GRADUATE REAL ANALYSIS

BRANDON HANSON

UNIVERSITY OF MAINE, ORONO

SPRING 2024

Table of Contents

1 FOUNDATIONS

1.1 Sets

A set S is a collection of objects s, called elements, and we write $s \in S$. To define a given set is to define the elements it contains. Sometimes we do it by listing elements explicitly; for instance the natural numbers are

$$\mathbb{N} = \{1, 2, 3, \ldots\},$$

and the integers are

$$\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$$

The initial segments are the sets

$$[n] = \{1, 2, 3, \ldots, n\}$$

where $n \in \mathbb{N}$. If X and Y are two sets, we write $Y \subseteq X$ if for each $y \in Y$ we have $y \in X$, and then we call Y a subset of X. We can also cut out subsets of a given set by imposing constraints. For instance, the perfect squares are the set

$$PS = \{n \in \mathbb{N} : n = m^2 \text{ for some } m \in \mathbb{N}\} = \{m^2 : m \in \mathbb{N}\}.$$

If X and Y are sets

$$X \cup Y = \{z : z \in X \text{ or } z \in Y\},\ X \cap Y = \{z : z \in X \text{ and } z \in Y\}$$

and more generally, if $\mathscr{X}$ is a collection of sets then

$$\bigcup_{X\in\mathscr{X}} X = \{x : x \in X \text{ for some } X \in \mathscr{X}\},\ \bigcap_{X\in\mathscr{X}} X = \{x : x \in X \text{ for all } X \in \mathscr{X}\}.$$

We also write

$$X \setminus Y = \{x \in X : x \notin Y\},\ X\Delta Y = (X \cup Y) \setminus (X \cap Y).$$

The cartesian product of X and Y is the set

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

We can define an arbitrary cartesian product of a collection $\mathscr{X}$ of sets X as

$$\bigtimes_{X\in\mathscr{X}} X = \{(x_X)_{X\in\mathscr{X}} : x_X \in X \text{ for each } X \in \mathscr{X}\}$$

where the point $(x_X)_{X\in\mathscr{X}}$ has one coordinate, x_X, for each set X in $\mathscr{X}$. However, in defining this cartesian product we may require the Axiom of Choice.

Axiom 1.1: Axiom of Choice

For any collection $\mathscr{X}$ of non-empty sets, we have

$$\bigtimes_{X\in\mathscr{X}} X \neq \emptyset.$$

We might also use an index set I, which is to say, suppose that for each $i \in I$, we have some set X_i. Then

$$\bigtimes_{i\in I} X_i = \{(x_i)_{i\in I} : x_i \in X_i \text{ for each } i \in I\}.$$

For example

$$\{0,1\}^{\mathbb{N}} = \bigtimes_{i\in\mathbb{N}} \{0,1\} = \{(e_1, e_2, \ldots) : e_i \in \{0,1\} \text{ for each } i \in \mathbb{N}\}.$$

This set is called the infinite boolean cube. Another interesting collection of sets are the direct sums. One such direct sum is

$$\bigoplus_{i\in\mathbb{N}} \mathbb{Z} = \{(n_1, n_2, \ldots) \in \bigtimes_{i\in\mathbb{N}} \mathbb{Z} : n_i \neq 0 \text{ for only finitely many values of } i\}.$$

Some other important sets are

$$\mathbb{Q} = \{a/b : a, b \in \mathbb{Z}, b > 0\}$$

and the real numbers $\mathbb{R}$, which we take as given.

1.2 Functions

A function f between sets X and Y is a map $f : X \to Y$ which takes an input from X and produces and output $f(x) \in Y$. Strictly speaking, we construct f as the graph $\Gamma_f \subseteq X \times Y$ given as

$$\Gamma_f = \{(x, y) \in X \times Y : y = f(x)\}$$

where, to be a graph, Γ_f passes the vertical line test: for each $x \in X$ there is a unique $y \in Y$ such that $(x, y) \in \Gamma_f$.

If $f : X \to Y$ is a function then we call X the domain and Y the co-domain. For a subset $X' \subseteq X$, the image of X' under f is the set

$$f(X') = \{f(x) : x \in X'\}$$

and for $Y' \subseteq Y$, the pre-image of Y' under f is the set

$$f^{-1}(Y') = \{x \in X : f(x) \in Y'\},$$

where, despite the notation, we make no guarantees about the invertibility of f. The function $f : X \to Y$ is said to be surjective (or a surjection) if $f(X) = Y$, or said differently, for each $y \in Y$ there is some $x \in X$ with $f(x) = y$. The function f is said to be injective if $f^{-1}(\{y\})$ is empty or a singleton (i.e. has one element) for each $y \in Y$, or said differently, $f(x_1) = f(x_2)$ if and only if $x_1 = x_2$. A function which is both injective and surjective is called bijective.

For example the function

$$\square : \mathbb{R} \to \mathbb{R},\ \square(x) = x^2$$

is neither injective nor surjective, but

$$\square : \mathbb{R} \to \mathbb{R},\ \square(x) = x^3$$

is bijective.

A set A is finite if there is an injection $f : A \to [n]$ for some $n \in \mathbb{N}$. We can just as well insist that A be in bijection with some $[m]$. Indeed, if $f : A \to [n]$ is an injection, then since $[n]$ is ordered, we may write

$$f(A) = \{i_1, \ldots, i_m\} \subseteq [n]$$

for some $i_1 < i_2 < \ldots < i_m$, listed in order. Since there is a unique element $a_l \in A$ such that $f(a_l) = i_l$, we define a bijection $e : A \to [m]$ by the rule

$$e(a_i) = i.$$

This map e is called an enumeration of A, and the number m is called the cardinality of A, denoted $|A|$.

Sets which are infinite are said to have the same cardinality if there is a bijection between them. A useful tool in this area is the Schröder-Bernstein theorem.

Theorem 1.1: Schröder-Bernstein

If there exist injections $f : X \to Y$ and $g : Y \to X$ then there is a bijection $h : X \to Y$.

Example. *There is a bijection from $\phi : \mathbb{N}^r \to \mathbb{N}$.*

Proof. There is an injection $f : \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ given by

$$f(j) = (j, 1, 1, \ldots, 1).$$

Going the other way, we can define

$$g(i_1, i_2, \ldots, i_r) = 2^i 3^j \ldots p_r^{i_r},$$

where $p_1 = 2, p_2 = 3, \ldots, p_r$ are the first r primes. □

Example. *There is a bijection from $\phi : \mathbb{Q} \to \mathbb{N}$.*

Proof. Again the injection from $\mathbb{N}$ to $\mathbb{Q}$ is easy. Going the other way, we can map $\mathbb{Q}$ to $\mathbb{N} \times \mathbb{N} \times \mathbb{N}$ by encoding $q \in \mathbb{Q}$ as a reduced fraction $q = ea/b$ where $a, b > 0$ and share no common factor, and $e = 1$, 0 or -1. The map $f(q) = (2 + e, a, b)$ defines in injection from $\mathbb{Q} \to \mathbb{N}^3$, and we already know there is a bijection from $\mathbb{N}^3$ to $\mathbb{N}$. □

A set which is in bijection with $\mathbb{N}$ is called **countable** (or countably infinite).

Example. *The set $\bigoplus_{i \in \mathbb{N}} \mathbb{Z}$ is countable.*

Proof. The map $n \mapsto (n, 0, 0, \ldots)$ defines an injection from $\mathbb{N}$ to $\bigoplus_{i \in \mathbb{N}} \mathbb{Z}$. Going the other way, we first enumerate the primes as

$$\mathscr{P} = \{2, 3, 5, \ldots\} = \{p_1, p_2, \ldots\}$$

and define a map

$$\phi : \bigoplus_{i \in \mathbb{N}} \mathbb{Z} \to \mathbb{Q};\ \phi(r_1, r_2, \ldots) = \prod_{i=1}^{\infty} p_i^{r_i}$$

where the product on the right hand side is really a finite one, since the condition of belonging to the direct sum is that only finitely many values of r_i are non-zero. Since prime factorization is unique, this map defines an injection. □

1.3 Relations

When A is a set, a subset $\mathscr{R} \subseteq A \times A$ is called a **relation**, and we write

$$a \sim_{\mathscr{R}} b \iff (a,b) \in \mathscr{R}.$$

There are a few properties which relations can have that often crop up.

1. **Reflexive**: The relation $\mathscr{R}$ is called reflexive if $a \sim_{\mathscr{R}} a$ for each $a \in A$.
2. **Symmetric**: The relation $\mathscr{R}$ is called symmetric if $a \sim_{\mathscr{R}} b \iff b \sim_{\mathscr{R}}$ for each $a, b \in A$.
3. **Anti-symmetric**: The relation $\mathscr{R}$ is called anti-symmetric if $a \sim_{\mathscr{R}} b$ and $b \sim_{\mathscr{R}} a$ imply $a = b$.
4. **Transitive**: The relation $\mathscr{R}$ is called transitive if $a \sim_{\mathscr{R}} b$ and $b \sim_{\mathscr{R}} c$ imply $a \sim_{\mathscr{R}} c$ for all $a, b, c \in A$.

A relation which is reflexive, symmetric and transitive is called an **equivalence relation**.

Example. *The relation $a \sim_{\mathscr{R}} b \iff a = b$ defines an equivalence relation on an set.*

Example. *The relation $a \sim_{\mathscr{R}} b \iff a = b = 0$ or $ab > 0$ defines an equivalence relation on $\mathbb{Z}$. Indeed, $aa = a^2 > 0$ unless $a = 0$ shows the relation is reflexive, $ab = ba$ yields symmetry, and if $a \sim_{\mathscr{R}} b$ and $b \sim_{\mathscr{R}} c$ then either $a = b = 0$ and $b = c = 0$ which means $a = c = 0$, or else $ab > 0$ and $bc > 0$ so that $ab^2c > 0$ and since $b^2 > 0$ we get $ac > 0$, giving transitivity.*

For the rest of this section, we'll assume we have a set A with an equivalence relation $\sim$ on it. For $a \in A$, the set

$$[a] = \{b \in A : b \equiv a\}$$

is called the equivalence class of a.

Lemma 1.1

If $a \sim b$ then $[a] = [b]$, while if $a \not\sim b$ then $[a] \cap [b] = \emptyset$.

Proof. Suppose $a \sim b$, then for $c \in [a]$, we have $c \sim a$ and so $c \sim b$ by transitivity. Thus $c \in [b]$ and hence $[a] \subseteq [b]$. By symmetry, $[b] \subseteq [a]$.

Now suppose that $a \not\sim b$. If $c \in [a] \cap [b]$ then $a \sim c$ and $c \sim b$ by transitivity, so $a \sim b$, which is not the case. □

From this, it follows that we can speak of an equivalence class without reference to an explicit element. Let $\mathscr{C}$ denote the collection of equivalence classes of A.

Lemma 1.2

The set $\mathscr{C}$ partitions A in the sense that the elements of $\mathscr{C}$ are pairwise disjoint, non-empty, and

$$\bigcup_{C \in \mathscr{C}} C = A.$$

Proof. Distinct classes are pairwise disjoint by the preceding lemma. Since $a \in [a]$ and $[a] \in \mathscr{C}$, we have

$$\bigcup_{C \in \mathscr{C}} C = A.$$

□

A set R is called a complete set of representatives for the relation if for each $C \in \mathscr{C}$ there exists a unique $r \in R$ with $r \in C$. Thus

$$A = \bigcup_{r \in R} [r]$$

is the partition from the above lemma.

Example. *Let A be a set with the trivial relation $a \sim b \iff a = b$. Then $R = A$ is the only possible set of representatives since $[a] = \{a\}$ for each $a \in A$.*

Example. *Let*

$$A = \{f : (0,\infty) \to (0,\infty)\}$$

and define the relation

$$f \sim g \iff \lim_{x \to \infty} \frac{f(x)}{g(x)} = 1.$$

A complete set of representatives for this relation is tough to pin down. That it exists comes from the Axiom of Choice.

Example. *Let*

$$A = \mathbb{R}$$

and define the relation

$$x \sim y \iff x - y \in \mathbb{Z}.$$

A complete set of representatives for this relation is explicit: it's the interval $[0,1)$, or any unit interval for that matter.

We say a function $f : A \to B$ is well-defined with respect to the equivalence relation $\sim$ on A provided $a \sim b \implies f(a) = f(b)$. This allows us to descend from a function on A to a function on the equivalence classes of A without running into ambiguity – we just set $\tilde{f}([a]) = f(a)$ and this doesn't change if we write $[a] = [b]$ for some other $b\tilde{a}$.

Example. *The functions which are well defined with respect to the relation from the previous example are the 1-periodic functions. Fourier analysis (and harmonic analysis more broadly) studies this topic in depth.*

Example. *This time we define $A = \mathbb{R}$ and let our equivalence relation be defined by $x \sim y \iff x - y \in \mathbb{Q}$. We have*

$$[x] = x + \mathbb{Q}$$

and the set $x + \mathbb{Q}$ intersects every unit interval. Thus we can find a complete set of representatives $\mathcal{V}$ which is completely contained in $(0,1]$. The set $\mathcal{V}$ is called a Vitali set.

1.4 Partial Orders

A relation which is antisymmetric and transitive is called a **partial order**. We usually denote such a relation by $a \preceq_{\mathscr{R}} b$ or just $a \preceq b$. The relation is called a partial order because it does not guarantee that any two elements are comparable. If the partial order further satisfies $a \preceq b$ or $b \preceq a$ for any a and b then it is called a **total** order.

Example. *We define a partial order on the natural numbers by $a \preceq b$ if a divides b. Indeed $a = 1 \cdot a$; if $a = mb$ and $b = na$ then $a = mna$ so $mn = 1$ which forces $m = n = 1$; and if $b = ma$ and $c = nb$ then $c = (mn)a$. The order is partial, not total, since 2 and 3 are incomparable.*

If we replace $\mathbb{N}$ by $\mathbb{Z}$, this ceases to be a partial order since $a|-a$ and $-a|a$.

Example. *Let X be a non-empty set. Then $\subseteq$ defines a partial order on 2^X. If X contains at least two elements, then the order is not total: the sets $\{a\}$ and $\{b\}$ with $a \neq b$ are not comparable.*

Example. *Let $\mathscr{X}$ be any collection of sets. We would like to define a partial order on $\mathscr{X}$ by $A \preceq B$ if there exists an injection from A to B. However this will fail to be antisymmetric: there are injections from $\{a\}$ to $\{b\}$ and vice-versa, but they are not the same set. This can be resolved by defining an equivalence relation on $\mathscr{X}$ by $A \equiv B$ if there is a bijection from A to B and defining the order on equivalence classes: $[A] \preceq [B]$ if there is an injection from A to B. Unfortunately this order appears to be dependent on A and B, which could be problematic if we choose other representatives from $[A]$ and $[B]$. This turns out not to be the case: if $A \sim A'$ and $B \sim B'$ then there*

is a bijection $f : A' \to A$, and injection $i : A \to B$ and a bijection $g : B \to B'$, so that $g \circ i \circ f : A' \to B'$ is an injection. What this means is that if we to represent $[A] = [A']$ and $[B] = [B']$ by different representatives, the order relation would still hold. Furthermore, if $[A] \preceq [B]$ and $[B] \preceq [A]$ then there are injections from A to B and from B to A. By the Schröder-Bernstein theorem, there must be a bijection from A to B, so that $[A] = [B]$. Thus on the level of equivalence classes, $\preceq$ does *define a partial order.*

Definition 1.1: Chain

If X is a set partially ordered by $\preceq$ and Y is a subset of X, then Y is also partially ordered by $\preceq$, just by restriction our field of vision to Y. If Y is totally ordered by this partial order, we call Y a chain.

Example. *Returning to $\mathbb{N}$ with the division ordering, we see that while the entirety of $\mathbb{N}$ is not totally ordered (remember, 2 and 3 are incomparable), the set $\{1,2,4,8,\ldots\} = \{2^j : j \geq 0\}$ is totally ordered with respect to division, and so forms a chain.*

Definition 1.2: Upper bound, maximal element

If X is a set partially ordered by $\preceq$ and Y is a subset of X, then an element u is said to be an upper bound for Y if $y \preceq u$ for each $y \in Y$. An element m is said to be maximal if $m \preceq x$ implies $m = x$. Note that maximal elements need not be upper bounds for things – the definition of maximality does not require m be comparable to anything.

Axiom 1.2: Zorn's Lemma

Let X be a partially ordered set with the following property: if Y is a chain in X, then there is a $u \in X$ which is an upper bound for Y. Then X contains a maximal element.

We now give an application of Zorn's lemma, which is a prototypical example of how to use it. One should think of this as a form of induction, however we aren't using the natural number to index our statements.

Theorem 1.2

Let A and B be non-empty sets. Then there is either an injection $i : A \to B$ or else an injection $j : B \to A$.

Proof. Let

$$\mathscr{I} = \{(A', f) : A' \subseteq A,\ f : A' \to B \text{ an injection}\}$$

be the set of *partial injections* from A to B. In other words, $\mathscr{I}$ consists of the subdomains $A' \subseteq A$ on which an injection $f : A' \to B$ exists. We'd like to show that there is an $(A, i) \in \mathscr{I}$, which would mean we could define an injection on all of A.

For now, $\mathscr{I}$ is not empty. Indeed, since A and B are non-empty, we can choose some $a \in A$ and $b \in B$ and define $f(a) = b$ so that $(\{a\}, f) \in \mathscr{I}$. This singleton set $\{a\}$ is a far cry from all of A, but its sole purpose is to show $\mathscr{I}$ is non-empty, which you should think of as the base case of our induction.

The set $\mathscr{I}$ is partially ordered as follows: if (A_1, i_1) and (A_2, i_2) belong to $\mathscr{I}$, we write $(A_1, i_1) \preceq (A_2, i_2)$ if $A_1 \subseteq A_2$ and the injection i_2 extends i_1 in the sense that where they are both defined, they do the exact same thing:

$$i_2(a) = i_1(a) \text{ if } a \in A_1.$$

We next show this really is a partial order.

To see transitivity, suppose $(A_1, i_1) \preceq (A_2, i_2)$ and $(A_2, i_2) \preceq (A_3, i_3)$. Then, by definition, $A_1 \subseteq A_2 \subseteq A_3$, the maps $i_j : A_i \to B$ are injections for $j = 1, 2, 3$, and if $a \in A_1$ then $i_2(a) = i_1(a)$ while if $a \in A_2$ then $i_3(a) = i_2(a)$. Hence if $a \in A_1$, we also have $a \in A_2$ and so

$$i_3(a) = i_2(a) = i_1(a)$$

which shows that i_3 extends i_1. This tells us $(A_1, i_1) \preceq (A_3, i_3)$.

For asymmetry, suppose $(A_1, i_1) \preceq (A_2, i_2)$ and vice-versa. Then $A_1 \subseteq A_2$ and $A_2 \subseteq A_1$, so in fact $A_1 = A_2$ as sets. Moreover, on A_1, $i_2(a) = i_1(a)$. But A_1 is all of A_2, so i_1 and i_2 agree everywhere. This shows $(A_1, i_1) = (A_2, i_2)$.

So we have a non-empty, partially ordered set $\mathscr{I}$ on our hands, and we'd like to apply Zorn's Lemma to it. In order to do so, we need to establish the chain condition. So let $\mathscr{C}$ be a chain in $\mathscr{I}$. This means that $\mathscr{C}$ is a subset of $\mathscr{I}$ and any two elements of $\mathscr{C}$ are comparable under the partial order $\preceq$. Our task is to find an element of $\mathscr{I}$ which is an upper bound for $\mathscr{C}$.. Each element of $\mathscr{C}$ is a pair, say (C, i_C) where $C \subseteq A$ and $i_C : C \to B$ is an injection. To find an upper bound for $\mathscr{C}$, we need a pair (U, i_U) where U is a domain that is bigger than every domain C coming from $\mathscr{C}$, and $i_U : U \to B$ is an injection extending each injection i_C from $\mathscr{C}$. Let

$$U = \bigcup_{(C, i_C) \in \mathscr{C}} C,$$

so U is the union of all the domains that come from $\mathscr{C}$. It immediately follows that $C \subseteq U$ whenever $(C, i_C) \in \mathscr{C}$. We have our domain, now we need to set about defining an injection. It is here that we will use that $\mathscr{C}$ is a chain. If $a \in U$, pick

some C to which a belongs. Define $i_U : U \to B$ by $i_U(a) = i_C(a)$. In this way we have defined a function from i_U. It remains to show it is an injection, and it extends each i_C

To see that i_U is an extension of i_C, suppose $a \in C$. We defined $i_U(a)$ to be $i_{C'}(a)$ where C' was some domain, possibly distinct from C. But (C, i_C) and $(C', i_{C'})$ are both elements of $\mathscr{C}$ and $\mathscr{C}$ is a *chain*, so either $C' \subseteq C$ and $i_{C'}$ extends i_C or vice-versa. In any case, $a \in C \cap C'$ and $i_C = i_{C'}$ on their intersection. So $i_U(a) = i_{C'(a)}$ by definition of i_U and $i_{C'}(a) = i_C(a)$. Thus $i_U(a) = i_C(a)$ and i_U indeed extends i_C.

Now let's show i_U is injective. To that end, suppose $i_U(a_1) = i_U(a_2)$ for some $a_1, a_2 \in U$. Thus there is some C_1 and C_2 (the sets we used to define i_U at each a_j) with $a_1 \in C_1$, $a_2 \in C_2$ and

$$i_{C_1}(a_1) = i_{C_2}(a_2).$$

Again because $\mathscr{C}$ is a chain, we can assume without loss of generality that $C_1 \subseteq C_2$ and i_{C_1} extends i_{C_2}. So $i_{C_2}(a_1) = i_{C_2}(a_2)$. However, i_{C_2} is an injection, whence $a_1 = a_2$. This concludes the proof that i_U is an injection. Hence $(U, i_U) \in \mathscr{I}$ and is an upper bound for $\mathscr{C}$, allowing us to apply Zorn's Lemma.

From a maximal element we want to conclude the proof of this theorem. You should think of this as proving the inductive step. Suppose (M, i) is a maximal element of $\mathscr{I}$. So M is a subset of A and $i : M \to B$ is an injection. If $M = A$ we're done. If i is surjective, then i is a bijection between M and B, and i^{-1} is an injection from B to A, which also completes the proof. So assume that neither holds: $M \neq A$ and $i(M) \neq B$. Choose $a \in A \setminus M$ and $b \in B \setminus i(M)$. Let $M' = M \cup \{a\}$ and define $i' : M' \to B$ by $i'(a) = b$ and $i'(a') = i(a')$ for $a \in M$. The function i' is injective on M, since i is, and if $a' \in M$ then

$$i'(a) = b \neq i(a') = i'(a').$$

So we have successfully extended the injection i to the injection i', contradicting the maximality of (M, i). □

This proof is long, and that's because there are a lot of details to check. None of those details is particularly tough to check however, and so you should think of the proof as this: what is the largest subset of A on which we *can* define an injection to B. Zorn's Lemma tells us such a subset must exist, but if it doesn't exhaust A, and its image doesn't exhaust B, then we can easily extend this injection by *just one more element*, and that is a violation of maximality.

2

ANALYSIS ON METRIC SPACES

2.1 Metric Spaces

Definition 2.1: Metric Space

A metric space (X, d) is a set X endowed with a metric function

$$d : X \times X \to [0, \infty)$$

which satisfies the following rules, for all $x, y, z \in X$

Positivity: $d(x, y) = 0 \iff x = y$,

Symmetry: $d(x, y) = d(y, x)$, and

Triangle inequality: $d(x, y) \leq d(x, z) + d(z, y)$.

We'll give some examples below, not always with a proof, just yet.

Example. *The most fundamental example is $\mathbb{R}$ (or a subset of $\mathbb{R}$) endowed with the distance*

$$d(x, y) = |x - y|.$$

The properties of the metric are easy to check, and you should recognize the triangle

inequality for d as merely the triangle inequality $|x-y| \le |x-z| + |z-y|$.

Example. *The space $\mathbb{R}^n$ with the l^p-metric is defined by*

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}.$$

Example. *The space $\mathbb{R}^n$ with the l^∞-metric is defined by*

$$d(x,y) = \max_{1 \le i \le n} |x_i - y_i|.$$

Indeed, $d(x,y) \ge 0$ and (1), and (2) of the metric properties are easy. For the triangle inequality, we have

$$\begin{aligned} d(x,y) = \max_{1\le i\le n} |x_i - y_i| &\le \max_{1\le i\le n} \left(|x_i - z_i| + |z_i - y_i| \right) \le \max_{1\le i\le n} \left(|x_i - z_i| + \max_{1\le i\le n} |z_i - y_i| \right) \\ &= d(x,z) + d(z,y). \end{aligned}$$

Examine the two inequalities in the above line and be sure you understand them.

Example. *The space of continuous function $f : [0,1] \to \mathbb{R}$ is denoted $\mathscr{C}[0,1]$. All such functions are bounded, since $[0,1]$ is compact. Thus it makes sense to define*

$$d(f,g) = \sup_{x \in [0,1]} |f(x) - g(x)|.$$

Check that this is a metric on the space of functions $\mathscr{C}[0,1]$. The proof is much the same as in the preceding example.

2.2 Inner product spaces and normed vector spaces

Let V be a vector space (or any dimension, possible infinite) over $\mathbb{R}$. It is often the case that V can be endowed with a notion of size that will allow us to measure distance.

Definition 2.2: Normed Space

A real vector space V is said to be a normed space if there is a function

$$\|\cdot\| : V \to [0, \infty)$$

with the properties

Positivity: $\|v\| = 0 \iff v = 0$,

Scaling: $\|c \cdot v\| = |c| \|v\|$, and

Triangle inequality: $\|u + v\| \le \|u\| + \|v\|$.

The relevance of normed spaces is that they give us nice metric spaces.

Lemma 2.1: Norms give metrics

If V is a normed space then V is also a metric space with the metric

$$d(u,v) = \|u - v\|.$$

Proof. We only verify the triangle inequality, the other properties are immediate from the definition of norm. Indeed, if $u, v, w \in V$ then

$$d(u,v) = \|u - v\| = \|(u - w) + (w - v)\| \leq \|u - w\| + \|w - v\| = d(u,w) + d(w,v).$$

□

A particularly nice type of normed space is a (real)-inner product space.

Definition 2.3: Real inner product space

A real vector space V is said to be an inner product space if there is a function

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$$

with the following properties

Positivity: $\langle v, v \rangle \geq 0$ with equality if and only if $v = 0$,

Symmetry: $\langle u, v \rangle = \langle v, u \rangle$,

Linearity: $\langle u + cv, w \rangle = \langle u, w \rangle + c\langle v, w \rangle$.

The following is left as an exercise.

Lemma 2.2

Let V be a real inner product space, suppose $u_1, \ldots, u_m, v_1, \ldots, v_n \in V$, and suppose $c_1, \ldots, c_m, d_1, \ldots, d_n \in \mathbb{R}$. Then

$$\left\langle \sum_{i=1}^{m} c_i u_i, \sum_{j=1}^{n} d_j v_j \right\rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} c_i d_j \langle u_i, v_j \rangle.$$

Theorem 2.1: The Cauchy-Schwarz Inequality

Suppose $u, v \in V$ for some real inner product space V. Then

$$\langle u, v\rangle^2 \leq \langle u, u\rangle\langle v, v\rangle.$$

Moreover, the inequality is strict unless $u = tv$ for some $t \in \mathbb{R}$.

Proof. Fix $u, v \in V$ and consider the function of $t \in \mathbb{R}$ defined by

$$Q(t) = \langle u - tv, u - tv\rangle = \langle u, u\rangle - 2t\langle u, v\rangle + t^2\langle v, v\rangle,$$

by the preceding lemma. On the one hand $Q(t)$ is a quadratic polynomial in t, and on the other, it's of the form $\langle w, w\rangle$ with $w = u - tv$ and so the inner product property (1) tells us $Q(t) \geq 0$, and $Q(t) = 0$ only if $u = tv$. Thus the discriminant of the polynomial Q is non-negative, and zero only if Q has a root. But Q has discriminant

$$(-2\langle u, v\rangle)^2 - 4(\langle v, v\rangle)(\langle u, u\rangle) \leq 0.$$

This rearranges to the claimed inequality, and equality can only only hold if $Q(t) = 0$ for some t whence $u = tv$ for that same t. □

Corollary 2.1

Let V be a real inner product space. Then $\|v\| = \sqrt{\langle v, v\rangle}$ is a norm on V.

Proof. We prove the triangle inequality for norms and leave the rest to the reader. By definition

$$\|u\|^2 = \langle u, u\rangle, \|v\|^2 = \langle v, v\rangle, \ \|u + v\|^2 = \langle u + v, u + v\rangle = \|u\|^2 + 2\langle u, v\rangle + \|v\|^2.$$

From the Cauchy-Schwarz inequality,

$$\|u+v\|^2 = \|u\|^2+2\langle u, v\rangle+\|v\|^2 \leq \|u\|^2+2|\langle u, v\rangle|+\|v\|^2 \leq \|u\|^2+2\|u\|\|v\|+\|v\|^2 = (\|u\|+\|v\|)^2.$$

□

This prove that $\mathbb{R}^n$ with the l^2 metric is indeed a metric space.

