcup_{k=1}^{K} \mathcal{N}_k\right) \leq 3 < \sum_{k=1}^{K} m^*(\mathcal{N}_k)$, and so finite additivity fails for the outer measure. □

15.4. **Measurable functions.** Now that we have defined the Lebesgue measure, we move to the notion of *measurable functions* $f : \Omega \to [-\infty, \infty]$ for measurable subsets Ω of $\mathbb{R}^n$. We allow for the possibility that a function f takes on the infinite values ∞ or $-\infty$, so that $-\infty \leq f(x) \leq \infty$. We say that f is finite-valued if $-\infty < f(x) < \infty$ for all $x \in \Omega$. In practice, we only need to worry about functions taking on infinite values on a set of measure zero.

Definition 15.23. Let $\Omega \subseteq \mathbb{R}^n$ be a measurable subset of $\mathbb{R}^n$. A function $f : \Omega \to [-\infty, \infty]$ is measurable if for all $a \in \mathbb{R}$, the set

$$f^{-1}([-\infty, a)) \coloneqq \{x \in \Omega : f(x) < a\}$$

is measurable.

For finite-valued functions, we may work more generally with open sets or closed sets.

Lemma 15.24. *A finite-valued function f is measurable if and only if $f^{-1}(E)$ is measurable for every open set $E \subseteq \mathbb{R}$. Similarly, a finite-valued function f is measurable if and only if $f^{-1}(F)$ is measurable for every closed set $F \subseteq \mathbb{R}$.*

Proof. If $f^{-1}(E)$ is measurable for every open set E and f is finite-valued, then this is true for $E = (-\infty, a)$ for each $a \in \mathbb{R}$, and so f is measurable.

Conversely, if f is measurable, then

$$f^{-1}((-\infty, a]) = \bigcap_{k=1}^{\infty} f^{-1}((-\infty, a + 2^{-k}))$$

is measurable for each $a \in \mathbb{R}$, and so is $f^{-1}([a, \infty)) = \mathbb{R} \setminus f^{-1}((-\infty, a))$, and so is $f^{-1}((a, b)) = f^{-1}((-\infty, b)) \setminus f^{-1}((-\infty, a]))$ for each $a < b$. Every open set in $\mathbb{R}$ is a countable union of open intervals, and so $f^{-1}(E)$ is measurable for every open set E. Finally, the result for closed sets follows by taking complements, since the complement of a closed set is open. □

Continuous functions are examples of measurable functions.

Lemma 15.25. *Let $\Omega \subseteq \mathbb{R}^n$ be measurable and let $f : \Omega \to \mathbb{R}$ be continuous. Then f is measurable.*

Proof. If $E \subseteq \mathbb{R}$ is open, then $f^{-1}(E)$ is open in Ω, so that there exists some open set $U \subseteq \mathbb{R}^n$ such that $f^{-1}(E) = U \cap \Omega$. This is the finite intersection of measurable sets, hence measurable. □

Corollary 15.26. *Let $\Omega \subseteq \mathbb{R}^n$ be measurable, let $f : \Omega \to \mathbb{R}$ be measurable, and let $g : \mathbb{R} \to \mathbb{R}$ be continuous. Then $g \circ f : \Omega \to \mathbb{R}$ is measurable. In particular, if f measurable, then so are $|f(x)|$, $\max\{f(x), 0\}$, and $\min\{f(x), 0\}$.*

Note that if g is merely measurable rather than continuous, it need not be the case that $g \circ f$ is measurable.

Proof. If $E \subseteq \mathbb{R}$ is open, then

$$(g \circ f)^{-1}(E) = \{x \in \Omega : g(f(x)) \in E\} = \{x \in \Omega : f(x) \in g^{-1}(E)\} = f^{-1}(g^{-1}(E))$$

is measurable, since g being continuous means that $g^{-1}(E)$ is open, and f being measurable means that $f^{-1}(g^{-1}(E))$ is measurable.

Now take g to be the continuous function $g(x) \coloneqq |x|$, $g(x) \coloneqq \max\{x, 0\}$, and $g(x) \coloneqq \min\{x, 0\}$ respectively to deduce the measurability of $|f(x)|$, $\max\{f(x), 0\}$, and $\min\{f(x), 0\}$. □

We often work with functions that agree outside of a set of measure zero.

Definition 15.27. Two functions $f, g : \Omega \to [-\infty, \infty]$ are equal *almost everywhere* or simply *a.e.* if $\{x \in \Omega : f(x) \neq g(x)\}$ has measure zero.

Clearly if $f(x) = g(x)$ a.e., then f is measurable if and only if g is measurable.

We also need to work with sequences of measurable functions.

Lemma 15.28. *Suppose that (f_k) is a sequence of measurable functions. Then the functions*

$$\sup_{k \in \mathbb{N}} f_k(x), \qquad \inf_{k \in \mathbb{N}} f_k(x), \qquad \limsup_{k \to \infty} f_k(x), \qquad \liminf_{k \to \infty} f_k(x)$$

are all measurable. In particular, if (f_k) converges pointwise to a function f, then f is measurable.

Some of these conditions can be weakened; for example, if (f_k) converges pointwise a.e. to f, then f is measurable.

Proof. For the first function, we note that

$$\left\{x \in \Omega : \sup_{k\in\mathbb{N}} f_k(x) > a\right\} = \bigcup_{k=1}^{\infty} \{x \in \Omega : f_k(x) > a\}.$$

Similarly, for the second function, we note that

$$\left\{x \in \Omega : \inf_{k\in\mathbb{N}} f_k(x) < a\right\} = \bigcup_{k=1}^{\infty} \{x \in \Omega : f_k(x) < a\}.$$

The measurability of the third and fourth functions holds from the first two, since

$$\limsup_{k\to\infty} f_k(x) = \inf_{j\in\mathbb{N}} \sup_{k\geq j} f_k(x), \qquad \liminf_{k\to\infty} f_k(x) = \sup_{j\in\mathbb{N}} \inf_{k\geq j} f_k(x).$$

Finally, if (f_k) converges pointwise to f, then $f(x) = \liminf_{k\to\infty} f_k(x) = \limsup_{k\to\infty} f_k(x)$. □

An important feature of measure theory is that one can strengthen pointwise convergence of sequences of functions to uniform convergence simply by excising a small subset of the domain.

Theorem 15.29 (Egorov's theorem)**.** *Let (f_k) be a sequence of measurable functions from a measurable set $\Omega \subseteq \mathbb{R}^n$ to $\mathbb{R}$. Suppose that (f_k) is pointwise convergent a.e. on Ω and that Ω has finite measure. Then for all $\varepsilon > 0$, there exists a closed set $F \subseteq \Omega$ such that $m(\Omega \setminus F) < \varepsilon$ and such that (f_k) is uniformly convergent on F.*

Heuristically, Egorov's theorem can be thought of as stating that every convergent sequence of measurable functions is *nearly* uniformly convergent.

Proof. By assumption, there exists a measurable subset $\Omega' \subseteq \Omega$ and a measurable function $f : \Omega' \to \mathbb{R}$ such that $m(\Omega \setminus \Omega') = 0$ and such that $\lim_{k\to\infty} f_k(x) = f(x)$ for all $x \in \Omega'$. For each $j, k \in \mathbb{N}$, let

$$\Omega_{j,k} \coloneqq \left\{x \in \Omega' : |f_\ell(x) - f(x)| < 2^{-j} \text{ for all } \ell > k\right\}.$$

Note that these measurable sets are contained in Ω' and satisfy the inclusion $\Omega_{j,k} \subseteq \Omega_{j,k+1}$ for all $j, k \in \mathbb{N}$ and are such that $\bigcup_{k=1}^{\infty} \Omega_{j,k} = \Omega'$ since $\lim_{\ell\to\infty} f_\ell(x) = f(x)$ for all $x \in \Omega'$. It follows that $m(\Omega') = \lim_{k\to\infty} m(\Omega_{j,k})$, so that there exists some $k = k(j) \in \mathbb{N}$ such that

$$m(\Omega') < m(\Omega_{j,k(j)}) + 2^{-j}.$$

By finite additivity, the left-hand side is equal to $m(\Omega' \setminus \Omega_{j,k(j)}) + m(\Omega_{j,k(j)})$. Since Ω' has finite measure, so does $\Omega_{j,k(j)}$, and hence $m(\Omega' \setminus \Omega_{j,k(j)}) < 2^{-j}$. We choose $J \in \mathbb{N}$ such that $2^{2-J} \leq \varepsilon$ and let $A \coloneqq \bigcap_{j=J}^{\infty} \Omega_{j,k(j)}$. Then by countable subadditivity,

$$m(\Omega' \setminus A) = m\left(\bigcup_{j=J}^{\infty} (\Omega' \setminus \Omega_{j,k(j)})\right) \leq \sum_{j=J}^{\infty} m(\Omega' \setminus \Omega_{j,k(j)}) < \sum_{j=J}^{\infty} 2^{-j} = 2^{1-J} \leq \frac{\varepsilon}{2}.$$

Since A is measurable, being the countable intersection of measurable sets, so is A^c, so that there exists an open set $E \supseteq A^c$ such that $m(E \setminus A^c) < \frac{\varepsilon}{2}$, and hence there exists a closed set $F \coloneqq E^c \subseteq A$ such that $m(A \setminus F) < \frac{\varepsilon}{2}$. Then by finite additivity, we have that

$$m(\Omega \setminus F) = m(\Omega \setminus \Omega') + m(\Omega' \setminus A) + m(A \setminus F) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

It remains to show that (f_k) converges uniformly to f on F. Given $\varepsilon > 0$, there exists some $j \geq J$ such that $2^{-j} \leq \varepsilon$, and hence whenever $\ell > k(j)$, we have that $|f_\ell(x) - f(x)| < 2^{-j} \leq \varepsilon$ for all $x \in F$ since $F \subseteq A \subseteq \Omega_{j,k(j)}$. □

15.5. **Simple functions.** We move on to approximating measurable functions by less complicated functions. We require the following types of functions.

Definition 15.30. The *indicator function* (or *characteristic function*) χ_Ω of a set $\Omega \subseteq \mathbb{R}^n$ is the function $\chi_\Omega : \mathbb{R}^n \to \mathbb{R}$ given by

$$\chi_\Omega(x) \coloneqq \begin{cases} 1 & \text{if } x \in \Omega, \\ 0 & \text{if } x \notin \Omega. \end{cases}$$

Next, we pass to the functions that are the building blocks of integration theory: these are simply linear combinations of indicator functions of measurable sets with finite measure.

Definition 15.31. A *simple function* on $\mathbb{R}^n$ is a function $\varphi : \mathbb{R}^n \to \mathbb{R}$ of the form

$$\varphi(x) = \sum_{k=1}^{K} a_k \chi_{\Omega_k}(x)$$

for some positive integer $K \in \mathbb{N}$, real constants $a_1, \dots, a_K \in \mathbb{R}$, and measurable sets $\Omega_1, \dots, \Omega_K \subset \mathbb{R}^n$ that are each of finite measure.

More generally, given a measurable subset $\Omega \subseteq \mathbb{R}^n$, we can define a simple function on Ω by taking a measurable function on $\mathbb{R}^n$ and restricting it to Ω, which has the effect of replacing each Ω_k with $\Omega \cap \Omega_k$.

Of crucial importance to is is the fact that simple functions can be used to approximate measurable functions.

Theorem 15.32. *Let $\Omega \subseteq \mathbb{R}^n$ be measurable and let $f : \Omega \to [-\infty, \infty]$ be a measurable function. Then there exists a sequence of simple functions (φ_m) that converges pointwise to f and satisfies $|\varphi_m(x)| \leq |\varphi_{m+1}(x)|$ for all $m \in \mathbb{N}$ and $x \in \Omega$.*

This result is of utmost importance in developing the theory of the Lebesgue integral, since we will *first* define the Lebesgue integral for simple functions and *then* define it for arbitrary measurable functions via this approximation.

Proof. We first prove this when f is nonnegative, in which case each φ_k is also nonnegative. For $j \in \mathbb{N}$, let $B_j(0)$ denote the open ball centred at the origin of radius j. Define

$$F_j(x) \coloneqq \begin{cases} f(x) & \text{if } x \in B_j(0) \cap \Omega \text{ and } f(x) \leq j, \\ j & \text{if } x \in B_j(0) \cap \Omega \text{ and } f(x) > j, \\ 0 & \text{otherwise.} \end{cases}$$

This is a sequence of measurable functions that converges pointwise to f as j tends to infinity.

Next, we partition the *range* of each function f_j. For fixed $j, k \in \mathbb{N}$, define for each nonnegative integer $\ell \in \{0, \dots, jk-1\}$ the set

$$\Omega_{j,k,\ell} \coloneqq \left\{ x \in B_j(0) : \frac{\ell}{k} < F_j(x) \leq \frac{\ell+1}{k} \right\},$$

so that if $x \in \Omega_{j,k,\ell}$, then $0 < F_j(x) - \frac{\ell}{k} \leq \frac{1}{k}$. Finally, define

$$F_{j,k}(x) \coloneqq \sum_{\ell=0}^{jk-1} \frac{\ell}{k} \chi_{\Omega_{j,k,\ell}}(x).$$

Each function $F_{j,k}$ is a simple function satisfying

$$0 \leq F_j(x) - F_{j,k}(x) \leq \frac{1}{k}$$

for all $x \in \mathbb{R}^n$.

We now choose $j = k = 2^m$ with $m \in \mathbb{N}$, and let $\varphi_m(x) \coloneqq F_{2^m,2^m}(x)$. This is a simple function satisfying $\varphi_m(x) \le \varphi_{m+1}(x)$ for all $x \in \Omega$ and

$$0 \le F_{2^m}(x) - \varphi_m(x) \le \frac{1}{2^m}.$$

Since F_{2^m} converges pointwise to f, φ_m also converges pointwise to f.

Now we consider the more general case where f need not be nonnegative. We write $f(x) = f^+(x) - f^-(x)$, where $f^+(x) \coloneqq \max\{f(x), 0\}$ and $f^-(x) \coloneqq \max\{-f(x), 0\}$. These functions are both nonnegative, so there exist sequences of simple functions (φ_m^+) and (φ_m^-) that converge pointwise to f^+ and f^- respectively and satisfy $0 \le \varphi_m^\pm(x) \le \varphi_{m+1}^\pm(x)$ for all $m \in \mathbb{N}$ and $x \in \Omega$. We define a sequence of simple functions (φ_m) by $\varphi_m(x) \coloneqq \varphi_m^+(x) - \varphi_m^-(x)$, so that $|\varphi_m(x)| = \varphi_m^+(x) + \varphi_m^-(x)$. This converges pointwise to $f(x) = f^+(x) - f^-(x)$ and satisfies $|\varphi_m(x)| \le |\varphi_{m+1}(x)|$ for all $m \in \mathbb{N}$ and $x \in \Omega$. □

16. The Lebesgue integral

Recommended reading: [Pug15, §6.6–6.7], [SS05, §2.1–2.3], [Tao16, §8.1–8.5].

To develop the Lebesgue integral, we proceed in a step-by-step fashion: we first develop this theory only for a highly restrictive family of measurable functions, and then iterate this process to slowly remove restrictions on this family. We begin with simple functions, proceed to bounded functions supported on a set of finite measure, then onto nonnegative functions, and finally finish with integrable functions.

16.1. **The Lebesgue integral for simple functions.** Recall that a simple function is a function $\varphi : \mathbb{R}^n \to \mathbb{R}$ of the form

$$\varphi(x) = \sum_{k=1}^{K} a_k \chi_{\Omega_k}(x)$$

for some positive integer $K \in \mathbb{N}$, real constants $a_1, \dots, a_K \in \mathbb{R}$, and measurable sets $\Omega_1, \dots, \Omega_K \subset \mathbb{R}^n$ that are each of finite measure. There is more than one representation of such a function: for example, it may be the case that there is some k, k' with $k \neq k'$ for which $a_k = -a_{k'}$ and $\Omega_k = \Omega_{k'}$. To avoid this, we work with simple functions that are written in their *canonical form*.

