From Measure to Probability

A probabilist's survey of measure-theoretic results

Feng Cheng*

Draft as of July 31, 2025

*Email: fecheng@math.washington.edu. Affiliation: Department of Mathematics, University of Washington, Seattle, WA 98195, USA.

Contents

Prologue

This is the most ambitious writing project undertaken by the author so far as a math student, and he hopes he can finish it in two years. The author, as a probability student, did not excel in his real analysis courses (MATH 202AB at UC Berkeley) during his senior year. To compensate, the author aims to write an extensive and detailed note that surveys through all the major measure theory results of interest to a rigorous-minded mathematical probabilist.

Part I of this note will be devoted to measure theory in a general setting, while Part II will discuss results in probability spaces built on top of Part I. The author hopes that his commentary and the overall structure of the survey can help the readers (and himself) truly understand both abstract measure theory and probability theory from a measure-theoretic point of view.

This entire survey will be based on multiple sources, listed in the bibliography page. As the old saying goes, "if you copy from one book that is plagiarism, but if you copy from ten books that is scholarship."

Shanghai, August 2024 F.C.

The prerequisite for this survey notes is a strong background in undergraduate real analysis and familiarity with elementary probability theory. Some key results about normed spaces, Hilbert spaces, and topology will be assumed, and these can usually be found on any first-year graduate analysis texts. Some rudimentary familiarity with weak topology on Banach spaces will contribute to the understanding of weak and vague convergence of measures. We have also included appendices at the end of the survey, which discuss some of these facts at a high level.

If you see any errors or typos, please inform the author via

fecheng@math.washington.edu.

Part I

Measure theory

Chapter 1 Measure spaces

1.A Basic setup

We let X be a nonempty set in Part I.

1.1 Definition. For $\{A_n\}_{n=1}^{\infty} \subseteq \wp(X)$, we define

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m \quad \text{and} \quad \liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m.$$

Note $\bigcap$ can be seen as "for all" and $\bigcup$ can be seen as "there exists". Therefore $\limsup_n A_n$ consists of elements that belong to infinitely many A_n's (spread out across $n \in \mathbf{N}$), while $\liminf_n A_n$ consists of elements that belong to all but finitely A_n (the n's at the beginning). To compare this with the lim sup and lim inf of a sequence of numbers, one may try the following exercise.

1.2 Exercise. Show that

$$\limsup_{n\to\infty} A_n = A \iff \limsup_{n\to\infty} \mathbf{1}_{A_n} = \mathbf{1}_A,$$
$$\liminf_{n\to\infty} A_n = A \iff \liminf_{n\to\infty} \mathbf{1}_{A_n} = \mathbf{1}_A.$$

Here $\mathbf{1}_A \colon X \to \{0, 1\}$ given by

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

is called the *indicator function* (*characteristic function* for analysts who choose to write χ_A).

If $\{A_n\}_{n=1}^{\infty}$ is an increasing sequence of sets, then

$$\liminf_n A_n = \limsup_n A_n = \bigcup_n A_n;$$

if the sequence is decreasing, then

$$\liminf_n A_n = \limsup_n A_n = \bigcap_n A_n.$$

Also remember that, by De Morgan's Law,

$$\limsup_n A_n^{\mathrm{c}} = \left(\liminf_n A_n\right)^{\mathrm{c}} \quad \text{and} \quad \liminf_n A_n^{\mathrm{c}} = \left(\limsup_n A_n\right)^{\mathrm{c}}.$$

Here is another exercise.

1.3 Exercise. Consider a sequence of functions f_n that convergences to f pointwise on some set E. If we define

$$E_{n,\epsilon} = \{x : |f_n(x) - f(x)| < \epsilon\}$$

for $\epsilon > 0$ and $n \in \mathbf{N}$, then

$$E = \bigcap_{k=1}^{\infty} \liminf_m E_m^{1/k} = \bigcap_{k=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcap_{n \geq m} E_n^{1/k}.$$

1.4 Definition. A nonempty collection $\mathcal{A}$ of subsets of X is an *algebra* if

(a) $\emptyset, X \in \mathcal{A}$;

(b) if $E \in \mathcal{A}$, then $E^c \in \mathcal{A}$; (closed under complement)

(c) if $E_1, E_2 \in \mathcal{A}$, then $E_1 \cup E_2, E_1 \cap E_2 \in \mathcal{A}$. (closed under finite unions and intersections)

Furthermore, $\mathcal{A}$ is called a σ*-algebra* if condition (c) asks for countable unions and intersections.

An algebra can be constructed from a more basic structure called semialgebra, which we define below.

1.5 Definition. A *semialgebra* $\mathcal{E}$ is a collection of sets such that

(a) $\emptyset \in \mathcal{E}$;

(b) closed under finite intersections;

(c) if $A \in \mathcal{E}$ then A^c is a finite disjoint union of elements in $\mathcal{E}$.

Some authors drop condition (a), while others add the condition that $X \in \mathcal{E}$. But of course there is no essential difference. Now comes the main result.

1.6 Proposition [Fol99, Proposition 1.7]. If $\mathcal{E}$ is a semialgebra[1], then all finite disjoint unions of sets in $\mathcal{E}$ form an algebra.

The most important example of a semialgebra consists of the empty set and all sets of the form

$$(a_1, b_1] \times \cdots \times (a_d, b_d] \subseteq \mathbf{R}^d,$$

where $-\infty \leq a_j < b_j \leq \infty$. The finite disjoint unions of half-open half-closed cubes should therefore form an algebra.

From now on we will assume $\mathcal{A}$ is by default a σ-algebra. Obviously the largest σ-algebra on X is the power set $\wp(X)$.

Given a σ-algebra $\mathcal{A}$ on X, the couplet $(X, \mathcal{A})$ is called a *measurable space*, a space on which we can possibly attach a measure. Given a measurable space $(X, \mathcal{A})$, we call a set E is $\mathcal{A}$-measurable if $E \in \mathcal{A}$.

Also in analysis, "σ" means countable union while "δ" means countable intersection. An F_σ *set* is a countable union[2] of closed[3] sets, while a G_δ *set* is a countable intersection[4] of open[5] sets.

We know that the preimage of a function $f : X \to Y$ is a mapping $f^{-1} : \wp(Y) \to \wp(X)$ that preserves unions, intersections, and complements, which are also operations in the definition of a σ-algebra. The next result makes the relationship between the two explicit. See Section 2.A for the use.

[1] Folland calls this elementary family.
[2] *somme* in French
[3] *fermé* in French
[4] *Durchschnitt* in German
[5] *Gebiet* in German

1.7 Proposition [Kal02, Lemma 1.3]. Consider $f\colon X \to Y$, and $\mathcal{M}$ and $\mathcal{N}$ be two respective σ-algebras on X and Y. The preimage f^{-1} induces two σ-algebras:

(a) $\mathcal{M}' = \{f^{-1}(A) : A \in \mathcal{N}\}$ on X, in the backward direction;

(b) $\mathcal{N}' = \{B \subseteq Y : f^{-1}(B) \in \mathcal{M}\}$ on Y, in the forward direction.

We will write $\mathcal{M}' = f^{-1}\mathcal{N}$ subsequently.

The following fact is left as an easy exercise to the reader. It shows these structures are nice to work with.

1.8 Fact. The intersection of a family of algebras/σ-algebras is an algebra/σ-algebra. Note that the union is not.

This fact holds for other set algebra structures as well, which include Dynkin's λ-system and the monotone class to be introduced in Section 1.B.

With this elementary fact in mind, we have the following definition.

1.9 Definition. Within X, given a family of subsets $\mathcal{E}$, the smallest σ-algebra containing $\mathcal{E}$, i.e., the intersection of all σ-algebras that contains $\mathcal{E}$, is called the *σ-algebra generated by $\mathcal{E}$*, denoted by $\sigma(\mathcal{E})$.

The same definitions apply to algebra and other set algebra structures, including Dynkin's λ-system and the monotone class to be introduced in Section 1.B.

Certainly the definitions of algebra and σ-algebra bear some resemblance to the definition of topology. The above fact and definition have just turned this connection even more evident. We will explore this connection further in Section 3.A, when discussing product σ-algebras.

Of course we need to endow a topology on the measure space X to make things interesting. If X is a topological space, then the *Borel σ-algebra* on X, which we denote by $\mathcal{B}_X$ or $\mathcal{B}(X)$, is the σ-algebra generated by all open sets. One can of course replace the "open" here by "closed".

If $X = \mathbf{R}$ with the standard Euclidean topology, then $\mathcal{B}(\mathbf{R})$ is generated

- by open intervals (or closed),
- by left-open right-closed intervals (or the other way around),
- by open rays $\{(a, \infty) : a \in \mathbf{R}\}$ (or the other way around),
- or by close rays $\{[a, \infty) : a \in \mathbf{R}\}$ (or the other way around).
- One may replace the endpoints of intervals by rationals as well.

The first bullet point boils down the fact that an open set in $\mathbf{R}$ can always be written into the disjoint union of a countable number of open intervals. The proof of this requires us to show that

1.10 Exercise. Given a set U open in $\mathbf{R}$. The relationship $\sim$ on U given by $x \sim y$ if $[x \wedge y, x \vee y] \subseteq U$ is an equivalence relation.

The theorem is of significant importance throughout measure theory, and is key to the construction of Lebesgue measure on the real line that we will see soon. The notations $x \wedge y$ and $x \vee y$ are shorthand for $\min\{x, y\}$ and $\max\{x, y\}$. We will use them later more often.

1.11 Definition. A *measure* μ on $(X, \mathcal{A})$ is a function $\mu\colon \mathcal{A} \to [0, \infty]$ such that

(a) $\mu(\emptyset) = 0$;

(b) μ is *countably additive/σ-additive*, i.e., let $\{E_n\}_{n=1}^{\infty}$ be any measurable partition of $E \in \mathcal{A}$, we have

$$\mu(E) = \mu\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mu(E_n).$$

For two different rearrangements of the same measurable partition of E, $\mu(E)$ should have the same value, because the sum of nonnegative values does not change under reordering. An easy way to see this is to note

$$\sum_{n=1}^{\infty} a_n = \sup\left\{\sum_{n\in I} a_n : I \text{ is a finite subset of } \mathbf{N}\right\}.$$

In fact the right hand side above is how we define generalized sums over possibly uncountable indices. Therefore condition (b) makes sense.