Example. *Again let $\mathscr{C}[0,1]$ denote the space of functions $f : [0,1] \to \mathbb{R}$. With pointwise addition and scalar multiplication, we can think of $\mathscr{C}[0,1]$ as a real vector space. In fact, it is an inner product space with*

$$\langle f, g\rangle = \int_0^1 f(x)g(x)dx.$$

The norm induced by this inner product is called the L^2 norm

$$\|f\|_{L^2} = \left(\int_0^1 |f(x)|^2 dx\right)^{1/2},$$

and the metric is

$$d(f,g) = \left(\int_0^1 |f(x) - g(x)|^2 dx\right)^{1/2},$$

which is an average of how far apart f and g tend to be.

2.3 Metric spaces, topologically

Definition 2.4: Topological space

A set X with a collection $\mathscr{U} \subseteq \mathscr{P}(X)$ of subsets of X is said to be a topological space if $\mathscr{U}$ has the following properties

1. $\emptyset \in \mathscr{U}$ and $X \in \mathscr{U}$,
2. for any collection $\mathscr{U}' \subseteq \mathscr{U}$ we have

$$\bigcup_{U \in \mathscr{U}'} U \in \mathscr{U},$$

   which is to say that $\mathscr{U}$ is closed under arbitrary unions, and
3. for $U_1, \ldots, U_n \in \mathscr{U}$, we have

$$U_1 \cap \cdots \cap U_n \in \mathscr{U}$$

   which is to say that $\mathscr{U}$ is closed under finite intersections.

We call the sets $U \in \mathscr{U}$ open.

We are going to define open sets in a metric space X so that it becomes a topological space.

Definition 2.5: Open ball

If (X,d) is a metric space, $x \in X$ and $\rho > 0$ then the open ball of radius ρ centred at x is the set

$$\mathscr{B}(x,\rho) = \{y \in X : d(x,y) < \rho\}.$$

Definition 2.6: Open set

If (X, d) is a metric space, a subset $U \subseteq X$ is said to be open if for every $x \in U$, there is some $\rho = \rho_x > 0$ (which can depend on x) such that $\mathscr{B}(x, \rho_x) \subseteq U$.

Lemma 2.3

The open sets in a metric space define a topology.

Proof. The fact that X is open is trivial and the fact that $\emptyset$ is open is vacuous. Suppose $\mathscr{U}$ is a collection of open sets and suppose $x \in \bigcup_{U \in \mathscr{U}} U$. Then $x \in U'$ for some U', and since U' is open,

$$\mathscr{B}(x, \rho) \subseteq U' \subseteq \bigcup_{U \in \mathscr{U}} U$$

for some $\rho > 0$. Meanwhile if $x \in U_1 \cap \cdots U_n$ for some open sets $U_1, \ldots, U_n$ then there are positive numbers ρ_j such that

$$\mathscr{B}(x, \rho_j) \subseteq U_j.$$

Letting $\rho = \min\{\rho_j : 1 \leq j \leq n\}$, we have $\rho > 0$ since the minimum is over a finite set. One should verify that $B(x, \rho) \subseteq B(x, \rho_j)$ for each j and then we deduce $B(x, \rho) \subseteq U_j$ for each j as well. Thus $B(x, \rho) \subseteq U_1 \cap \cdots \cap U_n$. □

With open sets in hand we can define a continuous function between metric spaces.

Definition 2.7: Continuity, uniform continuity

Let (X_1, d_1) and (X_2, d_2) be metric spaces and suppose $f : X_1 \to X_2$ is a function. We say f is continuous at $x \in X$ if for each $\varepsilon > 0$ there is some $\delta > 0$ such that $d_1(x, y) < \delta$ implies $d_2(f(x), f(y)) < \varepsilon$. We say that f is continuous if it's continuous at each $x \in X$. We say f is uniformly continuous if for $\varepsilon > 0$ there is some $\delta > 0$ such that $d_2(f(x), f(y))$ for all x, y with $d_1(x, y) < \delta$.

Note that, for vanilla continuity, δ depends both on the value of ε and the point x where f is continuous. For uniform continuity, δ is only allowed to depend on ε.

2.4 Convergence, Closed sets, and Completeness

Definition 2.8: Sequence, convergent sequence, Cauchy-Sequence

A sequence in a metric space (X,d) (or in a subset Y of X) is a function $x:\mathbb{N}\to X$ (or $x:\mathbb{N}\to Y$), but we will just write x_n for $x(n)$, and $\{x_n\}$ for the whole sequence. The sequence is said to converge to x if for any $\varepsilon>0$ there is a threshold N such that $d(x,x_n)<\varepsilon$ for $n\geq N$, and we write $x_n\to x$. The sequence is called Cauchy if for $\varepsilon>0$ there is a threshold N such that $d(x_m,x_n)<\varepsilon$ for $n,m\geq N$.

Definition 2.9: Closed set

A set F in a metric space (X,d) is closed if either of the following equivalent conditions holds: any sequence $\{x_n\}$ of points in F has a limit in F, or, F^c is open.

Lemma 2.4

Convergent sequences are Cauchy.

Proof. Suppose $x_n\to x$. Let $\varepsilon>0$ and choose N so large that $d(x_n,x)<\varepsilon/2$ for $n\geq N$. Then if $m,n\geq N$, we have

$$d(x_m,x_n)\leq d(x_m,x)+d(x,x_n)<\varepsilon.$$

□

A partial converse of the above lemma is that Cauchy sequences are guaranteed to converge once a potential limit has been identified.

Lemma 2.5

Suppose a Cauchy sequence $\{x_n\}$ has a subsequence converging to x. Then $x_n\to x$

In general metric spaces, Cauchy sequences may not converge.

Definition 2.10: Complete space

The metric space (X,d) is called complete if every Cauchy sequence in X converges.

Theorem 2.2

The space $\mathbb{R}$ with the usual metric $d(x,y) = |x-y|$ is complete.

Theorem 2.3

The metric space $\mathbb{R}^n$ with the l^2 metric $d(x,y) = \left(\sum_{i=1}^n (x_i - y_i)^2\right)^{1/2}$ is complete.

Proof. Let $\{x_k\}$ be a Cauchy sequence. Each x_k is a vector, which we write as

$$x_k = (x_k(1), \ldots, x_k(n)),$$

and the Cauchy condition tells us that

$$\left(\sum_{i=1}^n (x_k(i) - x_j(i))^2\right)^{1/2} < \varepsilon$$

provided j and k are sufficiently large. But then

$$\max_i |x_k(i) - x_j(i)| < \varepsilon$$

too, and this tells us that each sequence $\{x_k(i)\}_k$ is a Cauchy sequence in $\mathbb{R}$, and so converges to some $x(i)$. We claim $x_k \to x$, for if $\varepsilon > 0$ we can find some N such that $|x_k(i) - x(i)| < \varepsilon/\sqrt{n}$ whenever $k \geq N$, and from this

$$d(x_k, x) = \left(\sum_{i=1}^n (x_k(i) - x(i))^2\right)^{1/2} < \varepsilon.$$

□

The idea of the above theorem is to "bootstrap" the completeness of $\mathbb{R}$ to that of $\mathbb{R}^n$. The vectors $x_k = (x_k(1), \ldots, x_k(n))$ can just as well be thought of as functions $x_k : [N] \to \mathbb{R}$. In that spirit, we also have the following.

Theorem 2.4

The space $\mathscr{C}[0,1]$ with metric

$$d(f,g) = \sup_x |f(x) - g(x)|$$

is complete.

To prove this we'll need a bit of nomenclature concerning the convergence of functions.

Definition 2.11: Pointwise and uniform convergence

Let X be a subset of $\mathbb{R}$ and for each n, suppose $f_n : X \to \mathbb{R}$ is a function. We say $f_n \to f : X \to \mathbb{R}$ if for each $x \in X$, and for each $\varepsilon > 0$ there is an N such that $|f_n(x) - f(x)| < \varepsilon$ once $n \geq N$. In other words, for each $x \in X$, the sequence $\{f_n(x)\}_n$ of real numbers converges to $f(x)$. This convergence is called uniform if for $\varepsilon > 0$ there is an N such that $|f_n(x) - f(x)| < \varepsilon$ for all x, that is, N depends on ε, but not on x.

Lemma 2.6

If $f_n : X \to \mathbb{R}$ is a sequence of continuous (resp. uniformly continuous) functions converging uniformly to $f : X \to \mathbb{R}$, then $f : X \to \mathbb{R}$ is also continuous (resp. uniformly continuous).

Proof. Let $x, y \in X$. Let $\varepsilon > 0$ and suppose n is so large that $|f_n(z) - f(z)| < \varepsilon/3$ for all $z \in X$. Choose $\delta = \delta(x, \varepsilon)$ (resp. $\delta = \delta(\varepsilon)$) so that $|x - y| < \delta$ implies $|f_n(x) - f_n(y)| < \varepsilon/3$. Then

$$|f(x) - f(y)| \leq |f(x) - f_n(x)| + |f_n(x) - f_n(y)| + |f_n(y) - f(y)| < \varepsilon/3 + \varepsilon/3 + \varepsilon/3.$$

□

We can also upgrade Lemma 2.4 to the uniform convergence setting.

Lemma 2.7

Suppose $\{f_n : X \to \mathbb{R}\}_n$ is uniformly Cauchy sequence of functions in the sense that for $\varepsilon > 0$ and m, n sufficiently large

$$|f_n(x) - f_m(x)| < \varepsilon$$

holds for all $x \in X$. Furthermore, suppose there is a subsequence $\{f_{n_k}\}$ converging uniformly to f. Then $f_n \to f$ uniformly as well.

Proof. Let N be so large that $|f_{n_k}(x) - f(x)| < \varepsilon/2$ for all $x \in X$ once $k > N$ and furthermore, that $|f_n(x) - f_m(x)| < \varepsilon/2$ for all $x \in X$ once $m, n \geq N$. Then

$$|f(x) - f_n(x)| \leq |f(x) - f_{n_k}(x)| + |f_{n_k}(x) - f_n(x)| < \varepsilon$$

provided $k, n > N$ (using, implicitly, that $n_k \geq k$). □

Proof of Theorem 2.4. Let $\{f_n\}$ be a Cauchy sequence of functions. Then

$$\sup_x |f_n(x) - f_m(x)| < \varepsilon$$

provided m, n are sufficiently large. Thus for any x, if n, m are large enough, we know $|f_n(x) - f_m(x)| < \varepsilon$, which tells us the sequence $\{f_n(x)\}_n$ is Cauchy, and hence convergent to some number $f(x)$. Thus there is a function $f : [0,1] \to \mathbb{R}$ which we have identified as a potential limit of our sequence. However, it's hard to tell if f should be continuous (and hence in $\mathscr{C}[0,1]$) just yet. More to the point, we know that for each x, $f_n(x) \to f(x)$, which is pointwise convergence, but for $f_n \to f$ in our metric, we need

$$\sup_x |f_n(x) - f(x)| < \varepsilon$$

which is uniform convergence.

So, we would like $f_n(x)$ to converge uniformly. But actually, the metric on $\mathscr{C}[0,1]$ already tells us that a Cauchy sequence is uniformly Cauchy, so we need only identify a uniformly convergent subsequence and apply the preceding lemma. To that end, for $k \in \mathbb{N}$, let n_k be chosen in an increasing fashion so that

$$\sup_x |f_n(x) - f_m(x)| < \frac{1}{2^k}$$

when $n, m \geq n_k$, and in particular, so that

$$\sup_x |f_{n_k}(x) - f_{n_{k+1}}(x)| < \frac{1}{2^k}.$$

Now we apply the "summation trick"

$$f_{n_k}(x) = f_{n_1}(x) + \sum_{j=1}^{k-1} f_{n_{k+1}}(x) - f_{n_k}(x).$$

Because $f_n(x) \to f(x)$, we know $f_{n_k}(x) \to f(x)$ as well, and so, as a series

$$f(x) = f_{n_1}(x) + \sum_{j=1}^{\infty} f_{n_{k+1}}(x) - f_{n_k}(x)$$

and

$$|f(x) - f_{n_k}(x)| = \left| \sum_{j=k}^{\infty} f_{n_{k+1}}(x) - f_{n_k}(x) \right| \leq \sum_{j=k}^{\infty} \left| f_{n_{k+1}}(x) - f_{n_k}(x) \right| < \sum_{j=k}^{\infty} 2^{-j} = 2^{1-k},$$

which gives uniform convergence. □

Definition 2.12: Closure

Let Y be a set in a metric space (X, d). The closure of Y, denoted $\overline{Y}$, is the intersection of all closed sets containing Y, or equivalently,

$$\overline{Y} = \{z : \text{ there is some sequence } \{y_n\} \text{ in } Y \text{ with } y_n \to z\}.$$

The closure of Y is closed, and is the smallest closed set containing Y.

Definition 2.13: Denseness

A set B is said to be dense in A if $A \subseteq \overline{B}$.

Theorem 2.5: Existence of Completion

For any metric space (X, d), there is a complete metric space $(\tilde{X}, \tilde{d})$ and an injection

$$i : X \to \tilde{X}$$

such that

$$d(x, y) = \tilde{d}(i(x), i(y))$$

for $x, y \in X$ and $i(X)$ is dense in $\tilde{X}$.

We won't prove this here, but I'll give some homework problems that outline the construction. Instead, here's a classic theorem that will foreshadow some approximation theorems we'll see later in the course. The idea of the proof is a precursor to Monte Carlo methods, which uses some notion of randomness to approximate something deterministic.

Theorem 2.6: Weierstrass Approximation

The polynomial functions

$$\mathbb{R}[x] = \left\{ \sum_{j=0}^{d} c_j x^j : c_0, \ldots, c_d \in \mathbb{R} \right\}$$

are dense in $\mathscr{C}[0, 1]$.

Proof. We'll assume for this proof that we know $f \in \mathscr{C}[0, 1]$ is in fact uniformly continuous. This we shall prove in the next section. With that in mind, let $\varepsilon > 0$ and suppose δ is such that $|f(x) - f(y)| < \varepsilon/2$ whenever $|x - y| \leq \delta$. We'll also take for granted that $|f(x)| \leq M$ holds for all $x \in [0, 1]$, for some M.

We have to show that $f \in \mathscr{C}[0, 1]$ is the limit of some sequence of polynomials, or what is the same, that for every $\varepsilon > 0$, there is a polynomial $p(x)$ with $\sup_x |p(x) - f(x)| < \varepsilon$.

For $x \in [0, 1]$ imagine a biased coin c taking values 0 or 1 with probability distribution $\mathbb{P}(c = 1) = x$ and $\mathbb{P}(c = 0) = 1 - x$, and flip it n times, and denote the outcomes $c_1(x), \ldots, c_n(x)$. The basic statistics of this experiment follow a binomial distribution

$$\mathbb{P}(c_1(x) + \cdots + c_n(x) = i) = \binom{n}{i} x^i (1 - x)^{n-i},$$

the expected number of 1's is

$$\mathbb{E}(c_1(x)+\cdots+c_n(x)) = nx$$

and the variance is (by independence between each flip)

$$Var(c_1(x)+\cdots+c_n(x)) = nVar(c_1(x)) = nx(1-x) \le n.$$

So we don't expect $c_1(x)+\cdots+c_n(x)$ to deviate from nx by much more than a standard deviation, certainly $\sqrt{n}$. This can be formalized by Chebychev's inequality

$$\mathbb{P}(|c_1(x)+\cdots+c_n(x)-nx| \ge n^{2/3}) \le \frac{1}{n^{4/3}} Var(c_1(x)+\cdots+c_n(x)) \le n^{-1/3}.$$

Now split up the interval $[0,1]$ into $n+1$ pieces $\left[\frac{i}{n+1}, \frac{i+1}{n+1}\right]$, $i = 0,\ldots,n$. If x belongs to the i'th piece, then x is about $i/n+1$ and nx is about i. So we expect i 1's to show up for x. Because f is continuous, and because we expect nx to be pretty close to i, and we might bet on

$$\sigma(x) = f\left(\frac{c_1(x)+\cdots+c_n(x)}{n}\right) \approx f(x).$$

So we use this random function as a predictor for f. What do we expect from σ in actuality? Well

$$\mathbb{E}(\sigma(x)) = \sum_{i=0}^{n} \mathbb{P}(c_1(x)+\cdots+c_n(x)=i) f\left(\frac{i}{n}\right) = \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} f(i/n)$$

which is a polynomial function of x!

How good is this random prediction? Well, let's take n big enough so that $n^{-1/3} < \min\{\delta, \varepsilon/4M\}$. Then

$$|\mathbb{E}(\sigma(x)) - f(x)| \le \mathbb{E}(|\sigma(x)-f(x)|)$$

and we split the expectation on the right according to whether $|c_1(x)+\cdots+c_n(x)-nx|$ is big or small. When it's bigger than $n^{2/3}$, we bound

$$|\sigma(x)-f(x)| \le 2M$$

trivially, but this only happens with probability at most $n^{-1/3}$. In the complementary case,

$$|c_1(x)+\cdots+c_n(x)-nx| < n^{2/3}$$

so

$$\left|\frac{c_1(x)+\cdots+c_n(x)}{n} - x\right| < n^{-1/3} < \delta$$

and by continuity, we have $|\sigma(x)-f(x)| < \varepsilon/2$. So

$$\begin{aligned}|\mathbb{E}(\sigma(x)) - f(x)| \le{}& \mathbb{P}(|c_1(x)+\cdots+c_n(x)-nx| > n^{2/3}) 2M + \\ &+ \mathbb{P}(|c_1(x)+\cdots+c_n(x)-nx| \le n^{2/3}) \frac{\varepsilon}{2} \\ &\le \frac{\varepsilon}{4M}\cdot 2M + 1\cdot\frac{\varepsilon}{2} = \varepsilon.\end{aligned}$$

□

2.5 Compactness

Definition 2.14: Sequential compactness

A set C in (X,d) is called sequentially compact if any sequence $\{x_n\}$ in C has a subsequence which converges to a limit in C.

Definition 2.15: Open cover

Let (X,d) be a metric space (or a topological space, more broadly) and $C \subseteq X$. A set $\mathscr{U}$ of of open subsets of X is said to be an open cover of C if $C \subseteq \bigcup_{U\in\mathscr{U}} U$.

Definition 2.16: Compactness

A set C in a metric space (X,d) (or a topological space, more broadly) is called compact if for any open cover $\mathscr{U}$ of C, there are finitely many sets $U_1,\dots,U_n \in \mathscr{U}$ such that $C \subseteq U_1 \cap \cdots \cap U_n$.

Lemma 2.8

In a metric space (X,d), compact sets are closed, as are sequentially compact sets.

Proof. Let C be compact and $\{x_n\}$ a sequence converging to x. For $n \in \mathbb{N}$, let

$$U_n = \{y : d(y,x) > 1/n)\}$$

which is an open set. If $x \notin C$ then the sets U_r cover all of C and hence admit a finite subcover. But these sets are increasing, so U_N covers all of C for some N, and that means x_n can't converge to x.

Sequentially compact sets are closed almost by definition. If $\{x_n\}$ is a sequence in C with $x_n \to x$ then x_n has a subsequence which converges C, but the limit of this subsequence is also x. □

Lemma 2.9

Let $\mathscr{C}$ be a collection of compact subsets of a metric space (X,d) with the property that any finite intersection of sets from $\mathscr{C}$ is non-empty. Then $\bigcap_{C\in\mathscr{C}} C$ is non-empty too.

Proof. For $C \in \mathscr{C}$, let $U_C = X \setminus C$, which is an open set. If $\bigcap_{C\in\mathscr{C}} C = \emptyset$ then $\bigcup_{C\in\mathscr{C}} U_C = X$ and certainly covers any $C' \in \mathscr{C}$. So for any $C' \in C$, there are finitely many sets

$U_{C_1}, \ldots, U_{C_n}$ covering C' whence

$$C' \cap C_1 \cap \cdots C_n = \emptyset,$$

in violation of the finite intersection property. □

Lemma 2.10

If C is a compact set and $F \subseteq C$ is closed, then F is compact.

Proof. Let $\mathscr{U}$ be an open cover of F and let $\mathscr{U}' = \mathscr{U} \cup \{F^c\}$. Then $\mathscr{U}'$ is an open cover of X and hence of C, and so it has a finite subcover. The sets in this subcover distinct from F^c cover F. □

Lemma 2.11

Compact sets are sequentially compact.

Proof. Let $\{x_n\}$ be a sequence in some compact set C. Let $C_n = \overline{\{x_k : k \geq n\}}$, which is a closed, and hence compact subset of C. These sets have the finite intersection property and hence there is some $x \in C_n$ for all n, which mean for each n, there is some $x_{k(n)}$ with $k(n) > n$ and $d(x_{k(n)}, x) < 1/n$. Since $k(n) > n$, we can extract from the sequence $k(n)$ an infinite sequence $n_j \to \infty$ and the sequence $x_{n_j} \to x$ by construction. □

One convenient aspect of compact sets is they let us “discretize" things, in the sense that we can approximate C arbitrarily well by a finite set.

Definition 2.17: Total boundedness

A set C in a metric space (X, d) is said to be totally bounded if for any $\varepsilon > 0$, there are finitely many open ball of radius ε which cover C.

Lemma 2.12

If C is sequentially compact then it is totally bounded.

Proof. If C is not totally bounded then for some $\varepsilon > 0$, there is no covering of C by finitely many balls of radius ε. Let $x_1 \in C$. The ball $\mathscr{B}(x_1, \varepsilon)$ fails to cover C, so there is some x_2 in C with $d(x_1, x_2) > \varepsilon$. Inductively define x_j as follows: having defined $x_1, \ldots, x_{j-1}$, the balls $\mathscr{B}(x_i, \varepsilon)$ cannot cover all of C and so fail to cover some $y \in C$. Set $x_j = y$. By construction, the distance between distinct points in this sequence is at least ε, and so the sequence is not Cauchy, and hence not convergent. □

Compact sets are also useful because they guarantee that continuous functions are really nice.

Lemma 2.13: Continuity and compactness

Suppose C is a compact set in a metric space (X_1, d_1), (X_2, d_2) is some other metric space and $f : X_1 \to X_2$ is continuous. Then f is uniformly continuous on C and $f(C)$ is a compact subset of X_2. If C is only sequentially compact, then $f(C)$ is still sequentially compact.

Proof. For the first claim, let $\varepsilon > 0$. Then to each $x \in C$, there is an open ball $\mathscr{B}(x, \delta_x)$ such that $y \in B(x, \delta_x)$ implies $d_2(f(x), f(y)) < \varepsilon/2$. The sets $\mathscr{B}(x, \delta_x/2)$ form an open cover of C, and so have a finite subcover, say $\mathscr{B}(x_1, \delta_{x_1}/2), \ldots, \mathscr{B}(x_n, \delta_{x_n}/2)$. Let $\delta = \min\{\delta_{x_1}, \ldots, \delta_{x_n}\}/2$ and consider $x, y \in X_1$ with $d_1(x, y) < \delta$. We know $x \in B(x_i, \delta_{x_i}/2)$ for some i, and so

$$d_1(x_i, y) \leq \delta_{x_i}/2 + d_1(x, y) < \delta_{x_i}$$

and hence

$$d_2(f(x), f(y)) \leq d_2(f(x), f(x_i)) + d_2(f(x_i), f(y)) < \varepsilon.$$

For the second claim, let $\mathscr{U}_2$ be a covering of $f(C)$ by open sets. The sets $f^{-1}(U)$ with $U \in \mathscr{U}_2$ form an open cover of C and hence admit a finite subcover $f^{-1}(U_1), \ldots, f^{-1}(U_n)$. The sets $U_1, \ldots, U_n$ cover $f(C)$.

For the final claim, suppose $\{y_n\}$ is a sequence in $f(C)$. Then $y_n = f(x_n)$ and $\{x_n\}$ is a sequence in C with a convergent subsequence $\{x_{n_k}\}$ such that $x_{n_k} \to x$. Then $y_{n_k} \to f(y) \in f(C)$. □

Corollary 2.2

Let $f : X \to \mathbb{R}$ be a continuous function using the usual metric on $\mathbb{R}$. Then f achieves a maximum and minimum value on any compact set $C \subseteq X$.

Proof. We already know $f(C)$ is compact, which implies $f(C)$ is closed and totally bounded (and certainly just bounded). Let $\{x_n\}$ be a sequence in C such that $f(x_n) \to \sup f(C)$. Then $\{x_n\}$ has a subsequence $\{x_{n_k}\}$ converging to some $x \in C$, and by continuity $f(x_{n_k}) \to f(x)$, whence $f(x) = \sup f(C)$, achieving a maximum value. Applying the same argument to $-f$ achieves the minimum. □

Lemma 2.14: The Lebesgue Number

Let C be a sequentially compact set in a metric space (X,d), and let $\mathscr{U}$ be an open cover. Then there is a $\delta = \delta(\mathscr{U})$ such that for each $x \in C$, the ball $B(x,\delta)$ is contained in some U belonging to $\mathscr{U}$.

Proof. Let $V = \bigcup_{U\in\mathscr{U}} U$. Define the function $f : X \to \mathbb{R}$ by

$$f(x) = \begin{cases} \sup\{\rho : \rho \in [0,1],\ B(x,\rho) \subseteq U \text{ for some } U \in \mathscr{U}\} & x \in V \\ 0 & \text{otherwise.} \end{cases}$$

This function is continuous. Indeed, suppose x, y satisfy $d(x,y) < \delta$. If neither x nor y belong to V, then $f(x) = f(y) = 0$. Suppose next that exactly one of $x, y \in V$, say y. Then because $x \in B(y,2\delta)$ we have $B(y,2\delta) \not\subseteq U$ for any $U \in \mathscr{U}$, hence $f(y) < 2\delta$, so in this case f is continuous at x with $\delta = \varepsilon/2$. Otherwise $x \in V$, and the same argument applies. We are left to contend with the case that $x, y \in V$. Take $\delta = \varepsilon/2$, let $f(x) - \varepsilon/2 < \rho_x < f(x)$, and suppose $\rho_y = \rho_x - \varepsilon/2$. Then $f(x) - \varepsilon < \rho_y < f(x)$. There is some U with $\mathscr{B}(x,\rho_x) \subseteq U$ and if $z \in \mathscr{B}(y,\rho_y)$, then $d(z,x) \le \rho_y + d(x,y) < \rho_y + \varepsilon/2 < \rho_x$. Thus $\mathscr{B}(y,\rho_y) \subseteq U$ and we see $f(y) \ge \rho_y > f(x) - \varepsilon$. Similarly we see $f(x) < f(y) + \varepsilon$.

From continuity and compactness we see that f achieves a minimum on C and since f is strictly positive on C, that minimum, say 2δ, is positive too. Thus for each $x \in C$ we have $\delta < f(x)/2$ and hence $\mathscr{B}(x,\delta) \subseteq U$ for some $U \in \mathscr{U}$. □

Theorem 2.7: Borel-Lebesgue

In a metric space (X,d), compactness and sequential compactness are equivalent.

Proof. We've already seen that compactness implies sequential compactness. Now suppose that we have a sequentially compact set C and $\mathscr{U}$ is an open cover of C. Let δ be the Lebesgue number of $\mathscr{U}$, and because C is totally bounded, let $c_1, \ldots, c_n$ be centres of $\delta/2$ balls covering C. For each i, pick some $x_i \in \mathscr{B}(c_i, \delta/2)$. Then $B(x_i,\delta) \subseteq U_i$ for some $U_i \in \mathscr{U}$ and for any other $x \in C$, $d(x,c_j) < \delta/2$ whence $d(x,x_j) < \delta$ and hence $x \in B(x_j,\delta) \subseteq U_j$. Thus the sets $U_1, \ldots, U_n$ cover C. □

Having seen that sequentially compact and compact are equivalent we now relate compactness to a more concrete condition.

Theorem 2.8: Heine-Borel

Let (X,d) be a metric space. Then $C \subseteq X$ is compact if and only if C is closed, every Cauchy sequence in C converges, and C totally bounded.

Proof. We've already seen any compact C has to be closed and totally bounded. If $\{x_n\}$ is a Cauchy sequence in C, it has to have a convergent subsequence, and that in turn forces the whole sequence to converge to the same limit.