Definition 16.1. A simple function $\varphi(x) = \sum_{k=1}^K a_k \chi_{\Omega_k}(x)$ is in *canonical form* if $a_1, \dots, a_K$ are distinct and nonzero and the sets $\Omega_1, \dots, \Omega_K$ are disjoint.

Lemma 16.2. *Every simple function can be written uniquely in canonical form.*

Proof. If $\varphi : \mathbb{R}^n \to \mathbb{R}$ is a simple function, then for each $a \in \mathbb{R}$, the set $\{x \in \mathbb{R}^n : \varphi(x) = a\}$ is measurable and there are only finitely many nonzero $a \in \mathbb{R}$ for which this set is nonempty. We order this finite collection as $a_1, \dots, a_K$, which are distinct and nonzero, and set $\Omega_k \coloneqq \{x \in \mathbb{R}^n : \varphi(x) = a_k\}$, which are disjoint. Then $\varphi(x) = \sum_{k=1}^K a_k \chi_{\Omega_k}(x)$, as desired. To show that this form is unique, we note that this is clear if $K = 1$, and we simply proceed by induction on K. □

Definition 16.3. The *Lebesgue integral* of a simple function $\varphi : \mathbb{R}^n \to \mathbb{R}$ in canonical form $\varphi(x) = \sum_{k=1}^K a_k \chi_{\Omega_k}(x)$ is

$$\int_{\mathbb{R}^n} \varphi(x)\, dx \coloneqq \sum_{k=1}^{K} a_k m(\Omega_k).$$

Given a measurable set $\Omega \subseteq \mathbb{R}^n$, $\varphi\chi_\Omega$ is also a simple function, and we define

$$\int_\Omega \varphi(x)\,dx \coloneqq \int_{\mathbb{R}^n} \varphi(x)\chi_\Omega(x)\,dx.$$

While it is not immediate from the definition, we can in fact define the integral of a simple function even if it is not written in canonical form.

Lemma 16.4. *If $\varphi = \sum_{k=1}^K a_k\chi_{\Omega_k}$ is* any *representation of a simple function φ, then*

$$\int_{\mathbb{R}^n} \varphi(x)\,dx = \sum_{k=1}^K a_k m(\Omega_k).$$

This is known as the *independence of the representation* of the Lebesgue integral of a simple function.

Proof. We first prove this in the special case where the sets $\Omega_1, \dots, \Omega_K$ are disjoint. Suppose initially that $\varphi = \sum_{k=1}^K a_k\chi_{\Omega_k}$ with the sets $\Omega_1, \dots, \Omega_K$ disjoint but $a_1, \dots, a_K$ not necessarily distinct and nonzero. For each distinct nonzero value $a \in A$ with $A = \{a_1, \dots, a_K\}$, let $\Omega(a) \coloneqq \bigcup_{a_k=a} \Omega_k$. The sets $\{\Omega(a) : a \in A\}$ are disjoint and satisfy $m(\Omega(a)) = \sum_{\substack{k=1 \\ a_k=a}}^K m(\Omega_k)$. Then $\varphi = \sum_{a\in A} a\chi_{\Omega(a)}$ is the canonical form of φ and

$$\int_{\mathbb{R}^n} \varphi(x)\,dx = \sum_{a\in A} am(\Omega(a)) = \sum_{k=1}^K a_k m(\Omega_k).$$

Next suppose that $\varphi = \sum_{k=1}^K a_k\chi_{\Omega_k}$ where now we no longer assume the disjointness of the sets. We refine the decomposition $\bigcup_{k=1}^K \Omega_k$ by breaking up the collection $\Omega_1, \dots, \Omega_K$ into mutually disjoint measurable sets $\Omega'_1, \dots, \Omega'_L$ for which either $\Omega'_\ell \subseteq \Omega_k$ or $\Omega_{\ell'} \cap \Omega_k = \emptyset$, so that $\bigcup_{k=1}^K \Omega_k = \bigcup_{\ell=1}^L \Omega'_\ell$ and $\Omega_k = \bigcup_{\substack{\ell=1 \\ \Omega'_\ell \subseteq \Omega_k}}^L \Omega'_\ell$. For each $\ell \in \{1, \dots, L\}$, let $a'_\ell \coloneqq \sum_{\substack{k=1 \\ \Omega_k \supseteq \Omega'_\ell}}^K a_k$. Then $\varphi = \sum_{\ell=1}^L a'_\ell \varphi_{\Omega'_\ell}$, and since the sets Ω'_ℓ are mutually disjoint, we have that

$$\int_{\mathbb{R}^n} \varphi(x)\,dx = \sum_{\ell=1}^L a'_\ell m(\Omega'_\ell) = \sum_{\ell=1}^L \sum_{\substack{k=1 \\ \Omega_k \supseteq \Omega'_\ell}}^K a_k m(\Omega'_\ell) = \sum_{k=1}^K a_k m(\Omega_k). \qquad \square$$

We record the following properties of the Lebesgue integral of simple functions. These properties shall be shown to remain true for less restrictive families of functions.

Proposition 16.5. *The Lebesgue integral of simple functions satisfies the following properties:*

(1) *Linearity: if $c_1, c_2 \in \mathbb{R}$, then*

$$\int_{\mathbb{R}^n} (c_1\varphi_1(x) + c_2\varphi_2(x))\,dx = c_1 \int_{\mathbb{R}^n} \varphi_1(x)\,dx + c_2 \int_{\mathbb{R}^n} \varphi_2(x)\,dx.$$

(2) *Additivity: if Ω, Ω' are disjoint measurable subsets of $\mathbb{R}^n$, then*

$$\int_{\Omega\cup\Omega'} \varphi(x)\,dx = \int_\Omega \varphi(x)\,dx + \int_{\Omega'} \varphi(x)\,dx.$$

(3) *Monotonicity: if $\varphi_1(x) \le \varphi_2(x)$ for all $x \in \mathbb{R}^n$, then*

$$\int_{\mathbb{R}^n} \varphi_1(x)\,dx \le \int_{\mathbb{R}^n} \varphi_2(x)\,dx.$$

(4) *Triangle inequality: $|\varphi|$ is also a simple function, and*

$$\left|\int_{\mathbb{R}^n} \varphi(x)\,dx\right| \leq \int_{\mathbb{R}^n} |\varphi(x)|\,dx.$$

In the course of the proof, we will make use of some key facts about simple functions: linear combinations of simple functions are simple functions; the product of a simple function by the indicator function of a measurable set $\Omega \subseteq \mathbb{R}^n$ is a simple function; and the absolute value of a simple function is a simple function.

Proof.
(1) Linearity follows by taking representations $\varphi_1 = \sum_{k=1}^K a_k \chi_{\Omega_k}$ and $\varphi_2 = \sum_{\ell=1}^{K'} a'_\ell \chi_{\Omega'_\ell}$, for then $c_1\varphi_1 + c_2\varphi_2$ is also a simple function with representation $\sum_{k=1}^K c_1 a_k \chi_{\Omega_k} + \sum_{\ell=1}^{K'} c_2 a'_\ell \chi_{\Omega'_\ell}$; using the independence of the representation, we have that

$$\begin{aligned}\int_{\mathbb{R}^n} (c_1\varphi_1(x) + c_2\varphi_2(x))\,dx &= c_1 \sum_{k=1}^K a_k m(\Omega_k) + c_2 \sum_{\ell=1}^{K'} a'_\ell m(\Omega'_\ell) \\ &= c_1 \int_{\mathbb{R}^n} \varphi_1(x)\,dx + c_2 \int_{\mathbb{R}^n} \varphi_2(x)\,dx.\end{aligned}$$

(2) Additivity follows from the fact that if Ω, Ω' are disjoint, then $\chi_{\Omega\cup\Omega'} = \chi_\Omega + \chi_{\Omega'}$, so that $\varphi\chi_{\Omega\cup\Omega'} = \varphi\chi_\Omega + \varphi\chi_{\Omega'}$, and so the result now follows from linearity.
(3) For monotonicity, we note that if φ is a *nonnegative* simple function, then its canonical form $\sum_{k=1}^K a_k \chi_{\Omega_k}$ is everywhere nonnegative, as $a_1, \ldots, a_K > 0$, and from the definition of the integral of φ, we have that $\int_{\mathbb{R}^n} \varphi(x)\,dx \geq 0$, since each summand is nonnegative. Now if $\varphi_1 \leq \varphi_2$, then $\varphi_2 - \varphi_1$ is also a simple function, at which point linearity implies monotonicity.
(4) For the triangle inequality, we write $\varphi = \sum_{k=1}^K a_k \chi_{\Omega_k}$ in its canonical form, so that $|\varphi|$ is also a simple function with representation $|\varphi| = \sum_{k=1}^K |a_k| \chi_{\Omega_k}$ (though this need not be the canonical form of $|\varphi|$, since it could be the case that $a_k = -a'_k$ for some $k \neq k'$). By the triangle inequality for finite sums of real numbers and the independence of the representation of $|\varphi|$, we have that

$$\left|\int_{\mathbb{R}^n} \varphi(x)\,dx\right| = \left|\sum_{k=1}^K a_k m(\Omega_k)\right| \leq \sum_{k=1}^K |a_k| m(\Omega_k) = \int_{\mathbb{R}^n} |\varphi(x)|\,dx. \qquad \square$$

16.2. **The Lebesgue integral for bounded functions supported on sets of finite measure.** We now proceed to define the Lebesgue integral for a larger class of functions. These functions need not be simple, but they must be bounded and vanish outside of a set of finite measure.

Definition 16.6. Given a metric space X and a function $g : X \to \mathbb{R}$, the *support* of g is

$$\operatorname{supp}(g) \coloneqq \{x \in X : g(x) \neq 0\}.$$

We say that g is *supported* on a set $Y \subseteq X$ if $g(x) = 0$ whenever $x \notin Y$.

Note that the support of a measurable function is necessarily measurable.

We previously proved that if g is measurable, then there exists a sequence of simple functions (φ_m) that converges pointwise to g and whose absolute value is nondecreasing. In particular, if g is supported on a set Ω, then these simple functions are supported on Ω, and if $\sup_{x\in\mathbb{R}^n} |g(x)| \coloneqq M$, so that g is bounded by a constant $M \geq 0$, then these simple functions are also bounded by M.

Lemma 16.7. *Let g be a bounded measurable function with support $\Omega \subset \mathbb{R}^n$ of finite measure. If (φ_m) is a sequence of simple functions bounded by $M \coloneqq \sup_{x\in\mathbb{R}^n} |g(x)| \geq 0$, supported on Ω, and pointwise convergent a.e. to g, then:*
(1) *The limit $\lim_{m\to\infty} \int_{\mathbb{R}^n} \varphi_m(x)\,dx$ exists;*
(2) *If $g(x) = 0$ a.e., then the limit $\lim_{m\to\infty} \int_{\mathbb{R}^n} \varphi_m(x)\,dx$ is equal to 0;*
(3) *The limit is independent of the sequence (φ_m).*

Proof.
(1) We appeal to Egorov's theorem. Given $\varepsilon > 0$, there exists a closed (and hence measurable set) $F \subseteq \Omega$ such that $m(\Omega \setminus F) < \frac{\varepsilon}{4M}$ if $M > 0$ and $m(\Omega \setminus F) < \varepsilon$ otherwise and such that (φ_m) converges *uniformly* to g on F. Letting $a_m \coloneqq \int_{\mathbb{R}^n} \varphi_m(x)\,dx \in \mathbb{R}$, we have that for all $m, \ell \in \mathbb{N}$,

$$\begin{aligned} |a_m - a_\ell| &= \left| \int_{\mathbb{R}^n} \varphi_m(x)\,dx - \int_{\mathbb{R}^n} \varphi_\ell(x)\,dx \right| \\ &\leq \int_\Omega |\varphi_m(x) - \varphi_\ell(x)|\,dx \end{aligned}$$

by linearity and the triangle inequality for the Lebesgue integral,

$$= \int_F |\varphi_m(x) - \varphi_\ell(x)|\,dx + \int_{\Omega\setminus F} |\varphi_m(x) - \varphi_\ell(x)|\,dx$$

by additivity for the Lebesgue integral. If $m(F) = 0$, then the first integral is equal to zero, while if $m(F) > 0$, then as (φ_m) is uniformly convergent on F, there exists some $N = N(\frac{\varepsilon}{2m(F)}) \in \mathbb{N}$ such that the simple function $|\varphi_m(x) - \varphi_\ell(x)|\chi_F(x)$ is bounded above by the simple function $\frac{\varepsilon}{2m(F)}\chi_F(x)$ for all $m, \ell \geq N$. So by monotonicity, the first integral is bounded by $\frac{\varepsilon}{2}$ for $m, \ell \geq N$. Similarly, the simple function $|\varphi_m(x) - \varphi_\ell(x)|\chi_{\Omega\setminus F}(x)$ is bounded from above by the simple function $2M\chi_{\Omega\setminus F}(x)$, and so by monotonicity, the second integral is bounded by $2Mm(\Omega \setminus F)$, which is less than $\frac{\varepsilon}{2}$. Thus for all $\varepsilon > 0$, there exists some $N = N(\varepsilon) \in \mathbb{N}$ such that $|a_m - a_\ell| < \varepsilon$ whenever $m, \ell \geq N$, and so (a_m) is a Cauchy sequence of real numbers, and hence convergent. It follows that the limit $\lim_{m\to\infty} \int_{\mathbb{R}^n} \varphi_m(x)\,dx$ exists.
(2) If $g(x) = 0$ a.e., then we repeat the same argument as above for $|a_m|$ instead of $|a_m - a_\ell|$, noting that since (φ_m) converges uniformly to 0 on F, $\int_F |\varphi_m(x)|\,dx < \frac{\varepsilon}{2}$.
(3) Let (ψ_m) be another such sequence. Then the sequence $(\varphi_m - \psi_m)$ consists of simple functions that are bounded by $2M$, supported on a set of finite measure, and converge pointwise a.e. to 0. By (2), we have that $\lim_{m\to\infty} \int_{\mathbb{R}^n} (\varphi_m(x) - \psi_m(x))\,dx = 0$, and so by linearity, we deduce that $\lim_{m\to\infty} \int_{\mathbb{R}^n} \varphi_m(x)\,dx = \lim_{m\to\infty} \int_{\mathbb{R}^n} \psi_m(x)\,dx$. □

With this in hand, we can now proceed to the definition of the Lebesgue integral for a larger class of functions.

Definition 16.8. The *Lebesgue integral* of a bounded measurable function $g : \mathbb{R}^n \to \mathbb{R}$ supported on a set of finite measure is

$$\int_{\mathbb{R}^n} g(x)\,dx \coloneqq \lim_{m\to\infty} \int_{\mathbb{R}^n} \varphi_m(x)\,dx,$$

where (φ_m) is any sequence of simple functions that converge pointwise a.e. to g and satisfy $\operatorname{supp}(\varphi_m) \subseteq \operatorname{supp}(g)$ and $\sup_{x\in\mathbb{R}^n} |\varphi_m(x)| \leq \sup_{x\in\mathbb{R}^n} |g(x)|$ for all $m \in \mathbb{N}$.

Given a measurable set $\Omega \subseteq \mathbb{R}^n$, $g\chi_\Omega$ is also a bounded function supported on a set of finite measure, and we define

$$\int_\Omega g(x)\,dx \coloneqq \int_{\mathbb{R}^n} g(x)\chi_\Omega(x)\,dx.$$

The Lebesgue integral is well-defined since $\lim_{m\to\infty} \int_{\mathbb{R}^n} \varphi_m(x)\,dx$ exists and is independent of the choice of sequence (φ_m). We can think of the Lebesgue integral of g as being the area under the graph of g determined by breaking up this region into *horizontal* rectangles, namely $\int_{\mathbb{R}^n} \varphi_m(x)\,dx = \sum_{k=1}^{K} a_k m(\Omega_k)$, and then taking the limit as these rectangles get smaller and smaller, namely $\lim_{m\to\infty} \int_{\mathbb{R}^n} \varphi_m(x)\,dx$. This should be compared to the Riemann integral, where we take *vertical* rectangles.