From now on we assume by default that μ is a measure. The triplet $(X, \mathcal{A}, \mu)$ is called a *measure space*.

A measure μ on $(X, \mathcal{A})$ is a *probability measure*[6] if $\mu(X) = 1$; μ is *finite* if $\mu(X) < \infty$; and μ is *σ-finite* if X can be written as a countable union of measurable sets $A_n \in \mathcal{A}$, each of which is of finite measure. Note for a σ-finite measure, we can replace this countable collection of finite-measure sets that make up X by an increasing sequence of finite-measure sets. We may even further assume that the sets are mutually disjoint. These assumption can be handy in some proofs.

It is clear that any probability measure is a finite measure, which is in turn a σ-finite measure. The probability measure is the essential example of a finite measure, because mostly you can normalize the measure of the whole space to 1.

A σ-finite measure is a well-behaved kind of measure. The Lebesgue measure that we will rigorously see soon, for example, is σ-finite. Some major results in measure theory, for example the Fubini–Tonelli theorem (see Section 3.B), are only true for σ-finite measure spaces. A measure that is not σ-finite is considered, in some sense, a little pathological.

The following "restricted" measures will come up a couple of times.

1.12 Fact. Fix some $S \in \mathcal{A}$. The function $\nu : \mathcal{A} \to [0, \infty]$ given by $\nu(E) = \mu(E \cap S)$ is still a measure on $(X, \mathcal{A})$.

1.13 Fact. Fix $S \in \mathcal{A}$. By intersecting S we can get a sub-σ-algebra $\mathcal{A}|_S$ on S, where

$$\mathcal{A}|_S = \{E \cap S : E \in \mathcal{A}\}.$$

Such $(S, \mathcal{A}|_S)$ is called a *measurable subspace* of $(X, \mathcal{A})$. Note that μ restricted to the σ-algebra $\mathcal{A}|_S$ is a measure on $\mathcal{A}|_S$. We denoted this restricted measure on $(S, \mathcal{A}|_S)$ by $\mu|_S$, or simply μ when the context is clear.

Below are some important basic properties about measures that are used all the time.

1.14 Proposition. We have the following properties about a measure μ on $(X, \mathcal{A})$.

(a) monotonicity: for $A, B \in \mathcal{A}$,

$$A \subseteq B \implies \mu(A) \le \mu(B);$$

(b) inclusion-exclusion: for $A, B \in \mathcal{A}$ with $\mu(A \cap B) < \infty$, we have

$$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B).$$

[6] Why use this name? Because the probability of the entire sample space should be 1.

(c) σ-*subadditivity*: for possibly intersecting sets[7] $\{E_n\}_{n=1}^{\infty} \subseteq \mathcal{A}$,

$$\mu\Big(\bigcup_{n=1}^{\infty} E_n\Big) \leq \sum_{n=1}^{\infty} \mu(E_n).$$

(d) continuity from below: for a sequence of sets $\{E_n\}_{n=1}^{\infty} \subseteq \mathcal{A}$ that increases to E, we have

$$\mu(E_n) \uparrow \mu(E).$$

(e) continuity from above (when the first set is of finite measure): for a sequence of sets $\{E_n\}_{n=1}^{\infty} \subseteq \mathcal{A}$ with $\mu(E_1) < \infty$ and $E_n \downarrow E$, we have

$$\mu(E_n) \downarrow \mu(E).$$

All these properties above require the famous disjointification trick to prove: we partition the sets in question into pairwise disjoint pieces, and then use countable additivity of the measure.

Now we discuss two important examples of measure extremely useful in application[8].

The first one is the *counting measure*. Consider the measurable space $\big(X, \wp(X)\big)$. The function $\mu\colon \wp(X) \to [0,\infty]$ given by $\mu(E) = |E|$ is a measure. Basically it counts how many elements are in each subset of X.

The second one is the *Dirac point mass*. Given $(X, \mathcal{A})$ and some $x \in X$, we define the function $\delta_X : \mathcal{A} \to \{0,1\}$ given by

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

This is clearly a probability measure. Notice its difference from the indicator function. The point mass $\delta_x(A)$ takes in a set and spits out 1/0, while the corresponding indicator $\mathbf{1}_A(x)$ takes in a point and spits out 1/0.

A countable linear combination of Dirac point mass defines a measure μ on $\mathcal{A}$ called the *discrete measure* To be precise, given a countable set $Y \subseteq X$, and a function $c\colon y \mapsto [0,\infty]$ at each $y \in Y$, we can define $\mu : \mathcal{A} \to \infty$ by

$$\mu = \sum_{y \in Y} c(y)\delta_y.$$

The meticulous reader should notice that the function c here resembles the probability mass function on a discrete probability space; see Section 7.A.

We now introduce two additional elementary results about measures, which are simple consequences from Proposition 1.14. These two results are important in probability theory, but both are indeed purely measure-theoretic.

1.15 Corollary (Upper and lower semicontinuity of measures). For $\{E_n\}_n \subseteq \mathcal{A}$, we have

$$\mu\big(\liminf_n E_n\big) \leq \liminf_n \mu(E_n).$$

If in addition μ is finite, then

$$\limsup_n \mu(E_n) \leq \mu\big(\limsup_n E_n\big).$$

[7] Recall in σ-additivity the sets must be mutually disjoint.

[8] In this note we will avoid going deep into facts/examples/counterexmaples that are ultimately not very useful in practice. One such "useless" example that is often mentioned here is the countable-cocountable measure on an uncountable set. One may also list the collection of all countable and cocountable sets as an example of a σ-algebra earlier, but we have omitted for the same reason. Some results of greater generality and particular examples add further insight to the subject matter and help our understanding, but in many situations this is not the case.

1.16 Borel–Cantelli lemma I. For $\{E_n\}_n \subseteq \mathcal{A}$, assume $\sum_n \mu(E_n) < \infty$, then

$$\mu\big(\limsup_n E_n\big) = 0.$$

We will see that the above result are commonly used to prove almost everywhere convergence of (measurable) functions, a notion that will be introduced in Chapter 2.

One can skip the rest of this section for now, and come back after reading about the Lebesgue measure on the real line.

Given $(X, \mathcal{A}, \mu)$, a subset $E \subseteq X$ is called a *null set* if there is $B \in \mathcal{A}$ such that $E \subseteq B$ and $\mu(B) = 0$. If $\mathcal{A}$ contains all these null sets, then the measure space is *complete*. The *completion* $\mathcal{A}^\mu$ is the smallest σ-algebra containing $\mathcal{A}$ such that there exists a measure $\bar{\mu}$, which extends μ to $\mathcal{A}^\mu$, that makes $(X, \mathcal{A}^\mu)$ complete.

Why is a complete measure space sometimes desirable? In some cases we want to make all subsets of measure zero sets measurable to avoid some technical peculiarity, and meanwhile we can measure a larger collection of sets. However, it is important to remember that a larger σ-algebra can lead to more technical peculiarities as well. In many cases the additional measurable sets after completion may not be well-behaved with respected functions, which we will see in Section 2.A. In addition, even a complete measure space $(X, \mathcal{A}, \mu)$ may still not measure every subset of X.

The completion of a measure space is given explicitly, as stated in the following theorem.

1.17 Theorem [Fol99, Theorem 1.9]. The completion $\mathcal{A}^\mu$ is unique, which is given by

$$\mathcal{A}^\mu = \{E \cup F : E \in \mathcal{A} \text{ and } F \subseteq N, \text{ where } N \text{ is a null set}\}.$$

In addition, the measure $\bar{\mu}$ given by $\bar{\mu}(E \cup F) = \mu(E)$ not only completes $\mathcal{A}$, but also is the unique extension of μ from $\mathcal{A}$ to $\mathcal{A}^\mu$.

Proof. The first part of the proof is given in the reference. For the uniqueness part, suppose there is some other measure $\hat{\mu}$ on $\mathcal{A}^\mu$ such that $\hat{\mu}(E) = \mu(E)$ for all $E \in \mathcal{A}$. However, there exists some $D \subseteq N$, where $\mu(N) = 0$, such that $\hat{\mu}(E \cup D) \neq \mu(E) = \hat{\mu}(E)$. This implies $\hat{\mu}(D - E) > 0$. Yet $D - E \subseteq N$ where $\hat{\mu}(N) = 0$. This contradicts monotonicity. □

To similarly avoid peculiarities caused by null sets, we give the following definitions. Let μ be a finite measure. The set $A \in \mathcal{A}$ is called an *atom* of the measure μ if the set has measure $\mu(A) > 0$, but all its measurable subsets must be either of measure 0 or of measure $\mu(A)$. A measure is *atomless* if there are no atoms. A measure μ is *(purely) atomic* if the measure μ is concentrated on a countable union of atoms $\bigcup_{n=1}^\infty A_n$, i.e., $\mu(X - \bigcup_n A_n) = 0$.

1.B Two tools from set theory

1.18 Definition. A *π-system* on X is a nonempty collection of subsets of X that is closed under finite intersections.

A *λ-system* $\mathcal{L}$ on X is a collection of subsets of X such that

(a) $X \in \mathcal{L}$;

(b) if $A, B \in \mathcal{L}$ and $A \subseteq B$, then $B - A \in \mathcal{L}$; (closed under proper differences)

(c) if $A_n \in \mathcal{L}$ and $A_n \uparrow A$ then $A \in \mathcal{L}$. (closed under ascending countable unions)

1.19 Definition. A *monotone class* on X is a collection of subsets of X that is closed under ascending countable unions and descending countable intersections.

1.20 Dynkin's π-λ theorem. Within X, if $\mathcal{P}$ is a π-system that is contained in a λ-system $\mathcal{L}$, then $\sigma(\mathcal{P}) \subseteq \mathcal{L}$.

1.21 Monotone class theorem. Given an algebra $\mathcal{A}_0$ of sets, then the monotone class $\mathcal{M}$ generated[9] by $\mathcal{A}_0$ coincides with the σ-algebra $\sigma(\mathcal{A}_0)$ generated by $\mathcal{A}_0$.

We deferred the proofs of both theorems to Appendix J; they are somewhat involved and not too interesting in the end. "The structure generated from $\mathcal{E}$ is the smallest containing $\mathcal{E}$" is always the main proof idea behind results on generated σ-algebras (or other structures). We will see this proof idea also in our immediate result below.