Conversely, suppose C is closed and totally bounded, and Cauchy sequences converge in C. Let $\{x_n\}$ be any sequence in C. It suffices to find a convergent subsequence. Inductively, we produce a nested sequence of balls $\mathscr{B}(c_n, 1/n)$, each containing infinitely many terms of the sequence $\{x_n\}$. Indeed, we begin with $n = 1$, and since C is totally bounded, we can cover C by finitely many radius 1 balls, and so one such ball, say $\mathscr{B}(c_1, 1)$, contains infinitely many terms of our sequence. Let $I_1 = \{n : x_n \in \mathscr{B}(c_1, 1)\}$. At stage j, we have some infinite set $I_j = \{n : x_n \in \mathscr{B}(c_j, 1/j)\}$. Cover C by balls of radius $1/(j+1)$. One ball must contain infinitely terms of the sequence with $n \in I_j$, say $\mathscr{B}(c_{j+1}, 1/(j+1))$, and let $I_{j+1} = \{n \in I_j : x_n \in \mathscr{B}(c_{j+1}, 1/(j+1))\}$. Next select an increasing sequence n_j with $n_j \in I_j$. The sequence $\{x_{n_j}\}$ is Cauchy, since if $j > J$, each x_{n_j} belongs to $\mathscr{B}(c_J, 1/J)$ and any two points from this ball are at most $2/J$ apart. The sequence $\{x_{n_j}\}$ converges to some limit $x \in C$ by completeness and the fact that C is closed. □

3 MEASURES

3.1 Counting measures and abstract measures

The counting measure on a set X is a function

$$m : \mathscr{P}(X) \to [0, \infty]$$

which maps $A \subseteq X$ to $|A|$, where $m(A) = \infty$ is allowed too. It does *not* distinguish between cardinalities of infinite sets. Thus, it is a primitive tool which allows us to describe the size of the set A. It's called the counting measure, because it counts the number of elements of A.

We would like to explore other notions of measures of a set, in particular ones which apply to subsets of $\mathbb{R}^n$. Before doing so, it's worth taking stock of some familiar properties of the counting measure.

(*Positivity*) If A is a set, $m(A) \geq 0$ and $m(A) = 0$ if and only if $A = \emptyset$.

(*Additivity*) If A and B are disjoint sets, then $m(A \cup B) = m(A) + m(B)$.

Our goal is to come up with a notion of measure which is a little more fine-tuned in its capacity to handle infinite sets. It turns out our list of desired properties will change a little.

Definition 3.1: Abstract measure

In general, we call a function m a measure on a certain collection of sets Σ (which is called a σ-algebra) if it has the following properties

Non-negativity: If $A \in \Sigma$ is a set, $m(A) \geq 0$ and $m(A) = 0$ if $A = \emptyset$.

σ-additivity: If $A_1, A_2, \ldots$ are pairwise disjoint sets from Σ, then $m(\bigcup_i A_i) = \sum_{i=1}^{\infty} m(A_i)$.

Let's examine the changes. First, we no-longer have positivity. This is because we may try to assign a finite measure to an infinite set, and this may force us to think of small sets as *very small*.

Exercise. *Let X be uncountable and suppose that m is an abstract measure on some collection Σ of sets such that $X \in \Sigma$ and every $x \in X$ satisfies $\{x\} \in \Sigma$. Then either $m(\{x\}) = 0$ for some $x \in X$ or else $m(X) = \infty$.*

Another change is that we have upgraded the additivity to handle finitely many pairwise disjoint sets to countably many. This, it turns out, is really a necessary change in order to get a robust notion of measure.

Finally, we have only defined the measure on a certain collection of sets Σ, which we called a σ-algebra. These sets are the sets we're allowed to measure, naturally called the *measurable sets*.

Definition 3.2: σ-algebra

A σ-algebra on a set X is a set $\Sigma \subseteq \mathscr{P}(X)$ with the following properties:

Empty set: $\emptyset \in \Sigma$,

Closure under complement: $A \in \Sigma \implies A^c \in \Sigma$, and

Closure under countable union: If $A_i \in \Sigma$ for $i \in \mathbb{N}$ then $\bigcup_{i \in \mathbb{N}} A_i \in \Sigma$.

We record, before moving on, that abstract measures are monotone, in the sense that the measure of larger sets is indeed larger.

Exercise. *If m is an abstract measure on Σ and $A, B \in \Sigma$ with $A \subseteq B$ then $m(A) \leq m(B)$.*

3.2 In search of Lebesgue measure

If we treat counting measure as an abstract measure on $\mathbb{R}$, then it will assign to each infinite set the value ∞. This is too narrow-minded to me able to say anything interesting.

The simplest subsets of $\mathbb{R}$ are the intervals, $[a,b]$, and we already have a notion of size for them: their length $l([a,b]) = b-a$. Notice that this is translation invariant in the sense that $l([a+t,b+t]) = l([a,b])$. So perhaps we might try to extend these properties to a measure on subsets of $\mathbb{R}$. That is to say, we might like that $m(A+t) = m(A)$ where

$$A+t = \{a+t : a \in A\}.$$

Unfortunately, this cannot work.

Theorem 3.1

There is no abstract measure m, defined on all subsets of $\mathbb{R}$ with the properties that $m([a,b]) = b-a$ and $m(A+t) = m(A)$ for all sets A and translates $t \in \mathbb{R}$.

Proof. Let $\mathcal{V} \subseteq (0,1)$ be a Vitali set, see 1.3. Let $Q = \mathbb{Q} \cap [-1,1]$. Now each $x \in [0,1]$ can be written as

$$x = q_x + v_x$$

with $q_x \in \mathbb{Q}$ and $v_x \in \mathcal{V}$. In fact

$$|q_x| = |x - v_x| \leq 1$$

so $q_x \in Q$. Thus $[0,1] \subseteq \bigcup_{q\in Q} q+V$. The set Q is countable, and the translates $q+V$ are pairwise disjoint:

$$q_1 + v_1 = q_2 + v_2 \implies v_1 - v_2 = q_2 - q_1 \in \mathbb{Q} \implies v_1 = v_2 \text{ and } q_1 = q_2$$

where we have used that $\mathcal{V}$ is a complete set of representatives. By translation invariance and σ-additivity

$$m\left(\bigcup_{q\in Q} q+V\right) = \sum_{q\in Q} m(q+V) = \sum_{q\in Q} m(V)$$

so either

$$m(V) = m\left(\bigcup_{q\in Q} q+V\right) = 0$$

or $m\left(\bigcup_{q\in Q} q+V\right) = \infty$. The former cannot occur since

$$m\left(\bigcup_{q\in Q} q+V\right) \geq m([0,1]) \geq 1.$$

Meanwhile

$$-1+0 \leq q+v \leq 1+1$$

so

$$\bigcup_{q\in Q} q+V \subseteq [-1,2] \implies m\left(\bigcup_{q\in Q} q+V\right) \leq m([-1,2]) = 3.$$

□

This would appear to dash our hopes of defining a nice measure which satisfies our list of desired properties. All is not lost however – the Vitali set is pretty bizarre, and we might just as well ignore it. This is why we define measures on σ-algebras: we can choose to measure only well-behaved sets.

We already know that we want to be able to measure intervals, and we want their measure to be their length. Actually, we have only stated as much for closed intervals, but the length of an interval does not depend on the inclusion of either endpoint.

Exercise. *Suppose Σ contains all intervals and m is a measure on Σ. Show that if $m(I) = l(I)$ for all bounded intervals I then $m(I) = \infty$ when I is unbounded.*

We might also hope to measure other nice sets, like open sets and closed sets. As concerns open sets, we have the following lemma.

Lemma 3.1: Lindelöf's Lemma

If $U \subseteq \mathbb{R}$ is open then U is a disjoint union of countably many open intervals.

Proof. Define an equivalence relation $\sim$ on U by $x \sim y \iff [\min\{x,y\}, \max\{x,y\}] \subseteq U$. The only tricky bit about showing this is an equivalence relation is transitivity. If $x \sim y$ and $y \sim z$ then we are to show that (assuming $x \leq z$, without loss of generality)

$$[x,z] \subseteq U.$$

If $y \leq x$ then $[x,z] \subseteq [y,z] \subseteq U$. If $x \leq y \leq z$ then $[x,z] = [x,y] \cup [y,z] \subseteq U$ and if $y \geq z$ then $[x,z] \subseteq [x,y] \subseteq U$.

Now we show that the equivalence class of x, $[x]$, is the open interval

$$[x] = (\inf[x], \sup[x]).$$

Indeed, let I be said interval. First neither endpoint belongs to $[x]$. This is obvious if either endpoint is infinite, since $[x] \subseteq \mathbb{R}$. If, say $l = \inf[x]$ is finite and belongs to $[x]$, then $l \in U$, and because U is open $[l-\varepsilon, l] \subseteq U$ for some $\varepsilon > 0$, whence $l - \varepsilon \equiv l \equiv x$, contradicting that l is a lower bound for $[x]$. Next, For $t \in I$, we can find $y, z \in [x]$

with $\inf[x] < y < t < z < \sup[x]$, and $y \sim z$ (they belong to the same equivalence class, namely $[x]$) so $t \in [y,z] \in U$. Conversely, we plainly have $[x] \subseteq [\inf[x], \sup[x]]$ and since neither endpoint belongs to $[x]$, we have that in fact $[x] \subseteq (\inf[x], \sup[x])$.

Now, since U is the disjoint union of equivalence classes, each of which is an open interval, it remains to show there are only countably many. But each such open interval contains a rational number, and these must be distinct as the intervals are disjoint. Since there are only countably many such rationals, there are only countably many such intervals. □

Thus to measure an open set U, we might write

$$U = \bigcup_{n=1}^{\infty} I_n$$

for some (possibly empty) open intervals I_n and it would then follow that

$$m(U) = \sum_{n=1}^{\infty} l(I_n),$$

using the σ-additive property of m.

If U is not open, it may very well not be a disjoint union of open intervals. However, we could try to fit it in an ever-so-slightly larger open set and approximate its measure by that of the open set.

Definition 3.3: Lebesgue Outer Measure on $\mathbb{R}$

Let $A \subseteq \mathbb{R}$. Then the Lebesgue outer measure, m_*, on $\mathbb{R}$ is the (possibly infinite) value

$$m_*(A) = \inf\left\{ \sum_{n=1}^{\infty} l(I_n) : A \subseteq \bigcup_{n=1}^{\infty} I_n, \text{ each } I_n \text{ a closed interval} \right\}.$$

We have now come up with a definition of measure that is so general it applies to all sets. This brings about two issues. First, we know something has to fail, because we already said there is no notion of measure with all our desired properties that can measure the Vitali set. Second, the definition of outer measure does not tell us right off the bat that the measure of an interval is its length. Fortunately, this is still the case.

Lemma 3.2

Let $[a,b]$ be a finite interval. Then $m_*([a,b]) = b - a$.

Proof. Since $[a,b] \subseteq [a,b] \cup [a,a] \cup [a,a] \cup \cdots$ and $l([a,a]) = 0$ we can take $I_1 = [a,b]$ and $I_n = [a,a]$ for $n \geq 2$ to cover $[a,b]$ by closed intervals such that

$$\sum_{n=1}^{\infty} l(I_n) = b - a.$$

Hence

$$m_*([a,b]) \leq b - a.$$

Now we show that $m_*([a,b]) \geq b - a - \varepsilon$ for each $\varepsilon > 0$, which will conclude the proof. To that end, let $\{I_n\}$ be a collection of closed intervals with the property that

$$[a,b] \subseteq \bigcup_n I_n, \ \sum_{n=1}^{\infty} l(I_n) \leq m_*([a,b]) + \varepsilon.$$

This is purely a consequence of the definition of m_*. Because $[a,b]$ is compact, we may pass to a finite subcollection of intervals covering $[a,b]$, and we may further assume it's minimal, in the sense that no interval in the subcollection can be removed without failing to cover $[a,b]$. Thus we have

$$[a,b] \subseteq I_1 \cup \ldots \cup I_r$$

and

$$\sum_{n=1}^{r} l(I_n) \leq m_*([a,b]) + \varepsilon.$$

Let's assume that the left endpoints of the $I_n = [a_n, b_n]$ are increasing so that $a_1 \leq \ldots \leq a_r$. If, for any n, $b_n < a_{n+1}$, we would have

$$[a,b] \subseteq [a_1,b_1] \cup \cdots [a_n,b_n] \cup [a_{n+1},b_{n+1}] \cup \cdots \cup [a_r,b_r] \subseteq (-\infty, b_n] \cup [a_{n+1}, \infty).$$

Because each interval I_n must intersect $[a,b]$ (if not, we could remove it and get a smaller covering), there is an element from $[a,b]$ in each of the above intervals on the right. But the right hand side is disconnected, and $[a,b]$ is connected. This tells us that

$$a_{n+1} \leq b_n$$

for $n \leq r-1$. Consequently

$$\sum_{n=1}^{r} l([a_n,b_n]) = \sum_{n=1}^{r} b_n - a_n \geq b_r - a_r + \sum_{n=1}^{r-1} a_{n+1} - a_n = b_r - a_1.$$

But a_1 is the left-most point in all of the I_n, so must be smaller than a. Meanwhile

$$b_r \geq a_r \geq b_{r-1} \geq a_{r-1} \geq b_{r-2} \geq \ldots$$

so that b_r is the right-most point. Hence $b \leq b_r$. □

Theorem 3.2

If A is countable then $m_*(A) = 0$. Consequently, and interval of positive length is uncountable.

Proof. Enumerate A, and around a_n, put an interval $I_n = [a_n - \varepsilon/2^n, a_n + \varepsilon/2^n]$. These intervals cover A and hence

$$m_*(A) \leq \sum_{n=1}^{\infty} l(I_n) = 2\varepsilon.$$

□

3.3 Extending to $\mathbb{R}^n$

The Lindelöf lemma is no longer true in $\mathbb{R}^n$, so there is no longer an obvious way to define the outer measure. There are a few ways to extend the notions of interval and length. The most convenient are using rectangles and volume.

Definition 3.4: Rectangles and Cubes

A closed rectangle R in $\mathbb{R}^n$ is a cartesian product of closed (possibly unbounded) intervals

$$R = [a_1, b_1] \times \cdots \times [a_n, b_n].$$

And open rectangle is the same but with open intervals. The volume of R is

$$v(R) = \prod_{j=1}^{n} (b_j - a_j).$$

When $b_j - a_j = b_i - a_i$ for all i, j, we call R a cube.

A convenient collection of cubes are the dyadic ones.

Definition 3.5: Dyadic points, intervals and cubes

A dyadic point in $\mathbb{R}$ is a rational of the form $j/2^k$ with $j, k \in \mathbb{Z}$. We call k the height, or scale, of the point. A (closed) dyadic interval is one of the form $[j/2^k, (j+1)/2^k]$. A (closed) dyadic cube is a product of dyadic intervals of a fixed length.

Lemma 3.3

There are countably many dyadic cubes. If Q_1 and Q_2 are dyadic cubes then either one contains the other, or else their interiors are disjoint.

Proof. A dyadic cube has the form

$$Q = [(j_1)/2^k, (j_1+1)/2^k] \times \cdots \times [(j_n)/2^k, (j_n+1)/2^k].$$

Since $j_1, \ldots, j_n, k \in \mathbb{Z}$, there are as many dyadic cubes as there are elements $(j_1, \ldots, j_n, k) \in \mathbb{Z}^{n+1}$, which is countable. We prove the second claim for intervals only, and leave the rest as an exercise.

Let $Q_1 = [j_1/2^{k_1}, (j_1+1)/2^{k_1}]$ and $Q_2 = [j_2/2^{k_2}, (j_2+1)/2^{k_2}]$. Assume, without loss of generality that $k_1 - l = k_2$ for some $l \geq 0$. Then Q_2 can be broken into dyadic subintervals

$$Q_2 = [2^l j_2/2^{k_1}, 2^l(j_2+1)/2^{k_1}] = \bigcup_{i=0}^{2^l-1} [(2^l j_2 + i)/2^{k_1}, (2^l j_2 + i + 1)/2^{k_1}],$$

and these intervals are in line, with two adjacent intervals intersecting at only the endpoints. If Q_1 is one of these, then $Q_1 \subseteq Q_2$. Otherwise, Q_1 is some other dyadic interval at scale k and can at most intersect one of the above intervals at an endpoint. □

Two cubes are said to be almost disjoint if their interiors are disjoint.

Lemma 3.4

If $U \subseteq \mathbb{R}^n$ is open then it is the almost-disjoint union of at most countably many dyadic cubes.

Proof. For $x \in U$ there is a ball $\mathscr{B}(x, \varepsilon_x)$ around x which is contained in U. For any k, there is a dyadic cube

$$Q_{j_1,\ldots,j_n,k} = \prod_{i=1}^{n} [j_i/2^k, (j_i+1)/2^k]$$

which contains x. The distance from x to any point in this cube is at most $\sqrt{n2^{-k}}$, and hence such a cube is contained in $B(x, \varepsilon)$ for k sufficiently large. We call a dyadic cube good if $Q \subseteq U$ and no larger dyadic cube is a subset Q. That is, if Q is a cube at scale k, the cube at scale $k-1$ which contains Q is not a subset of U. For $x \in U$, let $k(x)$ denote the smallest k such that $x \in Q_x \subseteq U$ for some Q_x at scale $k(x)$. Clearly Q_x is a good cube, and by the preceding lemma, good cubes are almost disjoint. Hence the good cubes satisfy the conditions of the lemma. □

Definition 3.6: Lebesgue outer measure on $\mathbb{R}^n$

Let $A \subseteq \mathbb{R}^n$. Then the Lebesgue outer measure, m_*, on $\mathbb{R}^n$ is the (possibly infinite) value

$$m_*(A) = \inf\left\{\sum_{j=1}^{\infty} V(R_j) : A \subseteq \bigcup_{j=1}^{\infty} Q_j, \text{ each } Q_j \text{ a closed cube}\right\}.$$

We have not insisted that the cubes are dyadic here, and it doesn't make too much difference. We'll see a few techniques which allow us to replace coverings by closed cubes by coverings with other objects. For instance, closed, bounded rectangles will do.

Lemma 3.5

For any set A, and let

$$m'_*(A) = \inf\left\{\sum_{j=1}^{\infty} V(R_j) : A \subseteq \bigcup_{j=1}^{\infty} R_j, \text{ each } R_j \text{ a closed, bounded rectangle}\right\}.$$

Then

$$m_*(A) = m'_*(A).$$

Proof. Let $A \subseteq \bigcup Q_j$ be a covering of A by closed cubes. Since cubes are themselves rectangles,

$$m'_*(A) \leq \sum_j V(Q_j),$$

and hence

$$m'_*(A) \leq m_*(A).$$

Conversely, if $R = I_1 \times \cdots \times I_n$ is a rectangle, then we can cover each I_i with a union of almost disjoint intervals of length $\delta \min_j\{l(I_j)\}$, say $I_{i,k}$ where $\sum_k l(I_{i,k}) \leq l(I_i)(1+\delta)$. Hence the cubes $I_{1,k_1} \times \cdots \times I_{n,k_n}$ cover R_j, and

$$\sum_{k_1,\ldots,k_n} V(I_{1,k_1} \times \cdots \times I_{n,k_n}) = \prod_{i=1}^{n}\left(\sum_k l(I_{i,k})\right) \leq \prod_{i=1}^{n}(1+\delta)l(I_i) = (1+\delta)^n V(R).$$

Now cover A with rectangles R_j, with

$$\sum_j V(R_j) \leq m'_*(A) + \varepsilon$$

Do this for each j, taking δ_n so small that $(1+\delta_j)^n V(R_j) \leq V(R_j) + \varepsilon/2^j$. Then the total volume of all the cubes involved is at most

$$\sum_j V(R_j) + \varepsilon/2^j \leq m'_*(A) + 2\varepsilon,$$

whence $m_*(A) \le m'_*(A) + 2\varepsilon$, and we get the claimed result by letting $\varepsilon \to \infty$. □

We next extend Lemma 3.2 to $\mathbb{R}^n$.

Lemma 3.6

Let Q be a closed and bounded cube in $\mathbb{R}^n$. Then

$$m_*(Q) = V(Q).$$

Proof. It's immediate from the definition of outer measure that $m_*(Q) \le V(Q)$, since Q is itself a covering of Q. Now, we show $V(Q) \le m_*(Q) + \varepsilon$ for $\varepsilon > 0$.

Let $\bigcup_j Q_j$ cover Q and be such that $\sum_j V(Q_j) < m_*(Q) + \varepsilon/2$. We can expand Q_j to an open cube Q'_j containing Q_j with volume at most $V(Q'_j) + \varepsilon/2^{j+1}$. The cubes Q'_j cover Q too, and their total volume is at most $m_*(Q) + \varepsilon$. Because these cubes are open and Q is compact, we need only finitely many to cover Q, say $Q'_1, \ldots, Q'_N$. Finally let $R_j = Q'_j \cap Q$, which is a rectangle. Then $\bigcup_j R_j = Q$ and

$$\sum_j V(R_j) \le m*(Q) + \varepsilon.$$

Now let $\boldsymbol{a} = (a_1, \ldots, a_n) \in Q = I_1 \times \cdots \times I_n$. Then $\boldsymbol{a} \in R_j = I_1^j \times \cdots \times I_n^j$ for some j, and from this $a_i \in I_i^j$. This tells us that the intervals I_i^j with $j = 1, \ldots, N$ cover the interval I_j. Finally, let X_i denote the endpoints of the intervals I_i^j, and if $X_i = \{t_{i,1} < \ldots < t_{i,M_i}\}$ then $I_i \subseteq \bigcup_{k=1}^{M_i - 1} [t_{i,k}, t_{i,k+1}]$. From the one dimensional case, Lemma 3.2, we have

$$\sum_{k=1}^{M_i} t_{i,k+1} - t_{i,k} \ge l(I_i).$$

The rectangles $S_{k_1,\ldots,k_n} = \prod_{i=1}^n [t_{i,k_i}, t_{i,k_i+1}]$ cover Q, are almost disjoint, and each is contained in some R_j. Thus

$$\begin{aligned}
\sum_j V(R_j) &\ge \sum_{k_1,\ldots,k_n} V(S_{k_1,\ldots,k_n}) \\
&= \sum_{k_1,\ldots,k_n} (t_{1,k_1+1} - t_{1,k_1}) \cdots (t_{n,k_n+1} - t_{n,k_n}) \\
&= \prod_{i=1}^n (\sum_{k=1}^{M_i} t_{i,k+1} - t_{i,k}) \\
&\ge \prod_{i=1}^n l(I_j) = V(Q).
\end{aligned}$$

□

3.4 Properties of outer measure

Lemma 3.7: Monotonicity and subadditivity

The outer measure is increasing: if $A \subseteq B$ then $m_*(A) \leq m_*(B)$. If $A_1, A_2, \ldots$ is any collection of sets then $m_*(\bigcup_n A_n) \leq \sum_n m_*(A_n)$.

Proof. Let $\{Q_n\}$ be a countable collection of closed cubes covering B. Then it covers A as well, so $m_*(A) \leq \sum_n V(Q_n)$. It follows that $m_*(A) \leq m_*(B)$.

For the second claim, cover A_n by a family $\{Q_{n,j}\}_{j\in\mathbb{N}}$ of cubes subject to the constraint

$$m_*(A_n) \leq \sum_j V(Q_{n,j}) \leq m_*(A_n) + \varepsilon/2^n.$$

The family of all such cubes, $\{Q_{n,j}\}_{n,j\in\mathbb{N}}$ is still countable, covers the union of the A_n, and we have

$$\sum_{n,j} V(Q_{n,j}) \leq \sum_n \left(m_*(A_n) + \varepsilon 2^{-n}\right) \leq \left(\sum_n m_*(A_n)\right) + \varepsilon.$$

□

Corollary 3.1

We have

$$m_*(A) = \inf\{m_*(U) : A \subseteq U,\ U \text{ open}\}.$$

Proof. By the preceding lemma, for $A \subseteq U$, we have $m_*(A) \leq m_*(U)$ and hence

$$m_*(A) \leq \inf\{m_*(U) : A \subseteq U,\ U \text{ open}\}.$$

Conversely, given a covering of A by closed cubes $\{Q_n\}$, we can replace each Q_n with a larger open cube Q_n' of volume $V(Q_n') = V(Q_n) + \varepsilon/2^n$ by enlarging the sides of Q_n slightly. The open cubes Q_n' still cover A and their union. Furthermore, since Q_n' is covered by $\overline{Q_n'}$, we have $m_*(Q_n') \leq V(Q_n')$. Thus, by subadditivity, if $U = \bigcup_n Q_n'$ then

$$m_*(U) \leq \sum_n m_*(Q_n') \leq \sum_n V(Q_n') \leq \sum_n \left(V(Q_n) + \varepsilon 2^{-n}\right) \leq \left(\sum_n V(Q_n)\right) + \varepsilon.$$

If we choose the Q_n subject to $\sum_n V(Q_n) \leq m_*(A) + \varepsilon$, we have

$$m_*(U) \leq m_*(A) + 2\varepsilon.$$

Hence for any $\varepsilon > 0$

$$\inf\{m_*(U) : A \subseteq U,\ U \text{ open}\} \leq m_*(A) + 2\varepsilon,$$

which implies

$$\inf\{m_*(U) : A \subseteq U,\ U \text{ open}\} \leq m_*(A).$$

□

We would like to upgrade countable subadditivity from Lemma 3.4 to countable additivity when the sets are disjoint. This is a little bit tricky. For now, we'll settle for a slightly weaker result.

Lemma 3.8

Suppose

$$d(A,B) = \inf\{\|a-b\|_{l^2} : a \in A, b \in B\} > 0.$$

Then $m_*(A \cup B) = m_*(A) + m_*(B)$.

Proof. By subadditivity, we have

$$m_*(A \cup B) \leq m_*(A) + m_*(B).$$

Let $\delta = \inf\{\|a-b\|_{l^2} : a \in A, b \in B\}$ and let $\{Q_n\}$ be a collection of closed cubes covering $A \cup B$. Let l_n denote the sidelength of Q_n. By iteratively halving the sides of Q_n, we can replace Q_n with an almost disjoint union of cubes $Q_{n,j}$ with sidelengths $l_n/2^k$ for any $k \geq 0$, and total volume $V(Q_n)$. The reason for doing so is that if $x, y \in Q_{n,j}$ then (if we are working in $\mathbb{R}^d$)

$$\|x-y\|_{l^2} = \left(\sum_{k=1}^{d} (x_i - y_i)^2\right)^{1/2} \leq \left(d\, l_n 2^{-k}\right)^{1/2}$$

and this is smaller that δ for k sufficiently large. Thus each $Q_{n,j}$ intersects at most one of A and B. It follows that the set of all cubes $\{Q_{n,j}\}$ can be partitioned into two sets, $\mathcal{Q}_1$ and $\mathcal{Q}_2$, covering A and B respectively, and the total volume is still $\sum_n V(Q_n)$, so

$$m_*(A) + m_*(B) \leq \sum_{Q \in \mathcal{Q}_1} V(Q) + \sum_{Q \in \mathcal{Q}_2} V(Q) = \sum_n V(Q_n).$$

If the $\{Q_n\}$ are chosen to have total volume at most $m_*(A) + \varepsilon$, we have thus shown

$$m_*(A) + m_*(B) \leq m_*(A \cup B) + \varepsilon.$$

The result follows upon letting $\varepsilon \to 0$. □

Lemma 3.9

Let $\{Q_j\}$ be a countable collection of almost-disjoint cubes of finite volume. Then

$$m_*\left(\bigcup Q_j\right) = \sum_j V(Q_j).$$

Proof. That

$$m_*(\bigcup Q_j) \leq \sum_j V(Q_j)$$

follows from subadditivity and Lemma 3.3. Conversely, suppose $Q = [a_1, b_1] \times \cdots \times [a_n, b_n]$. Then $Q' = [a_1 + \delta, b_1 - \delta] \times \cdots \times [a_n + \delta, b_n - \delta]$ is a cube with volume tending to $V(Q)$ as $\delta \to 0$. We can choose δ small enough so as to have

$$V(Q') > V(Q) - \varepsilon.$$

Apply this procedure for each j to get a cube $Q'_j \subseteq Q_j$ with $V(Q'_j) > V(Q_j) - \varepsilon/2^j$. Then

$$\sum_{j=1}^{N} V(Q'_j) > \sum_{j=1}^{N} V(Q_j) - \varepsilon/2^j.$$

On the other hand the cubes $Q'_1, \ldots, Q'_N$ all have positive distance between them because the Q_j's are almost disjoint. Applying Lemma 3.4 iteratively,

$$\sum_{j=1}^{N} V(Q'_j) = \sum_{j=1}^{N} m_*(Q'_j) = m_*(\bigcup_{j=1}^{N} Q'_j) \leq m_*(\bigcup_j Q_j).$$

Taking $N \to \infty$, we get

$$\sum_j V(Q_j) - \varepsilon \leq m_*(\bigcup_j Q_j).$$

□

3.5 Measurability

We now turn our attention to finding suitably nice sets, such that the restriction of m_* to these sets *is* countably additive. The existence of the Vitali set necessitates this endeavour.

Definition 3.7: Lebesgue's Measurability Criterion

A set $E \subseteq \mathbb{R}^n$ is called Lebesgue measurable if for every $\varepsilon > 0$ there is an open set U such that $E \subseteq U$ and $m_*(U \setminus E) < \varepsilon$.