Proposition 16.9. *The Lebesgue integral of bounded measurable functions supported on sets of finite measure satisfies the following properties:*

(1) *Linearity: if $c_1, c_2 \in \mathbb{R}$, then*

$$\int_{\mathbb{R}^n} (c_1 g_1(x) + c_2 g_2(x))\,dx = c_1 \int_{\mathbb{R}^n} g_1(x)\,dx + c_2 \int_{\mathbb{R}^n} g_2(x)\,dx.$$

(2) *Additivity: if Ω, Ω' are disjoint measurable subsets of $\mathbb{R}^n$, then*

$$\int_{\Omega\cup\Omega'} g(x)\,dx = \int_{\Omega} g(x)\,dx + \int_{\Omega'} g(x)\,dx.$$

(3) *Monotonicity: if $g_1(x) \le g_2(x)$ for all $x \in \mathbb{R}^n$, then*

$$\int_{\mathbb{R}^n} g_1(x)\,dx \le \int_{\mathbb{R}^n} g_2(x)\,dx.$$

(4) *Triangle inequality: $|g|$ is also bounded and supported on a set of finite measure, and*

$$\left|\int_{\mathbb{R}^n} g(x)\,dx\right| \le \int_{\mathbb{R}^n} |g(x)|\,dx.$$

Proof. These all follow by approximating by simple functions and using the linearity, additivity, monotonicity, and triangle inequality for the Lebesgue integral of simple functions. □

A useful observation is the following.

Corollary 16.10. *Let $g : \mathbb{R}^n \to \mathbb{R}$ be a bounded measurable function supported on a set of finite measure. If $g(x) \ge 0$ for all $x \in \mathbb{R}^n$ and $\int_{\mathbb{R}^n} g(x)\,dx = 0$, then $g(x) = 0$ for a.e. $x \in \mathbb{R}^n$.*

Proof. Let $\Omega \coloneqq \operatorname{supp}(g)$. For each integer $k \in \mathbb{N}$, define

$$\Omega_k \coloneqq \{x \in \Omega : g(x) \ge 2^{-k}\}.$$

Then $g(x) \ge 2^{-k}\chi_{\Omega_k}(x)$ for *all* $x \in \mathbb{R}^n$, and hence by monotonicity, we have that

$$0 \le 2^{-k} m(\Omega_k) = \int_{\mathbb{R}^n} 2^{-k}\chi_{\Omega_k}(x)\,dx \le \int_{\mathbb{R}^n} g(x)\,dx = 0.$$

It follows that $m(\Omega_k) = 0$ for every integer $k \in \mathbb{N}$, and so by countable subadditivity,

$$0 \le m\left(\{x \in \mathbb{R}^n : g(x) > 0\}\right) = m\left(\bigcup_{k=1}^{\infty} \Omega_k\right) \le \sum_{k=1}^{\infty} m(\Omega_k) = 0.$$

So $m(\{x \in \mathbb{R}^n : g(x) > 0\}) = 0$. □

With these properties in hand, we can finally prove an important theorem regarding Lebesgue integration concerning the interchanging the order of limits and integration.

Theorem 16.11 (Bounded convergence theorem)**.** *Let (g_m) be a sequence of measurable functions that are all bounded by the same constant $M \geq 0$, supported on the same set $\Omega \subset \mathbb{R}^n$ of finite measure, and are pointwise convergent a.e. to a function g. Then g is measurable, bounded by M, supported a.e. on Ω, and satisfies*

$$\lim_{m\to\infty} \int_{\mathbb{R}^n} |g_m(x) - g(x)|\, dx = 0.$$

By the triangle inequality and linearity, this implies that

$$\lim_{m\to\infty} \int_{\mathbb{R}^n} g_m(x)\, dx = \int_{\mathbb{R}^n} \lim_{m\to\infty} g_m(x)\, dx = \int_{\mathbb{R}^n} g(x)\, dx.$$

Thus the bounded convergence theorem gives us conditions under which we may interchange the order of limits and integration.

Proof. Since (g_m) converges pointwise a.e. to g, g must be bounded by M a.e. and be supported on Ω except possibly on a set of measure zero. Now fix $\varepsilon > 0$. By Egorov's theorem, there exists a closed set $F \subseteq \Omega$ such that $m(\Omega \setminus F) < \frac{\varepsilon}{4M}$ if $M > 0$ and $m(\Omega \setminus F) < \varepsilon$ otherwise and such that (f_m) converges *uniformly* to g on F. By additivity and the fact that g_m and g vanish a.e. outside of Ω, we have that

$$\int_{\mathbb{R}^n} |g_m(x) - g(x)|\, dx = \int_F |g_m(x) - g(x)|\, dx + \int_{\Omega\setminus F} |g_m(x) - g(x)|\, dx.$$

If $m(F) = 0$, then the first integral is equal to zero, while if $m(F) > 0$, then as (g_m) converges uniformly to g on F, there exists some $N = N(\frac{\varepsilon}{2m(F)}) \in \mathbb{N}$ such that the bounded function supported on a set of finite measure $|g_m(x) - g(x)|\chi_F(x)$ is bounded above by the simple function $\frac{\varepsilon}{2m(F)}\chi_F(x)$. So by monotonicity, the first integral is bounded by $\frac{\varepsilon}{2}$. Similarly, $|g_m(x) - g(x)|\chi_{\Omega\setminus F}(x)$ is bounded from above by $2M\chi_{\Omega\setminus F}(x)$, and so by monotonicity, the second integral is bounded by $2Mm(\Omega \setminus F)$, which is less than $\frac{\varepsilon}{2}$. Thus for all $\varepsilon > 0$, there exists some $N = N(\varepsilon) \in \mathbb{N}$ such that $\int_{\mathbb{R}^n} |g_m(x) - g(x)|\, dx < \varepsilon$ whenever $m \geq N$, and so

$$\lim_{m\to\infty} \int_{\mathbb{R}^n} |g_m(x) - g(x)|\, dx = 0. \qquad \square$$

16.3. **The Riemann integral.** It is important to note that the Lebesgue integral is an *extension* of the Riemann integral, in the sense that a Riemann integrable function is Lebesgue integrable and both integrals are identical.

Theorem 16.12. *Let $f : [0,1] \to \mathbb{R}$ be Riemann integrable. Then f is Lebesgue integrable, and the Riemann integral of f is equal to the Lebesgue integral of f.*

Proof. As f is Riemann integrable, it is bounded, so that there exists some $M > 0$ such that $|f(x)| \leq M$ for all $x \in [0,1]$. Riemann integrability implies the existence of sequences of certain simple functions (φ_m) and (ψ_m) that satisfy the following:

- each φ_m and ψ_m is a linear combination of indicator functions of *boxes* (which, for $\mathbb{R}$, are simply closed bounded intervals), rather than arbitrary measurable sets of finite measure;
- $|\varphi_m(x)|, |\psi_m(x)| \leq M$;
- $\varphi_m(x) \leq \varphi_{m+1}(x)$ and $\psi_m(x) \geq \psi_{m+1}(x)$;
- $\varphi_m(x) \leq f(x) \leq \psi_m(x)$; and
- $\lim_{m\to\infty} \int_{[0,1]}^{\mathcal{R}} \varphi_m(x)\, dx = \lim_{m\to\infty} \int_{[0,1]}^{\mathcal{R}} \psi_m(x)\, dx =: \int_{[0,1]}^{\mathcal{R}} f(x)\, dx$.

Here $\int_{[0,1]}^{\mathcal{R}}$ denotes the Riemann integral, where the Riemann integrals of φ_m and ψ_m are simply the corresponding linear combinations of the lengths of the boxes.

Clearly the Riemann integral of these types of simple functions φ_m, ψ_m coincides with the Lebesgue integral. Letting $\widetilde{\varphi}(x) \coloneqq \lim_{m\to\infty} \varphi_m(x)$ and $\widetilde{\psi}(x) \coloneqq \lim_{m\to\infty} \psi_m(x)$, we obtain measurable functions $\widetilde{\varphi}, \widetilde{\psi}$ that satisfy $\widetilde{\varphi}(x) \le f(x) \le \widetilde{\psi}(x)$. The bounded convergence theorem implies that

$$\int_{[0,1]} \widetilde{\varphi}(x)\,dx = \lim_{m\to\infty} \int_{[0,1]} \varphi_m(x)\,dx = \lim_{m\to\infty} \int_{[0,1]}^{\mathcal{R}} \varphi_m(x)\,dx,$$
$$\int_{[0,1]} \widetilde{\psi}(x)\,dx = \lim_{m\to\infty} \int_{[0,1]} \psi_m(x)\,dx = \lim_{m\to\infty} \int_{[0,1]}^{\mathcal{R}} \psi_m(x)\,dx,$$

and hence that $\int_{[0,1]} (\widetilde{\psi}(x) - \widetilde{\varphi}(x))\,dx = 0$ by linearity. Thus $\widetilde{\varphi}(x) = \widetilde{\psi}(x)$ a.e., and hence $\widetilde{\varphi}(x) = f(x) = \widetilde{\psi}(x)$ a.e., so that f is Lebesgue measurable. By the definition of the Lebesgue integral, we have that

$$\int_{[0,1]} f(x)\,dx = \lim_{m\to\infty} \int_{[0,1]} \varphi_m(x)\,dx,$$

which yields the result. □

On the other hand, the converse does not hold: there are Lebesgue integrable functions that are not Riemann integrable.

Example 16.13. The function $f : [0,1] \to \mathbb{R}$ given by

$$f(x) \coloneqq \begin{cases} 1 & \text{if } x \in [0,1] \cap \mathbb{Q}, \\ 0 & \text{if } x \in [0,1] \cap \mathbb{Q}^c, \end{cases}$$

is not Riemann integrable. On the other hand, it is Lebesgue integrable, since it is a simple function, and its Lebesgue integral is simply $\int_{[0,1]} f(x)\,dx = m([0,1] \cap \mathbb{Q}) = 0$.

16.4. **The Lebesgue integral for nonnegative functions.** We continue to extend the size of the family of functions for which the Lebesgue integral is defined. We consider measurable functions that need not be bounded nor supported on sets on finite measure but that are nonnegative.

Definition 16.14. The *Lebesgue integral* of a nonnegative measurable function $f : \mathbb{R}^n \to [-\infty, \infty]$ is

$$\int_{\mathbb{R}^n} f(x)\,dx \coloneqq \sup_{0\le g\le f} \int_{\mathbb{R}^n} g(x)\,dx,$$

where the supremum is over all bounded measurable functions $g : \mathbb{R}^n \to \mathbb{R}$ supported on sets of finite measure that satisfy $0 \le g(x) \le f(x)$ for all $x \in \mathbb{R}^n$.

Given a measurable set $\Omega \subseteq \mathbb{R}^n$, $f\chi_\Omega$ is also a nonnegative measurable function, and we define

$$\int_\Omega f(x)\,dx \coloneqq \int_{\mathbb{R}^n} f(x)\chi_\Omega(x)\,dx.$$

Conversely, if $\Omega \subset \mathbb{R}^n$ is measurable and $f : \Omega \to [-\infty, \infty]$ is a nonnegative measurable function, then we can extend f to a nonnegative measurable function $h : \mathbb{R}^n \to [-\infty, \infty]$ by

$$h(x) \coloneqq \begin{cases} f(x) & \text{if } x \in \Omega, \\ 0 & \text{otherwise,} \end{cases}$$

and we define

$$\int_\Omega f(x)\,dx \coloneqq \int_{\mathbb{R}^n} h(x)\,dx.$$

Here there is no reason to expect that the Lebesgue integral of such a function be finite, and indeed we allow for this possibility. However, the integral of such a function must be nonnegative, so it is either a finite nonnegative number or ∞; in particular, it cannot be equal to $-\infty$.

Definition 16.15. A nonnegative measurable function $f : \mathbb{R}^n \to [-\infty, \infty]$ is *Lebesgue integrable* or simply *integrable* if $\int_{\mathbb{R}^n} f(x)\,dx$ is finite.

Once more, the Lebesgue integral satisfies several natural properties.

Proposition 16.16. *The Lebesgue integral of nonnegative measurable functions satisfies the following properties:*

(1) *Linearity: if $c_1, c_2 \geq 0$, then*

$$\int_{\mathbb{R}^n} (c_1 f_1(x) + c_2 f_2(x))\,dx = c_1 \int_{\mathbb{R}^n} f_1(x)\,dx + c_2 \int_{\mathbb{R}^n} f_2(x)\,dx.$$

(2) *Additivity: if Ω, Ω' are disjoint measurable subsets of $\mathbb{R}^n$ of finite measure, then*

$$\int_{\Omega \cup \Omega'} f(x)\,dx = \int_\Omega f(x)\,dx + \int_{\Omega'} f(x)\,dx.$$

(3) *Monotonicity: if $0 \leq f_1(x) \leq f_2(x)$ for all $x \in \mathbb{R}^n$, then*

$$\int_{\mathbb{R}^n} f_1(x)\,dx \leq \int_{\mathbb{R}^n} f_2(x)\,dx.$$

Note that for nonnegative functions, there is no need to state the triangle inequality!

Proof. Additivity and monotonicity follow by approximating by bounded measurable functions supported on sets of finite measure and using the additivity, monotonicity, and triangle inequality for the Lebesgue integral of bounded measurable functions supported on sets of finite measure. Linearity, on the other hand, does not follow immediately if both c_1, c_2 are positive. If g_1 and g_2 are bounded measurable functions supported on sets of finite measure that satisfy $0 \leq g_1(x) \leq f_1(x)$ and $0 \leq g_2(x) \leq f_2(x)$ for all $x \in \mathbb{R}^n$, then $c_1 g_1 + c_2 g_2$ is also of this form and satisfies $0 \leq c_1 g_1(x) + c_2 g_2(x) \leq c_1 f_1(x) + c_2 f_2(x)$. So

$$\begin{aligned}
c_1 \int_{\mathbb{R}^n} f_1(x)\,dx + c_2 \int_{\mathbb{R}^n} f_2(x)\,dx &= c_1 \sup_{0 \leq g_1 \leq f_1} \int_{\mathbb{R}^n} g_1(x)\,dx + c_2 \sup_{0 \leq g_2 \leq f_2} \int_{\mathbb{R}^n} f_2(x)\,dx \\
&= \sup_{0 \leq c_1 g_1 \leq c_1 f_1} \int_{\mathbb{R}^n} c_1 g_1(x)\,dx + \sup_{0 \leq c_2 g_2 \leq c_2 f_2} \int_{\mathbb{R}^n} c_2 g_2(x)\,dx \\
&\leq \sup_{0 \leq c_1 g_1 + c_2 g_2 \leq c_1 f_1 + c_2 f_2} \left(\int_{\mathbb{R}^n} c_1 g_1(x)\,dx + \int_{\mathbb{R}^n} c_2 g_2(x)\,dx \right) \\
&= \sup_{0 \leq c_1 g_1 + c_2 g_2 \leq c_1 f_1 + c_2 f_2} \int_{\mathbb{R}^n} (c_1 g_1(x)\,dx + c_2 g_2(x))\,dx \\
&= \int_{\mathbb{R}^n} (c_1 f_1(x)\,dx + c_2 f_2(x))\,dx
\end{aligned}$$

by the definition of the Lebesgue integral of a nonnegative measurable function and the linearity of the Lebesgue integral of bounded measurable functions supported on sets of finite measure. For the reverse inequality, let g be a bounded measurable function

supported on a set of finite measure satisfying $0 \le g(x) \le c_1 f_1(x) + c_2 f_2(x)$ for all $x \in \mathbb{R}^n$. Let

$$g_1(x) \coloneqq \min\left\{\frac{g(x)}{c_1}, f_1(x)\right\}, \qquad g_2(x) \coloneqq \frac{g(x) - c_1 g_1(x)}{c_2}.$$

These are both bounded measurable functions supported on sets of finite measure satisfying $0 \le g_1(x) \le f_1(x)$ and $0 \le g_2(x) \le f_2(x)$, so that by the linearity of the Lebesgue integral of bounded measurable functions supported on sets of finite measure and the definition of the Lebesgue integral of a nonnegative measurable function,

$$\begin{aligned}\int_{\mathbb{R}^n} g(x)\,dx &= \int_{\mathbb{R}^n} \left(c_1 g_1(x) + c_2 g_2(x)\right)\,dx \\ &= c_1 \int_{\mathbb{R}^n} g_1(x)\,dx + c_2 \int_{\mathbb{R}^n} g_2(x)\,dx \\ &\le c_1 \sup_{0\le g_1 \le f_1} \int_{\mathbb{R}^n} g_1(x)\,dx + c_2 \sup_{0 \le g_2 \le f_2} \int_{\mathbb{R}^n} g_2(x)\,dx \\ &= c_1 \int_{\mathbb{R}^n} f_1(x)\,dx + c_2 \int_{\mathbb{R}^n} f_2(x)\,dx.\end{aligned}$$

This inequality is valid for every g for which $0 \le g \le c_1 f_2 + c_2 f_2$; taking the supremum over all such g, we deduce the desired inequality. □

We also record some additional properties.