This next result is also of theoretical significance. It tells us a π-system that generates the σ-algebra identifies the measure.

Suppose we want to show some property holds on the entire $\mathcal{A}$. The way we apply the Dynkin's π-λ theorem usually looks like this. First we prove that the collection of sets with this property is a λ-system. If we have a π-system with this property that generates $\mathcal{A}$, then the entire $\mathcal{A}$ must agree with this λ-system.

from [ADM11, Proposition 1.15]

1.22 Coincidence criterion. Let μ_1 and μ_2 be two measures on $(X, \mathcal{A})$. Suppose we can find a π-system $\mathcal{P}$ on which the two measures agree, and $\sigma(\mathcal{P}) = \mathcal{A}$.

If $\mu_1(X) = \mu_2(X) < \infty$ (for example, both are probability measures), then the two measures agree on the entire $\mathcal{A}$.

More generally, if there exists $\{X_n\} \subseteq \mathcal{P}$ such that $X_n \uparrow X$ and

$$\mu_1(X_n) = \mu_2(X_n) < \infty \text{ for all } n \in \mathbf{N},$$

then the two measures agree on the entire $\mathcal{A}$.

Proof. Assume $\mu_1(X) = \mu_2(X) < \infty$. Define $\mathcal{D}$ to be the collection of all sets in $\mathcal{A}$ on which the two measures agree. It is easy to verify that $\mathcal{D}$ becomes a λ-system. Now invoke Dynkin's π-λ theorem and conclude that $\mathcal{D} = \mathcal{A}$. Without the finiteness assumption, we cannot verify condition (b) for a λ-system that makes $\mu(B) - \mu(A)$ computable.

Now consider the general assumption. We define for each n

$$\begin{aligned} \mathcal{A}_n &= \{E \cap X_n : E \in \mathcal{A}\}, \text{ which is a } \sigma\text{-algebra, and} \\ \mathcal{P}_n &= \{E \cap X_n : E \in \mathcal{P}\}, \text{ which is a } \pi\text{-system contained in } \mathcal{A}_n. \end{aligned}$$

Then μ_1 and μ_2 restricted to $(X_n, \mathcal{A}_n)$ is a finite measure. By the special case above, the two measures coincide on $\sigma(\mathcal{P}_n)$.

Now we prove $\mathcal{A}_n \subseteq \sigma(\mathcal{P}_n)$. Check that since $X_n \in \mathcal{P}$,

$$\{E \subseteq X : E \cap X_n \in \sigma(\mathcal{P}_n)\}$$

is a σ-algebra containing $\mathcal{P}$, and hence $\mathcal{A}$.

Now for each n and all $E \in \mathcal{A}$, the two measures agree on $E \cap X_n$. Now take $n \to \infty$ and we see that $\mu_1 = \mu_2$. □

σ-finite with sets in $\mathcal{P}$

[9] see Definition 1.9

1.C Extension theorems

1.23 Definition. The Carathéodory *outer measure* on X is a function $\mu^*\colon \wp(X) \to [0,\infty]$ such that

(a) $\mu^*(\emptyset) = 0$; (emptyset)

(b) if $A \subseteq B$, then $\mu^*(A) \leq \mu^*(B)$; (monotonicity)

(c) For subsets $A_1, A_2, \ldots$ of X, $\mu^*(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu^*(A_i)$. ($\sigma$-subadditivity)

A *null set* with respect to the outer measure μ^* is just a set with μ^*-value 0.

induced from additive set function

Let $\mathcal{C}$ be a collection of subsets of X such that $\emptyset \in \mathcal{C}$ and there are $D_1, D_2, \ldots$ in $\mathcal{C}$ such that $\bigcup_{i\in\mathbf{N}} D_i = X$. Suppose $\ell\colon \mathcal{C} \to [0,\infty]$ with $\ell(\emptyset) = 0$. Now if we define for all $E \in \wp(X)$

$$\mu^*(E) = \inf\left\{\sum_{i=1}^{\infty} \ell(A_i) : E \subseteq \bigcup_{i=1}^{\infty} A_i, \text{ where every } A_i \in \mathcal{C}\right\},$$

then μ^* is an outer measure on X. (Note that by assumption the infimum is taken over a nonempty set, and hence always exists. For simplicity one may just assume $X \in \mathcal{C}$ as well.) The proof is routine.

Here are some forewords to what we will construct.

- Let $X = \mathbf{R}$, $\mathcal{C}$ be the collection of all left-open right-closed intervals, and $\ell\big((a,b]\big) = b - a$. This gives the Lebesgue outer measure m^* used to construct the Lebesgue measure m.
- Let $f : \mathbf{R} \to \mathbf{R}$ be an increasing right-continuous[10] function. we let $\ell\big((a,b]\big) = f(b) - f(a)$. The μ^* that arises from this is used to construct the Lebesgue–Stieltjes measure.

1.24 Definition. For an outer measure μ^*, a set $A \subseteq X$ is μ^*-*measurable* if for all $E \subseteq X$,

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^{\mathrm{c}}).$$

This characterizes a collection of sets that are well-behaved under set operations, which leads to the next theorem. Note it A is μ^*-measurable if and only if for all E with $\mu^*(E) < \infty$,

$$\mu^*(E) \geq \mu^*(E \cap A) + \mu^*(E \cap A^{\mathrm{c}}).$$

1.25 Carathéodory's theorem. Given an outer measure μ^* on X, then the collection $\mathcal{A}$ of μ^*-measurable sets is in fact a σ-algebra on X. Let $\mu = \mu^*|_{\mathcal{A}}$, then μ is a measure. Also the σ-algebra $\mathcal{A}$ contains all the null sets, i.e., $(X, \mathcal{A}, \mu)$ is complete.

Proof. $\mathcal{A}$ is clearly closed under complements. We then check $\mathcal{A}$ is an algebra (the union of two sets in $\mathcal{A}$ is still in $\mathcal{A}$), and show μ^* is finitely additive on $\mathcal{A}$.

We wish to extend finite additivity to countable additivity. We let $B_n = \cup_{j=1}^{n} A_j$ and $B = \cup_{j=1}^{\infty} A_j$. For any E, we may conclude that

$$\mu^*(E \cap B_n) = \sum_{j=1}^{n} \mu(E \cap A_j).$$

[10] We will use "increasing" and "strictly increasing" in our note. Right-continuity at x means continuity from x^+.

It follows that $\mu^*(E) \geq \sum_{j=1}^n \mu^*(E \cap A_j) + \mu(E \cap B^c)$. Take $n \to \infty$ we may conclude

$$\begin{aligned}\mu^*(E) &\geq \sum_{j=1}^{\infty} \mu^*(E \cap A_j) + \mu^*(E \cap B^c) \\ &\geq \mu^*\Big(\bigcup_{j=1}^{\infty}(E \cap A_j)\Big) + \mu^*(E \cap B^c) \\ &= \mu^*(E \cap B) + \mu^*(E \cap B^c) \geq \mu^*(E).\end{aligned}$$

It follows that $B \in \mathcal{A}$, and if we let $E = B$, the first inequality (which is an equality) gives countable additivity.

It is easy to show $\mathcal{A}$ contains all μ^*-null sets: for N such that $\mu^*(N) = 0$, for any E we have

$$\mu^*(E) \leq \mu^*(E \cap N) + \mu^*(E \cap N^c) \leq \mu^*(E \cap N^c) \leq \mu(E). \qquad \square$$

1.26 Carathéodory extension theorem. For an algebra $\mathcal{A}_0$ on X and its premeasure μ_0, let

$$\mu^*(E) = \inf\Big\{\sum_{i=1}^{\infty} \mu_0(A_i) : E \subseteq \bigcup_{i=1}^{\infty} A_i, \text{ where every } A_i \in \mathcal{A}_0\Big\}$$

for all $E \subseteq X$. Then (1) μ^* is an outer measure on X, and hence by Carathéodory's theorem it gives a meausure space $(X, \sigma(\mathcal{A}_0), \mu)$; (2) $\mu^*|_{\mathcal{A}_0} = \mu_0$; (3) every set in $\mathcal{A}_0$ is μ^*-measurable; (4) if μ_0 is σ-finite, then μ in (1) is the unique extension of μ_0 from $\mathcal{A}_0$ to $\sigma(\mathcal{A}_0)$.

Proof. When proving $\mu^*(E) \geq \mu_0(E)$ in (2), consider the disjoint sets $B_n = E \cap (A_n - \cup_{i=1}^{n-1} A_i)$. Then $\cup_{n=1}^{\infty} B_n = E$, which implies $\sum_{n=1}^{\infty} \mu_0(A_n) \geq \sum_{n=1}^{\infty} \mu_0(B_n) = \mu_0(E)$. Then take infimum. (3) is fairly straightforward from definition.

To prove (4), let measure ν be another extension. Consider $E \in \sigma(\mathcal{A}_0)$ and $\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{A}_0$ that covers E, we have

$$\nu(E) \leq \sum_{i=1}^{\infty} \nu(A_i) = \sum_{i=1}^{\infty} \mu_0(A_i).$$

Take infimum and we get $\nu(E) \leq \mu(E)$.

Now let $A = \cup_{i=1}^{\infty} A_i$, then

$$\mu(A) = \lim_{n\to\infty} \mu(\cup_{i=1}^{n} A_i) = \lim_{n\to\infty} \nu(\cup_{i=1}^{n} A_i) = \nu(A).$$

If $\mu(E) < \infty$, then for any $\epsilon > 0$ we may choose $\{A_i\}_{i=1}^{\infty}$ such that $\mu(A - E) < \epsilon$. It follows that

$$\mu(E) \leq \mu(A) = \nu(A) = \nu(E) + \nu(A - E) < \nu(E) + \epsilon.$$

Therefore $\mu(E) = \nu(E)$.