Note that by Corollary 3.4, we can always find an open set U which contains E and satisfies $m_*(U) \leq m_*(E) + \varepsilon$. Of course we have $U = E \sqcup (U \setminus E)$, but we don't know that the measure is additive for disjoint unions. In this way, Lebesgue's criterion *forces the issue*. Another way to force the issue is called the Carathéodory Criterion.

Definition 3.8: Carathéodory's Measurability Criterion

A set $E \subseteq \mathbb{R}^n$ is called Carathéodory measurable if for every set $A \subseteq \mathbb{R}^n$,

$$m_*(A) = m_*(A \cap E) + m_*(A \cap E^c).$$

This notion of measurability also somehow forces disjoint sets to have their measures add. However, it's the measure of the arbitrary set A which is being broken down. The measurable set E is the set *doing the slicing*. In this sense, we are defining E to be measurable if it has a nice enough boundary so as to cleanly cut A up. The benefit of this criterion is that it makes no reference to topology, and is more algebraic. It works in more abstract measure theory settings. Fortunately, Lebesgue's criterion is the same.

Theorem 3.3

The Lebesgue measurable sets and the Carathéodory measurable sets are the same.

This theorem will take some working up to. We'll start by exploring some easy consequences of each definition. We begin with an obvious one.

Lemma 3.10

Open sets are Lebesgue measurable.

Building up from this, we have:

Lemma 3.11

Countable unions of Lebesgue measurable sets are measurable.

Proof. Let $E_1, E_2, \ldots$, be Lebesgue measurable and let U_j be the promised open sets with $E_j \subseteq U_j$ and $m_*(U_j \setminus E_j) \leq \varepsilon/2^j$. Let $U = \bigcup U_j$. Then U contains $\bigcup E_j$ and if $x \in U \setminus \bigcup E_j$ then $x \in U_j$ but $x \notin E_j$, so $x \in \bigcup(U_j \setminus E_j)$. Hence, by subadditivity

$$m_*(U \setminus \bigcup E_j) \leq \sum_j \varepsilon/2^j = \varepsilon.$$

□

Lemma 3.12

If $m_*(E) = 0$ then E is measurable.

Proof. We know that for any $\varepsilon > 0$ there is an open set $U \supset E$ with $m_*(U) < \varepsilon$. But $m_*(U \setminus E) \leq m_*(U)$. □

Lemma 3.13

Let X be a metric space and let C_1 and C_2 be two compact, disjoint subsets of X. Then

$$\inf\{d_X(x_1, x_2) : x_1 \in C_1,\ x_2 \in C_2\} > 0.$$

Proof. The set $X \times X$ with

$$d((x_1, x_2), (y_1, y_2)) = d_X(x_1, y_1) + d_X(x_2, y_2)$$

is a metric space, as the reader can, and should, verify.

Inside this space, the set $C_1 \times C_2$ is compact (we can find a convergent subsequence by asking the coordinates to converge one at a time), and the function $f : X \times X \to \mathbb{R}$ given by $f((x, y)) = d_X(x, y)$ is continuous. Indeed, for a given $(x_1, x_2) \in X \times X$, if $(y_1, y_2) \in X \times X$ then

$$d_X(x_1, x_2) \leq d_X(x_1, y_1) + d_X(y_1, y_2) + d_X(y_2, x_2),$$

and similarly

$$d_X(y_1, y_2) \leq d_X(x_1, y_1) + d_X(x_1, x_2) + d_X(y_2, x_2),$$

so

$$|d_X(x_1, x_2) - d_X(y_1, y_2)| \leq d_X(x_1, y_1) + d_X(x_2, y_2) = d((x_1, x_2), (y_1, y_2)).$$

From this, $|d_X(x_1, x_2) - d_X(y_1, y_2)| < \varepsilon$ provided $d((x_1, x_2), (y_1, y_2)) < \varepsilon$.

Thus the function f achieves a minimum on $C_1 \times C_2$, say

$$d_X(x_0, y_0) = \inf\{d_X(x_1, x_2) : x_1 \in C_1,\ x_2 \in C_2\}.$$

Since $x_0 \in C_1$, $y_0 \in C_2$ and $C_1 \cap C_2 = \emptyset$, we see $d_X(x_0, y_0) > 0$. □

Lemma 3.14

Closed sets are measurable.

Proof. We first show compact sets are measurable. The result will follow since if F is closed in $\mathbb{R}^n$ then for $N \in \mathbb{N}$, $F_N = F \cap [-N, N]^n$ is compact and $F = \bigcup_N F_N$ is then measurable by Lemma 3.5.

If C is a compact set, necessarily bounded, we may assume that $C \subseteq (-N, N)^n$ and let U be an open set containing C with $m_*(U) \leq m_*(C) + \varepsilon$. If need be, we can replace U with $U \cap (-N, N)^n$, which still covers C, is open, and has possibly smaller

measure. Thus we are free to assume $U \subseteq (-N,N)^n$ as well. Let $U' = U \setminus C = U \cap C^c$, which is open as well, since C is closed. The set U' is thus an almost disjoint union of cubes with finite volume closed : $U' = \bigcup_j Q_j$, and $m_*(U') = \sum_j V(Q_j)$ by Lemma 3.4. Thus

$$m_*(U') \geq \sum_{j \leq J} V(Q_j).$$

However, the set $C_J = Q_1 \cup \ldots \cup Q_J$ is compact (each cube is compact) and disjoint from C, whence there positive distance between C and C_J so that by Lemma 3.4,

$$m_*(U) \geq m_*(C \cup C_J) = m_*(C) + m_*(C_J) \geq m_*(U) - \varepsilon + m_*(C_J) = m_*(U) - \varepsilon + \sum_{j \leq J} V(Q_j),$$

which rearranges to

$$\sum_{j \leq J} V(Q_j) \leq \varepsilon.$$

Letting $J \to \infty$,

$$m_*(U') \leq \sum_j V(Q_j) \leq \varepsilon,$$

showing that C is measurable. □

Lemma 3.15

Let E be a Lebesgue measurable subset of $\mathbb{R}^n$. Then E^c is Lebesgue measurable.

Proof. Let $U_N \supset E$ be such that $m_*(U_N \setminus E) \leq 1/N$. Then U_N^c is a closed subset of E^c. Because the U_N^c are closed, they are measurable, and so $S = \bigcup_{N \in \mathbb{N}} U_N^c$ is measurable too, and still a subset of E^c. Now

$$E^c \setminus S = E^c \cap S^c = E^c \cap \bigcap_N U_N \subseteq U_M \setminus E$$

for any fixed M. Thus $m_*(E^c \setminus S) \leq 1/M$ for each M and hence $m_*(E^c \setminus S) = 0$ making $E^c \setminus S$ measurable. Thus $E^c = S \cup (E^c \setminus S)$ is measurable. □

Combining Lemma 3.5, Lemma 3.5, and Lemma 3.5, we have proved the following.

Theorem 3.4

The Lebesgue measurable sets form a Σ-algebra.

Next we will show that on the Σ-algebra of measurable sets, m_* is countably additive.

Lemma 3.16

For any integers $m_1, \ldots, m_n \in \mathbb{Z}$, the cubes

$$[m_1, m_1+1) \times \cdots \times [m_n, m_n+1)$$

partition $\mathbb{R}^n$ and are measurable.

Proof. That the cubes partition $\mathbb{R}^n$ is left as an exercise. Each cube can be written as

$$[m_1, m_1+1) \times \cdots \times [m_n, m_n+1) = \bigcup_{k\in\mathbb{N}} [m_1, m_1+1-1/k] \times \cdots \times [m_n, m_n+1-1/k]$$

which are unions of closed, and hence measurable, sets. □

Lemma 3.17

Let E be Lebesgue measurable with finite outer measure. Then, for $\varepsilon > 0$, there is a compact set $C \subseteq E$ with $m_*(E \setminus C) < \varepsilon$.

Proof. Let U be an open set containing E^c with $m_*(U \cap E) = m_*(U \setminus E^c) < \varepsilon/4$. Then $F = U^c$ is a closed subset of E with $m_*(E \setminus F) < \varepsilon/4$. Next let U' be an open set containing E with $m_*(U') < m_*(E) + \varepsilon/4$. We can write U' as an almost disjoint union of closed and bounded cubes, $U' = \bigcup_j Q_j$ and we know, from Lemma 3.4 that

$$m_*(U') = \sum_j V(Q_j).$$

Write $U_1' = \bigcup_{j\le J} Q_j$ and $U_2' = \bigcup_{j>J} Q_j$ so $U' = U_1' \cup U_2'$, and

$$m_*(U') = \sum_j V(Q_j) = \sum_{j\le J} V(Q_j) + \sum_{j>J} V(Q_j) = m_*(U_1') + m_*(U_2').$$

Suppose that J is so large that $\sum_{j\le J} V(Q_j) > m_*(E) - \varepsilon/4$, which means

$$m_*(U_2') = \sum_{j>J} V(Q_j) < m_*(U') - m_*(U_1') < (m_*(E) + \varepsilon/4) - (m_*(E) - \varepsilon/4) = \varepsilon/2.$$

Now let

$$C = \bigcup_{j\le J} Q_j \cap F = F \cap U_1'.$$

The sets $Q_j \cap F$ are compact since the sets Q_j are, and F is closed, and since C is a finite union, we know C is compact too. Since F is a subset of E, so is C. Finally,

$$E \setminus C \subseteq U' \setminus C = U' \cap (U_1' \cap F)^c = U' \cap (U_1'^c \cup F^c) = (U' \cap U_1'^c) \cup (U' \setminus F).$$

Now,

$$m_*(U' \cap U_1'^c) \le m_*(U_2') < \varepsilon/2$$

while

$$m_*(U' \setminus F) \leq m_*((U' \setminus E) \cup (E \setminus F)) \leq \varepsilon/2.$$

□

Theorem 3.5

Let $E_1, E_2, \ldots$ be Lebesgue measurable sets which are pairwise disjoint. Then

$$m_*\left(\bigcup_j E_j\right) = \sum_j m_*(E_j).$$

Proof. First, we may assume each E_j is bounded. Indeed, otherwise we can write (for $m_1, \ldots, m_n \in \mathbb{Z}$)

$$E_{j,(m_1,\ldots,m_n)} = E_j \cap [m_1, m_1+1) \times \cdots \times [m_n, m_n+1),$$

and partition

$$E_j = \bigcup_{m_1,\ldots,m_n} E_{j,(m_1,\ldots,m_n)}$$

into bounded, measurable sets.

Now, since each E_j is assumed to be bounded, we can find compact sets $C_j \subseteq E_j$ with $m_*(C_j) > m_*(E_j) - \varepsilon > 2^j$. The sets C_j are disjoint and compact, so by monotonicity, Lemma 3.5 and Lemma 3.4, we have

$$m_*\left(\bigcup_j E_j\right) \geq m_*\left(\bigcup_{j \leq J} C_j\right) = \sum_{j \leq J} m_*(C_j).$$

Letting $J \to \infty$, we get

$$m_*\left(\bigcup_j E_j\right) \geq m_*\left(\bigcup_{j \leq J} C_j\right) = \sum_j m_*(C_j) \geq \sum_j \left(m_*(E_j) - \varepsilon/2^j\right) = \sum_j m_*(E_j) - \varepsilon.$$

From this

$$m_*\left(\bigcup_j E_j\right) \geq \sum_j m_*(E_j).$$

The reverse inequality follows from subadditivity. □

Corollary 3.2

The function m_* restricted to the Lebesgue measurable sets is a measure in the sense of Definition 3.1.

When we restrict m_* to the Lebesgue measurable sets we call it the *Lebesgue measure*, and denote it by m rather than m_*.

One particularly nice property of m is that it obeys are certain type of continuity.

Theorem 3.6

Let $E_1, E_2, \ldots$ be a sequence of measurable sets.

1. If the sequence is increasing, $E_1 \subseteq E_2 \subseteq \ldots$ then $m\left(\bigcup_j E_j\right) = \lim_{j\to\infty} m(E_j)$.
2. If the sequence is decreasing, $E_1 \supseteq E_2 \supseteq \ldots$, and if $m(E_1) < \infty$ then $m\left(\bigcap_j E_j\right) = \lim_{j\to\infty} m(E_j)$.

Proof. For (1), we let

$$E'_j = E_j \setminus E_{j-1}.$$

Then the sets E'_j are subsets of E_j but now are disjoint since if $k > j$, we have $E'_k \cap E_j = \emptyset$. Moreover, E'_j is measurable, since it is obtained from measurable sets using the operations permitted by a σ-algebra, and

$$\bigcup_{j=1}^{J} E'_j = \bigcup_{j=1}^{J} E_j,$$

for all J, including $J = \infty$. So

$$m\left(\bigcup_{j=1}^{\infty} E_j\right) = m\left(\bigcup_{j=1}^{\infty} E'_j\right) = \sum_{j=1}^{\infty} m\left(E'_j\right) = \lim_{J\to\infty} \sum_{j\le J} m\left(E'_j\right) = \lim_{J\to\infty} m\left(\bigcup_{j\le J} E'_j\right) = \lim_{J\to\infty} m\left(E_J\right).$$

For (2), if $j \ge 2$ we set $E'_j = E_1 \setminus E_j$. Since E_1 has finite measure, so do E_j and E'_j and

$$m(E_1) = m(E_j) + m(E'_j) \implies m(E_j) = m(E_1) - m(E'_j).$$

The sets E'_j are increasing since the E_j are decreasing. By (1),

$$\lim_{j\to\infty} m(E_1) - m(E'_j) = m(E_1) - \lim_{j\to\infty} m(E'_j) = m(E_1) - m\left(\bigcup_{j=2}^{\infty} E'_j\right).$$

But

$$\bigcup_{j=2}^{\infty} E'_j = \bigcup_{j=2}^{\infty} E_1 \cap E_j^c = E_1 \cap \left(\bigcup_{j=2}^{\infty} E_j^c\right) = E_1 \cap \left(\bigcap_{j=2}^{\infty} E_j\right)^c$$

so

$$m(E_1) = m\left(E_1 \cap \left(\bigcap_{j=2}^{\infty} E_j\right)^c\right) + m\left(\bigcap_{j=1}^{\infty} E_j\right) = m\left(\bigcup_{j=2}^{\infty} E'_j\right) + m\left(\bigcap_{j=1}^{\infty} E_j\right)$$

and the theorem follows. □

We conclude this section with a proof of Theorem 3.5. We'll need one final lemma to do it.

Lemma 3.18

If E_1 and E_2 are both Carathéodory measurable, then so is $E_1 \cup E_2$. Consequently, $E_1 \cap E_2$ is Carathéodory measurable as well.

Proof. Let A be arbitrary. Then, applyling Carathéodory's criterion twice,

$$\begin{aligned} m_*(A) &= m_*(A \cap E_1) + m_*(A \cap E_1^c) \\ &= m_*(A \cap E_1 \cap E_2) + m_*(A \cap E_1 \cap E_2^c) + m_*(A \cap E_1^c \cap E_2) + m_*(A \cap E_1^c \cap E_2^c). \end{aligned}$$

Our aim is to show that

$$m_*(A) = m_*(A \cap (E_1 \cup E_2)) + m_*(A \cap (E_1 \cup E_2)^c) = m_*(A \cap (E_1 \cup E_2)) + m_*(A \cap E_1^c \cap E_2^c),$$

so we just need to show

$$m_*(A \cap (E_1 \cup E_2)) = m_*(A \cap E_1 \cap E_2) + m_*(A \cap E_1 \cap E_2^c) + m_*(A \cap E_1^c \cap E_2).$$

But, applying Carathéodory's criterion two more times,

$$\begin{aligned} m_*(A \cap (E_1 \cup E_2)) &= m_*(A \cap (E_1 \cup E_2) \cap E_1) + m_*(A \cap (E_1 \cup E_2) \cap E_1^c) \\ &= m_*(A \cap E_1) + m_*(A \cap E_2 \cap E_1^c) \\ &= m_*(A \cap E_1 \cap E_2) + m_*(A \cap E_1 \cap E_2^c) + m_*(A \cap E_2 \cap E_1^c). \end{aligned}$$

Since Carathéodory's criterion is symmetric in E and E^c, we know E_1^c, E_2^c, $E_1^c \cup E_2^c = (E_1 \cap E_2)^c$ and hence $E_1 \cap E_2$ are all Carathéodory measurable as well. □

Proof of Theorem 3.5. We first show that Lebesgue measurability implies Carathéodory measurability. To do so it suffices to show that for $\varepsilon > 0$ and any set $A \subseteq \mathbb{R}^n$, we have $m_*(A) + \varepsilon \geq m_*(A \cap E) + m_*(A \cap E^c)$. This will prove that $m_*(A) \geq m_*(A \cap E) + m_*(A \cap E^c)$, while subadditivity shows that $m_*(A) \leq m_*(A \cap E) + m_*(A \cap E^c)$.

So let E be Lebesgue measurable and let A be an arbitrary subset of $\mathbb{R}^n$. Let U be an open set containing A which has measure at most $m_*(A) + \varepsilon$. Then $U \cap E$ and $U \cap E^c$ are both Lebesgue measurable and disjoint, so $m_*(U \cap E) + m_*(U \cap E^c) = m_*(U) \leq m_*(A) + \varepsilon$.

Next suppose that E is Carathéodory measurable. Since $[-N, N]^d$ is Lebesgue measurable, it is Carathéodory measurable too, and hence $E_N = E \cap [-N, N]^d$ is Carathéodory measurable. But E_N has finite measure, if U is an open set containing E_N with $m_*(U) \leq m_*(E_N) + \varepsilon$, we have

$$\varepsilon + m_*(E_N) \geq m_*(U) = m_*(E_N) + m_*(U \setminus E_N) \implies m_*(U \setminus E_N) \leq \varepsilon,$$

since we can subtract $m_*(E_N)$ from both sides (using that it is finite). So E_N is Lebesgue measurable, and hence so is $E = \bigcup_N E_N$. □

3.6 Measurable Functions

Given a set X, we can think of a σ-algebra on X as defining some "measurable" sets, even if we don't have a measure defined on them yet. So if X and Y are sets with respective σ-algebras Σ_X and Σ_Y, we think of $f : X \to Y$ as being well-behaved with respect to these σ-algebras if $f^{-1}(E) \in \Sigma_X$ whenever $E \in \Sigma_Y$, which is reminiscent of the definition of continuity.

Definition 3.9: Measurable Function

Let X and Y be sets with respective σ-algebras Σ_X and Σ_Y. A function $f : X \to Y$ is called $(\Sigma_X \to \Sigma_Y)$-measurable (or just measurable, if the context is clear) if $f^{-1}(E) \in \Sigma_X$ whenever $E \in \Sigma_Y$.

Lemma 3.19

Let X and Y be sets with respective σ-algebras Σ_X and Σ_Y. Suppose Y has a second σ-algebra Σ'_Y which satisfies $\Sigma'_Y \subseteq \Sigma_Y$. Then if $f : X \to Y$ is $(\Sigma_X \to \Sigma_Y)$-measurable, it is also $(\Sigma_X \to \Sigma'_Y)$-measurable.

In our context, we have the Lebesgue measurable sets defined on $\mathbb{R}^d$ and we want to understand measurable functions $f : \mathbb{R}^n \to \mathbb{R}$. However, if we endow $\mathbb{R}$ with the σ-algebra of Lebesgue measurable sets, we will not recover enough measurable functions, and will even miss out on some continuous functions. There is a smaller σ-algebra on $\mathbb{R}$, called the *Borel* algebra, which is generated by the open sets.

Lemma 3.20: σ-algebra generated by $\mathscr{X}$

Let X be a non-empty set and suppose $\mathscr{X} \subseteq \mathscr{P}(X)$ is some collection of subsets. There is a unique minimal σ-algebra on X containing all the sets from $\mathscr{X}$ called the σ-algebra generated by $\mathscr{X}$.

Proof. Given any collection $\mathscr{C}$ of σ-algebras, $\bigcap_{\Sigma \in \mathscr{C}} \Sigma$ is also a σ-algebra. There is at least one σ-algebra containing $\mathscr{X}$, namely all of $\mathscr{P}(X)$. Then let

$$\mathscr{C} = \{\Sigma : \Sigma \text{ is a } \sigma\text{-algebra on } X \text{ containing } \mathscr{X}\},$$

and set

$$\Sigma_{\mathscr{X}} = \bigcap_{\Sigma \in \mathscr{C}} \Sigma,$$

which is the minimal σ-algebra containing $\mathscr{X}$. □

We want to consider functions which are Lebesgue measurable→Borel-measurable, and it so happens that it is enough to check this on intervals, which generate open sets and hence Borel sets. We will also want our functions to be allowed to take on the value ∞.

Definition 3.10: Extended real-valued measurable function

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a function. We say f is extended real-valued measurable if for any $a \in \mathbb{R} \cup \{\infty\}$, the set $f^{-1}((-\infty, a])$ is Lebesgue-measurable.

Lemma 3.21

An indicator function $\mathbf{1}_E$ is measurable if and only if the set $E \in \mathbb{R}^d$ is measurable.

Proof. Indeed $\mathbf{1}_E^{-1}((-\infty, a])$ is one of $\emptyset$, E^c, or $\mathbb{R}^d$, according to whether $a < 0$, $0 \le a < 1$ or $a \ge 1$. Thus $\mathbf{1}_E$ is measurable if and only if E^c is measurable which is equivalent to E being measurable. □

Lemma 3.22

The following are equivalent:

1. $f^{-1}((-\infty, a])$ is measurable for all $a \in \mathbb{R}$,
2. $f^{-1}((-\infty, a))$ is measurable for all $a \in \mathbb{R}$,
3. $f^{-1}((a, b))$ is measurable for all $a, b \in \mathbb{R}$,
4. $f^{-1}([a, b))$ is measurable for all $a, b \in \mathbb{R}$.

Proof. Assume (1), then

$$f^{-1}((-\infty, a)) = \bigcup_{n \in \mathbb{N}} f^{-1}((-\infty, a - 1/n])$$

which is a union of measurable sets, so we conclude (2). If (2) holds then

$$f^{-1}((a, b)) = \bigcup_{n \in \mathbb{N}} f^{-1}((-\infty, b)) \setminus f^{-1}((-\infty, a + 1/n)),$$

which proves (3). If (3) holds,

$$f^{-1}([a, b)) = \bigcap_{n \in \mathbb{N}} f^{-1}((a - 1/n, b)),$$

proving (4). Given (4),

$$f^{-1}((-\infty, a)) = \bigcup_{n \in \mathbb{N}} f^{-1}([-n, a)),$$

and we conclude (1). □

Lemma 3.23

If f is measurable and $c \in \mathbb{R}$ then cf is measurable.

Proof. If $c = 0$ then $f = 0 = \mathbf{1}_\emptyset$ which is measurable. If $c > 0$, for any a,

$$(cf)^{-1}((-\infty, a]) = \{x : cf(x) \le a\} = \{x : f(x) \le a/c\} = f^{-1}((-\infty, a/c]).$$

If $c < 0$

$$(cf)^{-1}((-\infty, a]) = \{x : cf(x) \le a\} = \{x : f(x) \ge a/c\} = f^{-1}((-\infty, a/c))^c.$$

□

Lemma 3.24

If f and g are measurable then so is $f + g$.

Proof. Observe that

$$f(x) + g(x) < a \iff f(x) < q \text{ and } g(x) < a - q \text{ for some } q \in \mathbb{Q}.$$

Indeed, given such a q,

$$f(x) + g(x) < q + (a - q) = a$$

and conversely, if $f(x) + g(x) < a$ let $q \in (f(x), a - g(x))$ and then

$$f(x) < q,\ g(x) < a - q.$$

It follows that

$$\begin{aligned}(f+g)^{-1}((-\infty, a)) &= \{x : f(x) + g(x) < a\} \\ &= \bigcup_{q \in \mathbb{Q}} \{x : f(x) < q,\ g(x) < a - q\} \\ &= \bigcup_{q \in \mathbb{Q}} \{x : f(x) < q\} \cap \{x : g(x) < a - q\} \\ &= \bigcup_{q \in \mathbb{Q}} f^{-1}((-\infty, q)) \cap g^{-1}((-\infty, a - q))\end{aligned}$$

which is a countable union of measurable sets and hence measurable. □

Lemma 3.25

If $\{f_n\}$ is a sequence of measurable functions, then so are the measurable (and possible infinite-valued) functions defined by

$$\inf(x) = \inf_n f_n(x),\ \sup(x) = \sup_n f_n(x), \liminf(x) = \liminf_n f_n(x),\ \limsup(x) = \limsup_n f_n(x).$$

Proof. We have

$$\overline{\sup}^{-1}((-\infty, a]) = \{x : \sup_n f_n(x) \le a\} = \bigcup_n \{x : f_n(x) \le a\}$$

which is measurable. Since $\inf(x) = \sup_n -f_n(x)$, we conclude that $\sup(x)$ is also measurable. Then

$$\liminf(x) = \sup_n \inf_{k \ge n} f_k(x)$$

is measurable by the first two parts. Similarly,

$$\limsup(x) = \inf_n \sup_{k \ge n} f_k(x)$$

is measurable. □

Theorem 3.7

If $f : \mathbb{R}^d \to \mathbb{R}$ is measurable and $g : \mathbb{R} \to \mathbb{R}$ is continuous then $g \circ f$ is measurable.

Proof. We have

$$g(f(x)) < a \iff f(x) \in g^{-1}((-\infty, a)).$$

But $g^{-1}((-\infty, a))$ is a countable union of intervals (a_n, b_n) so

$$g \circ f^{-1}((-\infty, a)) = \bigcup_n f^{-1}((a_n, b_n))$$

which is a measurable set. □

Theorem 3.8

If f is a non-negative measurable function, then there is an increasing sequence ϕ_n of non-negative, measurable simple functions ϕ_n converging to f.

Proof. Define

$$\phi_n(x) = \begin{cases} n & \text{if } f(x) \ge n, \\ m/n & \text{if } m/2^n \le f(x) < (m+1)/2^n \text{ for some } 0 \le m \le n2^n - 1. \end{cases}$$

The function ϕ_n rounds f down to n if f is too big, or else it rounds f down to the nearest fraction with denominator 2^n. In this way

$$\phi_n = n\mathbf{1}_{f^{-1}([n,\infty))} + \sum_{m=0}^{n2^n-1} \frac{m}{n} \mathbf{1}_{f^{-1}([m/2^n,(m+1)/2^n)}$$

is a measurable, simple function.

The function ϕ_n is bounded above by f pointwise, by its very definition. If $n' > n$ then, if $f(x) \geq n'$, we have

$$\phi_{n'}(x) = n' \geq n \geq \phi_n(x)$$

if $n \leq f(x) < n'$

$$\phi_{n'}(x) = m/2^{n'}$$

where $n \leq f(x) \leq (m+1)/2^{n'}$ and so $m/2^{n'} \geq n2^{n'}/2^{n'} = n$ too. Thus here $\phi_{n'}(x) \geq \phi_n(x)$.

Finally if $f(x) \leq n$ then since all fractions with denominator 2^n are also fractions $2^{n'-n}m/2^{n'}$ with denominator n', we would not round $f(x)$ down to $\phi_n(x)$ further than $\phi_{n'}(x)$. Thus ϕ_n is an increasing sequence.

Finally $|\phi_n(x) - f(x)| \leq 1/2^n$ for all n sufficiently large, and hence $\phi_n(x) \to f(x)$. □

Corollary 3.3

If $f : E \to \mathbb{R}$ is a non-negative measurable function on some measurable set $E \subseteq \mathbb{R}^d$, we can find non-negative simple functions ϕ_n increasing pointwise to f.

Proof. We can extend f to all of $\mathbb{R}^d$ by setting $f(x) = 0$ for $x \notin E$. This is still measurable. Then we apply the preceding theorem to f to get ϕ_n increasing to f. The functions ϕ_n, restricted to E, are still simple and still increase to f □

3.7 Littlewood's Principles

Littlewood's three principles are guiding heuristics that make some of the more technical parts of measure theory a bit more palatable. The first states that all measurable sets with finite measure are basically unions of cubes.

Theorem 3.9

Let E be a measurable set of finite measure. Then for $\varepsilon > 0$, there is a compact set C, which is a finite union of disjoint closed cubes, which satisfies $m(E \Delta C) < \varepsilon$.

Proof. Since E is measurable, there is an open set $U \subseteq \mathbb{R}^d$, containing E, and such that $m(U \setminus E) < \varepsilon/2$. The set U is an almost disjoint union of closed cubes Q_n with

$$m(U) = \sum_n V(Q_n).$$

We truncate this series so that

$$\sum_{n>N} V(Q_n) \leq \varepsilon/4.$$

Next, we replace the cubes Q_n (with $n \leq N$) with a slightly smaller cube $Q_n' \subseteq Q_n$, with the same center, and whose volume is $V(Q_n) - \varepsilon/2^{n+2}$. Our hope is that

$$C = \bigcup_{n\leq N} Q_n'$$

is a good approximation to E. If $x \in E\Delta C$ then $x \in E \setminus C$ or else $x \in U \setminus E$. So

$$m(E\Delta C) \leq m(U \setminus E) + m(E \setminus C) \leq \varepsilon/4 + m(E \setminus C).$$

Since $E \subseteq U$, $x \in E \setminus C$ only if $x \in Q_n$ for some $n > N$ or else $x \in Q_n \setminus Q_n'$ for some $n \leq N$. But

$$m\left(\bigcup_{n>N} Q_n\right) \leq \varepsilon/4$$

and

$$m\left(\bigcup_{n\leq N} Q_n \setminus Q_n'\right) \leq \sum_{n\leq N} V(Q_n) - V(Q_n') \leq \sum_{n\leq N} \frac{\varepsilon}{2^{n+2}} \leq \varepsilon/4.$$

All together, we have

$$m(E\Delta C) \leq 3\varepsilon/4 < \varepsilon.$$

Finally, C is compact as it is a finite union of closed cubes. □

The second principle tells us that all convergent sequences of functions are nearly uniformly convergent.