Proposition 16.17. *The Lebesgue integral of nonnegative measurable functions satisfies the following additional properties:*

(1) *If $0 \le f_1(x) \le f_2(x)$ for all $x \in \mathbb{R}^n$ and f_2 is integrable, then f_1 is integrable.*
(2) *If f is integrable, then $f(x)$ is finite a.e.*
(3) *If $\int_{\mathbb{R}^n} f(x)\,dx = 0$, then $f(x) = 0$ a.e.*

Proof.
(1) This is an immediate consequence of monotonicity.
(2) Let $\Omega_k \coloneqq \{x \in \mathbb{R}^n : f(x) \ge 2^k\}$ for each $k \in \mathbb{N}$ and let $\Omega_\infty \coloneqq \{x \in \mathbb{R}^n : f(x) = \infty\}$. Then by monotonicity,

$$\int_{\mathbb{R}^n} f(x)\,dx \ge \int_{\mathbb{R}^n} f(x)\chi_{\Omega_k}(x)\,dx \ge 2^k m(\Omega_k).$$

Since $\int_{\mathbb{R}^n} f(x)\,dx$ is finite, we must have that $\lim_{k\to\infty} m(\Omega_k) = 0$. Moreover, since $\Omega_k \supseteq \Omega_{k+1}$ for all $k \in \mathbb{N}$, $m(\Omega_1)$ is finite, and $\bigcap_{k=1}^\infty \Omega_k = \Omega_\infty$, we deduce that $m(\Omega_\infty) = \lim_{k\to\infty} m(\Omega_k) = 0$.
(3) This follows in the exact same way as the same result for bounded measurable functions supported on sets of finite measure. □

We consider the issue of interchanging the order of limits and integration. Some subtlety is required here, since it is *not* the case that we can always interchange these orders.

Example 16.18. Let (f_m) be the sequence of nonnegative measurable functions on $\mathbb{R}$ given by

$$f_m(x) \coloneqq \begin{cases} 0 & \text{if } x \le 2^{-m}, \\ 2^{2(m+1)}(x - 2^{-m}) & \text{if } 2^{-m} \le x \le \frac{3}{2}2^{-m}, \\ 2^{2(m+1)}(2^{1-m} - x) & \text{if } \frac{3}{2}2^{-m} \le x \le 2^{1-m}, \\ 0 & \text{if } x \ge 2^{1-m}. \end{cases}$$

This sequence converges pointwise to $f(x) \coloneqq 0$. However, $\int_{\mathbb{R}} f_m(x)\,dx = 1$ for all $m \in \mathbb{N}$, whereas $\int_{\mathbb{R}} f(x)\,dx = 0$, and so $\lim_{m\to\infty} \int_{\mathbb{R}} f_m(x)\,dx \neq \int_{\mathbb{R}} \lim_{m\to\infty} f_m(x)\,dx$.

Note, however, that in this example it is the case that $\int_{\mathbb{R}} f_m(x)\,dx$ is *at least as large* as $\int_{\mathbb{R}} f(x)\,dx$. This holds true for *all* such sequences of nonnegative measurable functions that are pointwise convergent a.e.

Lemma 16.19 (Fatou's lemma)**.** *Let (f_m) be a sequence of nonnegative measurable functions that is pointwise convergent a.e. to a function f. Then f is measurable and*

$$\int_{\mathbb{R}^n} f(x)\,dx \leq \liminf_{m\to\infty} \int_{\mathbb{R}^n} f_m(x)\,dx.$$

In particular, we do *not* exclude the possibilities that $\liminf_{m\to\infty} \int_{\mathbb{R}^n} f_m(x)\,dx = \infty$ or additionally that $\int_{\mathbb{R}^n} f(x)\,dx = \infty$.

Proof. The measurability of f follows from the fact that (f_m) converges pointwise a.e. to f. For the inequality, we take a bounded measurable function g supported on a set Ω of finite measure satisfying $0 \leq g(x) \leq f(x)$ for all $x \in \mathbb{R}^n$. We then define $g_m(x) \coloneqq \min\{g(x), f_m(x)\}$, which is a bounded measurable function supported on Ω for which $g_m(x) \leq f_m(x)$ for all $x \in \mathbb{R}^n$. Moreover, (g_m) converges pointwise a.e. to g, and so by the bounded convergence theorem, we have that

$$\int_{\mathbb{R}^n} g(x)\,dx = \lim_{m\to\infty} \int_{\mathbb{R}^n} g_m(x)\,dx.$$

By monotonicity, the right-hand side is at most $\liminf_{m\to\infty} \int_{\mathbb{R}^n} f_m(x)\,dx$. We obtain the desired inequality by taking the supremum over all such functions $g(x)$ and recalling the definition $\int_{\mathbb{R}^n} f(x)\,dx \coloneqq \sup_{0\leq g\leq f} \int_{\mathbb{R}^n} g(x)\,dx$. □

Equality holds in Fatou's lemma provided one assumes additional restrictions on the sequence (f_m).

Corollary 16.20. *Let (f_m) be a sequence of nonnegative measurable functions that is pointwise convergent a.e. to a function f and additionally satisfies $f_m(x) \leq f(x)$ for all $x \in \mathbb{R}^n$. Then*

$$\lim_{m\to\infty} \int_{\mathbb{R}^n} f_m(x)\,dx = \int_{\mathbb{R}^n} f(x)\,dx.$$

Proof. By monotonicity, we have that $\int_{\mathbb{R}^n} f_m(x)\,dx \leq \int_{\mathbb{R}^n} f(x)\,dx$, and so

$$\int_{\mathbb{R}^n} f(x)\,dx \geq \limsup_{m\to\infty} \int_{\mathbb{R}^n} f_m(x)\,dx.$$

In conjuction with Fatou's lemma, we obtain the desired equality. □

A useful consequence of Fatou's lemma is the monotone convergence theorem, which gives a useful condition for which the interchanging of the order of limits and integration is valid.

Theorem 16.21 (Monotone convergence theorem)**.** *Let (f_m) be a sequence of nonnegative measurable functions that is pointwise convergent a.e. to a function f and is pointwise nondecreasing a.e., so that $f_m(x) \leq f_{m+1}(x)$ for a.e. $x \in \mathbb{R}^n$ for all $m \in \mathbb{N}$. Then*

$$\lim_{m\to\infty} \int_{\mathbb{R}^n} f_m(x)\,dx = \int_{\mathbb{R}^n} f(x)\,dx.$$

Again, we do *not* exclude the possibility that both sides are equal to ∞. Note that the analogue of this result does *not* hold for the Riemann integral. This is one of the chief advantages that the Lebesgue integral has over the Riemann integral.

16.5. **The Lebesgue integral for Lebesgue integrable functions.** Finally, we proceed to the most general family of functions for which the Lebesgue integral is defined. These are simply Lebesgue integrable functions.

Definition 16.22. A measurable function $f : \mathbb{R}^n \to [-\infty, \infty]$ is *Lebesgue integrable* or simply *integrable* if $\int_{\mathbb{R}^n} |f(x)|\,dx$ is finite.

This integral is well-defined since $|f|$ is a nonnegative measurable function. For an integrable function f, we define

$$f^+(x) \coloneqq \max\{f(x), 0\},$$
$$f^-(x) \coloneqq \max\{-f(x), 0\},$$

so that $f(x) = f^+(x) - f^-(x)$ and $|f(x)| = f^+(x) + f^-(x)$. The functions f^+ and f^- are both nonnegative and measurable. Since $0 \le f^+(x) \le |f(x)|$ and $0 \le f^-(x) \le |f(x)|$, monotonicity implies that the integrals of f^+ and f^- are both finite, so that these are both integrable.

Definition 16.23. The *Lebesgue integral* of an integrable function $f : \mathbb{R}^n \to [-\infty, \infty]$ is

$$\int_{\mathbb{R}^n} f(x)\,dx \coloneqq \int_{\mathbb{R}^n} f^+(x)\,dx - \int_{\mathbb{R}^n} f^-(x)\,dx.$$

Given a measurable set $\Omega \subseteq \mathbb{R}^n$, $f\chi_\Omega$ is also an integrable function, and we define

$$\int_\Omega f(x)\,dx \coloneqq \int_{\mathbb{R}^n} f(x)\chi_\Omega(x)\,dx.$$

Conversely, if $\Omega \subset \mathbb{R}^n$ is measurable and $f : \Omega \to [-\infty, \infty]$ is a measurable function, then we can extend f to a measurable function $h : \mathbb{R}^n \to [-\infty, \infty]$ by

$$h(x) \coloneqq \begin{cases} f(x) & \text{if } x \in \Omega, \\ 0 & \text{otherwise,} \end{cases}$$

and we say f is integrable if h is also integrable and define

$$\int_\Omega f(x)\,dx \coloneqq \int_{\mathbb{R}^n} h(x)\,dx.$$

As the integrals of f^+ and f^- are both finite, the Lebesgue integral of an integrable function is well-defined. Once more, we have the following properties of the Lebesgue integral.

Proposition 16.24. *The Lebesgue integral of integrable functions satisfies the following properties:*

(1) *Linearity: if $c_1, c_2 \in \mathbb{R}$, then*

$$\int_{\mathbb{R}^n} (c_1 f_1(x) + c_2 f_2(x))\,dx = c_1 \int_{\mathbb{R}^n} f_1(x)\,dx + c_2 \int_{\mathbb{R}^n} f_2(x)\,dx.$$

(2) *Additivity: if Ω, Ω' are disjoint measurable subsets of $\mathbb{R}^n$, then*

$$\int_{\Omega \cup \Omega'} f(x)\,dx = \int_\Omega f(x)\,dx + \int_{\Omega'} f(x)\,dx.$$

(3) *Monotonicity: if $f_1(x) \le f_2(x)$ for all $x \in \mathbb{R}^n$, then*

$$\int_{\mathbb{R}^n} f_1(x)\,dx \le \int_{\mathbb{R}^n} f_2(x)\,dx.$$

(4) *Triangle inequality:*

$$\left|\int_{\mathbb{R}^n} f(x)\,dx\right| \leq \int_{\mathbb{R}^n} |f(x)|\,dx.$$

Proof. Linearity, additivity, and monotonicity follow from writing $f = f^+ - f^-$ and using the corresponding results for the Lebesgue integral for nonnegative functions. For the triangle inequality, we simply have via the triangle inequality for $\mathbb{R}$ that

$$\begin{aligned}\left|\int_{\mathbb{R}^n} f(x)\,dx\right| &= \left|\int_{\mathbb{R}^n} f^+(x)\,dx - \int_{\mathbb{R}^n} f^-(x)\,dx\right| \\ &\leq \left|\int_{\mathbb{R}^n} f^+(x)\,dx\right| + \left|\int_{\mathbb{R}^n} f^-(x)\,dx\right| \\ &= \int_{\mathbb{R}^n} (f^+(x) + f^-(x))\,dx \\ &= \int_{\mathbb{R}^n} |f(x)|\,dx.\end{aligned}$$

□

An important consequence of integrability is that integrable functions are *mostly* concentrated in sets of finite measure and away from sets of small measure.

Lemma 16.25. *Let $f : \mathbb{R}^n \to [-\infty, \infty]$ be an integrable function.*
(1) *For all $\varepsilon > 0$, there exists a measurable set $B \subset \mathbb{R}^n$ of finite measure for which*

$$\int_{B^c} |f(x)|\,dx < \varepsilon.$$

(2) *For all $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that*

$$\int_{\Omega} |f(x)|\,dx < \varepsilon$$

for all measurable sets $\Omega \subset \mathbb{R}^n$ satisfying $m(\Omega) < \delta$.
(3) *If $\int_{\mathbb{R}^n} |f(x)|\,dx = 0$, then $f(x) = 0$ for a.e. $x \in \mathbb{R}^n$.*

Part (2) is known as the *absolute continuity* of integrable functions.

Proof.
(1) Let $B_m(0)$ denote the open ball of radius m centred at the origin. The functions $g_m(x) \coloneqq |f(x)|\chi_{B_m(0)}(x)$ for $m \in \mathbb{N}$ are measurable, nonnegative, satisfy $g_m(x) \leq g_{m+1}(x)$ for all $m \in \mathbb{N}$ and $x \in \mathbb{R}^n$, and define a sequence (g_m) that converges pointwise to $|f(x)|$. Thus by the monotone convergence theorem, we have that $\lim_{m\to\infty} \int_{\mathbb{R}^n} g_m(x)\,dx = \int_{\mathbb{R}^n} |f(x)|\,dx$. By monotonicity, each g_m is integrable, so that $\int_{\mathbb{R}^n} g_m(x)\,dx$ is finite. By linearity, we deduce that for all $\varepsilon > 0$, there exists some $N = N(\varepsilon)$ such that

$$\int_{B_m(0)^c} |f(x)|\,dx = \int_{\mathbb{R}^n} |f(x)|\,dx - \int_{\mathbb{R}^n} g_m(x)\,dx < \varepsilon$$

for all $m \geq N$.
(2) Let $\Omega_m \coloneqq \{x \in \mathbb{R}^n : |f(x)| \leq m\}$, and let $g_m(x) \coloneqq |f(x)|\chi_{\Omega_m}(x)$, so that $g_m(x) \leq m\chi_{\Omega_m}(x)$ for all $x \in \mathbb{R}^n$. Once more, this satisfies the conditions of the monotone convergence theorem, so that $\lim_{m\to\infty} \int_{\mathbb{R}^n} g_m(x)\,dx = \int_{\mathbb{R}^n} |f(x)|\,dx$. Again, each g_m is integrable, and so for all $\varepsilon > 0$, there exists some $N = N(\frac{\varepsilon}{2}) \in \mathbb{N}$ such that

$$\int_{\mathbb{R}^n} (|f(x)| - g_m(x))\,dx < \frac{\varepsilon}{2}$$

for all $m \geq N$ by linearity. For any measurable set $\Omega \subset \mathbb{R}^n$ satisfying $m(\Omega) < \delta$ with $\delta = \delta(\varepsilon) > 0$ such that $\delta \leq \frac{\varepsilon}{2N}$, we have by linearity and monotonicity that

$$\begin{aligned}\int_\Omega |f(x)|\,dx &= \int_\Omega (|f(x)| - g_N(x))\,dx + \int_\Omega g_N(x)\,dx \\ &\leq \int_{\mathbb{R}^n} (|f(x)| - g_N(x))\,dx + Nm(\Omega) \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\ &= \varepsilon.\end{aligned}$$

(3) Since $\int_{\mathbb{R}^n} |f(x)|\,dx = \int_{\mathbb{R}^n} f^+(x)\,dx + \int_{\mathbb{R}^n} f^-(x)\,dx$, each of these integrals must also be zero, and now the same result for nonnegative measurable functions implies that $f^+(x)$ and $f^-(x)$ are both zero a.e., which means $f(x) = f^+(x) - f^-(x)$ is also zero a.e. □

We proceed to one last result concerning the issue of interchanging the order of limits and integration.