Now suppose we have $X = \cup_{j=1}^{\infty} B_j$ such that $\mu_0(B_j) < \infty$ and that the B_j's are pairwise disjoint. Then for $E \in \sigma(\mathcal{A}_0)$, we have

$$\mu(E) = \sum_{j=1}^{\infty} \mu(E \cap B_j) = \sum_{j=1}^{\infty} \nu(E \cap B_j) = \nu(E),$$

where the second equality follows from what we have previously. $\square$

Notice we have proved (4) from the first principle; however, this is also a direct consequence of the coincidence criterion.

measure approximation in symmetric difference

1.27 Theorem.

1.D The Lebesgue measure

1.28 Fact. Assuming the full axiom of choice, we can use Zorn's lemma to assert that $\mathcal{L} \neq \wp(\mathbf{R})$.

1.29 Fact. With the countable axiom of choice, we can explicitly show that $\mathcal{L} \neq \mathcal{B}$.

We know as a consequence of Proposition 1.6 that the finite disjoint unions of $(a, b]$, where $a, b \in \mathbf{R}$, form an algebra on $\mathbf{R}$. We refer to this algebra as $\mathcal{A}_0$ below.

1.30 Theorem. For an increasing right-continuous function $F : \mathbf{R} \to \mathbf{R}$, the function $\mu_0 : \mathcal{A}_0 \to [0, \infty]$ such that $\mu_0(\emptyset) = 0$ and

$$\mu_0\left(\bigcup_{j=1}^{n}(a_j, b_j]\right) = \sum_{j=1}^{n} F(b_j) - F(a_j) \quad \text{for disjoint } \{(a_j, b_j]\}_{j=1}^{n}$$

is countably additive, and hence a premeasure on $\mathcal{A}_0$.

1.31 Theorem [Fol99, Theorem 1.16].

(a) Let $F : \mathbf{R} \to \mathbf{R}$ be an increasing, right-continuous function, then there is a unique associated Borel measure μ_F on $\mathbf{R}$ such that

$$\mu_F(a, b] = F(b) - F(a) \quad \text{for all } a \leq b.$$

If G is another increasing, right-continuous function, then $\mu_F = \mu_G$ if and only if F and G differ by a constant.

(b) Conversely, if μ is a finite Borel measure on $\mathbf{R}$, then the function $F : \mathbf{R} \to \mathbf{R}$ given by $F(x) = \mu(-\infty, x]$ is increasing and right-continuous. Furthermore $\mu = \mu_F$, and the function has left limits, i.e., $F(x-) = \lim_{y \to x^-} F(y)$ exists at every $x \in \mathbf{R}$. More specifically,

$$F(x-) = \mu(-\infty, x). \tag{1.32}$$

The conclusion of part (a) indicates that we should quotient out the difference up to a constant from the collection of F's. In this way, we obtain a one-to-one correspondence between finite Borel measures on $\mathbf{R}$ with the "normalized" collection of increasing, right-continuous functions F with $F(-\infty) = 0$.

Regarding equation (1.32), it is customary to write $F(x-) = \lim_{y \to x^-} F(y)$ when the limit exists. Note that having left limits implies

$$\mu\{x\} = F(x) - F(x-)$$

for all $x \in \mathbf{R}$.

Proof.

(a) Following Theorem 1.30, we have a premeasure μ_0 on $\mathcal{A}_0$ given by

$$\mu_0(a, b] = F(b) - F(a).$$

Note μ_0 is σ-finite as $\mathbf{R} = \cup_{j \in \mathbf{Z}}(j, j + 1]$. Therefore by Carathéodory extension theorem, it has a unique extension to a measure on $\sigma(\mathcal{A}_0) = \mathcal{B}(\mathbf{R})$.

The $\mu_F = \mu_G$ if and only if $F - G$ is a constant part is easy.

(b) F is increasing because μ is a nonnegative function. Right-continuity follows from

$$\begin{aligned}\lim_{y\to x^+} F(y) &= \lim_{y\to x^+} \mu(-\infty, y] \\ &= \lim_{n\to\infty} \mu\Big(-\infty, x+\frac{1}{n}\Big] \\ &= \mu\Big(\bigcap_{n=1}^{\infty}\Big(-\infty, x+\frac{1}{n}\Big]\Big) \\ &= \mu(-\infty, x] = F(x).\end{aligned}$$

Note that the second equality is justified because both "$\geq$" and "$\leq$" hold.

To show $\mu = \mu_F$, we check for any $a \leq b$,

$$\begin{aligned}\mu(a,b] &= \mu(-\infty, b] - \mu(-\infty, a] \\ &= F(b) - F(a),\end{aligned}$$

and use part (a).

It remains to show for every $x \in \mathbf{R}$ that (1.32) holds:

$$\begin{aligned}\lim_{y\to x^-} F(y) &= \lim_{y\to x^-} \mu(-\infty, y] \\ &= \lim_{n\to\infty} \mu\Big(-\infty, x-\frac{1}{n}\Big] \\ &= \mu\Big(\bigcup_{n=1}^{\infty}\Big(-\infty, x-\frac{1}{n}\Big]\Big) \\ &= \mu(-\infty, x).\end{aligned}$$

□

For part (b), if μ is a Borel measure on $\mathbf{R}$ that is finite on all bounded Borel sets, then F can be instead defined by

$$F(x) = \begin{cases} \mu(0,x] & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -\mu(x,0] & \text{if } x < 0, \end{cases}$$

and all conclusions still hold.

In the context of part (a), the μ_F is called the *Lebesgue–Stieltjes measure* associated to F. When the function F is the identity function, μ_F is called the *Lebesgue measure* on $\mathbf{R}$, which we will denote by m in this note[11]. It generalizes the notion of length of intervals to a wide collection of subsets of $\mathbf{R}$, that is sufficient for application most of the time.

In some cases it is useful to consider the completion of $(\mathbf{R}, \mathcal{B}, \mu_F)$, so that we can measure more sets than the Borel sets. The completion of $\mathcal{B}$ with respect to the Lebesgue measure m is called the Lebesgue σ-algebra, which we denote by $\mathcal{L}$.

1.33 Theorem. The Lebesgue measure m on $(\mathbf{R}, \mathcal{B})$ is the only nontrivial measure, up to multiplicative constants, that is translation invariant and locally finite.

[11] Other common notations include $\lambda, \mathcal{L}, |\cdot|$.

1.E Regularity of measures

1.34 Definition. A measure μ on $(X, \mathcal{A})$ is *outer regular* if for all $E \in \mathcal{A}$,

$$\mu(E) = \inf\{\mu(G) : G \text{ is open in } X \text{ and } G \supseteq E\};$$

it is *closed inner regular* if

$$\mu(E) = \sup\{\mu(F) : F \text{ is closed in } X \text{ and } F \subseteq E\};$$

it is *compact inner regular* if

$$\mu(E) = \sup\{\mu(K) : K \text{ is compact in } X \text{ and } K \subseteq E\}.$$

We say a finite measure μ is *tight* if

$$\mu(X) = \sup\{\mu(K) : K \text{ is compact in } X \text{ and } K \subseteq X\}.$$

1.35 Proposition. Every finite measure on a topological space is outer regular if and only if it is closed inner regular.

The proof is obvious. If a set is outer regular, then its complement is inner regular.

1.36 Theorem [Bil99, Theorem 1.1]. For a finite measure μ on a metric space X with the Borel σ-algebra, μ is both outer regular and closed inner regular. It follows that a tight Borel measure is compact inner regular, by Proposition A.10.

Proof. Here is a common way to characterize the regularity of measures: for all $E \in \mathcal{B}(X)$, for all ϵ, there exist closed F and open G such that $F \subseteq E \subseteq G$ with $\mu(G - F) < \epsilon$. We will refer to this as the regularity condition in this problem.

If we can prove that 1) the above claim holds for all closed sets E, and then show that 2) the collection of all E's satisfying the regularity condition forms a σ-algebra, then we are done.

Let E be closed, and define $U_n = \{x : d(x, E) < \frac{1}{n}\}$.[12] These U_n's are open, since U_n^c is the continuous preimage of a closed set $[1/n, \infty)$. Also $U_n \downarrow \{x : d(x, E) = 0\}$, which is exactly E since E is closed. Therefore $\mu(U_n) \to \mu(E)$. This proves 1).

Now we show 2). Clearly if E is regular, then E^c is regular. It remains to prove that the regularity condition is closed under countable union. Let $E_1, E_2, \ldots$ be regular. Fix $\epsilon > 0$, then we can choose F_n and G_n such that $F_n \subseteq E_n \subseteq G_n$ and $\mu(G_n - F_n) < \epsilon/2^{n+1}$ for each n. Let $G = \bigcup_{n=1}^{\infty} G_n$, which is open, and

$$\mu\left(G - \bigcup_{n=1}^{\infty} F\right) \leq \mu\left(\bigcup_{n=1}^{\infty}(G_n - F_n)\right) \leq \sum_{n=1}^{\infty} \mu(G_n - F_n) < \epsilon/2.$$

Also let $F = \bigcup_{n=1}^{N} F_n$, a closed set, where the picked N forces $\mu(\bigcup_{n=1}^{\infty} F_n - F) < \epsilon/2$. (This is possible because μ is a finite measure.) Combining these two gives us $\mu(G - F) < \epsilon$, where $F \subseteq E \subseteq G$, as desired. □

If X is σ-compact, then X is compact inner regular.

1.37 Theorem. Lebesgue–Stieltjes measures are compact (and closed) inner regular and outer regular.

[12] See Proposition A.6 if you are not familiar with the definition of $d(\cdot, E)$.

Given a measure μ on a measurable space $(X, \mathcal{A})$, we can define its *induced outer measure* $\mu^*\colon \wp(X) \to [0, \infty]$ and *induced inner measure* $\mu_*\colon \wp(X) \to [0, \infty]$ respectively by

$$\mu^*(A) = \inf\{\mu(B) : A \subseteq B \in \mathcal{A}\} \quad \text{and} \quad \mu_*(A) = \sup\{\mu(B) : \mathcal{A} \ni B \subseteq A\}.$$

1.38 Fact. The measure μ is complete if and only if it contains all sets with zero induced outer measure.

Chapter 2 Measurable functions and integration

2.A Measurable functions

2.1 Definition. Given two measurable spaces $(X, \mathcal{M})$ and $(Y, \mathcal{N})$, a function $f : X \to Y$ is called a *measurable function* if $f^{-1}(A) \in \mathcal{M}$ for all $A \in \mathcal{N}$.

We would stress that the function is $\mathcal{M}/\mathcal{N}$-measurable if the context is not clear. When $(Y, \mathcal{N}) = (\mathbf{R}, \mathcal{B})$, we usually say f is $\mathcal{M}$-measurable[1]. Therefore when $\mathcal{M} = \mathcal{B}_X$ or $\mathcal{L}_X$, f would be called Borel or Lebesgue measurable, respectively.