Theorem 3.10: Egorov's Theorem

Let $\{f_n\}$ be a sequence of functions defined on a measurable set E with finite measure, and converging pointwise to f. Then for $\varepsilon > 0$ there is a set $B \subseteq E$ with $m(B) < \varepsilon$ and such that $f_n \to f$ uniformly on $E \setminus B$.

Proof. Let

$$E_{k,N} = \{x : |f_n(x) - f(x)| \leq 1/k \text{ for } n \geq N\}.$$

Then the sets $E_{k,N}$ are increasing in N and measurable, as

$$E_{k,N} = (|f_n - f|)^{-1}((-\infty, 1/k]).$$

We also have

$$E = \bigcup_{N\geq 1} E_{k,N}$$

for all k. Indeed, for any x there is some N for which $|f_n(x) - f(x)| \leq 1/k$, and for that N, and all $N' \geq N$, we have $x \in E_{k,N'}$. Since E has finite measure and

$$m(E_{k,N}) \to m(E),$$

there is some N_k such that $m(E_{k,N_k}) > m(E) - \varepsilon/2^k$. Let $B = \bigcup_k E \setminus E_{k,N_k}$ (which is measurable) so that

$$m(B) \leq \sum_{k=1}^{\infty} m(E \setminus E_{k,N_k}) \leq \sum_{k=1}^{\infty} \frac{\varepsilon}{2^k} \leq \varepsilon.$$

Now

$$B = \bigcup_k E \cap (E_{k,N_k})^c = E \cap \bigcup_k (E_{k,N_k})^c = E \cap \left(\bigcap_k E_{k,N_k} \right)^c.$$

Thus if $x \in E \setminus B$, x belongs to $\bigcap_k E_{k,N_k}$ and hence $|f_n(x) - f(x)| \leq 1/k$ for $n \geq N_k$, which implies uniform convergence. □

Littlewood's final principle tells us that all measurable functions are almost continuous. We begin with a lemma.

Lemma 3.26

Let f be a measurable function defined on a measurable set E of finite measure. Then there is a sequence of step functions

$$s_n = \sum_{j=1}^{M_n} c_j \mathbf{1}_{Q_{n,j}}$$

with $Q_{n,j}$ a collection of disjoint closed cubes, and such that $s_n \to f$ pointwise almost everywhere.

Proof. We already know that there are simple functions $\phi_n \to f$. Write

$$\phi_n = \sum_l d_l \mathbf{1}_{E_l}$$

for some disjoint sets $E_l \subseteq E$. These sets are measurable and have finite measure, and so can be approximated by a finite union of disjoint cubes $\bigcup_k Q_{l,k}$ by Littlewood's first principle, and such that

$$\sum_l m\left(E_l \Delta \bigcup_k Q_{l,k} \right) < 1/2^n.$$

We let

$$s_n = \sum_l d_l \sum_k \mathbf{1}_{Q_{l,k}}$$

and observe that we have only altered ϕ_n on the set

$$B_n = \bigcup_l E_l \Delta \bigcup_k Q_{l,k},$$

a set of measure at most $1/2^n$. Now $\phi_n(x) \to f(x)$ pointwise and the only way this can fail for s_n is if $s_n(x) \neq \phi_n(x)$ for infinitely many n, which means $x \in B_n$ for infinitely many n. But then

$$x \in \bigcap_{N=1}^{\infty} \bigcup_{n \geq N} B_n.$$

However

$$m\left(\bigcup_{n \geq N} B_n\right) \leq \sum_{n \geq N} m(B_n) \leq \sum_{n \geq N} \frac{1}{2^n} \leq \frac{1}{2^N}.$$

Thus

$$m\left(\bigcap_{N=1}^{\infty} \bigcup_{n \geq N}\right) \leq m\left(\bigcap_{n \geq M}\right) \leq \frac{1}{2^M}$$

for all M and hence $m\left(\bigcap_N \bigcup_{n \geq N}\right) = 0$. □

Theorem 3.11: Lusin's Theorem

Let $f : E \to \mathbb{R}$ be a measurable function defined on a set E which has finite measure. Then for any $\varepsilon > 0$, there is a set B with $m(B) < \varepsilon$ such that f, when restricted to $E \setminus B$, is continuous.

Proof. We know that f is a limit of step functions

$$s_n = \sum_{j=1}^{M_n} d_j \mathbf{1}_{Q_j}$$

by the preceding lemma. Such functions are locally constant, and hence continuous, unless x belongs to the boundary of some Q_j. So there is a measure zero set B_n off of which s_n is continuous. By Egorov's theorem, we can find a set B of measure at most ε such that off of B, $s_n \to f$ uniformly. Thus off of $B \bigcup_n B_n$ we have a sequence of continuous functions converging uniformly to f, and so f is continuous there too. Moreover, $m(B \cup \bigcup_n B_n) < \varepsilon$ as needed. □

4

INTEGRATION

4.1 Defining the integral

For a simple function

$$\phi = \sum_i c_i \mathbf{1}_{E_i}$$

supported on a sets of finite measure, we would like to define

$$\int \phi = \sum_i c_i m(E_i).$$

There is some cause for concern however, as it is not clear at first that this is well-defined. There may be more than one way to write a simple function as a linear combination of indicator functions, and we need to be sure that the integral is defined the same way regardless of said representation. It will be convenient to define the *canonical* form of a simple function. We say

$$\phi = \sum_i c_i \mathbf{1}_{E_i}$$

is in canonical form if $c_i \neq 0$ for any i, the numbers c_i are distinct, and the sets E_i are disjoint. In this way

$$\phi^{-1}(a) = \begin{cases} E_i & \text{if } a = c_i \\ (\bigcup_i E_i)^c & \text{if } a = 0 \\ \emptyset & \text{otherwise.} \end{cases}$$

Lemma 4.1

Let

$$\phi = \sum_{i=1}^{m} c_i \mathbf{1}_{E_i} = \sum_{j=1}^{n} d_j \mathbf{1}_{F_j}$$

be a simple function on $\mathbb{R}^d$, for some finite collections of measurable sets E_i and F_j. Then

$$\sum_{i=1}^{m} c_i m(E_i) = \sum_{j=1}^{n} d_j m(F_j).$$

Proof. We assume that $\sum_{i=1}^{m} c_i \mathbf{1}_{E_i}$ is the canonical representation. Let J be a subset of $[n]$ and set

$$G_J = \{x \in \mathbb{R}^d : x \in F_j \text{ for } j \in J,\ x \notin F_j \text{ for } j \notin J\} = \bigcap_{j \in J} F_j \cap \bigcap_{j \notin J} F_j^c,$$

so the G_J is always measurable. The sets G_J with $J \subseteq [n]$ partition $\mathbb{R}^d$. Moreover, $G_J \subseteq F_j$ whenever $j \in J$, so that

$$m(F_j) = \sum_{J \supseteq j} m(G_J),$$

and hence

$$\sum_j d_j m(F_j) = \sum_{J \subseteq [n]} m(G_J) \sum_{j \in J} d_j.$$

Next set $G_{J,i} = G_J \cap E_i$, so that

$$\sum_{J \subseteq [n]} m(G_{J,i}) = m(E_i),\ \sum_{i=1}^{m} m(G_{J,i}) = m(G_J)$$

and hence

$$\sum_{i=1}^{m} c_i m(E_i) = \sum_{J \subseteq [n]} \sum_{i=1}^{m} c_i m(G_{J,i}).$$

If there is some $x \in G_{J,i}$, we must have

$$\sum_{j \in J} d_j = \phi(x) = c_i.$$

Otherwise, $G_{J,i}$ is empty and $m(G_{J,i}) = 0$. In either case

$$m(G_{J,i})c_i = m(G_{J,i}) \sum_{j \in J} d_j$$

$$\sum_{J \subseteq [n]} \sum_{i=1}^{m} c_i m(G_{J,i}) = \sum_{J \subseteq [n]} \sum_{i=1}^{m} m(G_{J,i}) \sum_{j \in J} d_j = \sum_{J \subseteq [n]} m(G_J) \sum_{j \in J} d_j.$$

□

Definition 4.1: Lebesgue integral for simple functions

Let

$$\phi = \sum_i c_i \mathbf{1}_{E_i}$$

be a simple function on $\mathbb{R}^d$. Then we define

$$\int \phi = \sum_i c_i m(E_i),$$

provided there are no sets E_i, E_j with $m(E_i) = m(E_j) = \infty$ and $c_i c_j < 0$.

Lemma 4.2

If ϕ_1 and ϕ_2 are simple functions, and c is a real number, then

1. $\int c\phi_1 + \phi_2 = c \int \phi_1 + \int \phi_2$,
2. for any disjoint measurable sets E_1 and E_2, $\int_{E_1 \cup E_2} \phi_1 = \int_{E_1} \phi_1 + \int_{E_2} \phi_1$,
3. if $\phi_1 \leq \phi_2$ pointwise then $\int \phi_1 \leq \int \phi_2$, and
4. $\left| \int \phi_1 \right| \leq \int |\phi_1|$.

Proof. Write $\phi_1 = \sum_j c_j \mathbf{1}_{A_j}$ and $\phi_2 = \sum_k d_k \mathbf{1}_{B_k}$, in canonical form. Then

$$\int c\phi_1 + \phi_2 = \int \sim_j cc_j \mathbf{1}_{A_j} + \sum_k d_k \mathbf{1}_{B_k} = \sum_j cc_j m(A_j) + \sum_k d_k m(B_k) = c \int \phi_1 + \int \phi_2.$$

This shows (1). For (2), we have

$$\int_{E_1 \cup E_2} \phi_1 = \int \mathbf{1}_{E_1 \cup E_2} \phi_1 = \int (\mathbf{1}_{E_1} + \mathbf{1}_{E_2}) \phi_1 = \int \mathbf{1}_{E_1} \phi_1 + \int \mathbf{1}_{E_2} \phi_1 = \int_{E_1} \phi_1 + \int_{E_2} \phi_1,$$

where we've used disjointness in the second equality and (1) in the third.

For (3), we let $C_{jk} = A_j \cap B_k$, so

$$\sum_{j,k} c_j \mathbf{1}_{C_{jk}} = \phi_1 \leq \phi_2 = \sum_{j,k} d_k \mathbf{1}_{C_{jk}}$$

so, since the sets C_{jk} are disjoint (using that ϕ_1 and ϕ_2 were in canonical form), we must have $c_j \le d_k$ whenever C_{jk} is non-empty. In this way

$$\int \phi_2 - \int \phi_1 = \sum_{j,k} m(C_{jk})(d_k - c_j) \ge 0.$$

Finally, for (4), $-|\phi_1| \le \phi_1 \le |\phi_1|$ pointwise, so

$$-\int |\phi_1| \le \int \phi_1 \le \int |\phi_1|.$$

□

Since non-negative measurable functions are monotone limits of simple functions, we can immediately extend the definition. Recall that the support of f is the set $\operatorname{supp}(f) = \{x \in \mathbb{R}^d : f(x) \neq 0\}$, and that

$$f = f_+ - f_-, \; f_+ = \max\{f, 0\}, \; f_- = -\min\{f, 0\}.$$

Definition 4.2: Lebesgue integral for measurable functions

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a non-negative measurable function. Then we define

$$\int f = \sup_{\phi} \int \phi$$

where ϕ ranges over all simple functions with the property $0 \le \phi(x) \le f(x)$ for all x and $m(\operatorname{supp}(\phi)) < \infty$. If this integral is finite, we call f **integrable**. If f is not non-negative, we write $f = f_+ - f_-$ where f_+ and f_- are non-negative, and we set

$$\int f = \int f_+ - \int f_-$$

provided both f_+ and f_- are integrable, and in this case we call f integrable.

Note that we can safely omit the $m(\{x : \phi(x) \neq 0\}) < \infty$ condition just by replacing ϕ with $\mathbf{1}_{\|x\| \le N}\phi$, which is still a simple function, and then sending $N \to \infty$.

Theorem 4.1

If f_1 and f_2 are non-negative, measurable functions, then

1. for any disjoint measurable sets E_1 and E_2, $\int_{E_1 \cup E_2} f_1 = \int_{E_1} f_1 + \int_{E_2} f_1$,
2. if $f_1 \le f_2$ pointwise then $\int f_1 \le \int f_2$.

Proof. The proof of this theorem comes from the corresponding results for simple functions along with approximation. Property (2) is merely because any simple

function dominated by f_1 is also dominated by f_2.

$$\int f = \sup_{0\le\phi\le f_1} \int \phi \le \sup_{0\le\phi\le f_2} \int \phi = \int f_2.$$

For property (1), note that if ϕ is a simple function dominated by f_1, then

$$\mathbf{1}_{E_1\cup E_2}\phi = \mathbf{1}_{E_1}\phi + \mathbf{1}_{E_2}\phi$$

and $\mathbf{1}_{E_i}\phi$ is a simple function dominated by $\mathbf{1}_{E_i} f$ for $i = 1,2$. Thus

$$\int_{E_1\cup E_2} \phi = \int_{E_1} \phi + \int_{E_2} \phi \le \int_{E_1} f_1 + \int_{E_2} f_1$$

and so taking supremums,

$$\int_{E_1\cup E_2} f_1 \le \int_{E_1} f_1 + \int_{E_2} f_1.$$

Conversely, if ϕ_1 and ϕ_2 are non-negative, simple functions dominated by $\mathbf{1}_{E_1} f_1$ and $\mathbf{1}_{E_2} f_1$ then $\phi_1 + \phi_2$ is dominated by $\mathbf{1}_{E_1\cup E_2} f_1$ and so

$$\int_{E_1} \phi_1 + \int_{E_2} \phi_2 = \int_{E_1\cup E_2} \phi_1 + \phi_2 \le \int_{E_1\cup E_2} f_1$$

and taking supremums finishes the proof. □

Corollary 4.1

If f and g are integrable functions, then

1. for any disjoint measurable sets E_1 and E_2, $\int_{E_1\cup E_2} f = \int_{E_1} f + \int_{E_2} f$,
2. if $f \le g$ pointwise then $\int f \le \int g$, and
3. $\left|\int f\right| \le \int |f|$.

Proof. First

$$\int_{E_1\cup E_2} f = \int_{E_1\cup E_2} f_+ - \int_{E_1\cup E_2} f_- = \int_{E_1} f_+ + \int_{E_2} f_+ - \int_{E_1} f_- - \int_{E_2} f_- = \int_{E_1} f + \int_{E_2} f.$$

Next, if $f \le g$ then $f_+ \le g_+$ while $f_- \ge g_-$, so

$$\int f = \int f_+ - \int f_- \le \int g_+ - \int g_- = \int g.$$

Finally, we may assume that $\int f_+ \ge \int f_-$, then

$$\left|\int f\right| = \int f_+ - \int f_- \le \int f_+ + \int f_- = \int |f|.$$

□

Lemma 4.3

If $f = 0$ almost everywhere then $\int f = 0$.

Proof. Both f_- and f_+ are also zero almost everywhere, so we can assume f is non-negative. In that case, and simple function ϕ dominated by f has to vanish almost everywhere too, and so $\int \phi = 0$. □

Already, we have all the necessary tools to prove our first convergence theorem. A common theme is to break an integral into pieces, as in (1) of Corollary 4.1, and handle the pieces separately.

Theorem 4.2: Bounded Convergence Theorem

Suppose that $\{f_n\}$ is a sequence of functions all supported in a measurable set E of finite measure. Suppose that $\sup_n |f_n(x)| \leq M$ for almost all x, and that $f_n \to f$ pointwise almost everywhere. Then

$$\int |f_n - f| \to 0.$$

Proof. By Egorov's theorem, for any $\varepsilon > 0$, there is a measurable set $B_\varepsilon \subseteq E$ of measure at most $\varepsilon/4M$ and such that off of B_ε, $f_n \to f$ uniformly. If n is sufficiently large,

$$\int_{E \setminus B_\varepsilon} |f_n - f| \leq \int_{E \setminus B_\varepsilon} \varepsilon/2m(E) < \varepsilon/2.$$

Meanwhile, on B_ε, we still have $|f_n| \leq M$ almost everywhere, and $f_n \to M$ almost everywhere. So $B_\varepsilon = X \cup Y$ where Y has measure zero, and on X, we have $\sup_n |f_n| \leq M$ and $|f| \leq M$. By the preceding lemma

$$\int_Y |f_n - f| = 0$$

and on X, we have

$$\int_X |f_n - f| \leq \int_X 2M \leq \int_{B_\varepsilon} 2M \leq \varepsilon/2.$$

□

Corollary 4.2

Let ϕ_n be a sequence of uniformly bounded, non-negative simple functions supported on E of finite measure, converging pointwise to f from below. Then $\int \phi_n \to \int f$.

Proof. On the one hand, ϕ_n is dominated by f so

$$\int \phi_n \leq \int f.$$

If ψ is any non-negative simple function dominated by f with

$$\int f - \varepsilon \leq \int \psi \leq \int f,$$

then $\psi_n = \min\{\phi_n, \psi\}$ is a non-negative simple function bounded by f, and $\psi_n \to \psi$ pointwise. We have

$$\left| \int \psi_n - \int \psi \right| = \left| \int \psi_n - \psi \right| \leq \int |\psi_n - \psi| \to 0$$

by the Bounded Convergence Theorem. Thus for n sufficiently large

$$\int f - 2\varepsilon \leq \int \psi - \varepsilon \leq \int \psi_n \leq \int \phi_n$$

so

$$\int f - \int \phi_n \leq 2\varepsilon.$$

□

Theorem 4.3: Linearity of the integral

If f_1 and f_2 are non-negative, bounded and measurable functions with support of finite measure, and if $c \geq 0$ then

1. $\int c f_1 + f_2 = c \int f_1 + \int f_2$,

Proof. The proof of this theorem comes from the corresponding results for simple functions along with approximation.

First if $c \geq 0$, then ϕ is a simple function dominated by f_1 if and only if $c\phi$ is a simple function dominated by $c f_1$. so

$$\int cf = \sup_{0 \leq \phi \leq f} \int c\phi = c \sup_{0 \leq \phi \leq f} \int \phi = c \int f_1,$$

so we can assume $c = 1$.

Now suppose first that f_1 and f_2 have supports with finite measure. Let $\{\phi_n^{(1)}\}$ and $\{\phi_n^{(2)}\}$ be sequences of simple functions increasing to f_1 and f_2, respectively, which exist by Theorem 3.6. Then

$$\int f_1 + f_2 = \lim \int \phi_n^{(1)} + \phi_n^{(2)} = \lim \int \phi_n^{(1)} + \lim \int \phi_n^{(2)} = \int f_1 + \int f_2.$$

□

Theorem 4.4: Fatou's Lemma

Let f_n be a sequence of non-negative measurable functions converging to f almost everywhere. Then

$$\int f \leq \liminf_n \int f_n.$$

Proof. Let ϕ be a simple function dominated by f, and with support of finite measure, so that $\int \phi \leq \int f$. Let $\phi_n = \min\{\phi, f_n\}$. Then $\phi_n \to \phi$ almost everywhere, and the functions ϕ_n are bounded (since ϕ is) hand have support of finite measure. So, by linearity the Bounded Convergence Theorem

$$\left| \int \phi - \phi_n \right| \leq \int |\phi - \phi_n| \to 0.$$

It follows that

$$\int \phi \leq \liminf_f \int \phi_n \leq \liminf_n \int f_n.$$

This holds for all simple ϕ dominated by f and hence

$$\int f \leq \liminf_n \int f_n.$$

□

Corollary 4.3: Monotone Convergence Theorem

Let $\{f_n\}$ be a sequence of non-negative measurable functions increasing to a measurable function f pointwise. Then

$$\int f_n \to \int f.$$

Proof. Indeed,

$$\int f_n \leq \int f$$

for all n by monotonicity, so

$$\limsup_n \int f_n \leq \int f \leq \liminf_n \int f_n$$

and thus

$$\int f = \liminf_n \int f_n = \limsup_n \int f_n = \lim \int f_n.$$

□

Corollary 4.4

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a non-negative measurable function. Then

$$\int f = \lim_{N\to\infty} \int_{E_N} f$$

where

$$E_N = \{x : \|x\| \le N, |f(x)| \le N\}.$$

Proof. Since $\bigcup_N E_N = \mathbb{R}^d$ and the sets E_N are increasing, the functions $\mathbf{1}_{E_N} f$ increase to f monotonically. The result now follows from the Monotone Convergence Theorem. □

Corollary 4.5: Linearity of the integral, again

Suppose f and g are integrable functions, and c is a constant. Then

$$\int cf + g = c\int f + \int g.$$

Proof. First assume $c \ge 0$ and f and g are non-negative. We have $\int f = \lim \int f_n$ and $\int g = \lim \int g_n$ where f_n and g_n are increasing monotonically to f and g, and are bounded with support of finite measure. Then

$$c\int f + \int g = \lim c\int f_n + \lim \int g_n = \lim \int cf_n + g_n = \int cf_n + g_n,$$

again by the Monotone Convergence Theorem.

If f and g are not non-negative, but $c \ge 0$, we can apply the preceding result to f_+, g_+, f_- and g_-, using that $cf_+ = (cf)_+$. Finally if $c \le 0$, replace f with $-f$ and c with $-c$. □

We now proceed to the most powerful of all convergence theorems. First we need some integrability results.

Lemma 4.4

If f is an integrable function on $\mathbb{R}^d$ then $|f| < \infty$ almost everywhere,

$$\int_{\|x\|>N} |f| < \varepsilon$$

for N sufficiently large, and if δ is sufficiently small, then

$$\int_E |f| < \varepsilon$$

for any measurable set E of measure at most δ. In particular, if N is suffciently large,

$$\int_{|f|\geq N} |f| < \varepsilon.$$

Proof. For the first claim, we have

$$\int_{\|x\|\leq N} |f| \to \int |f|$$

by the Monotone Convergence Theorem, so if N is sufficiently large, we must have

$$\int_{\|x\|>N} |f| = \int |f| - \int_{\|x\|\leq N} f < \varepsilon.$$

Next $|f|$ is non-negative, and if $|f| = \infty$ on E then

$$\int |f| \geq \int_E |f| \geq Nm(E)$$

for any N so $m(E) = 0$.

Finally, assuming $|f|$ is bounded (as we may, since it is finite almost everywhere), from the Monotone Convergence Theorem, we know

$$\int |f| = \lim_N \int_{E_N} |f|$$

where

$$E_N = \{x : \|x\| < N, |f(x)| < N\},$$

so it suffices to prove the second claim when f is bounded and has support with finite measure. If $|f| \leq N$ everywhere, and E has measure at most δ then

$$\int_E |f| \leq N\delta < \varepsilon$$

if $\delta < \varepsilon/N$. Again

$$\int_{|f|\geq N} |f| \geq Nm(\{x : |f(x)| \geq N\})$$

and so $m(\{x : |f(x)| \geq N\}) < \delta$ for N sufficiently large, and the final claim follows. □

Theorem 4.5: Dominated Convergence Theorem

Suppose $\{f_n\}$ is a sequence of integrable functions, such that $|f_n| \leq g$ a.e. for some integrable g and each n. If $f_n \to f$ pointwise almost everywhere, then $\int |f_n - f| \to 0$.

Proof. Since $f_n \to f$ almost everywhere, $|f| \leq g$ almost everywhere too. Ignoring the set of measure zero where this fails (which does not contribute to the integrals), we have $|f_n - f| \leq |f_n| + |f| \leq 2|g|$. Thus if $E_1 = \{x : \|x\| \geq N\}$ and $E_2 = \{x : |g(x)| \geq N\}$, then for N sufficiently large. Then

$$\int_{E_i} |f_n - f| \leq \int_{E_i} 2|g| < \varepsilon$$

for $i = 1, 2$, by the preceding lemma. On the remaining set, E, which has finite measure, f_n and f are bounded and so by the Bounded Convergence Theorem,

$$\int_E |f_n - f| \to 0.$$

□

4.2 The Differentiation Theorem

Lemma 4.5: Vitali's Covering Lemma

Let $\mathscr{B}$ be a finite set of balls in $\mathbb{R}^d$. Then there is a finite subset $\mathscr{B}'$ of $\mathscr{B}$ such that the balls in $\mathscr{B}'$ are disjoint and

$$\bigcup_{B \in \mathscr{B}'} 3B \supseteq \bigcup_{B \in \mathscr{B}} B,$$

where $3B$ denotes the dilation of B by a factor of 3.

Proof. Iteratively apply the following rule, beginning with $\mathscr{B}' = \emptyset$: if $B \in \mathscr{B} \setminus \mathscr{B}'$ is such that B is disjoint from each ball in $\mathscr{B}'$ and has maximal radius among all such balls in $\mathscr{B}$ then add B to $\mathscr{B}'$.

The process has to terminate in finitely many steps since $\mathscr{B}$ is finite, and the balls in $\mathscr{B}'$ are disjoint by construction. If x belongs to some ball B' from $\mathscr{B} \setminus \mathscr{B}'$ then B' cannot be disjoint from all balls in $\mathscr{B}'$, or else we could add it to $\mathscr{B}'$. Let B be the first ball added to $\mathscr{B}'$ which intersected B'. At this stage we could have added B' to $\mathscr{B}'$ instead, but we opted not to, and this can only have happened if the radius of B', say r', is smaller than that of B, say r. So B' intersects B and has smaller radius. If c is the centre of B and c' the centre of B', and if y is a point in their intersection,

then

$$\|x-c\| \le \|x-c'\| + \|c'-y\| + \|y-c\| \le r' + r' + r \le 3r.$$

This shows $x \in 3B$. □

Definition 4.3: The Hardy-Littlewood Maximal Function

Let f be an integrable function. Then we define the function

$$(\mathscr{M}f)(x) = \sup_B \frac{1}{m(B)} \int_B f$$

where the supremum is taken over all balls B such that $x \in B$.

Lemma 4.6

If f is integrable then $\mathscr{M}f$ is measurable.

Proof. Let $a \in \mathbb{R}$, we then need to show that $E_a = \{x : (\mathscr{M}f)(x) > a\}$ is measurable. If $x \in E_a$ then there is a ball B for which

$$\frac{1}{m(B)} \int_B f > a.$$

The same is true for any other $y \in B$ since the left hand side is unchanged, and so $(\mathscr{M}f)(y) > a$ too. Thus in fact E_a is open. □

Lemma 4.7

Let f be integrable. Then $\mathscr{M}f$ is finite almost everywhere, and

$$m(\{x : |(\mathscr{M}f)(x)| > \lambda\}) \le \frac{3^d}{\lambda} \int |f|.$$

Proof. We prove the second claim, the first will follow immediately. Set

$$E_\lambda = \{x : |(\mathscr{M}f)(x)| > \lambda\}$$

and $E_\lambda^N = E_\lambda \cap \{x : \|x\| \le N\}$. Now E_λ^N has finite measure and so can be approximated by a compact set C to within ε:

$$m(E_\lambda^N) - \varepsilon < m(C).$$

For each $x \in C$, there is a ball B_x containing x and with

$$\int_{B_x} f > \lambda m(B_x).$$

Since C is compact, it can be covered by finitely many B_xs, say those with $x \in X$ for some finite set X. Applying the Vitali Covering Lemma, we can find a subset $X' \subseteq X$ such that the balls B_x with $x \in X'$ are disjoint, and their dilations by 3 still cover C. By disjointness,

$$\lambda \sum_{x \in X'} m(B_x) \leq \sum_{x \in X'} \int_{B_x} |f| \leq \int |f|.$$

However, $m(3B_x) = 3^d m(B_x)$, and

$$C \subseteq \bigcup_{x \in X'} 3B_x \implies m(C) \leq \sum_{x \in X'} m(3B_x) = 3^d \sum_{x \in X'} m(B_x).$$

So

$$m(E_\lambda^N) - \varepsilon \leq \frac{3^d}{\lambda} \int |f|.$$

Taking $\varepsilon \to 0$ and then $N \to \infty$ concludes the proof. □

Theorem 4.6: Lebesgue's Differentiation Theorem

et f be an integrable function. Then for almost every x

$$\lim_{m(B) \to 0} \frac{1}{m(B)} \int_B f \to f(x)$$

where the limit is taken over all balls B containing x and with radius tending to 0.