Theorem 16.26 (Dominated convergence theorem)**.** *Let (f_m) be a sequence of measurable functions that is pointwise convergent a.e. to a function $f : \mathbb{R}^n \to [-\infty, \infty]$. Suppose that there exists an integrable function $g : \mathbb{R}^n \to [-\infty, \infty]$ such that $|f_m(x)| \leq g(x)$ for all $m \in \mathbb{N}$ and $x \in \mathbb{R}^n$. Then f is integrable and satisfies*

$$\lim_{m\to\infty} \int_{\mathbb{R}^n} |f_m(x) - f(x)|\,dx = 0.$$

By the triangle inequality and linearity, this implies that

$$\lim_{m\to\infty} \int_{\mathbb{R}^n} f_m(x)\,dx = \int_{\mathbb{R}^n} \lim_{m\to\infty} f_m(x)\,dx = \int_{\mathbb{R}^n} f(x)\,dx.$$

Thus the dominated convergence theorem gives us conditions under which we may interchange the order of limits and integration. This is significantly stronger than the bounded convergence theorem. On the other hand, it is neither stronger nor weaker than the monotone convergence theorem, which involves integrals that need not be finite. Note that the analogue of this result does *not* hold for Riemann integration. This is another chief advantage that the Lebesgue integral has over the Riemann integral.

Proof. Since $|f_m(x)| \leq g(x)$ for all $m \in \mathbb{N}$ and $x \in \mathbb{R}^n$, we also have that $|f(x)| \leq g(x)$ a.e., and so the integrability of f follows from the integrability of g. Now for each $k \in \mathbb{N}$, let $\Omega_k \coloneqq \{x \in B_k(0) : g(x) \leq k\}$, where $B_k(0)$ denotes the open ball of radius k centred at the origin, and let $h_k(x) \coloneqq g(x)\chi_{\Omega_k}(x)$. Since (h_k) is a sequence of pointwise nondecreasing nonnegative measurable functions converging pointwise to g, the monotone convergence theorem implies that

$$\lim_{k\to\infty} \int_{\Omega_k} g(x)\,dx = \lim_{k\to\infty} \int_{\mathbb{R}^n} h_k(x) = \int_{\mathbb{R}^n} g(x)\,dx.$$

Thus for all $\varepsilon > 0$, there exists some $N = N(\frac{\varepsilon}{3}) \in \mathbb{N}$ such that

$$\int_{\Omega_k^c} g(x)\,dx = \int_{\mathbb{R}^n} g(x)\,dx - \int_{\mathbb{R}^n} g(x)\chi_{\Omega_k}(x)\,dx < \frac{\varepsilon}{3}$$

for all $k \geq N$ by additivity. Since $|f_m(x)| \leq g(x)$, it follows that the functions $f_m(x)\chi_{\Omega_N}(x)$ are bounded by N and supported on Ω_N, which has finite measure as it is contained in $B_N(0)$, and converge pointwise to $f(x)\chi_{\Omega_N(x)}$. The bounded convergence theorem there

implies that $\lim_{m\to\infty}\int_{\Omega_N}|f_m(x)-f(x)|\,dx=0$. Thus for all $\varepsilon>0$, there exists some $M=M(\frac{\varepsilon}{3})$ such that

$$\int_{\Omega_N}|f_m(x)-f(x)|\,dx<\frac{\varepsilon}{3}$$

for all $m\geq M$. By additivity together with the fact that f_m and f are both bounded by g, we therefore have that for all $m\geq M$,

$$\begin{aligned}\int_{\mathbb{R}^n}|f_m(x)-f(x)|\,dx&=\int_{\Omega_N}|f_m(x)-f(x)|\,dx+\int_{\Omega_N^c}|f_m(x)-f(x)|\,dx\\&\leq\int_{\Omega_N}|f_m(x)-f(x)|\,dx+2\int_{\Omega_N^c}g(x)\,dx\\&<\frac{\varepsilon}{3}+\frac{2\varepsilon}{3}\\&=\varepsilon.\end{aligned}$$

□

16.6. **Fubini's theorem.** We end with a consideration of the problem of reducing integration on $\mathbb{R}^n$ to smaller dimensions. Since $\mathbb{R}^n=\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}$ for any positive integers n_1,n_2 satisfying $n_1+n_2=n$, we may write any point in $\mathbb{R}^n$ in the form (x,y), where $x\in\mathbb{R}^{n_1}$ and $y\in\mathbb{R}^{n_2}$.

Definition 16.27. The *slice* of a function $f:\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}\to[-\infty,\infty]$ corresponding to a point $y\in\mathbb{R}^{n_2}$ is the function $f^y:\mathbb{R}^{n_1}\to[-\infty,\infty]$ given by $f^y(x)\coloneqq f(x,y)$. Similarly, the *slice* of f corresponding to a point $x\in\mathbb{R}^{n_1}$ is the function $f_x:\mathbb{R}^{n_2}\to[-\infty,\infty]$ given by $f_y(x)\coloneqq f(x,y)$.

In general, if $f:\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}\to[-\infty,\infty]$ is a measurable function, then it need not be the case that the slices $f^y:\mathbb{R}^{n_1}\to[-\infty,\infty]$ and $f_x:\mathbb{R}^{n_2}\to[-\infty,\infty]$ are measurable for each point $y\in\mathbb{R}^{n_2}$ and $x\in\mathbb{R}^{n_1}$. Nonetheless, measurability holds for *almost all* slices.

Theorem 16.28 (Fubini's theorem)**.** *Let $f:\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}\to[-\infty,\infty]$ be integrable. Then for a.e. $y\in\mathbb{R}^{n_2}$, the slice $f^y:\mathbb{R}^{n_1}\to[-\infty,\infty]$ is integrable and the function $\Phi:\mathbb{R}^{n_2}\to[-\infty,\infty]$ given by $\Phi(y)\coloneqq\int_{\mathbb{R}^{n_1}}f^y(x)\,dx$ is integrable. Similarly, for a.e. $x\in\mathbb{R}^{n_1}$, the slice $f_x:\mathbb{R}^{n_2}\to[-\infty,\infty]$ is integrable and the function $\Psi:\mathbb{R}^{n_1}\to[-\infty,\infty]$ given by $\Psi(x)\coloneqq\int_{\mathbb{R}^{n_2}}f_x(y)\,dy$ is integrable. Finally, we have that*

$$\iint_{\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}}f(x,y)\,dx\,dy=\int_{\mathbb{R}^{n_2}}\left(\int_{\mathbb{R}^{n_1}}f(x,y)\,dx\right)dy=\int_{\mathbb{R}^{n_1}}\left(\int_{\mathbb{R}^{n_2}}f(x,y)\,dy\right)dx.$$

Note that we can write this as

$$\iint_{\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}}f(x,y)\,dx\,dy=\int_{\mathbb{R}^{n_2}}\Phi(y)\,dy=\int_{\mathbb{R}^{n_1}}\Psi(x)\,dx.$$

Fubini's theorem shows that the integral of f on $\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}=\mathbb{R}^n$ can be computed by iterating lower-dimensional integrals and that these integrals can be taken in any order.

We defer the proof of Fubini's theorem. We first show some consequences.

Theorem 16.29 (Tonelli's theorem)**.** *Let $f:\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}\to[-\infty,\infty]$ be a nonnegative measurable function. Then for a.e. $y\in\mathbb{R}^{n_2}$, the slice $f^y:\mathbb{R}^{n_1}\to[-\infty,\infty]$ is measurable and the function $\Phi:\mathbb{R}^{n_2}\to[-\infty,\infty]$ given by $\Phi(y)\coloneqq\int_{\mathbb{R}^{n_1}}f^y(x)\,dx$ is measurable. Similarly, for a.e. $x\in\mathbb{R}^{n_1}$, the slice $f_x:\mathbb{R}^{n_2}\to[-\infty,\infty]$ is measurable and the function $\Psi:\mathbb{R}^{n_1}\to[-\infty,\infty]$ given by $\Psi(x)\coloneqq\int_{\mathbb{R}^{n_2}}f_x(y)\,dy$ is measurable. Finally, we have that*

$$\iint_{\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}}f(x,y)\,dx\,dy=\int_{\mathbb{R}^{n_2}}\left(\int_{\mathbb{R}^{n_1}}f(x,y)\,dx\right)dy=\int_{\mathbb{R}^{n_1}}\left(\int_{\mathbb{R}^{n_2}}f(x,y)\,dy\right)dx.$$

Again, we can write this as

$$\iint_{\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}} f(x,y)\,dx\,dy = \int_{\mathbb{R}^{n_2}} \Phi(y)\,dy = \int_{\mathbb{R}^{n_1}} \Psi(x)\,dx.$$

The key difference between Tonelli's theorem and Fubini's theorem is that we do not assume integrability in the latter, merely measurability (and additionally nonnegativity), so that the integral of f may be infinite; nonetheless, this equality of integrals and iterated integrals holds even if one of these is infinite (in which case they all are infinite).

Proof. Let $\Omega_m \coloneqq \{(x,y) \in B_m(0) : f(x,y) < m\}$, and define $f_m(x,y) \coloneqq f(x,y)\chi_{\Omega_m}(x,y)$ for each $m \in \mathbb{N}$. Each such function is measurable, and so by Fubini's theorem, the slice f_m^y is measurable for each y outside a set $A_m \subset \mathbb{R}^{n_1}$ of measure zero. Thus the slice f^y, being the pointwise limit of (f_m^y), is measurable outside the set $A \coloneqq \bigcup_{m=1}^{\infty} A_m$ of measure zero. For all such $y \notin A$, we have by the monotone convergence theorem that

$$\lim_{m\to\infty} \int_{\mathbb{R}^{n_1}} f_m^y(x)\,dx = \int_{\mathbb{R}^{n_1}} f^y(x)\,dx =: \Phi(y).$$

The right-hand side is measurable for all $y \notin A$ by Fubini's theorem, and the sequence of functions (Φ_m) given by $\Phi_m(y) \coloneqq \int_{\mathbb{R}^{n_1}} f_m(x,y)\,dx$ is pointwise nondecreasing a.e. and converges pointwise a.e. to $\Phi(y) \coloneqq \int_{\mathbb{R}^{n_1}} f(x,y)\,dx$. By the monotone convergence theorem once more,

$$\begin{aligned}\lim_{m\to\infty} \int_{\mathbb{R}^{n_2}} \left(\int_{\mathbb{R}^{n_1}} f_m(x,y)\,dx \right) dy &= \lim_{m\to\infty} \int_{\mathbb{R}^{n_2}} \Phi_m(y)\,dy \\ &= \int_{\mathbb{R}^{n_2}} \Phi(y)\,dy \\ &= \int_{\mathbb{R}^{n_2}} \left(\int_{\mathbb{R}^{n_1}} f(x,y)\,dx \right) dy.\end{aligned}$$

By Fubini's theorem,

$$\int_{\mathbb{R}^{n_2}} \left(\int_{\mathbb{R}^{n_1}} f_m(x,y)\,dx \right) dy = \iint_{\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}} f_m(x,y)\,dx\,dy,$$

and so once more by the monotone convergence theorem,

$$\lim_{m\to\infty} \iint_{\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}} f_m(x,y)\,dx\,dy = \iint_{\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}} f(x,y)\,dx\,dy.$$

Thus

$$\iint_{\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}} f(x,y)\,dx\,dy = \int_{\mathbb{R}^{n_2}} \left(\int_{\mathbb{R}^{n_1}} f(x,y)\,dx \right) dy = \int_{\mathbb{R}^{n_2}} \Phi(y)\,dy.$$

The same argument shows that the functions $f_x : \mathbb{R}^{n_2} \to [-\infty,\infty]$ and $\Psi : \mathbb{R}^{n_1} \to [-\infty,\infty]$ given by $\Psi(x) \coloneqq \int_{\mathbb{R}^{n_2}} f_x(y)\,dy$ are measurable for a.e. $x \in \mathbb{R}^n$ and that

$$\iint_{\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}} f(x,y)\,dx\,dy = \int_{\mathbb{R}^{n_1}} \left(\int_{\mathbb{R}^{n_2}} f(x,y)\,dy \right) dx = \int_{\mathbb{R}^{n_1}} \Psi(x)\,dx. \qquad \square$$

Tonelli's theorem is extremely useful in that it can be used to *justify* the use of Fubini's theorem. In order to apply Fubini's theorem, we must already know that $f : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to [-\infty,\infty]$ is integrable as well as measurable. Instead, we could replace f with $|f|$ and then apply Tonelli's theorem; if the iterated integral (in either order) of $|f|$ is finite, then Tonelli's theorem tells us that $|f|$ is integrable, and hence so is f, at which point the conditions of Fubini's theorem are met, so that we can apply Fubini's theorem to write the integral of f as an iterated integral.

Example 16.30. We shall use Tonelli's theorem to determine a formula for $m(B_r(x))$ in terms of the gamma function $\Gamma(z) \coloneqq \int_0^\infty y^{z-1}e^{-y}\,dy$. We let $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ be given by

$$\begin{aligned} f(x,y) &\coloneqq \begin{cases} e^{-y} & \text{if } y > x_1^2 + \cdots + x_n^2, \\ 0 & \text{otherwise,} \end{cases} \\ &= \chi_{[0,\infty)}(y)e^{-y}\chi_{B_{\sqrt{y}}(0)}(x) \\ &= \chi_{[x_1^2+\cdots+x_n^2,\infty)}(y)e^{-y}. \end{aligned}$$

Then by Tonelli's theorem and the fact that $m(B_r(0)) = r^n m(B_1(0))$ by the dilation equivariance of the Lebesgue measure,

$$\begin{aligned} \iint_{\mathbb{R}^n\times\mathbb{R}} f(x,y)\,dx\,dy &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}^n} \chi_{[0,\infty)}(y)e^{-y}\chi_{B_{\sqrt{y}}(0)}(x)\,dx \right) dy \\ &= \int_0^\infty e^{-y} m(B_{\sqrt{y}}(0))\,dy \\ &= m(B_1(0)) \int_0^\infty y^{\frac{n}{2}} e^{-y}\,dy \\ &= m(B_1(0))\Gamma\left(\frac{n}{2}+1\right). \end{aligned}$$

On the other hand,

$$\begin{aligned} \iint_{\mathbb{R}^n\times\mathbb{R}} f(x,y)\,dx\,dy &= \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}} \chi_{[x_1^2+\cdots+x_n^2,\infty)}(y)e^{-y}\,dy \right) dx \\ &= \int_{\mathbb{R}^n} e^{-(x_1^2+\cdots+x_n^2)}\,dx \\ &= \prod_{j=1}^n \int_{\mathbb{R}} e^{-x_j^2}\,dx_j \\ &= \pi^{n/2}. \end{aligned}$$

Here we have used the fact that $\int_{\mathbb{R}} e^{-x_j^2}\,dx_j = \sqrt{\pi}$, which follows from squaring both sides and then passing to polar coordinates on the left-hand side: letting $x_1 = r\cos\theta$ and $x_2 = r\sin\theta$, so that $dx_1\,dx_2 = r\,dr\,d\theta$, and then making the change of variables $r \mapsto \sqrt{r}$, we see that

$$\left(\int_{\mathbb{R}} e^{-x^2}\,dx\right)^2 = \iint_{\mathbb{R}\times\mathbb{R}} e^{-x_1^2}e^{-x_2^2}\,dx_1\,dx_2 = \int_0^{2\pi}\int_0^\infty e^{-r^2} r\,dr\,d\theta = \frac{1}{2}\int_0^{2\pi}\int_0^\infty e^{-r}\,dr\,d\theta = \pi.$$

By the translation invariance of the Lebesgue measure, we deduce that

$$m(B_r(x)) = r^n m(B_1(0)) = \frac{\pi^{n/2}r^n}{\Gamma\left(\frac{n}{2}+1\right)}.$$

We can go further by explicitly determining $\Gamma(\frac{n}{2}+1)$. For $z > 0$, we have via integration by parts that

$$\Gamma(z+1) = \int_0^\infty x^z e^{-x}\,dx = z\int_0^\infty x^{z-1}e^{-x}\,dx = z\Gamma(z).$$

Furthermore, $\Gamma(1) = \int_0^\infty e^{-x}\,dx = 1$. Finally, by making the change of variables $x \mapsto x^2$, we have that

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty \frac{1}{\sqrt{x}} e^{-x}\,dx = 2\int_0^\infty e^{-x^2}\,dx = \int_{\mathbb{R}} e^{-x^2}\,dx = \sqrt{\pi}.$$

By induction, we deduce that

$$\Gamma\left(\frac{n}{2}+1\right) = \begin{cases} \frac{n}{2}\cdot\left(\frac{n}{2}-1\right)\cdots 1 & \text{if } n \text{ is even,} \\ \frac{n}{2}\cdot\left(\frac{n}{2}-1\right)\cdots \frac{1}{2}\sqrt{\pi} & \text{if } n \text{ is odd.} \end{cases}$$

Another application of Tonelli's theorem is to take f to be the indicator function of a measurable set $\Omega \subseteq \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$. The resulting theorem involves slices of sets.