Check on your own that compositions of measurable functions is measurable.

To check measurability, it suffices to just check preimage condition for a collection of subsets that generates the image σ-algebra $\mathcal{N}$. This is the content of the next proposition, and is a direct consequence of Proposition 1.7(b).

2.2 Proposition. If $\mathcal{N}$ is generated by $\mathcal{E}$, then $f : X \to Y$ is $\mathcal{M}/\mathcal{N}$-measurable if and only if $f^{-1}(E) \in \mathcal{M}$ for all $E \in \mathcal{E}$.

With this sufficient condition in mind, it is easy to check that

(a) continuous functions between topological spaces are Borel measurable;

(b) increasing/decreasing functions from $\mathbf{R}$ to $\mathbf{R}$ are Borel measurable.

2.3 Fact. For a *countable* sequence of measurable functions $f_n : X \to \mathbf{R}$, we have $\sup_n f_n$ and $\inf_n f_n$ measurable. It follows that $\limsup_n f_n = \inf_n(\sup_{j \geq n} f_n)$ and $\liminf_n f_n$ are both measurable as well, and hence $\lim_n f_n$ is measurable if it exists.

2.4 Exercise. Lower and upper semicontinuous functions are measurable (in the extended sense).

Given a set X, a measurable space $(Y, \mathcal{N})$, and a function $f : X \to Y$, then by Proposition 1.7(b) we know

$$\{f^{-1}(A) : A \in \mathcal{N}\}$$

is the smallest σ-algebra on X that makes f measurable. We call it the *σ-algebra generated by f*, denoted by $\sigma(f)$.

More generally, consider a collection of measurable spaces $(Y_\alpha, \mathcal{N}_\alpha)$ over all $\alpha \in I$. Suppose we are given $f_\alpha : X \to Y_\alpha$ for all α. The *σ-algebra generated by the class of functions $\{f_\alpha\}_{\alpha \in I}$* on X is defined to be

$$\sigma(\{f_\alpha\}_{\alpha \in I}) = \sigma\big(\cup_{\alpha \in I}\{f^{-1}(A_\alpha) : A_\alpha \in \mathcal{N}_\alpha\}\big).$$

(Recall that union of σ-algebras is not necessarily a σ-algebra.)

[1] Now be aware that either a set or a function may be called $\mathcal{M}$-measurable.

2.5 Proposition. For any $\sigma(f)/\mathcal{B}(\mathbf{R})$ measurable function φ, there is a Borel-measurable function g such that $\varphi = g \circ f$.

2.6 Simple function approximation. Given $f \in L^+(X, \mathcal{A})$, there exists a sequence of nonnegative simple functions $\{s_n\}_{n=1}^{\infty}$ such that $s_n \uparrow f$ pointwise. Furthermore $s_n \to f$ uniformly on any set on which f is bounded.

Note that the "furthermore" part essentially means that every nonnegative bounded measurable function is the increasing uniform limit of nonnegative simple functions.

Folland Ex 2.9

Baire σ-algebra

2.B Nonnegative Lebesgue integrals

Repartition function is cadlag

2.7 Monotone convergence theorem. If $\{f_n\} \subseteq L^+$ such that $f_n \uparrow f$, then

$$\int f = \lim_n \int f_n$$

2.8 Fatou's lemma. Let $\{f_n\} \subseteq L^+$, then

$$\int \left(\liminf_n f_n\right) \leq \liminf_n \int f_n$$

Fatou's lemma is usually useful when one of the two lim inf's is attained.

We see an example when the equality is not achieved. Let the measure space be $(\mathbf{R}, \mathcal{B}, m)$, and set $f_n = n\mathbf{1}_{(0,1/n]}$. Then $\lim f_n = 0$, while $\liminf \int f_n = 1$.

2.C Signed Lebesgue integrals

2.9 Lebesgue dominated convergence theorem. If $f_n \to f$ pointwise a.e. [limit], and there exists some nonnegative $g \in L^1$ such that $|f_n| \leq g$ a.e. for all n, [bound] then $f \in L^1$ with the L^1 convergence

$$\lim_n \int |f - f_n| = 0.$$

(The type of convergence above is known as L^1 convergence; see Section 2.E.) In particular, we have

$$\int f = \lim_n \int f_n.$$

2.10 Bounded convergence theorem. When the measure space is finite, it is clear that we can set g in the theorem above to be a nonnegative real number M.

Aside from showing convergence of integrals, the above theorems are used to establish the continuity of integrals of parametrized function, and allow us to perform differentiation under the integral sign; see Section 2.H for precise statements.

2.11 Markov's inequality. Let $f : X \to \mathbf{R}$ be measurable and $\varphi : \mathbf{R} \to [0, \infty)$ be increasing (and hence measurable). Then for any $a \in \mathbf{R}$ with $\varphi(a) \neq 0$, we have

$$\mu\{x : f(x) \geq a\} \leq \frac{1}{\varphi(a)} \int \varphi \circ f \, d\mu.$$

The above statement still holds if we replace all $\mathbf{R}$ above by $[0, \infty)$.

Proof. Fix a with $\varphi(a) \neq 0$. Using φ is increasing and nonnegative, we have

$$\begin{aligned} \varphi(a)\mu\{x : f(x) \geq a\} &\leq \int_{\{x : f(x) \geq a\}} \varphi(a) \, d\mu(x) \\ &\leq \int_{\{x : f(x) \geq a\}} \varphi\big(f(x)\big) \, d\mu(x) \\ &\leq \int \varphi\big(f(x)\big) \, d\mu(x). \end{aligned}$$

□

If we let $\varphi(y) = y^p$ $(0 < p < \infty)$, and use $|f|$ in place of $f : X \to \mathbf{R}$, then we get for any $a > 0$,

$$\mu\{x : |f| \geq a\} \leq \frac{1}{a^p} \int |f|^p \, d\mu. \tag{2.12}$$

2.13 Jensen's Inequality. Let μ be a probability measure, and $f \in L^1$. Suppose I is an interval containing the range of f, and we have a convex function $\varphi : I \to \mathbf{R}$. Then

$$\varphi\left(\int f \, d\mu\right) \leq \int \varphi \circ f \, d\mu. \tag{2.14}$$

We do not ask $\varphi \circ f \in L^1$. When $\varphi \circ f \notin L^1$, the integral attains $+\infty$.

Equality condition

2.D Connections to the Riemann theory

2.15 Bounded convergence theorem (Riemann integration).

We use $\int_a^b f(x) \, dx$ for Riemann integrals, and $\int_{[a,b]} f(x) \, dm(x)$ for Lebesgue integrals.
Improper Riemann integral
An improper Riemann integral is Lebesgue integrable if it is absolutely convergent.
but $\frac{\sin x}{x} \mathbf{1}_{[0,\infty)}$ is not Lebesgue integrable.

2.16 Dirichlet integral. Let us show $\int_0^\infty \frac{\sin x}{x} \, dx = \pi/2$. The easiest solution is to use the double integral trick.

2.E Modes of convergence

2.17 Definition. For a sequence of measurable functions f_n, we say f_n converges to some function f

- *almost everywhere* (a.e.) if
$$\mu\{x : \lim_n f_n(x) = f(x)\}^c = 0.$$
- *in* L^p ($1 \le p < \infty$), if $\int |f_n|^p < \infty$ for all n, and
$$\int |f_n - f|^p \to 0.$$
In Section 5.A we will show that the limiting function f also has $\int |f|^p < \infty$, along with other basic facts about L^p spaces.
- *in measure* if for any $\epsilon > 0$,
$$\lim_n \mu\{x : |f_n(x) - f(x)| > \epsilon\} = 0. \tag{2.18}$$

We say $\{f_n\}$ is

- *Cauchy/fundamental in measure* if for any $\epsilon > 0$, there exists $N \in \mathbf{N}$ such that for all $m > n \ge N$,
$$\mu\{x : |f_n(x) - f_m(x)| > \epsilon\} < \epsilon \tag{2.19}$$

Note that the ">" in both (2.18) and (2.19) can be replaced by "$\ge$", obviously. It suffices to use only one ϵ in (2.19) because we can always choose the smaller of two distinct ϵ's.

2.20 Theorem (relationships between different modes of convergence).

(a) The a.e.-limit, L^p-limit, and limit-in-measure are all unique a.e.

(b) $f_n \to f$ in measure implies $\{f_n\}$ is Cauchy in measure; and $\{f_n\}$ being Cauchy in measure implies $f_n \to f$ in measure for some f.

(c) $f_n \to f$ in measure implies there exists a subsequence $\{f_{n_k}\}$ that converges a.e. to f as $k \to \infty$.

(d) Convergence in L^p implies convergence in measure.

(e) If the measure space is finite, then convergence a.e. implies convergence in measure. (Hence in a finite measure space, if a function converges a.s./in measure and in L^p, then the two limits should agree.)

(f) $f_n \to f$ in measure if and only if for every subsequence f_{n_k} there exists a further subsequence $f_{n_{k_j}}$ that converges in measure to f.

Proof.

(a) The first is obvious. The second follows from Minkowski's inequality; in particular when $p = 1$ we may just use the triangular inequality.

For the third one, suppose f and g are both limits-in-measure. Then for any $\epsilon > 0$, it holds that
$$\lim_n \mu\big\{x : |f_n(x) - f(x)| > \epsilon/2 \text{ or } |f_n(x) - g(x)| > \epsilon/2\big\} = 0.$$
This implies
$$\mu\{x : |f(x) - g(x)| > \epsilon\} = 0.$$

The result follows by ϵ being arbitrary.

We emphasize that the containment relation

$$|f(x)-g(x)|>\epsilon \implies |f(x)-h(x)|>\epsilon/2 \text{ or } |h(x)-g(x)|>\epsilon/2 \tag{2.21}$$

for some appropriate functions f, g, h, is the common trick used to prove convergence in measure.

(b) The first claim is easy and left to the readers, again by the containment relation (2.21). For the second one, the idea is to construct a subsequence that converges pointwise a.e. to some function, which we prove is our f.

For each $k \in \mathbf{N}$, define $g_k = f_{n_k}$, where n_k is the smallest integer such that

$$\mu\{x : |f_n(x)-f_m(x)|>2^{-k}\} < 2^{-k} \quad \text{for all } m \geq n \geq n_k. \tag{2.22}$$

We claim this appropriately picked sequence $g_k = f_{n_k}$ converges for a.e. x. This is equivalent to proving that g_k is a.e. Cauchy.