Proof. First, by replacing f with $f \mathbf{1}_{\{\|x\| < N, |f(x)| < N\}}$, and letting $N \to \infty$ we can assume that f has support of finite measure and is bounded.

In this case, we know that there are step functions $s_n \to f$ pointwise almost everywhere by Lemma 3.7. If

$$s_n = \sum_j c_j \mathbf{1}_{Q_j}$$

where the Q_j are disjoint cubes, we can just as well assume that $c_j \leq \sup |f| = N$. Indeed, we just replace c_j by N if $c_j > 0$ or $-N$ if $c_j < 0$. Then

$$|f(x) - N| \leq |f(x) - c_j|$$

for $x \in Q_j$ and we still have a step function.

Now, we will prove the theorem by showing that

$$X_t = \left\{ x : \limsup_B \left| \frac{1}{m(B)} \int_B f - f(x) \right| > t \right\} = 0$$

for each $t > 0$. Given $x \in X_t$ and some ball B containing x, we have

$$\frac{1}{m(B)} \int_B f - f(x) = \left(\frac{1}{m(B)} \int_B f - s_n \right) + \left(\frac{1}{m(B)} \int_B s_n - s_n(x) \right) + (s_n(x) - f(x)).$$

The first term is at most

$$\frac{1}{m(B)}\int_B |f-s_n| \le \mathscr{M}(|f-s_n|)(x),$$

the second is at most

$$\frac{1}{m(B)}\int_B |s_n - s_n(x)|$$

and the third is at most $|s_n(x)-f(x)|$. Thus we have

$$t < \limsup_B \left|\frac{1}{m(B)}\int_B f - f(x)\right| < \mathscr{M}(|f-s_n|)(x) + \limsup_B \frac{1}{m(B)}\int_B |s_n - s_n(x)| + |s_n(x)-f(x)|$$

and so for $x \in X_t$, we have on of the following:

$$\mathscr{M}(|f-s_n|)(x) > t/3,\ \limsup_B \frac{1}{m(B)}\int_B |s_n - s_n(x)| > t/3,\ \text{or } |s_n(x)-f(x)| > t/3.$$

We let

$$\begin{aligned} U &= \{x : \mathscr{M}(|f-s_n|)(x) > t/3\},\\ V &= \{x : \limsup_B \frac{1}{m(B)}\int_B |s_n - s_n(x)| > t/3\},\\ W &= \{x : |s_n(x)-f(x)| > t/3\}. \end{aligned}$$

By the preceding lemma,

$$m(U) \le \frac{3^d}{t/3}\int |f-s_n|,$$

and by Chebyshev's inequality

$$m(V) \le \frac{1}{t/3}\int |f-s_n|.$$

To estimate V, note that once $B \subseteq Q_j$ for some j,

$$\int_B |s_n - s_n(x)| = 0$$

since s_n is constant on Q_n. So if $x \in Q_j^o$ for some j then

$$\limsup_B \int_B \frac{1}{m(B)}|s_n - s_n(x)| = 0.$$

The same is true if $x \notin Q_j$, since for B sufficiently small containing x $s_n(y) = s_n(x) = 0$ for $y \in B$. So

$$m(V) \le \sum_j m(\partial Q_j) = 0.$$

It thus suffices to show that $\int |f-s_n| \to 0$, but this now follows from the Dominated Convergence Theorem. □

5

INTRO. TO DISCRETE ANALYSIS

5.1 The complex exponential

Recall that for a complex number $z = x + iy$ we have the polar form

$$z = |z|e(2\pi i \arg z)$$

where, in our case, we have normalized the argument to lie between 0 and 1. Thus $\arg z$ is the proportion of the full circumference of the unit circle taken up by the angle z makes with the positive x-axis. This polar form is made possible by Euler's identity

$$e^{i\theta} = \cos(\theta) + i\sin(\theta).$$

Since we will prefer the normalized argument, we define the related function

$$e(\theta) = e^{2\pi i\theta}$$

so that now, polar coordinates take the form

$$z = |z|e(\arg(z)).$$

Lemma 5.1

The function e possesses the following qualities:

1. $|e(\theta)| = 1$ for all real numbers θ,
2. $e(n) = 1$ for all integers n, and consequently e is periodic with period 1,
3. $e(\theta + \tau) = e(\theta)e(\tau)$, for all real numbers θ and τ, and
4. $e(n\theta) = (e(\theta))^n$ for all real numbers θ and integers n.

Proof. Properties (1) and (2) follow immediately from Euler's identity. Property (3) follows from the fact that e is an exponential. Property (4) follows from (3) for all non-negative integers, while if $n < 0$ is an integer, then

$$e(\theta)^n e(-n\theta) = e(n\theta)e(-n\theta) = 1$$

shows that $e(-n\theta) = 1/(e(\theta))^n$. □

The next property of the function e is one of the most vital and will be fundamental to almost everything in the course.

Theorem 5.1: The orthogonality relations

Let $n \in \mathbb{Z}$. Then

$$\int_0^1 e(n\theta)\,d\theta = \begin{cases} 1 & n = 0, \\ 0 & n \neq 0. \end{cases}$$

We'll give two proofs.

First proof. If $n = 0$ then we just have $\int_0^1 1\,d\theta = 1$. If $n \neq 0$, then we consider a Riemann sum approximation to the integral

$$\int_0^1 e(n\theta)\,d\theta = \lim_{k\to\infty} \frac{1}{k} \sum_{j=0}^{k-1} e(nj/k) = \lim_{k\to\infty} \frac{1}{k} \sum_{j=0}^{k-1} (e(n/k))^j.$$

But the sum on the right is over a geometric progression, so

$$\frac{1}{k} \sum_{j=0}^{k-1} e(nj/k) = \frac{1}{k} \frac{e(n/k)^k - 1}{e(n/k) - 1} = 0.$$

The latter is because $e(n/k)^k = e(n) = 1$ while if $k > n$ then $e(n/k) \neq 1$. □

Second proof. Again the case $n = 0$ is trivial. If $n \neq 0$ consider the function $F(x, y) = \frac{-sx+ty}{2\pi n}$ (where we leave s and t as parameters to choose later) and its gradient field

$$\nabla F = \left(\frac{-s}{2\pi n}, \frac{t}{2\pi n} \right).$$

This vector field is conservative and so integrates to zero over any closed curve. We use the closed curve parametrized by $r(\theta) = (\cos(2\pi n\theta), \sin(2\pi n\theta))$ with $0 \leq \theta \leq 1$, which wraps around the unit circle n times. Then

$$\begin{aligned} 0 = \int \nabla F \cdot dr = \int_0^1 \left(\frac{-s}{2\pi n}, \frac{t}{2\pi n} \right) \cdot (-2\pi n \sin(2\pi n\theta), 2\pi n \cos(2\pi n\theta)) d\theta \\ = s \int_0^1 \sin(2\pi n\theta) d\theta + t \int_0^1 \cos(2\pi n\theta) d\theta. \end{aligned}$$

Now choose $s = \int_0^1 \sin(2\pi n\theta) d\theta$ and $t = \int_0^1 \cos(2\pi n\theta) d\theta$ to get

$$0 = s^2 + t^2 = |t + is|^2.$$

So

$$\int_0^1 e(n\theta) d\theta = t + is = 0.$$

□

The term *orthogonality relations* is perhaps a bit confusing at first. It does come from an inner product, however. Since we will use the periodicity properties of e, we write $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ for the set of real numbers modulo the equivalence relation $x \equiv y$ if $x - y \in \mathbb{Z}$. A complete set of representatives for this relation is $[0, 1)$.

Definition 5.1: The space $L^2(\mathbb{T})$

Consider the set $\mathscr{L}^2 = \{f : [0,1] \to \mathbb{C} : \int_0^1 |f|^2 < \infty\}$ consisting of (complex-valued) Lebesgue square-integrable functions. We define an equivalence relation on $\mathscr{L}^2$ by

$$f \sim g \iff f - g = 0 \text{ a.e.}$$

Then $L^2(\mathbb{T})$ consists of equivalence classes of this relation.

It turns out that L^2 is endowed which the structure of a Hilbert space (a complex inner product space which is complete as a metric space).

Definition 5.2: Complex inner product space

A complex vector space V is an inner product space if there is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{C}$ such that

Positivity: $\langle v, v \rangle \geq 0$ and $\langle v, v \rangle = 0$ if and only if $v = 0$,

Linearity: $\langle v + cu, w \rangle = \langle v, w \rangle + c\langle u, w \rangle$,

Conjugate-symmetry: $\langle v, u \rangle = \overline{\langle u, v \rangle}$

holds for all vectors $u, v, w \in V$ and scalars $c \in \mathbb{C}$. We then define a norm on V by $\|v\| = \sqrt{\langle v, v \rangle}$. If this norm induces a complete metric space, then V is called a Hilbert space.

The reason for not defining L^2 on functions, rather than equivalence classes, is that if $g = 0$ almost everywhere then $\int_0^1 fg = 0$ regardless of f. This is incompatible with (1), as we shall see.

Lemma 5.2

The space $L^2(\mathbb{T})$ is a complex inner product space.

Proof. Let $[f]$ and $[g]$ be two equivalence classes, represented by functions f and g respectively. We set

$$\langle [f], [g] \rangle = \int_0^1 f(\theta)\overline{g(\theta)}d\theta.$$

This is well-defined: if $u, v \in \mathscr{L}^2$ are 0 almost everywhere then

$$\int_0^1 (f(\theta) + u(\theta))\overline{(g(\theta) + v(\theta))}d\theta$$
$$= \int_0^1 f(\theta)\overline{g(\theta)}d\theta + \int_0^1 u(\theta)\overline{g(\theta)}d\theta + \int_0^1 f(\theta)\overline{v(\theta)}d\theta + \int_0^1 u(\theta)\overline{v(\theta)}d\theta$$

and since $f\overline{v}$, $u\overline{g}$ and $u\overline{v}$ are 0 almost everywhere, those integrals vanish.

Now properties (2), (3) and (4) are straightforward from the linearity properties of integral, and we just check (1). Since $|f(\theta)|^2 \geq 0$ we have

$$\langle [f], [f] \rangle \geq 0.$$

From measure theory, we know that if $F \geq 0$ is an integrable function with $\int F = 0$ then $F = 0$ almost everywhere. So in our case $|f|^2 = 0$ almost everywhere, which means $f = 0$ almost everywhere (since $f = 0 \iff |f|^2 = 0$), and so $[f] = [0]$. □

Strictly speaking, when we speak of elements of L^2, we aren't talking about functions, but about equivalence classes. In practice, however, we will just refer to L^2

functions with the caveat that anything said is meant to be interpreted almost everywhere.

Theorem 5.2: T

e set of functions $\{e(n\theta)\}_{n\in\mathbb{Z}}$ is orthonormal in $L^2(\mathbb{T})$.

Proof. Indeed,

$$\langle e(m\theta), e(n\theta)\rangle = \int_0^1 e(m\theta)\overline{e(n\theta)}\,d\theta = \int_0^1 e((m-n)\theta)\,d\theta = \begin{cases} 1 & n = m \\ 0 & n \neq m. \end{cases}$$

□

5.2 Absract measure spaces

Taking inspiration from the Lebesgue measure we define a list of axioms that will allow us to reproduce, *mutatis mutandis*, a number of the theorems we proved about the Lebesgue measure and integral. You will recall that we had to establish these facts, with a fair amount of work, when we defined the Lebesgue measure. Here we take them for granted.

Definition 5.3: Abstract measure space

An abstract measure space consists of three pieces of information, stored as a triple (X, Σ, μ), namely a set of points X, a σ-algebra of *measurable sets* Σ and a measure $\mu : \Sigma \to [0, \infty]$ satisfying the properties

1. $\mu(\emptyset) = 0$, and
2. $\mu(\bigsqcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$ for any countable collection of pairwise disjoint sets $\{A_n\} \subseteq \Sigma$.

The measure space is called *finite* if $\mu(X) < \infty$ and is called a *probability space* if $\mu(X) = 1$.

As is so often the case in math, when we have a nicely defined object, we are wont to define the nice maps between said objects.

Definition 5.4: Measurable function

If (X, Σ_X, μ_X) and (Y, Σ_Y, μ_Y) are two measure spaces, then a measurable function is a function $f : X \to Y$ such that $f^{-1}(B) \in \Sigma_X$ for $B \in \Sigma_Y$.

Notice that we don't actually need the measures μ_X and μ_Y to define a measurable function, or a measurable set for that matter – only the σ-algebras are needed. For that reason, we will refer to the pair (X, Σ_X) as a *measurable space*, and reserve the freedom to choose any measure on this space.

Example. *The triple* $([0,1], \mathscr{M}, m)$ *consisting of* $[0,1] \subseteq \mathbb{R}$,

$$\mathscr{M} = \{A \subseteq [0,1] : A \text{ is Lebesgue measurable}\}$$

and m *the Lebesgue measure restricted to* $[0,1]$ *is a probability space.*

Example. *Let* $f : [0,1] \to [0,1]$ *be a measurable function. The triple* $([0,1], \mathscr{M}, \mu_f)$ *consisting of* $[0,1] \subseteq \mathbb{R}$,

$$\mathscr{M} = \{A \subseteq [0,1] : A \text{ is Lebesgue measurable}\}$$

and the measure μ_f *defined by*

$$\mu_f(A) = \int_A f(x)\,dx$$

is a finite measure space, and the function f *is called the density function.*

We will see later on that this way of producing new measures using integrals of measurable functions extends to abstract measure spaces as well.

Example. *Since constant functions are always be measurable, we can use the above example to replace a finite measure* μ *by a probability measure. Indeed, suppose* (X, Σ, μ) *is a finite measure space and* $\mu(X) = C$. *Let* $f : X \to \mathbb{R}$ *be the function* $f(x) = 1/C$. *In line with the previous example, we define*

$$\mu_f(A) = \int_A f(x)\,dx = \frac{1}{C}\int_A dx = \frac{\mu(A)}{C},$$

and so in particular

$$\mu_f(X) = \frac{\mu(X)}{C} = 1$$

and we have a probability space.

5.3 Finite L^2 spaces

Suppose now that X is a finite set, which we want to turn into a probability space. We first need to define a Σ-algebra which, being a subset of $\mathscr{P}(X)$, is necessarily finite and hence just an algebra.

Lemma 5.3

Any algebra Σ on a finite set X corresponds to a partition of X, say

$$X = \bigsqcup_{i=1}^{n} X_i$$

of non-empty sets called *atoms*. Any element of Σ is then a union of atoms. The (Borel) measurable functions $f : X \to \mathbb{C}$ are those which are constant on atoms.

Proof. If we are presented with a partition of X as above, then we can just as well define

$$\Sigma = \left\{ \bigcup_{i \in I} X_i : I \subseteq \{1, \ldots, n\} \right\},$$

and this is an algebra. Conversely, for $x \in X$ we define

$$[x] = \bigcap_{A \in \Sigma : x \in A} A.$$

Then $[x] \in \Sigma$, being a finite intersection of sets from Σ, and moreover, the sets $[x]$ and $[y]$ are either disjoint or equal. Thus we have defined a partition of X into sets from Σ. By construction, $[x]$ is the smallest set in Σ which contains x, and so if $B \in \Sigma$ is measurable and contains x then it also contains $[x]$, and consequently

$$B = \bigcup_{x \in B} [x]$$

is a union of atoms.

Now suppose $f : X \to \mathbb{C}$ is measurable. Then $f^{-1}(z) = \{x \in X : f(x) = z\}$ is measurable and hence a union of atoms. If $f(x) = z$, then $x \in f^{-1}(z)$ whence $[x] \subseteq f^{-1}(z)$ and hence $f(y) = z$ for all $y \in [x]$. □

We now have everything we need to define one of the most important spaces in analysis. Here we are only focusing on the discrete world, so our ambient set X is finite. The definition extends in much the same way to infinite sets X but there are some technical hurdles that need to be overcome. The reader familiar with inner product spaces will probably see that there is little new here, but we are emphasizing an analytic viewpoint rather than the algebraic one.

Definition 5.5: The space $L^2(X,\Sigma,\mu)$

Let (X,Σ,μ) be a probability space with X finite. Then the space $L^2(X,\Sigma,\mu)$ consists of measurable functions $f : X \to \mathbb{C}$ endowed with the structure of a Hilbert space as follows. Suppose $X = \bigsqcup_{i=1}^n X_i$ is a decomposition of X into atoms. Then f is constant on atoms, say $f(X_i) = z_i$, and we define the integral of f as

$$\int f d\mu = \sum_{i=1}^n z_i \mu(X_i) = \sum_{i=1}^n f(X_i)\mu(X_i).$$

This in turn induces an inner product

$$\langle f, g\rangle = \int f\overline{g} d\mu = \sum_{i=1}^n f(X_i)\overline{g(X_i)}\mu(X_i).$$

An orthonormal basis for $L^2(X,\Sigma,\mu)$ is

$$\mathscr{B} = \left\{ \frac{1}{\sqrt{\mu(X_i)}} \mathbf{1}_{X_i} : i = 1,\dots,n \right\}.$$

Indeed,

$$\left\langle \frac{1}{\sqrt{\mu(X_j)}}\mathbf{1}_{X_j}, \frac{1}{\sqrt{\mu(X_k)}}\mathbf{1}_{X_k} \right\rangle = \sum_{i=1}^n \frac{1}{\sqrt{\mu(X_j)\mu(X_j)}} \mathbf{1}_{X_j}(X_i)\mathbf{1}_{X_k}(X_i)\mu(X_i)$$

and notice that $\mathbf{1}_{X_j}(X_i)\mathbf{1}_{X_k}(X_i) = 0$ unless $i = j = k$, in which case it is 1.

Corollary 5.1

Any function $f \in L^2(X,\Sigma,\mu)$ decomposes as a simple function of the form

$$f = \sum_{i=1}^n f(X_i)\mathbf{1}_{X_i}.$$

Proof. One can see this immediately by plugging in any $x \in X$. Indeed, if $x \in X_j$ then the left hand side is $f(x) = f(X_j)$ while the right hand side is

$$\sum_{i=1}^n f(X_i)\mathbf{1}_{X_i}(x_j) = f(X_j).$$

Alternatively, one can use that if $v_1,\dots,v_n$ form an orthonormal basis for V then any vector u can be expressed in the form

$$u = \sum_{i=1}^n \langle u, v_i\rangle v_i,$$

which in our case tells us that

$$f = \sum_{i=1}^{n} \left\langle f, \frac{\mathbf{1}_{X_i}}{\mu(X_i)^{1/2}} \right\rangle \frac{\mathbf{1}_{X_i}}{\mu(X_i)^{1/2}} = \sum_{i=1}^{n} \langle f, \mathbf{1}_{X_i} \rangle \frac{\mathbf{1}_{X_i}}{\mu(X_i)} = \sum_{i=1}^{n} \sum_{j=1}^{n} f(X_j) \mathbf{1}_{X_i}(X_j) \mu(X_j) \frac{\mathbf{1}_{X_i}}{\mu(X_i)}$$

and the inner summands all vanish but when $i = j$, giving

$$f = \sum_{i=1}^{n} f(X_i) \mathbf{1}_{X_i}.$$

□

It gets a bit tedious to have to recall the atoms of X every time we want to define an integral. We can avoid this by writing $\mu(x) = \mu(X_i)/|X_i|$ and then writing

$$\int f d\mu = \sum_{x} f(x)\mu(x).$$

Nothing is lost here, since both f and μ are constant on the atoms:

$$\sum_{x} f(x)\mu(x) = \sum_{i=1}^{n} \sum_{x \in X_i} f(x)\mu(x) = \sum_{i=1}^{n} \sum_{x \in X_i} f(X_i) \frac{\mu(X_i)}{|X_i|} = \sum_{i=1}^{n} f(X_i)\mu(X_i).$$

Recall that a *linear operator* is a function $T : V \to V$ which maps a vector space back into itself and is linear in the sense that $T(cu + v) = cT(u) + T(v)$.

Definition 5.6: Integral kernel operators

Suppose $K : X \times X \to \mathbb{C}$ is a function, and suppose that for each $x \in X$, the function $K_x(y) = K(x, y)$ is measurable, and that for each $y \in X$, the function $K^y(x) = K(x, y)$ is measurable. We define the integral kernel operator T_K on $L^2(X, \Sigma, \mu)$ by

$$[T_K f](x) = \int f K_x d\mu = \sum_{y \in X} f(y)K(x, y)\mu(y).$$

Strictly speaking, we need to show that the above is really an *operator* – namely that the output of the operator is also measurable.

Lemma 5.4

Suppose $K : X \times X \to \mathbb{C}$ is a function, and suppose that for each $x \in X$, the function $K_x(y) = K(x, y)$ is measurable, and that for each $y \in X$, the function $K^y(x) = K(x, y)$ is measurable.Then $T_K f$ is measurable when f is.

Proof. We need only to show that $T_K f$ is constant on atoms. So let x_1 and x_2 lie in a common atom. Thus, for each y, $K(x_1, y) = K^y(x_1) = K^y(x_2) = K(x_2, y)$ since

the function K^y is measurable. But then, for each y, $K_{x_1}(y) = K_{x_2}(y)$ and so the functions K_{x_1} and K_{x_2} are identical. Thus

$$[T_K f](x_1) = \int f K_{x_1} d\mu = \int f K_{x_2} d\mu = [T_K f](x_2).$$

□

Proposition 5.1

All linear operators on $L^2(X, \Sigma, \mu)$ are integral kernel operators.

Proof. Indeed, suppose T is a linear operator. Then for $x, y \in X$ with $x \in X_i$ and $y \in Y_j$, and set

$$K(x, y) = \frac{\langle T\mathbf{1}_{X_i}, \mathbf{1}_{X_j} \rangle}{\mu(X_i)\mu(X_j)}$$

which clearly satisfies the measurability conditions since it depends only on the atoms X_i and X_j. We claim $T = T_K$. Indeed, since K is constant on atoms, we have

$$K(X_i, X_j) = \frac{\langle T\mathbf{1}_{X_i}, \mathbf{1}_{X_j} \rangle}{\mu(X_i)\mu(X_j)} = \frac{1}{\mu(X_i)\mu(X_j)} \sum_{k=1}^{n} [T\mathbf{1}_{X_i}](X_k)\overline{\mathbf{1}_{X_j}(X_k)}\mu(X_k) = \frac{[T\mathbf{1}_{X_i}](X_j)}{\mu(X_i)}$$

and thus.

$$[T_K f](X_j) = \sum_{i=1}^{n} f(X_i)K(X_i, X_j)\mu(X_i) = \sum_{i=1}^{n} f(X_i)\frac{[T\mathbf{1}_{X_i}](X_j)}{\mu(X_i)}\mu(X_i) = \sum_{i=1}^{n} f(X_i)[T\mathbf{1}_{X_i}](X_j).$$

On the other hand, by linearity of T

$$[Tf](X_j) = \sum_{i=1}^{n} f(X_i)[T\mathbf{1}_{X_i}](X_j).$$

□

The fact that all operators have kernels, states as in the above theorem, does not extend to the case when X is no longer finite.

Theorem 5.3

Let I be the identity operator on $L^2([0,1], \mathscr{M}, m)$. Then I is not an integral kernel operator for any integrable function $K : [0,1] \times [0,1] \to \mathbb{C}$ with measurable slices.

Proof. Suppose there were a function $K : [0,1] \times [0,1] \to \mathbb{C}$ which acts as a kernel for I. Then for any L^2 function f,

$$f(x) = \int_0^1 f(y)K(x, y)dy$$

for almost every x. Thus if $f(x) = \mathbf{1}_{[0,t]}(x)$ for $t \in [0,1]$ then

$$\mathbf{1}_{[0,t]}(x) = \int_0^t K(x,y)\,dy$$

for almost all x, say on a set $A_t \subseteq [0,1]$ with $m(A_t) = 1$. It follows that if $h > 0$,

$$\frac{1}{h}\left(\mathbf{1}_{[0,t+h]}(x) - \mathbf{1}_{[0,t]}(x)\right) = \frac{1}{h}\int_t^{t+h} K(x,y)\,dy = \frac{1}{h}\int_t^{t+h} K_x(y)\,dy$$

holds for almost all x too, namely on $A_t \cap A_{t+h}$. By the Lebesgue differentiation theorem, as $h \to 0$,

$$\frac{1}{h}\int_t^{t+h} K_x(y)\,dy \to K_x(t)$$

holds for almost all t and x. To make this precise, we set $h = 1/n$ and

$$B_n = A_t \cap A_{t+1/2} \cap A_{t+1/3} \cap \ldots \cap A_{t+1/n}$$

so that $m(B_n) = 1$ for all n. Moreover B_n is decreasing and measurable so $B_t = \bigcap_n B_n$ has measure 1 too, and for $x \in B_t$ we have

$$\frac{1}{1/n}\left(\mathbf{1}_{[0,t+1/n]}(x) - \mathbf{1}_{[0,t]}(x)\right) = \frac{1}{1/n}\int_t^{t+1/n} K(x,y)\,dy = \frac{1}{1/n}\int_t^{t+1/n} K_x(y)\,dy.$$

Let $S = \{(x,t) : x \in B_t\}$ and for $x \in [0,1]$ let $S_x = \{t : (x,t) \in S\}$. Let us assume that for almost all x, S_x is measurable and has measure 1. Then By the Lebesgue differentiation theorem, there is a set D_x of measure 1 such that for $t \in D_x$,

$$\frac{1}{1/n}\int_t^{t+1/n} K_x(y)\,dy \to K_x(t).$$

On the other hand,

$$\mathbf{1}_{[0,t+h]}(x) - \mathbf{1}_{[0,t]}(x) = \mathbf{1}_{[t,t+h]}(x)$$

so as long as $x \neq t$, this tends to 0 as $h \to 0$. Hence, given $x \in B$, we know that for almost all $t \in [0,1]$, $K_x(t) = 0$. But then for such x,

$$f(x) = \int_0^1 f(y)K(x,y)\,dy = 0.$$

In other words, for almost all x, $f(x) = 0$ holds for every L^2 function f. This is obviously false.

The assumption on S_x is still left unjustified. We shall see that (later in the course) that it follows from the Fubini-Tonelli Theorem. □

Finally, given an operator, we want to know how it interacts with the inner product.

Definition 5.7: Adjoint operator

Let $K : X \times X \to \mathbb{C}$ define an integral kernel operator. Then the function $K^* : X \times X \to \mathbb{C}$ defined by

$$K^*(x,y) = \overline{K(y,x)}$$

is called the adjoint function of K, and T_{K^*} is called the adjoint operator of T_K.

Lemma 5.5

The adjoint satisfies $\langle T_K f, g\rangle = \langle f, T_{K^*} g\rangle$.

Proof. We have

$$\begin{aligned}
\langle T_K f, g\rangle &= \sum_{x\in X} [T_K f](x)\overline{g(x)}\mu(x) \\
&= \sum_{x\in X}\sum_{y\in X} f(y)K(x,y)\overline{g(x)}\mu(y)\mu(x) \\
&= \sum_{y\in X} f(y)\overline{\sum_{x\in X} K^*(y,x)g(x)\mu(x)}\mu(y) \\
&= \sum_{y\in X} f(y)\overline{[T_{K^*} g](y)}\mu(y) \\
&= \langle f, T_{K^*} g\rangle.
\end{aligned}$$

□

5.4 Some examples

Before moving on, we can endow a set X with various σ-algebras and measures. When these are understood, we might just write $L^2(X)$ as opposed to $L^2(X,\Sigma,\mu)$ for the sake of keeping notation brief.

5.4.1 Graphs

Our first example comes from combinatorics. Recall that a (simple, undirected) graph $G = (V,E)$ consists of a set V of vertices (not to be confused with vector space) and a set E of two-element subsets of V called the edges. Let $X = V$, $\Sigma = \mathscr{P}(V)$ and let $\mu(\{v\}) = \frac{1}{|V|}$ for any vertex v (so we are using the *uniform probability measure* on V). Then $L^2(V,\Sigma,\mu)$ just consists of all possible complex functions on V. This, so far, uses nothing of the graph structure of V. However, the is a very important operator, the *adjacency* operator $A : L^2(V,\Sigma,\mu) \to L^2(V,\Sigma,\mu)$ defined by

$$[Af](v) = \sum_{\{u,v\}\in E} f(u).$$

This operator replaces $f(v)$ by the "average" value of f at the neighbours of v in the graph. For instance, if $f = 1$ is the constant function, then

$$[A1](v) = \sum_{\{u,v\}\in E} 1 = \deg(v)$$

is the degree function. More generally, if $f = \mathbf{1}_U$ for some set U of vertices,

$$[A\mathbf{1}_U](v) = \sum_{\{u,v\}\in E} \mathbf{1}_U(u)$$

counts how many neighbours v has in the set U. In particular, if $U = \{u\}$, then $[A\mathbf{1}_U](v) = \mathbf{1}_E(\{u,v\})$ tests whether $\{u,v\}$ is an edge in the graph.

A particularly nice quality of A is its behaviour under iteration.