Definition 16.31. The *slice* of a set $\Omega \subseteq \mathbb{R}^{n_1}\times\mathbb{R}^{n_2}$ corresponding to a point $y \in \mathbb{R}^{n_2}$ is the set $\Omega^y \subseteq \mathbb{R}^{n_1}$ given by $\Omega^y \coloneqq \{x \in \mathbb{R}^{n_1} : (x,y) \in \Omega\}$. Similarly, the *slice* of Ω corresponding to a point $x \in \mathbb{R}^{n_1}$ is the set $\Omega_x \subseteq \mathbb{R}^{n_2}$ given by $\Omega_x \coloneqq \{y \in \mathbb{R}^{n_2} : (x,y) \in \Omega\}$.

In general, if $\Omega \subseteq \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, it need not be the case that the slices $\Omega^y \subseteq \mathbb{R}^{n_1}$ and $\Omega_x \subseteq \mathbb{R}^{n_2}$ are measurable. Nonetheless, measurability holds for *almost all* slices.

Corollary 16.32. *Let $\Omega \subseteq \mathbb{R}^{n_1}\times\mathbb{R}^{n_2}$ be a measurable set. Then for a.e. $y \in \mathbb{R}^{n_2}$, the slice $\Omega^y \subseteq \mathbb{R}^{n_1}$ is measurable and the function $\Phi : \mathbb{R}^{n_2} \to [-\infty,\infty]$ given by $\Phi(y) \coloneqq m(\Omega^y)$ is measurable. Similarly, for a.e. $x \in \mathbb{R}^{n_1}$, the slice $\Omega_x \subseteq \mathbb{R}^{n_2}$ is measurable and the function $\Psi : \mathbb{R}^{n_1} \to [-\infty,\infty]$ given by $\Psi(x) \coloneqq m(\Omega_x)$ is measurable. Finally, we have that*

$$m(\Omega) = \int_{\mathbb{R}^{n_2}} m(\Omega^y)\,dy = \int_{\mathbb{R}^{n_1}} m(\Omega_x)\,dx.$$

With this in hand, we may interpret the Lebesgue integral of an integrable function f in terms of the area underneath the graph of f. To show this, we first require the following.

Lemma 16.33. *If $\Omega_1 \subseteq \mathbb{R}^{n_1}$ and $\Omega_2 \subseteq \mathbb{R}^{n_2}$, then*

$$m^*(\Omega_1 \times \Omega_2) \le m^*(\Omega_1)m^*(\Omega_2),$$

where $\Omega_1 \times \Omega_2 \coloneqq \{(x,y) \in \mathbb{R}^{n_1}\times\mathbb{R}^{n_2} : x \in \Omega_1,\ y \in \Omega_2\}$.

Proof. This is immediate if either $m^*(\Omega_1)$ or $m^*(\Omega_2)$ is ∞. Otherwise, fix $\varepsilon > 0$. There exists covers $\{B_k\}$ and $\{B'_\ell\}$ of Ω_1 and Ω_2 by boxes such that

$$\sum_{k=1}^{\infty}\operatorname{vol}(B_k) \le m^*(\Omega_1) + \frac{\varepsilon}{3\max\{m^*(\Omega_2),1\}}, \quad \sum_{\ell=1}^{\infty}\operatorname{vol}(B'_\ell) \le m^*(\Omega_2) + \frac{\varepsilon}{3\max\{m^*(\Omega_1),1\}}.$$

Then $\{B_k \times B'_\ell\}$ is a cover of $\Omega_1 \times \Omega_2$ by boxes, so that

$$m^*(\Omega_1\times\Omega_2) \le \sum_{k,\ell=1}^{\infty}\operatorname{vol}(B_k\times B'_\ell) = \sum_{k=1}^{\infty}\operatorname{vol}(B_k)\sum_{\ell=1}^{\infty}\operatorname{vol}(B'_\ell) \le m^*(\Omega_1)m^*(\Omega_2)+\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, the result follows. □

Corollary 16.34. *If $\Omega_1 \subseteq \mathbb{R}^{n_1}$ and $\Omega_2 \subseteq \mathbb{R}^{n_2}$ are measurable, then $\Omega_1 \times \Omega_2 \subseteq \mathbb{R}^{n_1+n_2}$ is measurable, and*

$$m(\Omega_1\times\Omega_2) = m(\Omega_1)m(\Omega_2).$$

Proof. Since Ω_1, Ω_2 are both measurable, there exist G_δ sets $G_1 \supseteq \Omega_1$ and $G_2 \supseteq \Omega_2$ for which $m^*(G_1 \setminus \Omega_1) = m^*(G_2 \setminus \Omega_2) = 0$. Clearly $G_1 \times G_2 \subseteq \mathbb{R}^{n_1+n_2}$ is measurable, and

$$(G_1\times G_2)\setminus(\Omega_1\times\Omega_2) \subseteq ((G_1\setminus\Omega)\times G_2)\cup(G_1\times(G_2\setminus\Omega_2)).$$

It follows that

$$m^*((G_1\times G_2)\setminus(\Omega_1\times\Omega_2)) = 0,$$

which implies that $\Omega_1 \times \Omega_2$ is measurable. The identity $m(\Omega_1\times\Omega_2) = m(\Omega_1)m(\Omega_2)$ then follows from Tonelli's theorem applied to $\chi_{\Omega_1\times\Omega_2}$. □

Corollary 16.35. *If f is a measurable function on $\mathbb{R}^{n_1}$, then so is the function $\widetilde{f}$ on $\mathbb{R}^{n_1+n_2}$ defined by $\widetilde{f}(x,y) \coloneqq f(x)$.*

Proof. If $a \in \mathbb{R}$, then $\Omega_1 \coloneqq \{x \in \mathbb{R}^{n_1} : f(x) < a\}$ is measurable as f is measurable. Then

$$\{(x,y) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} : \widetilde{f}(x,y) < a\} = \Omega_1 \times \mathbb{R}^{n_2}$$

is measurable, and hence $\widetilde{f}$ is measurable. □

Theorem 16.36. *Let $f : \mathbb{R}^n \to [-\infty, \infty]$ be an integrable function and let*

$$\Omega^+ \coloneqq \left\{(x,y) \in \mathbb{R}^n \times \mathbb{R} : 0 \le y \le f^+(x)\right\},$$
$$\Omega^- \coloneqq \left\{(x,y) \in \mathbb{R}^n \times \mathbb{R} : 0 \le y \le f^-(x)\right\}.$$

Then Ω^+ and Ω^- are both measurable and of finite measure. Moreover,

$$\int_{\mathbb{R}^n} f(x)\,dx = m(\Omega^+) - m(\Omega^-).$$

Proof. We define $g^\pm : \mathbb{R}^n \times \mathbb{R} \to [-\infty, \infty]$ by $g^\pm(x,y) \coloneqq f^\pm(x) - y$. These are measurable functions, being the differences of two measurable functions, as $(x,y) \mapsto f^\pm(x)$ and $(x,y) \mapsto y$ are measurable. It follows that

$$\Omega^\pm = \{(x,y) \in \mathbb{R}^n \times \mathbb{R} : y \ge 0\} \cap \left\{(x,y) \in \mathbb{R}^n \times \mathbb{R} : g^\pm(x,y) \ge 0\right\}$$

are measurable sets, being the finite intersection of measurable sets. Then by Fubini's theorem,

$$m(\Omega^\pm) = \int_{\mathbb{R}^n \times \mathbb{R}} \chi_{\Omega^\pm}(x,y)\,dx\,dy = \int_{\mathbb{R}^n} m(\Omega^\pm_x)\,dx,$$

and

$$m(\Omega^\pm_x) = m\left(\{y \in \mathbb{R} : (x,y) \in \Omega^\pm\}\right) = m(\{y \in \mathbb{R} : 0 \le y \le f^\pm(x)\}) = f^\pm(x).$$

Thus $\int_{\mathbb{R}^n} f^\pm(x)\,dx = m(\Omega^\pm)$, and so

$$\int_{\mathbb{R}^n} f(x)\,dx = \int_{\mathbb{R}^n} f^+(x)\,dx - \int_{\mathbb{R}^n} f^-(x)\,dx = m(\Omega^+) - m(\Omega^-).$$ □

We return to the proof of Fubini's theorem. The proof that we give is similar to the construction of the Lebesgue integral, in the sense that we first prove Fubini's theorem for a particular class of integrable functions and then steadily increase the class of functions for which Fubini's theorem holds until we eventually arrive at arbitrary integrable functions. We begin as follows.

Lemma 16.37. *Fubini's theorem holds for $f = \chi_S$, where $S \subset \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ is an open box.*

Proof. We write $S = S_1 \times S_2$, where $S_1 \subset \mathbb{R}^{n_1}$ and $S_2 \subset \mathbb{R}^{n_2}$ are also open boxes. Then for each $y \in \mathbb{R}^{n_2}$, the function $f^y(x) = \chi_S(x,y)$ is measurable and integrable with integral

$$\Phi(y) = \int_{\mathbb{R}^{n_1}} \chi_S(x,y)\,dx = \begin{cases} \operatorname{vol}(S_1) & \text{if } y \in S_2, \\ 0 & \text{otherwise.} \end{cases}$$

It follows that $\Phi(y) = \operatorname{vol}(S_1)\chi_{S_2}(y)$ is also measurable and integrable with

$$\int_{\mathbb{R}^{n_2}} \Phi(y)\,dy = \operatorname{vol}(S_1)\operatorname{vol}(S_2) = \operatorname{vol}(S) = \iint_{\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}} \chi_\Omega(x,y)\,dx\,dy.$$

The same argument shows that $\int_{\mathbb{R}^{n_1}} \Psi(x)\,dx = \operatorname{vol}(S)$. □

Lemma 16.38. *Fubini's theorem holds for $f = \chi_L$, where $L \subset \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ is a subset of the boundary of a closed box.*

Proof. As the boundary of a closed box has measure zero in $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, we have that

$$\iint_{\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}} \chi_L(x, y)\, dx\, dy = 0.$$

Moreover, the slice $L^y \subset \mathbb{R}^{n_1}$ has measure zero for a.e. $y \in \mathbb{R}^{n_2}$, so that

$$\Phi(y) = \int_{\mathbb{R}^{n_1}} \chi_L(x, y)\, dx = 0$$

for a.e. $y \in \mathbb{R}^{n_2}$, and hence

$$\int_{\mathbb{R}^{n_2}} \Phi(y)\, dy = 0 = \iint_{\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}} \chi_L(x, y)\, dx\, dy.$$

The same argument shows that $\int_{\mathbb{R}^{n_1}} \Psi(x)\, dx = 0$. □

Lemma 16.39. *Fubini's theorem holds for f equal to the indicator function of a finite union of closed boxes $B_1, \ldots, B_K \subset \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ with disjoint interiors.*

Proof. We may write f as a finite linear combination of the indicator functions $\chi_{\mathrm{Int}(B_k)}$ together with indicator functions of the form χ_{L_k}, where each L_k is a subset of the boundary of B_k. The result then holds by linearity using the previous two lemmata. □

Lemma 16.40. *Fubini's theorem holds for f equal to the indicator function of an open set $E \subset \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ of finite measure.*

Proof. We may write E as a countable union of closed boxes B_j with disjoint interiors, so that $f(x, y) = \sum_{j \in J} \chi_{B_j}(x, y)$ for a.e. $(x, y) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$. Then the functions $f_k \coloneqq \sum_{\substack{j=1 \\ j \in J}}^{k} \chi_{B_j}$ are simple functions that are pointwise nondecreasing and converge pointwise a.e. to f. Moreover, by the previous lemma, we know that Fubini's theorem holds for each f_k. We claim that this implies that Fubini's theorem holds for f. Indeed, by the monotone convergence theorem,

$$\lim_{k \to \infty} \iint_{\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}} f_k(x, y)\, dx\, dy = \iint_{\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}} f(x, y)\, dx\, dy.$$

Similarly, the functions f_k^y are simple functions that are integrable outside a set A_k of measure zero that are pointwise nondecreasing and converge pointwise to f^y outside a set $A = \bigcup_{k=1}^{\infty} A_k$ of measure zero, so that for $y \notin A$,

$$\lim_{k \to \infty} \Phi_k(y) = \lim_{k \to \infty} \int_{\mathbb{R}^{n_1}} f_k^y(x)\, dx = \int_{\mathbb{R}^{n_1}} f^y(x)\, dx = \Phi(y)$$

by the monotone convergence theorem. Each Φ_k is integrable, is pointwise nondecreasing, and converges pointwise a.e. to Φ, so that by the monotone convergence theorem once more,

$$\lim_{k \to \infty} \int_{\mathbb{R}^{n_2}} \Phi_k(y)\, dy = \int_{\mathbb{R}^{n_2}} \Phi(y)\, dy.$$

It remains to note that $\int_{\mathbb{R}^{n_2}} \Phi_k(y)\, dy = \iint_{\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}} f_k(x, y)\, dx\, dy$, so that the left-hand side is equal to $\iint_{\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}} f(x, y)\, dx\, dy$. The same argument shows that $\int_{\mathbb{R}^{n_1}} \Psi(x)\, dx = \iint_{\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}} f(x, y)\, dx\, dy$. □

Lemma 16.41. *Fubini's theorem holds for f equal to the indicator function of a G_δ set $G \subset \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ of finite measure.*

Proof. Since G is a G_δ set, there exist a countable collection of open sets $\{E_j : j \in J\}$ for which $G = \bigcap_{j\in J} E_j$. Moreover, since G has finite measure, there exists an open set E of finite measure for which $G \subseteq E$. We let $\Omega_k \coloneqq E \cap \bigcap_{\substack{j=1\\ j\in J}}^{k} E_k$, which is open as it is the finite intersection of open sets. Moreover, Ω_1 has finite measure and $\Omega_k \supseteq \Omega_{k+1}$ for all $k \in \mathbb{N}$. We then define $f_k \coloneqq \chi_{\Omega_k}$, so that these are measurable functions that are pointwise nonincreasing and converge pointwise to f. By the previous lemma, we know that Fubini's theorem holds for each f_k. The same argument as in the previous lemma then implies that Fubini's theorem holds for f. □

Lemma 16.42. *Fubini's theorem holds for f equal to the indicator function of a set $\Omega \subset \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ of measure zero.*

Proof. As measure zero sets are measurable, Ω is measurable, and there exists a G_δ set G for which $\Omega \subseteq G$ and $m(G \setminus \Omega) = 0$; thus $m(G) = m(G \setminus \Omega) + m(\Omega) = 0$. By the previous lemma, Fubini's theorem holds for χ_G, so that

$$\int_{\mathbb{R}^{n_2}} \left(\int_{\mathbb{R}^{n_1}} \chi_G(x,y)\,dx \right) dy = \iint_{\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}} \chi_G(x,y) = 0,$$

and hence $\int_{\mathbb{R}^{n_1}} \chi_G(x,y)\,dx = 0$ for a.e. $y \in \mathbb{R}^{n_2}$. Thus the slice G^y has measure zero for a.e. $y \in \mathbb{R}^{n_2}$, and since $\Omega^y \subseteq G^y$, the same is true for Ω^y, so that

$$\int_{\mathbb{R}^{n_2}} \Phi(y)\,dy = 0 = \iint_{\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}} f(x,y).$$

The same argument implies that $\int_{\mathbb{R}^{n_1}} \Psi(x)\,dx = 0$. □

Lemma 16.43. *Fubini's theorem holds for f equal to the indicator function of a measurable set $\Omega \subset \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ of finite measure.*

Proof. Since Ω is measurable and of finite measure, there exists a G_δ set $G \subset \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ of finite measure for which $\Omega \subseteq G$ and $m(G \setminus \Omega) = 0$. Since $f = \chi_\Omega = \chi_G - \chi_{G\setminus\Omega}$, the result follows from the previous two lemmata together with linearity. □

Finally, we are able to prove Fubini's theorem in full generality.