Note that g_k is exactly the desired subsequence in part (c), by our claim that convergence in measure implies Cauchy in measure.

Define

$$E_j = \{x : |g_j(x)-g_{j+1}(x)| \geq 2^{-j}\}.$$

This gives

$$\mu\left(\bigcup_{j=k}^{\infty} E_j\right) \leq \sum_{j=k}^{\infty} 2^{-j} = 2^{-k+1},$$

which goes to 0 as $k \to \infty$. Hence $\mu(\limsup_k E_k) = 0$, that is, a.e. x falls in $\{E_k\}_{k=1}^{\infty}$ eventually.[2] To be precise, there is this $N \in \mathbf{N}$ such that for all $k \geq N$, for all $m > n \geq k$, it holds for a.e. x that

$$\begin{aligned} |g_n(x)-g_m(x)| &\leq \sum_{j=n}^{m-1} |g_j(x)-g_{j+1}(x)| \\ &\leq 2^{-n+1} \leq 2^{-k+1}. \end{aligned}$$

Hence we have a pointwise a.e. limit f of $\{g_k\} = \{f_{n_k}\}$. In fact g_k converges in measure to f as well. (If the measure space is finite we may use part (e), but this is true in general.)

Fix k, we have proved already that $\mu(\bigcup_{j=k}^{\infty} E_j) \leq 2^{-k+1}$; and for $x \notin \bigcup_{j=k}^{\infty} E_j$, for $m > n \geq k$,

$$|g_n(x)-g_m(x)| \leq 2^{-k+1}.$$

Take $m \to \infty$ in the inequality above, and we have for $x \notin \bigcup_{j=k}^{\infty} E_j$, there is k such that for all $n \geq k$,

$$|g_n(x)-f(x)| \leq 2^{-k+1},$$

This yields $g_k \to f$ in measure.

[2] The reader might notice that we have implicitly proved and used Borel–Cantelli lemma I here. This is how convergence a.e. is usually proved, and we will see more applications of this when discussing probability. The main reason we have not invoked Borel–Cantelli directly is that we will use the inequality again in the next section of the proof.

The final step is to use this to show $f_n \to f$ in measure. We again resort to the containment relation (2.21):

$$|f_n(x) - f(x)| > \epsilon \implies \underbrace{|f_n(x) - g_k(x)| > \epsilon/2}_{\text{terms in a Cauchy sequence}} \text{ or } \underbrace{|g_k(x) - f(x)| > \epsilon/2}_{\substack{\text{terms in a sequence} \\ \text{that converges in measure}}}.$$

Hence $f_n \to f$ in measure, as desired.

(c) Contained in the previous part.

(d) This is clearly a consequence of (2.12).

(e) Fix $\epsilon > 0$, define $E_n = \{x : |f_n(x) - f(x)| < \epsilon\}$. Recall $\liminf_n E_n$ consists of all x such that $|f_n(x) - f(x)| < \epsilon$ eventually.

Since ϵ has been fixed, we have $\liminf_n E_n$ should contain all x such that $f_n(x) \to f(x)$. By assumption

$$\mu(X) = \mu\{x : f_n \to f\} \le \mu\big(\liminf_n E_n\big) \le \liminf_n \mu(E_n),$$

which now implies $\mu(X) = \liminf_n \mu(E_n) = \lim_n \mu(E_n)$. This exactly means $f_n \to f$ in measure.

(f) The "only if" direction is trivial. The "if" direction, on the other hand, clearly resembles Proposition A.2: fix $\epsilon > 0$ and consider $y_n = \mu\{x : |f_n(x) - f(x)| > \epsilon\}$. □

2.23 Example. Part (e) is not true in general for infinite measure spaces: let μ be Lebesgue measure on $\mathbf{R}$, the sequence of functions specified by $f_n = \mathbf{1}_{[n,n+1)}$ converges to 0 a.e., but not in measure.

Convergence in L^p (and hence in measure) does not imply convergence a.e.: specify $f_n = \mathbf{1}_{[j/2^k,(j+1)/2^k)}$, where $n = 2^k + j$ with $0 \le j < 2^k$. The sequence dyadically moves across $[0, 1)$, in the sense that $f_1 = \mathbf{1}_{[0,1)}$, $f_2 = \mathbf{1}_{[0,1/2)}$, $f_3 = \mathbf{1}_{[1/2,1)}$, $f_4 = \mathbf{1}_{[0,1/4)}$, $f_5 = \mathbf{1}_{[1/4,1/2)}$, and so on. The sequence converges to 0 in L^1, but not a.e. This is a very important example to remember.

Pointwise, a.e., and uniform convergence does not give L^p convergence: consider $f_n = \frac{1}{n}\mathbf{1}_{[n,n+1)}$, $n\mathbf{1}_{[0,1/n)}$, and $\frac{1}{n}\mathbf{1}_{[0,n)}$ respectively, which converges pointwise, a.e., and uniformly to 0 but not in L^1.

2.24 Exercise. Give a proof of Theorem 2.20(e) using the bounded convergence theorem.

2.25 Fact. Convergence a.e. is preserved under continuous composition: given $f_n \to f$ a.e. and a continuous function $\Psi\colon \mathbf{R} \to \mathbf{R}$, then $\Psi(f_n) \to \Psi(f)$ a.e.

2.26 Corollary. Let μ be finite, and $f_n \to f$ and $g_n \to g$ in measure. Say $\Psi\colon \mathbf{R}^2 \to \mathbf{R}$ is a continuous function, then $\Psi(f_n, g_n) \to \Psi(f, g)$ in measure. In particular, $f_n + g_n \to f + g$ and $f_n g_n \to fg$ in measure.

Proof. The measurabilities of $\Psi(f_n, g_n)$ and $\Psi(f, g)$ are left to the readers. Suppose by contradiction that $\Psi(f_n, g_n) \not\to \Psi(f, g)$ in measure, then for some $\epsilon > 0$ and a subsequence $\{(f_{n_k}, g_{n_k})\}_k$ of $\{(f_n, g_n)\}_n$ we have

$$\mu\big\{x : \big|\Psi\big(f_{n_k}(x), g_{n_k}(x)\big) - \Psi\big(f(x), g(x)\big)\big| > \epsilon\big\} \ge \epsilon. \tag{2.27}$$

Recall the construction of the subsequence in Theorem 2.20(c). An obvious modification of n_k there, or n_{k_j} in our context, gives us a subsequence $\{n_{k_j}\}$ of $\{n_k\}$ such that simultaneously

$$f_{n_{k_j}} \to f \quad \text{and} \quad g_{n_{k_j}} \to g \quad \text{a.e.}$$

It follows that

$$\Psi\big(f_{n_{k_j}}(x), g_{n_{k_j}}(x)\big) \to \Psi\big(f(x), g(x)\big) \quad \text{a.e.},$$

and hence in measure. But this contradicts our pick of $\{n_k\}$ specified by (2.27). □

This proof shows the power of both part (c) and (e). Remember that extracting an a.e. convergent can be helpful in many proofs involving convergence in measure.

2.28 Remark. One can prove directly that two most important cases, $f_n + g_n \to f + g$ and $f_n g_n \to fg$ in measure above, without using proof by contradiction. One will also see that it is unnecessary to assume finite measure space when proving $f_n + g_n \to f + g$ in measure. We leave these as an exercise to the interested readers.

2.29 Exercise. Use Theorem 2.20(c) to prove the monotone convergence theorem and Fatou's lemma with convergence in measure.

2.F Littlewood's second and third principles

2.30 Egoroff's theorem. Say $\mu(X) < \infty$. Let $\{f_n\}$ be a sequence of $\mathcal{A}$-measurable functions from X to $\mathbf{R}$ (or $\mathbf{C}$) that converges to f a.e. Then for all $\epsilon > 0$, there exists some measurable set E such that

$$\mu(E^c) < \epsilon, \quad \text{while } f_n \to f \text{ uniformly on } E.$$

We call this conclusion f_n converges to f *almost uniformly*.

We mention that it is a good exercise to prove the bounded convergence theorem using this result.

2.31 Classical Luzin's theorem. Let $f : [a, b] \to \mathbf{R}$ (or $\mathbf{C}$) be a Borel measurable function. Then for every $\epsilon > 0$, there exists a closed set $F \subseteq [a, b]$ such that $f|_F$ is continuous while $m([a, b] - F) < \epsilon$.

when f takes values in a separable metric space, the reason will become

Santambrogio [San15, Box 1.6] mentions two types of Luzin's theorem: the *weak* Luzin's theorem only cares about the continuity of $f : A \to Y$ restricted to a closed/compact subset, while the *strong* Luzin's theorem also considers whether we may find a continuous function $g : A \to Y$ that coincides with f on this closed/compact subset.

The proof for finite measure μ and f defined on general (topological) spaces is given in the aforementioned source. In addition, the strong Luzin's theorem for Lebesgue measure with a slick proof is given in [RF23, Section 3.3]:

2.32 Theorem. Let $A \in \mathcal{L}(\mathbf{R})$, and let $f : A \to \mathbf{R}$ be Borel measurable. For any $\epsilon > 0$, there is a continuous function $g : \mathbf{R} \to \mathbf{R}$ and a set $F \subseteq A$ that is closed in $\mathbf{R}$, that satisfies

$$m(A - F) < \epsilon \quad \text{and} \quad f|_F = g|_F.$$

2.G Uniformly integrable functions

Use the material we have discussed so far to prove the following result.

2.33 Exercise [RF23]. Let $f \in L^1(\mu)$. Then

(a) for all $\epsilon > 0$, there is a $\delta > 0$ such that

$$\mu(E) < \delta \implies \int_E |f|\, d\mu < \epsilon;$$

(b) moreover, for each $\epsilon > 0$, there is some X_0 with $\mu(X_0) < \infty$ such that

$$\int_{X - X_0} |f| < \epsilon.$$

Notice that

$$\left|\int_E f\, d\mu\right| \leq \int_E |f|\, d\mu = \left|\int_{E\cap\{f\geq 0\}} f\, d\mu\right| + \left|\int_{E\cap\{f<0\}} -f\, d\mu\right|.$$

Hence conclusion (a) is equivalent to $\forall\, \epsilon > 0$, $\exists\, \delta > 0$ such that

$$\mu(E) < \delta \implies \left|\int_E f\, d\mu\right| < \epsilon.$$

This motivates the next definition, which requires (a) to hold uniformly for a class of integrable functions.