Lemma 5.6

Let $u \in V$ be a vertex and let $n \geq 0$ be an integer. Then $[A^n\mathbf{1}_{\{u\}}](v)$ counts the number of paths of length n from u to v in the graph.

Proof. If $n = 0$ then A^n is the identity, and so $[A^n\mathbf{1}_{\{u\}}](v) = \mathbf{1}_{\{u\}}(v)$ just tests whether $u = v$, which is the same as counting paths of length 0 (not moving) between u and v. We have seen that $[A\mathbf{1}_{\{u\}}](v)$ tests whether or not $\{u,v\}$ is an edge, which is the same as counting paths of length 1. Now we proceed by induction on n.

$$[A^{(n+1)}\mathbf{1}_{\{u\}}](v) = [A(A^n\mathbf{1}_{\{u\}})](v) = \sum_{\{w,v\}\in E} [A^n\mathbf{1}_{\{u\}}](w).$$

Now $[A^n\mathbf{1}_{\{u\}}](w)$ counts the number of paths of length n from u to w. But then, if $\{w,v\}$ is an edge, any path of length n from u to w can be extended to a path of length $n+1$ from u to v by traversing this edge. Conversely, any path of length $n+1$ from u to v has some penultimate vertex w so that $\{w,v\}$ is an edge and we have traversed n edges to get from u to w. □

5.4.2 Finite cyclic groups

Let $N \geq 2$ be an integer and let $\mathbb{Z}/N\mathbb{Z}$ be the cyclic group of integers modulo N. We set $X = \mathbb{Z}/N\mathbb{Z}$, $\Sigma = \mathscr{P}(\mathbb{Z}/N\mathbb{Z})$ and $\mu(n) = \frac{1}{N}$ for any residue class n. Consider the kernel $K : \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$ defined by

$$K(m,n) = e(-nm/N).$$

Then the associated integral transform satisfies

$$[T_K(f)](m) = \frac{1}{N} \sum_{n\in\mathbb{Z}/N\mathbb{Z}} f(n)e(-mn/N)$$

and is called the *discrete Fourier transform*. It is so significant that, rather than write $T_K f$, we write

$$\widehat{f}(m) = \frac{1}{N} \sum_{n \in \mathbb{Z}/N\mathbb{Z}} f(n) e(-mn/N).$$

The adjoint function $K^*(m,n)$ is defined to be $K^*(m,n) = e(mn/N)$. In fact, we claim that the adjoint satisfies $T_{K^*} = \frac{1}{N} T_K^{-1}$. Indeed,

$$[T_K^*[T_K f]](m) = \frac{1}{N} \sum_{n \in \mathbb{Z}/N\mathbb{Z}} \widehat{f}(n) e(mn/N) = \frac{1}{N^2} \sum_{n \in \mathbb{Z}/N\mathbb{Z}} \sum_{l \in \mathbb{Z}/N\mathbb{Z}} f(l) e(-nl/N) e(mn/N).$$

Interchanging the order of summation, we get

$$\frac{1}{N^2} \sum_{l \in \mathbb{Z}/N\mathbb{Z}} f(l) \sum_{n \in \mathbb{Z}/N\mathbb{Z}} e(n(l-m)/N).$$

Then innermost sum, upon unraveling notation, is

$$\sum_{n=1}^{N} (e^{2\pi i \frac{l-m}{N}})^n = \begin{cases} N & l = m \\ 0 & l \neq m. \end{cases}$$

Thus we are left with a factor of N, but only for the summand $l = m$ and we get

$$[T_K^*[T_K f]](m) = \frac{f(m)}{N} = \frac{1}{N}[T_K^{-1}[T_K f]](m).$$

The inversion identity in the other direction is similar. Thus we have in fact shown the *Fourier inversion formula*

$$f(m) = \sum_{n \in \mathbb{Z}/N\mathbb{Z}} \widehat{f}(n) e(mn/N).$$

Suppose instead that $g : \mathbb{Z}/N\mathbb{Z}$ is another function and define the *convolve with* g operator whose kernel is $K_g(m,n) = g(m-n)$. Thus

$$[T_{K_g} f](m) = \frac{1}{N} \sum_{n \in \mathbb{Z}/n\mathbb{Z}} f(n) g(m-n).$$

This operator is so significant that we instead write $[f * g](m)$ or just $f * g(m)$.

The convolution operation will play a hefty part in this course, as will the Fourier transform. This is because they are linked by the following beautiful identity.

Theorem 5.4: Convolution-to-product

Let f and g be functions in $L^2(\mathbb{Z}/N\mathbb{Z}, \Sigma, \mu)$. Then

$$\widehat{f * g}(n) = \widehat{f}(n)\widehat{g}(n).$$

Proof. By definition

$$\widehat{f*g}(n) = \frac{1}{N}\sum_{m\in\mathbb{Z}/N\mathbb{Z}} f*g(m)e(-nm/N),$$

and expanding out the definition of $f*g$, we get

$$\frac{1}{N}\sum_{m\in\mathbb{Z}/N\mathbb{Z}}\frac{1}{N}\sum_{l\in\mathbb{Z}/N\mathbb{Z}} f(l)g(m-l)e(-nm/N).$$

Now we interchange the order of summation to get

$$\frac{1}{N}\sum_{l\in\mathbb{Z}/N\mathbb{Z}} f(l)\frac{1}{N}\sum_{m\in\mathbb{Z}/N\mathbb{Z}} g(m-l)e(-nm/N).$$

Let's focus on the inner sum. It ranges over all residue classes $m\in\mathbb{Z}/N\mathbb{Z}$. There are terms that involve l, but if we only look at the inner sum, we can consider l as being fixed. Now we make a change of variable: $u=m-l$, so that $m=l+u$. As m ranges over all residues in $\mathbb{Z}/N\mathbb{Z}$, so does u. In these new variables

$$\begin{aligned}\frac{1}{N}\sum_{m\in\mathbb{Z}/N\mathbb{Z}} g(m-l)e(-nm/N) &= \frac{1}{N}\sum_{u\in\mathbb{Z}/N\mathbb{Z}} g(u)e(-n(l+u)/N)\\ &= e(-nl/N)\frac{1}{N}\sum_{u\in\mathbb{Z}/N\mathbb{Z}} g(u)e(-nu/N)\end{aligned}$$

and this is nothing more than $e(-nl/N)\widehat{g}(n)$. So if we plug that in, we get

$$\widehat{f*g}(n) = \frac{1}{N}\sum_{l\in\mathbb{Z}/N\mathbb{Z}} f(l)e(-nl/N)\widehat{g}(n) = \widehat{f}(n)\widehat{g}(n).$$

□

The last example of an interesting operator is the first one where we'll explore the role of the σ-algebra. If $\Sigma_1\subseteq\Sigma_2$ are σ-algebras, then the atoms of Σ_1 need to be unions of those from Σ_2. In other words, the partition of X into the atoms of Σ_1 is a *coarser* partition than the partition into the Σ_2-atoms.

Lemma 5.7

Suppose $\Sigma_1\subseteq\Sigma_2$ are σ-algebras on X with $X=\bigsqcup_{i=1}^{N} X_i$ being the partition into atoms of Σ_1 then the atoms of Σ_2 come from a further partition $X_i=\bigsqcup_{j=1}^{n_j} X_{i,j}$. The space $L^2(X,\Sigma_1,\mu)$ is a linear subspace of $L^2(X,\Sigma_2,\mu)$.

Proof. The only part not already observed is the final claim. But a function belongs to $L^2(X,\Sigma_1,\mu)$ if any only if it is constant on the atoms X_i, in which case it is certainly constant on the smaller atoms $X_{i,j}\subseteq X_i$ from Σ_2. Thus as sets, $L^2(X,\Sigma_1,\mu)\subseteq$

$L^2(X,\Sigma_2,\mu)$. But if $f,g \in L^2(X,\Sigma_1,\mu)$ are constant on atoms X_i, then for any constant c and $x,y \in X_i$,

$$[f+cg](x) = f(x)+cg(x) = f(y)+cg(y) = [f+cg](y)$$

so $f+cg$ is constant on X_i too. □

Recall that if W is a subspace of a vector space V then

$$W^\perp = \{v : \langle w, v\rangle = 0 \text{ for all } w \in W\}$$

is called the orthogonal complement of W and each vector $v \in V$ can be written uniquely as

$$v = \text{proj}_W(v) + (w - \text{proj}_W(v))$$

where $\text{proj}_W(v) \in W$ is the *orthogonal projection* of v onto W and $w - \text{proj}_W(v) \in W^\perp$.

Definition 5.8: Conditional expectation

If we set $V = L^2(X,\Sigma_2,\mu)$ and $W = L^2(X,\Sigma_1,\mu)$ where $\Sigma_1 \subseteq \Sigma_2$ are σ-algebras on X then the Σ_1-measurable function $\text{proj}_W f$ is now denoted by $\mathbb{E}(f|\Sigma_1)$ and is called the *conditional expectation* of f given Σ_1.

Lemma 5.8

Let $f \in L^2(X,\Sigma_2,\mu)$ and suppose $\Sigma_1 \subseteq \Sigma_2$ are σ-algebras on X. Then for $A \in \Sigma_1$, $\mathbb{E}(f|\Sigma_1)$ has the property that

$$\int_A \mathbb{E}(f|\Sigma_1)\,d\mu = \int_A f\,d\mu.$$

In particular, the value of $\mathbb{E}(f|\Sigma_1)$ on any atom X_i is

$$[\mathbb{E}(f|\Sigma_1)](X_i) = \frac{1}{\mu(X_i)}\int_{X_i} f\,d\mu.$$

Proof. We *define* $g \in L^2(X,\Sigma_1,\mu)$ by the rule

$$g(x) = \frac{1}{\mu(X_i)}\int_{X_i} f\,d\mu$$

for $x \in X_i$ and $i = 1,\ldots,N$, which is then constant on atoms by definition. Then if $A = X_{i_1} \cup \cdots X_{i_m}$ is a union of atoms from Σ_1, then g is a simple function, and by definition,

$$\int_A g\,d\mu = \sum_{k=1}^m g(X_{i_k})\mu(X_{i_k}) = \sum_{k=1}^m \frac{1}{\mu(X_{i_k})}\left(\int_{X_{i_k}} f\,d\mu\right)\mu(X_{i_k}) = \sum_{k=1}^m \int_{X_{i_k}} f\,d\mu = \int_A f\,d\mu.$$

So $g \in L^2(X, \Sigma_1, \mu)$ has the desired property, and we need only show that $g = \mathbb{E}(f|\Sigma_1)$, which in turn will be true if $\langle f - g, h\rangle = 0$ for any $h \in L^2(X, \Sigma_1, \mu)$, because $f = \mathbb{E}(f|\Sigma_1) + \big(f - \mathbb{E}(f|\Sigma_1)\big)$ is a *unique* decomposition into something from $L^2(X, \Sigma_1, \mu)$ and its complement. It is then enough to show that $\langle f - g, \mathbf{1}_{X_i}\rangle = 0$ since the indicator functions $\mathbf{1}_{X_i}$ span $L^2(X, \Sigma_1, \mu)$. Now

$$\langle f - g, \mathbf{1}_{X_i}\rangle = \int_{X_i} f - g\, d\mu = \int_{X_i} f\, d\mu - \int_{X_i} g\, d\mu = \int_{X_i} f\, d\mu - \int_{X_i} f\, d\mu = 0$$

where the second-to-last equality uses that $X_i \in \Sigma_1$ and the property we have already shown g to have. □

Example. *Suppose $C_1, \ldots, C_N$ are $\{1, -1\}$ independently at random with $\mathbb{P}(C_i = 1) = \mathbb{P}(C_i = -1) = \frac{1}{2}$. The underlying probability space is $X = \{1, -1\}^N$, $\Sigma_2 = \mathscr{P}(X)$, and the measure of any vector is $1/2^N$. We think of this space as the result of betting a dollar on a coin flip, the game being played N times. Suppose we have played the first $N-1$ games, so we know the results of those flips, but the N'th flip remains to be determined. We might represent this as $(C_1, \ldots, C_{N-1}, ?)$. If one were to ask us about the results, we can tell them anything we like about the first $N-1$ flips, but not about the last. This can be encoded by the σ-algebra Σ_1, whose atoms have the form*

$$\{(C_1, \ldots, C_{N-1}, 1), (C_1, \ldots, C_{N-1}, -1)\}.$$

This is because, by locking the last entries 1 and -1 into a common atom, we cannot distinguish between them, and so are uncertain as to the the final outcome. A function $f \in L^2(X, \Sigma_2, \mu)$ associates a number to the result of the N games. For instance, our winnings (or debts) are

$$f(C_1, \ldots, C_N) = C_1 + \ldots + C_N.$$

The value of $\mathbb{E}(f|\Sigma_1)$ represents what we expect f to be given what we know from Σ_1 (which is what we've learned from the first $N-1$ games). According to the preceding lemma, we have

$$\begin{aligned}[\mathbb{E}(f|\Sigma_1)](C_1, \ldots, C_{N-1}, ?) &= 2^{N-1}\left(f(C_1, \ldots, C_{N-1}, 1)\frac{1}{2^N} + f(C_1, \ldots, C_{N-1}, -1)\frac{1}{2^N}\right) \\ &= \frac{f(C_1, \ldots, C_{N-1}, 1) + f(C_1, \ldots, C_{N-1}, -1)}{2}\end{aligned}$$

and so again, if f is our winnings, then

$$[\mathbb{E}(f|\Sigma_1)](C_1, \ldots, C_{N-1}, ?) = \frac{(C_1 + \cdots + C_{N-1} + 1) + (C_1 + \cdots + C_{N-1} - 1)}{2} = C_1 + \cdots + C_{N-1}$$

is is our best guess at our winnings after N games given what we know for the first $N-1$ games – they don't change. Of course, they will go up or down after the N'th game, but each is equally likely.

6

FUNCTIONAL ANALYSIS

Functional analysis is linear algebra from a quantitative perspective that allows us to extend ideas from finite dimensional spaces to infinite dimensional spaces. There is an emphasis on approximation. We will focus primarily on finite dimensional spaces, but the flavour will be very quantitative.

6.1 The spectral theorem

Let T be a linear operator on a complex inner product space V, of finite dimension. Then we call T self-adjoint if $T = T^*$.

Our goal is to prove the following.

Theorem 6.1: Spectral Theorem

Suppose T is a self-adjoint operator on V. Then T admits and orthonormal basis of eigenvectors each with a real eigenvalue.

In the context of $L^2(X, \Sigma, \mu)$, we call eigenvectors *eigenfunctions* and we have

$$\lambda_f f(x) = [T_K f](x) = \sum_{j=1}^{n} f(X_j) K(X_i, X_j) \mu(X_j).$$

Lemma 6.1

Any operator T has an eigenvector.

Proof. The characteristic polynomial $\det(T - \lambda I)$ is a complex polynomial in λ and hence has a root $\lambda \in \mathbb{C}$, which is to say, $T - \lambda I$ is not invertible. Thus there is some vector $v \in \mathrm{Ker}(T - \lambda I)$ which is an eigenvector. □

Theorem 6.2: Schur decomposition

If V is a complex inner product space and T is a linear operator on V, then V admits an upper-triangular basis. That is, there is a basis $\{v_1, \ldots, v_n\}$ of V such that

$$T(v_k) \in \mathrm{Span}\{v_1, \ldots, v_k\}.$$

Furthermore, this basis can be taken to be orthonormal.

Proof. We proceed by induction on n, the case $n = 1$ being trivial.

Let v_1 be a an eigenvector for T with eigenvalue λ, normalized so that $\|v_1\| = 1$, and let $E = v_1^{\perp}$. We define $T' : E \to E$ by

$$T'(u) = T(u) - \langle T(u), v_1 \rangle v_1.$$

Then T' is linear and

$$\langle T'(u), v_1 \rangle = \langle T(u) - \langle T(u), v_1 \rangle v_1, v_1 \rangle = \langle T(u), v_1 \rangle - \langle T(u), v_1 \rangle \|v_1\|^2 = 0$$

so that T' really does map E to E. Moreover E has dimension $n-1$ so by induction, we know that there is a basis $\{v_2, \ldots, v_n\} \subseteq V$ such that $T'(v_k) \subseteq \mathrm{Span}\{v_2, \ldots, v_k\}$. It follows that

$$T(v_k) = \langle T(v_k), v_1 \rangle v_1 + T'(v_i) \subseteq \mathrm{Span}\{v_1, v_2, \ldots, v_k\}.$$

For the claim of orthonormality, recall that the Gram-Schmidt process takes the basis $\{v_1, \ldots, v_n\}$ and iteratively replaces v_k with u_k defined by the rule

$$u_1 = v_1$$

and

$$u_{k+1} = v_{k+1} - \sum_{j=1}^{k} \frac{\langle v_{k+1}, u_j \rangle}{\langle u_j, u_j \rangle} u_j$$

so that $\mathrm{Span}\{u_1, \ldots, u_k\} = \mathrm{Span}\{v_1, \ldots, v_k\}$. The resulting vectors u_k are orthogonal and still upper-triangular, by induction:

$$T(u_k) = T(v_k) - \sum_{j=1}^{k-1} \frac{\langle v_k, u_j \rangle}{\langle u_j, u_j \rangle} T(u_j) \in \mathrm{Span}\{v_1, \ldots, v_k\} + \mathrm{Span}\{u_1, \ldots, u_{k-1}\} = \mathrm{Span}\{u_1, \ldots, u_k\}.$$

Finally, we renormalize to make everything orthonormal. □

Proof of the Spectral Theorem. Suppose T is self-adjoint, and let $\{v_1, \ldots, v_n\}$ be the basis from the Schur decomposition. We can further assume that these vectors are orthonormal. We show by induction that v_k is an eigenvector, the case $k = 1$ being immediate. For the induction step

$$T(v_{k+1}) = \sum_{j=1}^{k+1} \langle T(v_{k+1}), v_j \rangle v_j = \sum_{j=1}^{k+1} \langle v_{k+1}, T(v_j) \rangle v_j = \sum_{j=1}^{k} \overline{\lambda_j} \langle v_{k+1}, v_j \rangle v_j + \langle v_{k+1}, T(v_{k+1}) \rangle v_{k+1}$$

and the sum vanishes by orthogonality. Thus

$$T(v_{k+1}) = \langle v_{k+1}, T(v_{k+1}) \rangle v_{k+1}$$

which is to say that v_{k+1} is an eigenvector too.

Finally if v is an eigenvector (normalized to have length 1)

$$\lambda = \langle \lambda v, v \rangle = \langle Tv, v \rangle = \langle v, Tv \rangle = \langle v, \lambda v \rangle = \overline{\lambda},$$

so that $\lambda \in \mathbb{R}$. □

6.2 Operator norms

Let V and W be normed vector spaces with respective norms $\|\cdot\|_V$ and $\|\cdot\|_W$. Then, for a linear map $T : V \to W$ we define $\|T\| = \sup_{v \neq 0} \frac{\|T(v)\|_W}{\|v\|_V}$. We say T is bounded if this number is finite.

Lemma 6.2

The linear map $T : V \to W$ is continuous if and only if it is bounded.

Proof. Suppose T is bounded. Then if $v_1 \neq v_2$,

$$\|T(v_1) - T(v_2)\|_W = \|T(v_1 - v_2)\|_W \leq \frac{\|T(v_1 - v_2)\|_W}{\|v_1 - v_2\|_V} \|v_1 - v_2\|_V \leq \|T\| \|v_1 - v_2\|_V$$

and so if $\varepsilon > 0$ we set $\delta = \varepsilon / \|T\|$ to establish continuity. Conversely, if T is continuous, then continuity at 0_V tells us that there is a δ such that if $\|v\|_V \leq \delta$ then $\|T(v)\|_W \leq 1$. Hence for any vector v

$$\frac{\delta}{\|v\|_V} \|T(v)\|_W = \left\| T\left(\frac{\delta}{\|v\|_V} v \right) \right\|_W \leq 1$$

so that

$$\|T\| \leq \frac{1}{\delta}.$$

□

Eigenvalues are particularly helpful with understanding $\|T\|$ when working with the L^2 norm.

Lemma 6.3

Suppose $T : V \to V$ is a self-adjoint linear operator with an orthonormal basis $\{v_1, \ldots, v_n\}$ of eigenvectors, say $T(v_i) = \lambda_i v_i$. Then $\|T\| = \max_i |\lambda_i|$

Proof. We have

$$T(v) = T\left(\sum_{i=1}^{n} \langle v, v_i \rangle v_i\right) = \sum_{i=1}^{n} \langle v, v_i \rangle \lambda_i v_i$$

so

$$\|T(v)\|_{L^2} = \left(\sum_{i=1}^{n} |\lambda_i|^2 |\langle v, v_i \rangle|^2\right)^{1/2} \leq \max_i |\lambda_i| \left(\sum_{i=1}^{n} |\langle v, v_i \rangle|^2\right)^{1/2} = \max_i |\lambda_i| \|v\|_{L^2}$$

and we note that equality holds if $v = v_j$ where $|\lambda_j| = \max |\lambda_i|$. □

7 APPLICATIONS

7.1 The expander mixing lemma

Let G be a graph with vertex-set V and edge-set E. The adjacency operator of G is self-adjoint and so has an orthonormal basis of eigenfunctions $f_\lambda : V \to \mathbb{C}$ with real eigenvalues λ. Because there are $|V| = \dim L^2(V, \mathscr{P}(V), 1/|V|)$ such functions, we can order the eigenvalues as $\lambda_{|V|} \le \cdots \le \lambda_1$. Recall that the degree function, $\deg : V \to \mathbb{Z}$ is defined as

$$\deg(v) = |\{u \in V : \{u, v\} \in E\}|.$$

We say G is d-regular if $\deg(v) = d$ for all $v \in V$.

Lemma 7.1

If G is a d-regular graph then $\mathbf{1}_V$ is an eigenfunction for the adjacency operator A with eigenvalue d. All other eigenvalues λ of A satisfy $|\lambda| \le d$.

Proof. We have

$$[A\mathbf{1}_V](v) = \sum_{u:\{u,v\}\in E} \mathbf{1}_V(u) = d.$$

On the other hand, if f_λ is any other eigenfunction then

$$|\lambda f_\lambda(v)| = |[Af_\lambda](v)| \le \sum_{u:\{u,v\}\in E} |f_\lambda(u)| \le d \max_u |f_\lambda(u)|,$$

so choosing v such that $|f_\lambda(v)|$ is as big as possible, we get

$$|\lambda f_\lambda(v)| \le d|f_\lambda(v)|.$$

Since f_λ is an eigenfunction, it is non-zero, and so $f_\lambda(v) \neq 0$ and the proof is complete. □

For $X, Y \subseteq V$ we define

$$E(X,Y) = |\{\{u,v\} \in E : u \in X,\ v \in Y\}|$$

to be the number of edges with one endpoint in X and the other in Y.

Theorem 7.1: Expander mixing lemma

If G is a d-regular graph and each eigenvalue λ of A, other than d, satisfies $|\lambda| \le T$. Then

$$\left| E(X,Y) - \frac{d}{|V|}|X||Y| \right| \le T\sqrt{|X||Y|}.$$

Corollary 7.1

If G is a d-regular graph and each eigenvalue λ of A, other than d, satisfies $|\lambda| \le T$. If $X, Y \subseteq V$ are such that $\frac{d}{|V|}|X||Y| > T\sqrt{|X||Y|}$ then there is an edge from X to Y. In particular, if $d > \sqrt{T|V|}$ then every $v \in V$ belongs to a triangle.

Proof. By the expander mixing lemma

$$E(X,Y) \ge \frac{d}{|V|}|X||Y| - T\sqrt{|X||Y|} > 0.$$

Thus there is an edge from X to Y. For any vertex v, let $N(v)$ denote the set of the d neighbours of v in G. Taking $X = Y = N(v)$ we get

$$E(N(v), N(v)) \ge \frac{d^3}{|V|} - Td = d\left(\frac{d^2}{|V|} - T\right) > 0$$

so there is an edge between some $x, y \in N(v)$. But then $\{v,x\}, \{x,y\}, \{y,v\}$ makes a triangle. □

Proof of the Expander Mixing Lemma. Let X and Y be the sets in question. Then

$$E(X,Y) = \sum_{\{u,v\}\in E} \mathbf{1}_X(u)\mathbf{1}_Y(v) = \sum_{v\in V}\left(\sum_{u\in V:\{u,v\}\in E} \mathbf{1}_X(u)\right)\mathbf{1}_Y(v) = |V| \cdot \langle A\mathbf{1}_X, \mathbf{1}_Y\rangle.$$

Now we estimate the inner product by expanding $A\mathbf{1}_X$ and $\mathbf{1}_Y$ according to the basis of eigenfunctions f_λ of A. Recall that one such eigenfunction is $\mathbf{1}_V$ and its eigenvalue is d. So, since our eigenfunctions are orthonormal,

$$\mathbf{1}_Y = \sum_\lambda \langle \mathbf{1}_Y, f_\lambda \rangle f_\lambda = \langle \mathbf{1}_Y, \mathbf{1}_V \rangle \mathbf{1}_V + \sum_{\lambda \neq d} \langle \mathbf{1}_Y, f_\lambda \rangle f_\lambda,$$

where in the last step we merely pulled out the term corresponding to $\lambda = d$. Now

$$\langle \mathbf{1}_Y, \mathbf{1}_V \rangle = \frac{1}{|V|} \sum_v \mathbf{1}_Y(v) \mathbf{1}_V(v) = \frac{|Y|}{|V|}.$$

Thus we have

$$\mathbf{1}_Y = \frac{|Y|}{|V|} \mathbf{1}_V + \sum_{\lambda \neq d} \langle \mathbf{1}_Y, f_\lambda \rangle f_\lambda.$$

Now we apply the same procedure to $A\mathbf{1}_X$:

$$A\mathbf{1}_X = \sum_\lambda \langle A\mathbf{1}_X, f_\lambda \rangle f_\lambda = \sum_\lambda \langle \mathbf{1}_X, Af_\lambda \rangle f_\lambda = \sum_\lambda \langle \mathbf{1}_X, \lambda f_\lambda \rangle f_\lambda = \sum_\lambda \overline{\lambda} \langle \mathbf{1}_X, f_\lambda \rangle f_\lambda,$$

using self-adjointness and the fact that $Af_\lambda = \lambda f_\lambda$. Again we extract the contribution from $\lambda = d$ (which comes with a factor d this time), to get

$$A\mathbf{1}_X = \frac{d|X|}{|V|} \mathbf{1}_V + \sum_{\lambda \neq d} \overline{\lambda} \langle \mathbf{1}_X, f_\lambda \rangle f_\lambda.$$

At this point we can take inner products and use orthonormality again to get

$$\langle A\mathbf{1}_X, \mathbf{1}_Y \rangle = \frac{d|X||Y|}{|V|^2} + \sum_{\lambda \neq d} \overline{\lambda} \langle \mathbf{1}_X, f_\lambda \rangle \overline{\langle \mathbf{1}_Y, f_\lambda \rangle}.$$

Thus

$$\left| E(X,Y) - \frac{d|X||Y|}{|V|} \right| = \left| |V| \langle A\mathbf{1}_X, \mathbf{1}_Y \rangle - \frac{d|X||Y|}{|V|} \right| \leq |V| \left| \sum_{\lambda \neq d} \overline{\lambda} \langle \mathbf{1}_X, f_\lambda \rangle \overline{\langle \mathbf{1}_Y, f_\lambda \rangle} \right|$$

and by the triangle inequality, the right hand side is at most

$$|V| \sum_{\lambda \neq d} |\lambda| |\langle \mathbf{1}_X, f_\lambda \rangle| |\langle \mathbf{1}_Y, f_\lambda \rangle|.$$

Each instance of $|\lambda|$ is, by our hypothesis, at most T. So using this, and the Cauchy-Schwarz inequality,

$$\begin{aligned} |V| \sum_{\lambda \neq d} |\lambda| |\langle \mathbf{1}_X, f_\lambda \rangle| |\langle \mathbf{1}_Y, f_\lambda \rangle| &\leq T|V| \sum_\lambda |\langle \mathbf{1}_X, f_\lambda \rangle| |\langle \mathbf{1}_Y, f_\lambda \rangle| \\ &\leq T|V| \left(\sum_\lambda |\langle \mathbf{1}_X, f_\lambda \rangle|^2 \right)^{1/2} \left(\sum_\lambda |\langle \mathbf{1}_Y, f_\lambda \rangle|^2 \right)^{1/2}. \end{aligned}$$

Since the L^2-norm of the coefficients of a function f expanded in terms of any orthonormal basis is always the same (Parseval's identity) we have

$$\sum_\lambda |\langle \mathbf{1}_X, f_\lambda\rangle|^2 = \sum_{v\in V} |\langle \mathbf{1}_X, |V|^{1/2}\mathbf{1}_{\{v\}}\rangle|^2.$$

Here we are using that $|V|^{1/2}\mathbf{1}_{\{v\}}$ also forms an orthonormal basis, and next we observe that

$$\langle \mathbf{1}_X, |V|^{1/2}\mathbf{1}_{\{v\}}\rangle = \frac{1}{|V|}\sum_{u\in V}\mathbf{1}_X(u)|V|^{1/2}\mathbf{1}_{\{v\}}(u) = \frac{\mathbf{1}_X(v)}{|V|^{1/2}}.$$

From this,

$$\sum_{v\in V} |\langle \mathbf{1}_X, |V|^{1/2}\mathbf{1}_{\{v\}}\rangle|^2 = \sum_{v\in V}\frac{\mathbf{1}_X(v)}{|V|} = \frac{|X|}{|V|}.$$

The same calculation applies with Y in place of X and the proof is complete. □

7.2 Cayley graphs, Paley graphs, and sums and products

The following will work with any abelian group, but we'll stick to $\mathbb{F}_p$, the residue classes mod p, with addition. Let $S \subseteq \mathbb{F}_p$ be a set with $0 \in S$ and $s \in S \implies -s \in S$. A Cayley graph is one whose vertices are $\mathbb{F}_p$ and whose edges are all pairs the form $\{x, x+s\}$.