Lemma 16.44. *Fubini's theorem holds for integrable functions $f : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to [-\infty, \infty]$.*

Proof. We write $f = f^+ - f^-$, where $f^\pm$ are nonnegative and integrable. There exist sequences of simple functions $(\varphi_m^\pm)$ that are pointwise nondecreasing and converge pointwise to $f^\pm$. Each such simple function $\varphi_k^\pm$ is a finite linear combination of indicator functions of measurable sets of finite measure. By the previous lemma, Fubini's theorem holds for such indicator functions, and by linearity, Fubini's theorem holds for each $\varphi_k^\pm$. Now the same monotone convergence theorem as used previously implies that Fubini's theorem holds for $f^\pm$, and so once more by linearity, Fubini's theorem holds for f. □

16.7. **The space of integrable functions.** We now discuss how we may view the space of integrable functions as a Banach space.

Definition 16.45. For functions $f, g : \mathbb{R}^n \to [-\infty, \infty]$, let $\sim$ denote the equivalence relation for which $f \sim g$ if and only if $f(x) = g(x)$ for a.e. $x \in \mathbb{R}^n$. We let $\mathcal{E}_f$ denote the equivalence class of f with respect to this equivalence relation. The space $L^1(\mathbb{R}^n)$ is the set

$$L^1(\mathbb{R}^n) \coloneqq \{\mathcal{E}_f : f \text{ is integrable}\}.$$

Theorem 16.46. *The space $L^1(\mathbb{R}^n)$ is a normed space with respect to the L^1-norm*

$$\|\mathcal{E}_f\|_{L^1} \coloneqq \int_{\mathbb{R}^n} |f(x)|\,dx.$$

In particular, $L^1(\mathbb{R}^n)$ is a metric space with respect to the L^1-metric $d_1(f,g) \coloneqq \|f - g\|_{L^1} = \int_{\mathbb{R}^n} |f(x) - g(x)|\,dx$.

Proof. We first note that this norm is well-defined, since if $f \sim g$, then $\int_{\mathbb{R}^n} |f(x)|\,dx = \int_{\mathbb{R}^n} |g(x)|\,dx$. Moreover, $L^1(\mathbb{R}^n)$ may be made into a vector space: if f, g are integrable and $\alpha \in \mathbb{R}$, then so is $f+g$ and αf, and we define $\mathcal{E}_f + \mathcal{E}_g \coloneqq \mathcal{E}_{f+g}$ and $\alpha\mathcal{E}_f \coloneqq \mathcal{E}_{\alpha f}$. This shows that $L^1(\mathbb{R}^n)$ is a vector space.

Next, we show that $\|\cdot\|_{L^1}$ is a norm, so that positivity, homogeneity, and subadditivity hold. Clearly $\|\mathcal{E}_f\|_{L^1} \geq 0$ for all integrable functions f; moreover, if $\|\mathcal{E}_f\|_{L^1} = 0$, then $f(x) = 0$ for a.e. $x \in \mathbb{R}^n$. Next, linearity of the Lebesgue integral implies that $\|\alpha\mathcal{E}_f\|_{L^1} = |\alpha|\|\mathcal{E}_f\|_{L^1}$ for all $\alpha \in \mathbb{R}$ and integrable functions f. Finally, the triangle inequality for $\mathbb{R}$ implies that $\|\mathcal{E}_f + \mathcal{E}_g\|_{L^1} \leq \|\mathcal{E}_f\|_{L^1} + \|\mathcal{E}_g\|_{L^1}$. □

In practice, we usually think of elements of $L^1(\mathbb{R}^n)$ as being functions rather than being equivalence classes of functions, with the usual understanding that we think of two functions of $L^1(\mathbb{R}^n)$ as being equal if they are equal a.e.

Theorem 16.47 (Riesz–Fischer)**.** *The space $L^1(\mathbb{R}^n)$ is a Banach space.*

Proof. We must show that $L^1(\mathbb{R}^n)$ is complete. Let (f_m) be a Cauchy sequence in $L^1(\mathbb{R}^n)$, so that for all $\varepsilon > 0$, there exists $M = M(\varepsilon) \in \mathbb{N}$ such that $\|f_m - f_\ell\|_{L^1} < \varepsilon$ whenever $m, \ell \geq M$. We take $\varepsilon = 2^{-k}$, which implies the existence of a subsequence (f_{m_k}) for which $\|f_{m_{k+1}} - f_{m_k}\|_{L^1} < 2^{-k}$ for all $k \in \mathbb{N}$, namely $m_k = M(2^{-k})$. We then define

$$f(x) \coloneqq f_{m_1}(x) + \sum_{k=1}^{\infty} \left(f_{m_{k+1}}(x) - f_{m_k}(x)\right),$$

$$g(x) \coloneqq |f_{m_1}(x)| + \sum_{k=1}^{\infty} \left|f_{m_{k+1}}(x) - f_{m_k}(x)\right|.$$

Since

$$\begin{aligned}\int_{\mathbb{R}^n} |f_{m_1}(x)|\,dx + \sum_{k=1}^{\infty} \int_{\mathbb{R}^n} \left|f_{m_{k+1}}(x) - f_{m_k}(x)\right| dx &\leq \int_{\mathbb{R}^n} |f_{m_1}(x)|\,dx + \sum_{k=1}^{\infty} \frac{1}{2^k} \\ &= \|f_{m_1}\|_{L^1} + 1,\end{aligned}$$

which is finite, the monotone convergence theorem implies that g is integrable; since $|f(x)| \leq g(x)$, f is also integrable. In particular, the series defining f converges a.e., and as the k-th partial sum of this series is f_{m_k}, we deduce that the series (f_{m_k}) converges pointwise a.e. to f. Moreover, $|f_{m_k}(x) - f(x)| \leq g(x)$ for all $k \in \mathbb{N}$ and for a.e. $x \in \mathbb{R}^n$, and so by the dominated convergence theorem,

$$\lim_{k\to\infty} \|f_{m_k} - f\|_{L^1} = \lim_{k\to\infty} \int_{\mathbb{R}^n} |f_{m_k}(x) - f(x)|\,dx = 0.$$

In particular, for all $\varepsilon > 0$, there exists $K = K(\frac{\varepsilon}{2}) \in \mathbb{N}$ such that $\|f_{m_k} - f\|_{L^1} < \frac{\varepsilon}{2}$ whenever $k \geq K$.

Finally, since (f_n) is Cauchy, for all $\varepsilon > 0$, there exists $M = M(\frac{\varepsilon}{2}) \in \mathbb{N}$ such that $\|f_m - f_{m_k}\| < \frac{\varepsilon}{2}$ whenever $m \geq M(\frac{\varepsilon}{2})$ and $m_k \geq \max\{M(\frac{\varepsilon}{2}), M(2^{-K(\frac{\varepsilon}{2})})\}$, and so by subadditivity,

$$\|f_m - f\|_{L^1} \leq \|f_m - f_{m_k}\|_{L^1} + \|f_{m_k} - f\|_{L^1} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Thus the Cauchy sequence (f_n) in $L^1(\mathbb{R}^n)$ converges to the integrable function $f \in L^1(\mathbb{R}^n)$, and hence $L^1(\mathbb{R}^n)$ is complete. □

The method of proof also yields the following.

Corollary 16.48. *Let (f_m) be a sequence in $L^1(\mathbb{R}^n)$ that converges in the L^1-norm to $f \in L^1(\mathbb{R}^n)$. Then there exists a subsequence $(f_{j(m)})$ that converges pointwise a.e. to f.*

The space $L^1(\mathbb{R}^n)$ contains plenty of interesting functions.

Proposition 16.49. *The following families of functions are dense in $L^1(\mathbb{R}^n)$:*

(1) *Simple functions $\varphi = \sum_{k=1}^{K} a_k \chi_{\Omega_k}$, where $a_k \in \mathbb{R}$ and $\Omega_k \subset \mathbb{R}^n$ is measurable and of finite measure;*
(2) *Simple functions $\psi = \sum_{\ell=1}^{L} b_\ell \chi_{B_\ell}$, where $b_\ell \in \mathbb{R}$ and $B_\ell \subset \mathbb{R}^n$ is a box;*
(3) *Continuous functions of compact support.*

Proof.
(1) We must show that given $f \in L^1(\mathbb{R}^n)$, there exists a sequence of simple functions (φ_m) for which $\lim_{m\to\infty} \|f - \varphi_m\|_{L^1} = 0$. This follows from the fact that there exists a sequence of simple functions (φ_m) that converges pointwise to f and satisfies $|\varphi_m(x)| \le |\varphi_{m+1}(x)|$ along with the dominated convergence theorem.
(2) It suffices to show that given a measurable set $\Omega \subset \mathbb{R}^n$ of finite measure and given $\varepsilon > 0$, there exists a simple function $\psi = \sum_{\ell=1}^{L} b_\ell \chi_{B_\ell}$, where $b_\ell \in \mathbb{R}$ and $B_\ell \subset \mathbb{R}^n$ is a box, for which $\|\chi_\Omega - \psi\|_{L^1} < \varepsilon$; by (1), linearity, and the triangle inequality, this will imply the desired result. To see this, as Ω is measurable, there exists a countable collection of almost disjoint boxes $\{B_j\}$ such that $\Omega \subseteq \bigcup_{j\in J} B_j$ and $\sum_{j\in J} \operatorname{vol}(B_j) \le m(\Omega) + \frac{\varepsilon}{2}$. As Ω has finite measure, the series on the left-hand side converges, so that there exists some $N = N(\frac{\varepsilon}{2})$ such that $\sum_{\substack{j=N+1 \\ j\in J}}^{\infty} \operatorname{vol}(B_j) < \frac{\varepsilon}{2}$. Then

$$\begin{aligned} m\left(\Omega \setminus \bigcup_{\substack{j=1 \\ j\in J}}^{N} B_j\right) + m\left(\bigcup_{\substack{j=1 \\ j\in J}}^{N} B_j \setminus \Omega\right) &\le m\left(\bigcup_{\substack{j=N+1 \\ j\in J}}^{\infty} B_j\right) + m\left(\bigcup_{j\in J} B_j \setminus \Omega\right) \\ &\le \sum_{\substack{j=N+1 \\ j\in J}}^{\infty} \operatorname{vol}(B_j) + \sum_{j\in J} \operatorname{vol}(B_j) - m(\Omega) \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\ &= \varepsilon. \end{aligned}$$

It follows that χ_Ω and $\psi \coloneqq \sum_{\substack{j=1 \\ j\in J}}^{N} \chi_{B_j}$ are equal except on a set of measure less than ε, and so $\|\chi_\Omega - \psi\|_{L^1} < \varepsilon$.
(3) It suffices to show that given a box $B \subset \mathbb{R}^n$ and given $\varepsilon > 0$, there exists a continuous compactly supported functions f for which $\|\chi_B - f\|_{L^1} < \varepsilon$; by (2), linearity, and the triangle inequality, this will imply the desired result. We write $B = [a_1, b_1] \times \cdots \times [a_n, b_n]$

and take $f(x) = f_1(x_1) \cdots f_n(x_n)$ with

$$f_j(x_j) \coloneqq \begin{cases} 0 & \text{if } x_j \leq a_j - \delta, \\ \dfrac{x_j - a_j}{\delta} + 1 & \text{if } a_j - \delta \leq x_j \leq a_j, \\ 1 & \text{if } a_j \leq x_j \leq b_j, \\ \dfrac{b_j - x_j}{\delta} + 1 & \text{if } b_j \leq x_j \leq b_j + \delta, \\ 0 & \text{if } x_j \geq b_j + \delta. \end{cases}$$

Then f is continuous and compactly supported, and $\|\chi_B - f\|_{L^1} < \varepsilon$ provided that $\delta > 0$ is sufficiently small with respect to ε. □

We briefly mention some standard results concerning change of variables for the Lebesgue integral.

Lemma 16.50. *Given $h \in \mathbb{R}^n$ and $f \in L^1(\mathbb{R}^n)$, define $f_h : \mathbb{R}^n \to \mathbb{R}$ by $f_h(x) \coloneqq f(x-h)$. Then f_h is integrable and satisfies*

$$\int_{\mathbb{R}^n} f(x-h)\,dx = \int_{\mathbb{R}^n} f(x)\,dx.$$

Similarly, given $\delta \in \mathbb{R} \setminus \{0\}$, the function $f(\delta x)$ is integrable and satisfies

$$\int_{\mathbb{R}^n} f(\delta x)\,dx = |\delta|^{-n} \int_{\mathbb{R}^n} f(x)\,dx.$$

These can be thought of as linear change of variables for the Lebesgue integrable. We do not give the proof; the method is to first prove these identities when f is the indicator function of a measurable set, then for f a simple function, then for integrable functions.

We additionally have the following.

Lemma 16.51. *For $f \in L^1(\mathbb{R}^n)$, we have that $\lim_{h\to 0} \|f_h - f\|_{L^1} = 0$.*

The proof takes advantage of the density of continuous compactly supported functions in $L^1(\mathbb{R}^n)$.

Proof. Since continuous compactly supported functions are dense in $L^1(\mathbb{R}^n)$, given $\varepsilon > 0$, there exists a continuous compactly supported function $g \in L^1(\mathbb{R}^n)$ for which $\|g - f\|_{L^1} < \frac{\varepsilon}{3}$. By the dominated convergence theorem and the uniform continuity of g, we have that $\lim_{h\to 0} \|g_h - g\|_{L^1} = 0$, so that there exists some $\delta = \delta(\frac{\varepsilon}{3}) > 0$ such that $\|g_h - g\|_{L^1} < \frac{\varepsilon}{3}$ whenever $|h| < \delta$. Thus for all $\varepsilon > 0$, there exists some $\delta > 0$ such that if $|h| < \delta$, then

$$\|f_h - f\|_{L^1} \leq \|f_h - g_h\|_{L_1} + \|g_h - g\|_{L_1} + \|g - f\|_{L_1} = \|g_h - g\|_{L_1} + 2\|g - f\|_{L_1} < \frac{\varepsilon}{3} + \frac{2\varepsilon}{3} = \varepsilon. \ \square$$

We can similarly define L^1 spaces for other subsets of $\mathbb{R}^n$.