2.34 Definition. A set of functions $\mathcal{F} \subseteq L^1(\mu)$ has *uniformly absolutely continuous integrals* if for every $\epsilon > 0$, there exists $\delta > 0$ such that

$$\mu(E) < \delta \implies \int_E |f|\, d\mu < \epsilon \text{ for all } f \in \mathcal{F},$$

or equivalently,

$$\left|\int_E f\, d\mu\right| < \epsilon \text{ for all } f \in \mathcal{F}.$$

The term "absolutely continuous" that appear in the definition above is related the notion of an absolutely continuous pair of measures we will discuss in Section 4.A. Since for $f \in L^1(X, \mathcal{A}, \mu)$, $\nu(E) = \int_E |f|\, d\mu$ defines a finite positive measure ν on $\mathcal{A}$ that is absolutely continuous with respect to μ. This immediately proves conclusion (a) in Exercise 2.33.

2.35 Definition. A set of functions $\mathcal{F} \subseteq L^1(\mu)$ is *uniformly integrable* if

$$\lim_{C\to\infty} \sup_{f\in\mathcal{F}} \int_{\{|f|>C\}} |f|\, d\mu = 0.$$

These two definitions are quite obviously related, as stated by the next proposition.

2.36 Proposition. Let μ be finite, then $\mathcal{F}$ is uniformly integrable if and only if it is bounded in L^1 and also has uniformly absolutely continuous integrals.

2.37 Fact. Any finite collection of L^1 functions is uniformly integrable. Any collection of bounded functions is uniformly integrable.

The following proposition gives an easy sufficient condition for uniform integrability. Note that this $p > 1$ will come back later

2.38 Proposition. Suppose there exists some $p > 1$ such that the collection $\mathcal{F}$ of functions is L^p bounded (i.e., $\sup_{f\in\mathcal{F}} \int |f|^p \, d\mu < \infty$) then the collection $\mathcal{F}$ is uniformly integrable.

Proof. This might as well be left as an exercise, but we write out the proof due to its importance.

Let $C > 0$, we first observe that

$$\int_{\{|f|>C\}} |f|^p \geq C^{p-1} \int_{\{|f|>C\}} |f|.$$

Hence

$$0 \leq \sup_f \int_{\{|f|>C\}} |f| \leq \frac{1}{C^{p-1}} \sup_f \int_{\{|f|>C\}} |f|^p.$$

Now with the assumption and $p > 1$, by the squeeze theorem we conclude that the collection $\mathcal{F}$ is uniformly integrable. □

2.39 Vitali convergence theorem. Suppose μ is finite. Let $\{f_n\} \subseteq L^1(X, \mathcal{A}, \mu)$, then the following are equivalent:

(a) $f \in L^1$ with $f_n \to f$ in L^1.

(b) $f_n \to f$ in measure, and $\{f_n\}$ is uniformly integrable.

2.H Continuity and differentiability of parametrized functions

2.40 Corollary [Jos05, Theorem 16.10]. Let $y_0 \in A$, a metric space[3], and $f : X \times A \to \mathbf{R}$. Assume

(a) for every $y \in A$, the function $x \mapsto f(x, y)$ is integrable;

(b) for a.e. $x \in X$, the function $y \mapsto f(x, y)$ is continuous;

(c) there exists $g \in L^1(X)$ such that for every $y \in A$,

$$|f(x, y)| \leq g(x) \quad \text{for a.e. } x \in X.$$

We may then conclude that the integrated function

$$F : y \mapsto \int_X f(x, y) \, d\mu(x)$$

is continuous at $y_0 \in A$.

We need to check if $\{y_n\} \subseteq A$ converges to y_0, then $F(y_n) \to F(y)$. The proof is then a straightforward application of Lebesgue dominated convergence theorem to $f(x, y_n)$.

perform differentiation under the integral sign

2.41 Corollary [Jos05, Theorem 16.11]. Let $I \subseteq \mathbf{R}$ be an open interval, and $f : X \times I \to \mathbf{R}$. Assume

(a) for every $t \in I$ we have $x \mapsto f(x, t)$ is integrable;

(b) for a.e. $x \in X$, $t \mapsto \partial f/\partial t$ exists for all $t \in I$;

(c) there exists $g \in L^1(X)$ such that for every $t \in I$,

$$\left| \frac{\partial}{\partial t} f(x, t) \right| \leq g(x) \quad \text{for a.e. } x \in X.$$

[3] first countable is already enough; see Theorem A.3

We may then conclude that the function

$$F : t \mapsto \int_X f(x,t)\, d\mu(x)$$

is differentiable on I, with

$$F'(t) = \int_X \frac{\partial}{\partial t} f(x,t)\, d\mu(x).$$

Proof. We need to show for any sequence $\{h_n\} \subseteq \mathbf{R} - \{0\}$ converging to 0 that

$$\lim_n \int \frac{f(x,t+h_n) - f(x,t)}{h_n}\, d\mu = \int \frac{\partial}{\partial t} f(x,t)\, d\mu.$$

Set

$$\varphi_n(x) = \frac{f(x,t+h_n) - f(x,t)}{h_n} \text{ and } \varphi(x) = \frac{\partial}{\partial t} f(x,t).$$

For each n, by the mean value theorem, we know for some θ_n between 0 and h_n that

$$|\varphi_n(x)| = \left| \frac{\partial}{\partial t} f(x, t+\theta_n) \right| \le g(x) \text{ a.e.}$$

Now apply Lebesgue dominated convergence theorem to $\varphi_n \to \varphi$. □

can replace differentiable by almost everywhere differentiable ?

2.I Image measures

Consider a measure space $(X, \mathcal{M}, \mu)$ and a measurable space $(Y, \mathcal{N})$. If we have an $(\mathcal{M}, \mathcal{N})$-measurable function $\varphi\colon X \to Y$, then we can define a function $\mu_*\colon \mathcal{N} \to [0,\infty]$ given by

$$\mu_*(E) = \mu(\varphi^{-1} E)$$

for all $E \in \mathcal{N}$. This turns out to a measure on $(Y, \mathcal{N})$, and we call this the *image/pushforward measure* of μ by φ, denoted by $\varphi_*\mu$ or $\varphi_\#\mu$.

Image measure characterizes change of variables, which is of basic importance in mathematics. We will use image measures later in Sections 3.C, 3.E and 7.B.

We state the main result below.

2.42 Proposition. Under the conditions stated above, let $g \in L^+(Y, \mathcal{N})$ or $g \circ \varphi \in L^1(X, \mathcal{M}, \mu)$. Then

$$\int_X g\big(\varphi(x)\big)\, d\mu(x) = \int_Y g(y)\, d\mu_*(y).$$

Proof. When $g = \mathbf{1}_E$ for $E \in \mathcal{N}$, we have

$$\text{LHS} = \mu\{x : \varphi(x) \in E\} = \mu(\varphi^{-1} E) \quad \text{and} \quad \text{RHS} = \mu_*(E).$$

Now extend this to simple functions, then nonnegative functions, and then integrable functions. □

Chapter 3 Product spaces

3.A Product σ-algebras

We start with a comparison between product topologies and product σ-algebras.

For topological spaces $(X_\alpha, \mathcal{T}_\alpha)$ $(\alpha \in I)$, recall that the *product topology* $\mathcal{T}$ on $X = \prod_{\alpha\in I} X_\alpha$ is the topology generated by all coordinate projections $\pi_\alpha \colon X \to X_\alpha$ (i.e., the smallest topology on X that makes all these maps continuous). Explicitly $\mathcal{T}$ is generated by the collection of subbasic sets

$$\{\pi_\alpha^{-1}(U_\alpha) : U_\alpha \in \mathcal{T}_\alpha, \alpha \in I\}. \tag{3.1}$$

For measurable spaces $(X_\alpha, \mathcal{A}_\alpha)$ $(\alpha \in I)$, the *product σ-algebra* $\mathcal{A} = \bigotimes_{\alpha\in I} \mathcal{A}_\alpha$ on $X = \prod_{\alpha\in I} X_\alpha$ is the σ-algebra generated by all coordinate projections π_α. Explicitly $\mathcal{A}$ is generated by the collection of sets

$$\{\pi_\alpha^{-1}(E_\alpha) : E_\alpha \in \mathcal{A}_\alpha, \alpha \in I\}. \tag{3.2}$$

Define the general *cylinder sets*[1] on the product of topological spaces $(X_\alpha, \mathcal{T}_\alpha)$ and measurable spaces $(X_\alpha, \mathcal{A}_\alpha)$ to be the sets of form

$$\bigcap_{j=1}^{n} \pi_{\alpha_j}^{-1}(U_{\alpha_j}) \quad \text{and} \quad \bigcap_{j=1}^{n} \pi_{\alpha_j}^{-1}(E_{\alpha_j}),$$

for any $n \in \mathbf{N}$, respectively. To put them into simple words, they are finite intersections of preimages of the projections. The collection of sets in (3.1) and (3.2) are 1-dimensional cylinders.

The general cylinder sets on the product of topological spaces, as finite[2] intersections of subbasic sets in (3.1), form a basis for the product topology $\mathcal{T}$. However, it is a well-known fact that σ-algebras, unlike topologies, cannot be written out explicitly from the elementary sets they are generated from.

Looking back at (3.2), you may expect a smaller collection of cylinder sets generates the product σ-algebra. Yet the proof is a little weird, like most arguments involving algebras of sets.

3.3 Proposition. Suppose each $\mathcal{A}_\alpha$ is generated by $\mathcal{E}_\alpha$. Then $\bigotimes_\alpha \mathcal{A}_\alpha$ is generated by the collection

$$\mathcal{K} = \{\pi_\alpha^{-1}(E_\alpha) : E_\alpha \in \mathcal{E}_\alpha, \alpha \in I\}.$$

Proof. Let the collection in (3.2) be $\mathcal{J}$. Clearly $\mathcal{K} \subseteq \mathcal{J}$. To see the other inclusion, consider the induced σ-algebra on X_α

$$\{E \subseteq X_\alpha : \pi_\alpha^{-1}(E) \in \sigma(\mathcal{K})\},$$

which contains $\mathcal{E}_\alpha$ and hence $\mathcal{A}_\alpha$. This means $\pi_\alpha^{-1}(E) \in \sigma(\mathcal{K})$ for all $\alpha \in I$ and $E \in \mathcal{A}_\alpha$. Hence $\mathcal{J} \subseteq \sigma(\mathcal{K})$. The proof is now complete. □

[1] This definition similarly holds for other set-collection pairs.