Lemma 7.2

A Cayley graph is d-regular where $d = |S|$. The eigenfunctions of the adjacency operator are the functions ψ_k defined as $\psi_k(x) = e(kx/p)$ where $k \in \{0,\dots,p-1\}$, and the eigenvalues are

$$\sum_{s\in S} e(ks/p) = p\widehat{\mathbf{1}_S}(-k).$$

Proof. The edges eminating from $x \in \mathbb{F}_p$ all have the form $\{x, x+s\}$, and there is exactly one such edge for $s \in S$. For the second claim,

$$[A\psi_k](x) = \sum_{\{x,y\}\in E}\psi_k(y) = \sum_{s\in S}\psi_k(x+s) = \psi_k(x)\sum_{s\in S}\psi_k(s).$$

There are $p = \dim L^2(\mathbb{F}_p, \mathscr{P}(\mathbb{F}_p), 1/p)$ functions pf the form ψ_k, and each is an eigenfunction, so we have accounted for all of them. □

If $p = 1 \pmod 4$ then the set of quadratic residues $S = \{x^2 : x \in \mathbb{F}_p,\ x \neq 0\}$ is a symmetric set of size $(p-1)/2$. Forming the Cayley graph with this particular set S makes a graph with a special name: the Paley graph mod p. The eigenvalues of this graph are of the form

$$\sum_{s\in S} e(ks/p) = \frac{1}{2}\left(\sum_{x\in\mathbb{F}_p} e(kx^2/p) - 1\right).$$

The above indentity is because we need to extract $x = 0$ from the sum on the right ($0 \notin S$) and every square $s \in S$ gets counted twice – as x^2 and $(-x)^2$.

Theorem 7.2: Gauss' sum

For $k \in \mathbb{F}_p$ with $k \neq 0$, we have

$$\left| \sum_{x \in \mathbb{F}_p} e(kx^2/p) \right| = \sqrt{p}.$$

As a consequence, the eigenvalues of the Payley graph, other than $(p-1)/2$, all have size at most

$$\left| \frac{1}{2} \left(\sum_{x \in \mathbb{F}_p} e(kx^2/p) - 1 \right) \right| \leq \frac{1}{2}(\sqrt{p}+1) \leq \sqrt{p}.$$

Corollary 7.2: Expander mixing lemma for the Paley graph

If $X, Y \subseteq \mathbb{F}_p$ and

$$E(X,Y) = |\{\{x,y\} : x \in X,\ y \in Y\ x - y \in S\}|,$$

then

$$\left| E(X,Y) - \frac{|X||Y|}{2} \right| \leq 2\sqrt{p|X||Y|}.$$

In particular, if $|X||Y| > 16p$ then there is an edge from X to Y.

Proof. We can apply the expander mixing lemma with $d = \frac{p-1}{2}$ and $T = \sqrt{p}$. We get

$$\left| E(X,Y) - |X||Y| \frac{\frac{p-1}{2}}{p} \right| \leq \sqrt{p|X||Y|}.$$

To complete the proof we need only note that

$$\left| E(X,Y) - \frac{|X||Y|}{2} \right| \leq \left| E(X,Y) - |X||Y| \frac{\frac{p-1}{2}}{p} \right| + \frac{|X||Y|}{2p}.$$

But since $|X| \leq p$ and $|Y| \leq p$,

$$\frac{|X||Y|}{2p} \leq \sqrt{p|X||Y|}.$$

For the final conclusion, we observe that

$$E(X,Y) \geq \frac{|X||Y|}{2} - 2\sqrt{p|X||Y|} = \sqrt{|X||Y|}\left(\frac{\sqrt{|X||Y|}}{2} - 2\sqrt{p} \right)$$

and the right hand side is positive since $|X||Y| > 16p$. □

Theorem 7.3

Let $A \subseteq \mathbb{F}_p$ be a set of size at least $10\sqrt{p}$. Then there are two elements $x, y \in \mathbb{F}_p$ such that $x + y \in A$ and $xy \in A$.

Proof. For $a, b \in A$, we consider the polynomial $t^2 - at + b$. If this polynomial can be factored, we get

$$t^2 - at + b = (t - x)(t - y) = t^2 - (x + y)t + (xy)$$

which, upon comparing coefficients, gives $a = x + y$ and $b = xy$. So to prove the theorem, it is enough to show that for some $a, b \in A$, the polynomial $t^2 - at + b$ factors. By the quadratic formula, this amounts to showing that the discriminant $a^2 - 4b$ is a square. In other words, we want to show that there is an edge between $X = \{a^2 : a \in A\}$ and $Y = \{4b : b \in A\}$ in the Paley graph. Now $|X| \geq |A|/2$ since squaring is at most 2-to-1, and $|Y| = |A|$ since the map $b \mapsto 4b$ is invertible. So $|X||Y| \geq |A|^2/4 \geq 25p$ which is enough to guarantee an edge in the Paley graph. □

7.3 Roth's theorem

Recall that an arithmetic progression of length k is a sequence of k terms having the form $\{a, a + d, \ldots, a + (k - 1)d\}$. In 1953, Klaus Roth proved the following theorem.

Theorem 7.4: Roth

Let $r_3(N)$ denote the largest cardinality of a set $A \subseteq \{1, \ldots, N\}$ such that A contains no three distinct elements forming an arithmetic progression. Then $r_3(N)/N \to 0$ as $N \to \infty$.

In other words, given a proportion $\delta > 0$, and provided N is sufficiently large, then any subset $A \subseteq \{1, \ldots, N\}$ containing at least δN elements automatically contains three numbers in (non-trivial) arithmetic progression. Here non-trivial means we do not count the arithmetic progression (a, a, a). In this section, we present a proof of Roth's theorem due to Croot and Sisask. Let's write, for a set $A \subseteq \mathbb{Z}$, $T_3(A)$ for the number of 3-term progressions in A.

We begin with a lemma, more or less due to Varnavides, which states that once one has passed the threshold $r_3(N)$ needed to guarantee a 3-term progression, then there are in fact many – i.e. $T_3(A)$ is big.

Lemma 7.3: Varnavides

Let M and N be positive integers with $1 \le M \le N$. Then for $A \subseteq \{1,\dots,N\}$ we have

$$T_3(A) \ge \frac{N^2}{M^4}\left(\frac{|A|}{N} - \frac{r_3(M)+2}{M}\right).$$

Let's digest this a bit. Suppose we take N larger than M^4. Then the N^2/M^4 factor is at least N, and so we will deduce a positive lower bound for $T_3(A)$ provided $|A|/N$ is larger than $\frac{r_3(M)+1}{M}$, which is roughly the density needed for a set $B \subseteq \{1,\dots,M\}$ to contain a 3-term progression. So if the density of A exceeds the threshold density that works for M, then we'll have lots of 3-term progressions.

Proof. For a positive integer k, let $\mathbf{AP}(N,M,k)$ denote the set of all M-term arithmetic progressions in $\{1,\dots,N\}$ with step at most k. In other words, $\mathbf{AP}(N,M,k)$ denotes the number of sequences of the form

$$a, a+d, \dots, a+(M-1)d$$

where $1 \le d \le k$ and $1 \le a \le a+(M-1)d \le N$. We warm up by counting the number of progressions with a fixed step d. This amounts to choosing the starting point a, since once a has been decided, and since d is fixed, we have no other choices to make. The only thing we need to bear in mind when choosing a is that we complete the full M term progression before hitting N, which is to say $a+(M-1)d \le N$, or

$$1 \le a \le N - \frac{M-1}{d}.$$

Meanwhile any a in this range will do, so we have about $N-(M-1)/d$ progressions of step d.

Next, given a fixed 3-term progression $b, b+e, b+2e$, how many progressions from $\mathbf{AP}(N,M,k)$ can contain $\{b, b+e, b+2e\}$? Suppose that

$$\{b, b+e, b+2e\} \subseteq \{a, a+d, \dots, a+(M-1)d\}.$$

Then d divides e, since

$$e = (b+e) - b = (a+md) - (a+m'd) = d(m-m')$$

for some m, m'. So write $e = dd'$. At the same time

$$2e = (b+2e) - (b) \le (a+(M-1)d) - a = (M-1)d$$

so $d' \le (M-1)/2$. By the same idea, b can equal $a+kd$ only if $(M-1)d - kd \ge 2e$, which is to say,

$$k \le M - 2e/d = M - 2d'.$$

So $\{b, b+e, b+2e\}$ is contained in at most $M-2d'$ progressions with step d, and only provided d divides $e = dd'$ and $d' \leq (M-1)/2$. Said differently, given $d' \leq (M-1)/2$ there at most $M-2d'$ progressions of step e/d' which contain $\{b, b+e, b+2e\}$. Thus we get a total of

$$\sum_{d' \leq (M-1)/2} M - 2d' \leq M^2/4$$

progressions which contain $\{b, b+e, b+2e\}$.

The key idea is that if $B \subseteq \{a, a+d, \ldots, a+(M-1)d\}$ has size at least $r_3(M)+1$ then B has to contain a 3-term progression. So, to count 3-term progressions in A, we consider

$$\sum_{P \in \mathbf{AP}(N,M,k)} T_3(P \cap A) = \sum_{b, b+e, b+2e \in A} \sum_{P \in \mathbf{AP}(N,M,k)} \mathbf{1}_P(b)\mathbf{1}_P(b+e)\mathbf{1}_P(b+2e).$$

We've already seen that there are at most $M^2/4$ possible $P \in \mathbf{AP}(N, M, k)$ which contain $\{b, b+e, b+2e\}$ so

$$\sum_{P \in \mathbf{AP}(N,M,k)} T_3(P \cap A) \leq \frac{M^2}{4} T_3(A).$$

On the other hand, we know that if $|P \cap A| \geq r_3(M)+1$ then $T_3(P \cap A) \geq 1$. So

$$\sum_{P \in \mathbf{AP}(N,M,k)} T_3(P \cap A) \geq |\{P \in \mathbf{AP}(N, M, k) : |A \cap P| \geq r_3(M)+1\}|.$$

It remains to understand the right hand side. Consider

$$\sum_{P \in \mathbf{AP}(N,M,k)} |A \cap P|.$$

We write $P = P_d$ if P has step P, so the above splits as

$$\sum_{1 \leq d \leq k} \sum_{P_d} |A \cap P_d| = \sum_{1 \leq d \leq k} \sum_{a \in A} \sum_{P_d} \mathbf{1}_{P_d}(a).$$

If $a \in A$ belongs to the interval $I = [(M-1)k, N-(M-1)k]$ then a belongs to precisely M progressions of length M and step d, since it can occur at any position in the progression. Hence

$$\sum_{1 \leq d \leq k} \sum_{a \in A} \sum_{P_d} \mathbf{1}_{P_d}(a) \geq \sum_{1 \leq d \leq k} \sum_{a \in A \cap I} \sum_{P_d} \mathbf{1}_{P_d}(a) = kM|A \cap I|.$$

But

$$|A \cap I| \geq |A| - 2(M-1)k.$$

Hence

$$\sum_{P \in \mathbf{AP}(N,M,k)} |A \cap P| \geq Mk(|A| - 2Mk).$$

We split $\mathbf{AP}(N,M,k)$ into to sets, say X and Y, where X consists of those P with $|A\cap P|\geq r_3(M)+1$ and Y consists of those P with $|A\cap P|\leq r_3(M)$. What we have already shown amounts to $T_3(A)\geq |X|/M^2$. Now clearly,

$$|Y|\leq \sum_{P\in Y}\frac{|A\cap P|}{r_3(M)}.$$

Hence

$$M|X|\geq \sum_{P\in X}|A\cap P|\geq \sum_{P\in \mathbf{AP}(N,M,k)}|A\cap P|-\sum_{P\in Y}|A\cap P|\geq Mk(|A|-2Mk)-|Y|r_3(M),$$

On the other hand $|Y|\leq |\mathbf{AP}(N,M,k)|\leq Nk$, so

$$M|X|\geq Mk(|A|-2Mk)-(Nk)r_3(M)$$

and if we take $k=\lfloor N/M^2\rfloor$, this gives

$$M|X|\geq \frac{N}{M}|A|-2\frac{N^2}{M^2}-\frac{N^2}{M^2}r_3(M)=N^2M^3\left(\frac{|A|}{NM^4}-\frac{r_3(M)+1}{M^5}\right)$$

□

It will be convenient to work modulo a prime p, so for the moment let's suppose we are trying to count 3-term progressions in a set $A\subseteq \mathbb{F}_p$. To begin, suppose $f:\mathbb{F}_p\to\mathbb{C}$ is a function. We define

$$\Lambda(f)=\frac{1}{p^2}\sum_{a,d\in\mathbb{F}_p}f(a)f(a+d)f(a+2d).$$

Then, if $A\subseteq\mathbb{F}_p$, we have

$$\Lambda(\mathbf{1}_A)=|\{(a,d):a,a+d,a+2d\in A\}|,$$

the number of 3-term progressions in A. Now, it's generally hard to understand $\Lambda(f)$ if f is mysterious. So we next use the Fourier expansion of f to give a new expression for Λ.

Lemma 7.4

We have

$$\Lambda(f)=\sum_{x\in\mathbb{F}_p}\widehat{f}(x)\widehat{f}(x)\widehat{f}(-2x).$$

Proof. The Fourier expansion of f is

$$f(t)=\sum_{x\in\mathbb{F}_p}\widehat{f}(x)e(xt/p).$$

Plugging this into the definition of $\Lambda(f)$, we get

$$\Lambda(f) = \frac{1}{p^2} \sum_{a,d \in \mathbb{F}_p} \sum_{x_1,x_2,x_3} \widehat{f}(x_1)\widehat{f}(x_2)\widehat{f}(x_3) e(x_1 a/p) e(x_2(a+d)/p) e(x_3(a+2d)/p).$$

We bring the sums over a and d inside to get

$$\Lambda(f) = \sum_{x_1,x_2,x_3} \widehat{f}(x_1)\widehat{f}(x_2)\widehat{f}(x_3) \frac{1}{p^2} \sum_{a,d \in \mathbb{F}_p} e(x_1 a/p) e(x_2(a+d)/p) e(x_3(a+2d)/p).$$

Now we rewrite

$$e(x_1 a/p) e(x_2(a+d)/p) e(x_3(a+2d)/p) = e(a(x_1+x_2+x_3)/p) e(d(x_2+2x_3)/p)$$

and when we plug this in and sum over a and d, the orthogonality relations tell us that the sums vanish unless $x_1 + x_2 + x_3 = 0$ and $x_2 + 2x_3 = 0$, which is a system of equations whose only solutions are $(x_1, x_2, x_3) = (x, -2x, x)$. This proves the lemma. □

Theorem 7.5: Dirichlet Approximation

Let R be a subset of $\mathbb{F}_p$ of size at most $\log_2 p$. Denote by I_k the interval $[-k, k]$ modulo p, that is $I_k = \{-k, -k+1, \dots, k-1, k\}$. Then there is a $d \in \mathbb{F}_p$ with $d \neq 0$ and such that $d \cdot R \subseteq I_k$ for some $k \leq 4p^{1-1/R}$.

Proof. Let k be fixed for the time being. Then $\mathbb{F}_p$ can be covered by at most $p/k+1$ translates of $\{1, \dots, k\}$. Hence, if $n = |R|$, then $\mathbb{F}_p^n$ can be covered by at most $(p/k+1)^n$ translates of $\{1, \dots, k\}^n$. Write $R = \{r_1, \dots, r_n\}$ and consider the points $(tr_1, \dots, tr_n)$ with $t \in \mathbb{F}_p$. There are p such points, and so if $(p/k+1)^n < p$, then one translate of $\boldsymbol{x} + \{1, \dots, k\}^n$ contains two of these points, say $(tr_1, \dots, tr_n)$ and $(t'r_1, \dots, t'r_n)$. Then, the difference

$$((t-t')r_1, \dots, (t-t')r_n) \in I_k^n.$$

So we set $d = t - t'$ and since $t \neq t'$ we have $d \neq 0$. Now we just need to choose k, and we want to do that with k as small as possible. The sole condition we need to satisfy is $(p/k+1)^n < p$ which rearranges as

$$p/k + 1 < p^{1/n} \iff p + k < kp^{1/n} \iff p < k(p^{1/n} - 1).$$

Since $p^{1/n} - 1 > p^{1/n}/2$ so long as $p > 2^n$, which we have assumed, our constraint is satisfied provided $k \geq 2p^{1-1/n}$, so any integer k in the range $2p^{1-1/n} \leq k \leq 4p^{1-1/n}$ will do. □

The reason for using an approximation theorem as above is that the fractions dr/p will be very close to an integer when $r \in R$. Indeed, $dr = q_r p + z_r$ by long division, where $-k \leq z_r \leq k$. It follows that $e(dr/p) = e(z_r/p)$.

Lemma 7.5

Let $0 \le \theta \le 1$ and write $\|\theta\| = \min\{\theta, 1-\theta\}$. Then

$$|e(\theta) - 1| \le 2\pi\|\theta\|.$$

Proof. The quantity $2\pi\|\theta\|$ is the length of the arc on the unit circle in the complex plane which joins 1 to $e(\theta)$. This is longer than the straight line which joins them, which is $|1 - e(\theta)|$. □

Corollary 7.3

Let $z \in \mathbb{F}_p$ belong to the interval $I_k = \{-k, \dots, k\}$ modulo p. Then $|e(z/p) - 1| \le \frac{2\pi k}{p}$.

Proof. The fraction $\theta = z/p$ has $\|\theta\| \le k/p$ so the above lemma finishes the proof. □

Lemma 7.6

Let $f : \mathbb{F}_p \to \mathbb{C}$ be a function such that $\max_r |\widehat{f}(r)| \le 1$. Suppose

$$R = \{r : |\widehat{f}(r)| \ge t\}$$

and $d \in \mathbb{F}_p$ is non-zero and such that $d \cdot R \subseteq I_k = \{-k, \dots, k\}$. Then

$$|\widehat{f}(r)||1 - e(xr/p)| \le \max\{t, k/p\}$$

for all $r \in \mathbb{F}_p$.

Proof. If $r \notin R$ then $|\widehat{f}(r)| \le t$ by definition. Otherwise $|1 - e(dr/p)| \le k/p$ by the preceding corollary and the fact that $|\widehat{f}(r)| \le 1$. □

We will combine this lemma with the expression for $\Lambda(\mathbf{1}_A)$ from Lemma 7.3 to show that one can replace $\mathbf{1}_A$ with a slightly smoother function g which has much larger support larger than A.

Proposition 7.1

Suppose $A \subseteq \mathbb{F}_p$ has no non-trivial three-term progressions. Then there is a function $g : \mathbb{F}_p \to \{0, 1/3, 2/3\}$ with $|\Lambda(\mathbf{1}_A) - \Lambda(g)| \le 10^5 (\log\log p)^{1/2} / (\log p)^{1/2}$. Moreover, $\mathrm{supp}(g) \subseteq A \cup (A-d) \cup (A-2d)$ for some d with $|d| \le 4p/\log p$.

Proof. Let

$$R = \{r : |\widehat{\mathbf{1}}_A(r)| \geq \left(2\log\log p/\log p\right)^{1/2}\}.$$

From Chebyshev's inequality,

$$|R| \leq \frac{\log p}{2\log\log p}\sum_{r\in R}|\widehat{\mathbf{1}}_A(r)|^2 \leq \frac{\log p}{2\log\log p}\sum_{r\in\mathbb{F}_p}|\widehat{\mathbf{1}}_A(r)|^2$$

and by Parseval's identity, the right hand side is exactly

$$\frac{\log p}{2\log\log p}\sum_{r\in\mathbb{F}_p}|\widehat{\mathbf{1}}_A(r)|^2 = \frac{\log p}{2\log\log p}\frac{1}{p}\sum_{x\in\mathbb{F}_p}|\mathbf{1}_A(x)|^2 \leq \frac{\log p}{2\log\log p}.$$

By Dirichlet approximation, we can find a non-zero d such that $d\cdot(R\cup\{1\}) \subseteq \{-k,\dots,k\}$ with $k \leq 4\frac{p}{\log p}$. Now let

$$h = \frac{p}{3}\left(\mathbf{1}_{0,d,2d}\right)$$

and let

$$g(t) = \mathbf{1}_A * h(t) = \frac{1}{3}\left(\mathbf{1}_A(t) + \mathbf{1}_A(t+d) + \mathbf{1}_A(t+2d)\right).$$

The second equality is pretty easily verified by expanding out the definition of convolution. Note the claim on the support of g is immediate from this and the fact that $d = d\cdot 1 \in I_k$. Finally, since A has no three-term progressions, g can only take the values $0, 1/3$ and $2/3$. Now since $\widehat{g}(x) = \widehat{\mathbf{1}}_A(x)\widehat{h}(x)$ we have

$$\Lambda(\mathbf{1}_A) - \Lambda(g) = \sum_{x\in\mathbb{F}_p}\widehat{\mathbf{1}}_A(x)^2\widehat{\mathbf{1}}_A(-2x)(1-\widehat{h}(x)^2\widehat{h}(-2x)).$$

From the triangle inequality, we have

$$|1-\widehat{h}(x)^2\widehat{h}(-2x)| \leq |1-\widehat{h}(x)| + |\widehat{h}(x)||1-\widehat{h}(x)| + |\widehat{h}(x)||1-\widehat{h}(-2x)|.$$

But

$$\widehat{h}(x) = \frac{1}{3}(1 + e(dx/p) + e(2dx/p))$$

so $|\widehat{h}(x)| \leq 1$,

$$|1-\widehat{h}(x)| \leq \frac{1}{3}|1-e(-dx/p)| + \frac{1}{3}|1-e(-2dx/p)|,$$

and similarly

$$|1-\widehat{h}(-2x)| \leq \frac{1}{3}|1-e(2dx/p)| + \frac{1}{3}|1-e(4dx/p)|,$$

By the preceding corollary, since $\pm dr, \pm 2dr, \pm 4dr \in \{-4k,\dots,4k\}$ we have

$$|\widehat{\mathbf{1}}_A(x)||1-\widehat{h}(x)^2\widehat{h}(-2x)| \leq 1000\max\left\{\frac{(\log\log p)^{1/2}}{(\log p)^{1/2}}, \frac{16}{\log p}\right\} \leq \frac{10000(\log\log p)^{1/2}}{(\log p)^{1/2}}.$$

Putting all of this together,

$$|\Lambda(\mathbf{1}_A) - \Lambda(g)| \leq \frac{10000(\log\log p)^{1/2}}{(\log p)^{1/2}}\sum_x|\widehat{\mathbf{1}}_A(x)||\widehat{\mathbf{1}}_A(-2x)| \leq \frac{10000(\log\log p)^{1/2}}{(\log p)^{1/2}}\sum_x|\widehat{\mathbf{1}}_A(x)|^2.$$

In the last inequality we used that $|a||b| \le (|a|^2+|b|^2)/2$ and that as x varies over $\mathbb{F}_p$, so does $-2x$. Finally, by Parseval,

$$\sum_x |\widehat{\mathbf{1}}_A(x)|^2 \le \frac{|A|}{p} \le 1.$$

□

If the function g above where an indicator function, $g = \mathbf{1}_B$, then $\Lambda(\mathbf{1}_B)$ would be close to $\lambda(\mathbf{1}_A)$, and so B would have about as many three term progressions as A. If B were larger than A this would create a lot of tension, since if A is maximal without three-term progressions, then B would have to have some. However, g is not an indicator function, but we *will* be able to relate it to one.

Lemma 7.7

Let A and g be as in Lemma 7.3. Then the set $T = \mathrm{supp}(g)$ has size at least $3|A|/2$ and $\Lambda(\mathbf{1}_T) \le 27\Lambda(g)$.

Proof. We have

$$|T| \ge \sum_{x\in T} \frac{3}{2} g(x) = \frac{1}{2}\sum_x \mathbf{1}_A(x) + \mathbf{1}_A(x+d) + \mathbf{1}_A(x+2d) = \frac{3|A|}{2}.$$

On the other hand $\mathbf{1}_T(x) \le 3g(x)$, and so

$$\Lambda(\mathbf{1}_T) = \frac{1}{p^2}\sum_{x,d} \mathbf{1}_T(x)\mathbf{1}_T(x+d)\mathbf{1}_T(x+2d) \le 27\Lambda(g).$$

□

Let's take stock of what we have done. We started with $A \subseteq [N]$ which has no three-term progressions. We reduced the problem to counting three-term progressions in $\mathbb{F}_p$ where $2N < p < 4N$, and found a new set T which is larger than A by half, and has relatively few three-term progressions. We summarize this below.

Corollary 7.4

Let A be a subset of $[N]$ containing no non-trivial three-term progressions. Then there is a subset $T' \subseteq [N]$ with $|T'| \ge 4/3|A|$ such that T' contains at most $10^8 N^2 (\log\log N/\log N)^{1/2}$ three-term progressions, or else $|A| \le 1000N/\log N$.

Proof. The set $T = \mathrm{supp}(g)$ is contained in

$$A \cup A-d \cup A-2d \subseteq [-8\log p/\log\log p, N + 8\log p/\log\log p]$$
$$\subseteq [-16\log N/\log\log N, N + 16\log N/\log\log N]$$

where we recall that our choice of p satisfied $p < 4N$. Thus

$$|T \cap [1,N]| \geq |T| - 34\log N/\log\log N.$$

Since $|T| \geq 3/2|A|$,

$$|T \cap [1,\dots,N]| \geq 3|A|/2 - 32\log N/\log\log N \geq 4|A|/3$$

unless $|A| \leq 1000\log N/\log\log N$. We just need to estimate the number of three-term progressions in T'. It is trivially at most

$$N^2\Lambda(\mathbf{1}_T) \leq 27N^2\Lambda(g)$$

and from Proposition 7.3,

$$\Lambda(g) \leq \Lambda(\mathbf{1}_A) + 10^5(\log\log p)^{1/2}/(\log p)^{1/2} \leq \frac{|A|}{N^2} + 10^6(\log\log N)^{1/2}/(\log N)^{1/2},$$

the second inequality being because A has no three-term progressions so $\Lambda(\mathbf{1}_A)$ includes only the $|A|$ trivial ones. But $|A|/N^2 \leq 1/N$ is negligible, so in fact we have

$$N^2\Lambda(\mathbf{1}_T) \leq 10^8N^2(\log\log N)^{1/2}/(\log N)^{1/2}.$$

□

We can now conclude the proof of Roth's theorem. Suppose that N is given and $A \subseteq [N]$ is a set of size $|A| = r_3(N)$ which is free of three-term progressions. Then either

$$\frac{|A|}{N} = \frac{r_3(N)}{N} \leq \frac{1000}{\log N},$$

and we're done, or else Corollary 7.3 tells us there is a set $T' \subseteq [N]$ with

$$T_3(T') \leq 10^8N^2\left(\frac{\log\log N}{\log N}\right)^{1/2}$$

and of size at least $4/3r_3(N)$. From Varnavides' Lemma, we get

$$T_3(T') \geq \frac{N^2}{M^4}\left(\frac{|T'|}{N} - \frac{r_3(M)+2}{M}\right)$$

and so comparing the two,

$$10^8M^4\left(\frac{\log\log N}{\log N}\right)^{1/2} \geq \frac{4}{3}\frac{r_3(N)}{N} - \frac{r_3(M)+2}{M}.$$

Now we choose N and M so that $M = (\log N/\log\log N)^{1/16}$ and the left hand side becomes $10^8(\log\log N)^{1/4}/(\log N)^{1/4}$ and hence

$$\frac{r_3(N)}{N} \leq 10^9\frac{(\log\log N)^{1/4}}{(\log N)^{1/4}} + \frac{3}{4}\frac{r_3(M)}{M}.$$

Again, if the first term dominates, we're done. Otherwise, we can just write

$$\frac{r_3(N)}{N} \leq \frac{99}{100}\frac{r_3(M)}{M}.$$

But this tells us that when we increase from M to N, the ratio $r_3(x)/x$ drops by at least $1/100$. From this, $r_3(N)/N \to 0$ as $N \to \infty$.