Definition 16.52. Given a measurable set $\Omega \subseteq \mathbb{R}^n$ for which $m(\Omega) > 0$, the space $L^1(\Omega)$ is the set of all equivalence classes of integrable functions $f : \Omega \to [-\infty, \infty]$.

This is again a Banach space and continuous compactly supported functions are dense in this space. In particular, if Ω is compact, then the space $C(\Omega)$ of continuous functions $f : \Omega \to \mathbb{R}$ is dense in $L^1(\Omega)$. Taking $n = 1$ and $\Omega = [0, 1]$, we deduce the following.

Corollary 16.53. *The completion of the normed space $(C([0,1]), \|\cdot\|_{L^1})$ is $L^1([0,1])$.*

One can also define Banach spaces with respect to the L^p-norm.

Definition 16.54. Let $\Omega \subseteq \mathbb{R}^n$ be measurable and satisfy $m(\Omega) > 0$. For functions $f, g : \Omega \to [-\infty, \infty]$, let $\sim$ denote the equivalence relation for which $f \sim g$ if and only if $f(x) = g(x)$ for a.e. $x \in \Omega$. We let $\mathcal{E}_f$ denote the equivalence class of f with respect to this equivalence relation. For $p \in [1, \infty)$, the space $L^p(\Omega)$ is the set

$$L^p(\Omega) \coloneqq \left\{ \mathcal{E}_f : \int_\Omega |f(x)|^p \, dx \text{ is finite} \right\}.$$

We have the following results that extend the analogous results for $L^1(\Omega)$.

Theorem 16.55. *Let $\Omega \subseteq \mathbb{R}^n$ be measurable and satisfy $m(\Omega) > 0$. Fix $p \in [1, \infty)$.*

(1) *The space $L^p(\Omega)$ is a normed spaced with respect to the L^p-norm*

$$\|\mathcal{E}_f\|_{L^p} \coloneqq \left(\int_\Omega |f(x)|^p \, dx \right)^{\frac{1}{p}}.$$

In particular, the space $L^2(\Omega)$ is an inner product space with respect to the inner product

$$\langle \mathcal{E}_f, \mathcal{E}_g \rangle \coloneqq \int_\Omega f(x) g(x) \, dx.$$

(2) *The space $L^p(\Omega)$ is a Banach space. In particular, $L^2(\Omega)$ is a Hilbert space.*

(3) *The space of continuous functions of compact support is dense in $L^p(\Omega)$. In particular, the completion of the normed space $(C([0,1]), \| \cdot \|_{L^p})$ is $L^p([0,1])$.*

16.8. **The Fourier transform.** We end with a brief discussion of the Fourier transform on $\mathbb{R}^n$. This requires us to extend the definition of the Lebesgue integral to allow for *complex-valued* functions.

Definition 16.56. Let $f : \mathbb{R}^n \to \mathbb{C}$. Write $f(x) = u(x) + iv(x)$, where $u, v : \mathbb{R}^n \to \mathbb{R}$. The function f is said to be *measurable* if both u and v are measurable. The function f is said to be *integrable* if the real-valued function $|f(x)| = (u(x)^2 + v(x)^2)^{1/2}$ is integrable. The *Lebesgue integral* of f is

$$\int_{\mathbb{R}^n} f(x) \, dx \coloneqq \int_{\mathbb{R}^n} u(x) \, dx + i \int_{\mathbb{R}^n} v(x) \, dx.$$

We can now extend the definition of $L^1(\mathbb{R}^n)$ to allow for *complex-valued* functions. This is a *complex* vector space (i.e. we define $\alpha \mathcal{E}_f \coloneqq \mathcal{E}_{\alpha f}$ for $\alpha \in \mathbb{C}$ and f an integrable complex-valued function). Moreover, it is a Banach space, where we extend the notion of homogeneity of the norm to mean that for all $\alpha \in \mathbb{C}$ and $f \in L^1(\mathbb{R}^n)$, we have that $\|\alpha f\|_{L^1} = |\alpha| \|f\|_{L^1}$. With this in hand, we can define the Fourier transform of a function $f \in L^1(\mathbb{R}^n)$.

Definition 16.57. The *Fourier transform* of a function $f \in L^1(\mathbb{R}^n)$ is the function $\widehat{f} : \mathbb{R}^n \to \mathbb{C}$ given by

$$\widehat{f}(\xi) \coloneqq \int_{\mathbb{R}^n} f(x) e^{-2\pi i x \cdot \xi} \, dx.$$

Here $x \cdot \xi = x_1 \xi_1 + x_2 \xi_2 + \cdots + x_n \xi_n$ denotes the usual dot product on $\mathbb{R}^n$.

Lemma 16.58. *The Fourier transform $\widehat{f}$ of a function $f \in L^1(\mathbb{R}^n)$ is continuous and bounded.*

In general, $\widehat{f}$ need not lie in $L^1(\mathbb{R}^n)$. (Recall that continuous *compactly supported* functions lie in $L^1(\mathbb{R}^n)$, but $\widehat{f}$ need not be compactly supported.)

Proof. We first note that $\widehat{f}(\xi)$ exists for all $\xi \in \mathbb{R}^n$, as the fact that $f \in L^1(\mathbb{R}^n)$ implies that the integral defining $\widehat{f}(\xi)$ converges for all $\xi \in \mathbb{R}^n$. Moreover, as $|e^{i\theta}| = 1$ for all $\theta \in \mathbb{R}$, we have by the triangle inequality for the Lebesgue integral that

$$\sup_{\xi \in \mathbb{R}^n} |\widehat{f}(\xi)| = \sup_{\xi \in \mathbb{R}^n} \left| \int_{\mathbb{R}^n} f(x) e^{-2\pi i x \cdot \xi} \, dx \right| \leq \int_{\mathbb{R}^n} |f(x)| \, dx = \|f\|_{L^1}.$$

This shows that f is bounded. Finally, for any sequence (ξ_m) in $\mathbb{R}^n$ that converges to ξ, we have that $\lim_{m\to\infty} f(x) e^{-2\pi i x \cdot \xi_m} = f(x) e^{-2\pi i x \cdot \xi}$, and so the dominated convergence theorem implies that $\lim_{m\to\infty} \widehat{f}(\xi_m) = \widehat{f}(\xi)$. It follows that $\widehat{f}$ is continuous. □

We have the following duality between products of integrable functions and their Fourier transforms.

Lemma 16.59 (Multiplication formula for the Fourier transform)**.** *If $f, g \in L^1(\mathbb{R}^n)$, then*

$$\int_{\mathbb{R}^n} \widehat{f}(\xi) g(\xi) \, d\xi = \int_{\mathbb{R}^n} f(y) \widehat{g}(y) \, dy.$$

Proof. Both integrals converge since $f, g \in L^1(\mathbb{R}^n)$ and $\widehat{f}, \widehat{g}$ are bounded. We define the function $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{C}$ by $F(\xi, y) \coloneqq g(\xi) f(y) e^{-2\pi i \xi \cdot y}$. By Tonelli's theorem,

$$\iint_{\mathbb{R}^n \times \mathbb{R}^n} |F(\xi, y)| \, d\xi \, dy = \|f\|_{L^1} \|g\|_{L^1},$$

which is finite, and so F is integrable. We may therefore apply Fubini's theorem in order to see that

$$\int_{\mathbb{R}^n} \widehat{f}(\xi) g(\xi) \, d\xi = \iint_{\mathbb{R}^n \times \mathbb{R}^n} F(\xi, y) \, d\xi \, dy = \int_{\mathbb{R}^n} f(y) \widehat{g}(y) \, dy. \qquad \square$$

In particular cases, the Fourier transform $\widehat{f}$ of f has a nice closed form. One such case is when f is a modulated Gaussian.

Lemma 16.60. *Fix $x \in \mathbb{R}^n$ and $\delta > 0$ and define $g : \mathbb{R}^n \to \mathbb{R}$ to be the modulated Gaussian $g(\xi) \coloneqq e^{-\pi\delta|\xi|^2} e^{2\pi i x \cdot \xi}$. Then*

$$\widehat{g}(y) = \delta^{-n/2} e^{-\frac{\pi}{\delta}|y-x|^2}.$$

Here we write $|\xi|^2 \coloneqq \xi \cdot \xi$.

Proof. We note that

$$\widehat{g}(y) = \int_{\mathbb{R}^n} e^{-\pi\delta|\xi|^2} e^{-2\pi i (y-x) \cdot \xi} \, d\xi,$$

so that by making the change of variables $\xi \mapsto \delta^{-1/2}\xi$ and replacing y with $\delta^{-1/2} y + x$, we obtain

$$\widehat{g}(\delta^{-1/2} y + x) = \delta^{-n/2} \int_{\mathbb{R}^n} e^{-\pi|\xi|^2} e^{-2\pi i y \cdot \xi} \, d\xi.$$

It therefore suffices to show that $f(\xi) \coloneqq e^{-\pi \xi \cdot \xi}$ satisfies $\widehat{f}(\xi) = f(\xi)$. Since

$$\int_{\mathbb{R}^n} e^{-\pi|\xi|^2} e^{-2\pi i y \cdot \xi} \, d\xi = \prod_{j=1}^{n} \int_{\mathbb{R}} e^{-\pi\xi_j^2} e^{-2\pi i y_j \xi_j} \, d\xi_j,$$

it suffices to show this when $n = 1$.

Via differentiation under the integral sign (which is justified via the dominated convergence theorem) and integration by parts,

$$\frac{d}{dy}\widehat{f}(y) = \int_{\mathbb{R}} e^{-\pi\xi^2} (-2\pi i \xi) e^{-2\pi i y \xi} \, d\xi = -2\pi y \int_{\mathbb{R}} e^{-\pi\xi^2} e^{-2\pi i y \xi} \, d\xi = -2\pi y \widehat{f}(y).$$

We set $F(y) \coloneqq e^{\pi y^2} \widehat{f}(y)$, so that

$$F'(y) = e^{\pi y^2} \frac{d}{dy} \widehat{f}(y) + 2\pi y e^{\pi y^2} \widehat{f}(y) = 0,$$

and hence F is constant. Since $\widehat{f}(0) = \int_{\mathbb{R}} e^{-\pi \xi^2} \, d\xi = 1$, we deduce that $F(y) = 1$ for all $y \in \mathbb{R}$, which yields the desired identity. □

With this in hand, we may show that when $\widehat{f}$ *does* lie in $L^1(\mathbb{R}^n)$, we can *invert* the Fourier transform.

Theorem 16.61. *Suppose that the Fourier transform $\widehat{f}$ of a function $f \in L^1(\mathbb{R}^n)$ is integrable. Then for a.e. $x \in \mathbb{R}^n$, we have the Fourier inversion formula*

$$f(x) = \int_{\mathbb{R}^n} \widehat{f}(\xi) e^{2\pi i x \cdot \xi} \, d\xi.$$

Proof. Clearly this holds if $f(x) = 0$ for a.e. $x \in \mathbb{R}^n$. Otherwise, fix $x \in \mathbb{R}^n$ and $\delta > 0$ and define $g : \mathbb{R}^n \to \mathbb{R}$ to be the modulated Gaussian $g(\xi) \coloneqq e^{-\pi \delta |\xi|^2} e^{2\pi i x \cdot \xi}$. By the multiplication formula for the Fourier transform,

$$\int_{\mathbb{R}^n} \widehat{f}(\xi) e^{-\pi \delta |\xi|^2} e^{2\pi i x \cdot \xi} \, d\xi = \int_{\mathbb{R}^n} f(y) \delta^{-n/2} e^{-\frac{\pi}{\delta} |x-y|^2} \, dy.$$

As $\widehat{f} \in L^1(\mathbb{R}^n)$, the dominated convergence theorem implies that for each $x \in \mathbb{R}^n$, the left-hand side converges to $\int_{\mathbb{R}^n} \widehat{f}(\xi) e^{2\pi i x \cdot \xi} \, d\xi$ as δ converges to 0. Let $\Delta_\delta(x)$ denote the difference of the right-hand side and $f(x)$ We shall show that $\lim_{\delta \to 0} \|\Delta_\delta\|_{L^1} = 0$, which implies that there exists a subsequence (Δ_{δ_m}) that converges pointwise a.e. to 0, as desired.

To begin, we note that

$$\int_{\mathbb{R}^n} \delta^{-n/2} e^{-\frac{\pi}{\delta} |x-y|^2} \, dy = \int_{\mathbb{R}^n} e^{-\pi |y|^2} \, dy = \prod_{j=1}^n \int_{\mathbb{R}} e^{-\pi y_j^2} \, dy_j = 1$$

via the change of variables $y \mapsto x - y$. It follows that

$$\begin{aligned} |\Delta_\delta(x)| &= \left| \int_{\mathbb{R}^n} \left(f(x-y) - f(x) \right) \delta^{-n/2} e^{-\frac{\pi}{\delta} |y|^2} \, dy \right| \\ &\leq \int_{\mathbb{R}^n} |f(x-y) - f(x)| \, \delta^{-n/2} e^{-\frac{\pi}{\delta} |y|^2} \, dy \end{aligned}$$

We integrate over $x \in \mathbb{R}^n$. By Fubini's theorem, we deduce that

$$\|\Delta_\delta\|_{L^1} \leq \int_{\mathbb{R}^n} \|f_y - f\|_{L^1} \delta^{-n/2} e^{-\frac{\pi}{\delta} |y|^2} \, dy,$$

where $f_y(x) \coloneqq f(x-y)$.

Next, given $\varepsilon > 0$, there exists some $\eta = \eta(\frac{\varepsilon}{2}) > 0$ such that $\|f_y - f\|_{L^1} < \frac{\varepsilon}{2}$ whenever $|y| < \eta$. Moreover, for each $\eta > 0$, we have that

$$\lim_{\delta \to 0} \int_{B_\eta(0)^c} \delta^{-n/2} e^{-\frac{\pi}{\delta} |y|^2} \, dy = 0,$$

so that for all $\varepsilon > 0$, there exists $\lambda = \lambda(\frac{\varepsilon}{4\|f\|_{L^1}}, \eta) > 0$ such that if $\delta < \lambda$, then

$$\int_{B_\eta(0)^c} \delta^{-n/2} e^{-\frac{\pi}{\delta} |y|^2} \, dy < \frac{\varepsilon}{4\|f\|_{L^1}}.$$

It follows that

$$\|\Delta_\delta\|_{L^1} \leq \int_{\mathbb{R}^n} \|f_y - f\|_{L^1} \delta^{-n/2} e^{-\frac{\pi}{\delta} |y|^2} \, dy$$

$$
\begin{aligned}
&= \int_{B_\eta(0)} \|f_y - f\|_{L^1} \delta^{-n/2} e^{-\frac{\pi}{\delta}|y|^2}\, dy + \int_{B_\eta(0)^c} \|f_y - f\|_{L^1} \delta^{-n/2} e^{-\frac{\pi}{\delta}|y|^2}\, dy \\
&\leq \frac{\varepsilon}{2} \int_{B_\eta(0)} \delta^{-n/2} e^{-\frac{\pi}{\delta}|y|^2}\, dy + 2\|f\|_{L^1} \int_{B_\eta(0)^c} \delta^{-n/2} e^{-\frac{\pi}{\delta}|y|^2}\, dy \\
&< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\
&= \varepsilon.
\end{aligned}
$$

□

References

[Pug15] Charles Chapman Pugh, *Real Mathematical Analysis (Second Edition)*, Undergraduate Texts in Mathematics, Springer, Cham, 2015.

[SS05] Elias M. Stein and Rami Shakarchi, *Real Analysis. Measure Theory, Integration, and Hilbert Spaces*, Princeton Lectures in Analysis **III**, Princeton University Press, Princeton, 2005.

[Tao16] Terence Tao, *Analysis II (Third Edition)*, Texts and Readings in Mathematics **38**, Springer, Singapore, 2016.