[2] As another reminder, if the intersection is allowed to be arbitrary, then we get a larger topology called the *box topology*. The box topology is generated by arbitrary products of open sets. When the product is finite, the box topology and the product topology coincide.

We have introduced very general definitions above. The reader should verify on their own that in the case where I is countable, $\mathcal{A} = \bigotimes_{k=1}^{\infty} \mathcal{A}_k$ is generated by

$$\left\{ \prod_{k=1}^{\infty} E_k : E_k \in \mathcal{A}_k \right\}.$$

Also, for measurable spaces $(X_1, \mathcal{A}_1), (X_2, \mathcal{A}_2), \ldots$, the product σ-algebra $\mathcal{A}$ is clearly generated from cylinder sets of the form

$$I_{n,B} = B \times \prod_{k=n+1}^{\infty} X_n, \text{ where } B \in \bigotimes_{k=1}^{n} \mathcal{A}_k.$$

This turns out to be clean to work with.

3.5.1 3.5.2 Bogachev

Since the Borel σ-algebra is the σ-algebra generated by open set, while the topological space consists of all the open sets. With our above detailed comparisons between product σ-algebras and product topological spaces, the Borel σ-algebra from the product topology and the product Borel σ-algebra from individual spaces should be the same, under some conditions.

3.4 Theorem [Bog07, Lemma 6.4.2]. For any second countable spaces $X_1, X_2, \ldots$ (finite or countably infinite), we have

$$\mathcal{B}(X) = \mathcal{B}(X_1) \otimes \mathcal{B}(X_2) \otimes \cdots, \tag{3.5}$$

where $X = X_1 \times X_2 \times \cdots$ with product topology $\mathcal{T}$.

Proof. We follow the proof in [Kal02][3].

Let $\mathcal{J}$ be the class of 1-dimensional cylinder sets

$$X_1 \times \cdots \times X_{k-1} \times U_k \times X_{k+1} \times \cdots$$

over all $k \in \mathbf{N}$ and $U_k \in \mathcal{T}_k$.

Since $\mathcal{J}$ consists entirely of open sets, and RHS $= \sigma(\mathcal{J})$ by Proposition 3.3, we have LHS $\supseteq$ RHS. *Note that this inclusion does not use any topological assumptions on the X_n's.*

If we can now show that $\mathcal{T} \subseteq \sigma(\mathcal{J})$, the proof will be complete. Now $(X, \mathcal{T})$, as a countable product of second countable spaces, is still second countable. Here we use a result from topology, included as Fact A.18 in the appendix:

> Every collection of open sets in a second countable space contains a countable subcollection with the same union.

Therefore every open set in X is a countable union of basic open sets. Since a topological basis is given by finite intersections of the cylinder sets in $\mathcal{J}$, we then have $\mathcal{T} \subseteq \sigma(\mathcal{J})$. □

Specifically, the result above holds for separable metric spaces; in particular, we have $\mathcal{B}(\mathbf{R}^d) = \bigotimes^d \mathcal{B}(\mathbf{R}^1)$. This theorem overall shows the fundamental importance of Borel σ-algebra in measure theory and its applications: it connects measurability to the underlying topological spaces.

As an exercise, use Proposition 2.2 to show the following:

3.6 Exercise [Fol99, Proposition 2.4]. Given measurable spaces $(X, \mathcal{M})$ and $(Y_\alpha, \mathcal{N}_\alpha)$ over all $\alpha \in I$. Let $Y = \prod Y_\alpha$ and $\mathcal{N} = \bigotimes \mathcal{N}_\alpha$. Then $f : X \to Y$ is $\mathcal{M}/\mathcal{N}$-measurable if and only if each $f_\alpha = \pi_\alpha \circ f$ is $\mathcal{M}/\mathcal{N}_\alpha$-measurable.

[3] We cite the second edition of the famous book here. The new proof in the third edition is misleading.

3.7 Remark. Say we are given a metric space (X, ρ) with the Borel σ-algebra. Fact A.1 says $\rho\colon X \times X \to [0, \infty)$ is continuous. It will appear later that we need to integrate this metric function, and therefore we need to ensure measurability of ρ with respect to the product σ-algebra $\mathcal{B}(X) \otimes \mathcal{B}(X)$.

If we assume X is separable, then by Theorem 3.4 we know $\mathcal{B}(X) \otimes \mathcal{B}(X) = \mathcal{B}(X \times X)$, which contains all the open sets in $X \times X$. Hence ρ becomes a measurable function.

One can already find an application of the above remark in Egoroff's theorem, albeit not in the context of integration. We may assume in general f_n and f to take value in a separable metric space there: the measurability of $x \mapsto d\big(f_n(x), f(x)\big)$ suffices for the proof to work.

3.B Integration on product spaces

Let $(X, \mathcal{M}, \mu)$ and $(Y, \mathcal{N}, \nu)$ be two measure spaces. We need to define a measure on the product space $(X \times Y, \mathcal{M} \otimes \mathcal{N})$. It is obvious that such a measure π should satisfy the condition that for any pair $A \in \mathcal{M}$ and $B \in \mathcal{N}$,

$$\lambda(A \times B) = \mu(A)\nu(B). \tag{3.8}$$

In this way, the idea of the area of a rectangle carries over our desired measure on the product space.

In fact sets $A \times B$ are often given the name *measurable rectangles*, and note that the collection $\mathcal{R}$ of all such measurable rectangles is a π-system.

Henceforth we will make the assumption that μ and ν are σ-finite. We need to establish that first, it is possible to extend the definition of λ to the entire $\mathcal{M} \otimes \mathcal{N}$, and get a unique *product measure*, denoted by $\mu \times \nu$. Second, we may compute the integral of a function $f : X \times Y \to \mathbf{R}$ (or $\mathbf{C}$) on the product space by doing a double integration with respect to the marginals dx and dy, whether you choose to integrate $f(x, y)\, dx$ or $f(x, y)\, dy$ first.

Define $E_x = \{y \in Y : (x, y) \in E\}$, and similarly define $E^y = \{x \in X : (x, y) \in E\}$. To understand the definition of E_x, imagine drawing a line $\{x\} \times Y$, and the proportion that hits E is exactly E_x.

For $E \in \mathcal{M} \times \mathcal{N}$, we have for all $x \in X$ and $y \in Y$,

$$E_x \in \mathcal{N} \quad E^y \in \mathcal{M}.$$

folland exercise 12

We reserve the discussion of an extremely important existence results about probability measures on product spaces to Appendix K. The first of the three results () tells us that there is a *natural* extension of product probability measures over all finite cylinder sets to a product probability measure over the entire product σ-algebra. The second and third results () say that if a sequence of probability measures are specified in a *consistent way*, then there is a natural extension of them to a product measure on the entire product σ-algebra.

Note that it makes sense to only discuss the countable product of *probability* measures, so that both the coordinate measures, the finite-dimensional product measures. and the countable product measures are all *normalized*. Because of this, and the significance of the existence theorems for product measures in probability, we delay our discussion of these two results despite their purely measure-theoretic statements and proofs.

Many books in probability only include and applies it as a special case

3.9 Fubini–Tonelli theorem.

3.C Change of variables

3.10 Proposition. Lipschitz functions maps Lebesgue null sets to Lebesgue null sets. Hence the Lipschitz image of Lebesgue measurable sets is Lebesgue measurable.

3.11 Sard's theorem. Let A be an open subset of $\mathbf{R}^n$. If $\varphi\colon A \to \mathbf{R}^m$ is a C^{n-m+1} map, then the set of critical values of φ has measure 0 in $\mathbf{R}^m$.

3.12 Change of variables for injective C^1 functions. Let A be an open subset of $\mathbf{R}^n$ and $\varphi\colon A \to \mathbf{R}^n$ be an injective C^1 mapping. Then for any $g \in L^+(A)$ or $L^1(A)$, we have

$$\int_{\varphi(A)} g(y)\,dy = \int_A g\big(\varphi(x)\big)\big|\det D_\varphi(x)\big|\,dx.$$

For Lebesgue measurable subset E of A, $G(E)$ is also Lebesgue measurable, with

$$m\big(\varphi(A)\big) = \int_A |\det D_\varphi(x)|\,dx.$$

See [Tay06, Appendix F] for the case when G is not even assumed to be injective, and references to further generalizations.

3.D Properties of the product Lebesgue measure

3.13 Brunn–Minkowski inequality. For two compact sets in $\mathbf{R}^n$, we have

(a) (additive ver.) $m(A)^{1/n} + m(B)^{1/n} \le m(A+B)^{1/n}$.

(b) (multiplicative ver.) $m(A)^{1-\lambda} m(B)^{\lambda} \le m\big((1-\lambda)A + \lambda B\big)$ for any $0 < \lambda < 1$.

If we substitute A by $(1-\lambda)A$ and B by λB, and use the logarithm convexity inequality, then we get the multiplicative version from the additive version. (One can in fact show the two versions are equally strong, which we leave as an exercise.)

The inequality is a direct consequence of the concavity of logarithm.

3.14 Logarithm convexity inequality. For $0 < \lambda < 1$ and $a, b \ge 0$, we have

$$a^{1-\lambda} b^{\lambda} \le (1-\lambda)a + \lambda b,$$

which attains equality if and only if $a = b$.

3.15 Prékopa–Leindler inequality.

log-concave measure
Gaussian measure

3.E The Gamma function and polar coordinates

$\mathbf{R}^n$ and S^{n-1}

Cauchy formula for repeated integration

Let $z \in \mathbf{C}$ with $\operatorname{Re} z > 0$, and we define $f_z\colon (0,\infty) \to \mathbf{C}$ by

$$f_z(t) = t^{z-1} e^{-t} = \exp\big((z-1)\log t\big) \cdot e^{-t}.$$

Since

3.16 Theorem. There is a unique Borel measure σ on S^{n-1} such that $m_* = \rho \times \sigma$. If $f \in L^+$ or L^1, then we have

$$\int_{\mathbf{R}^n} f(x)\,dx = \int_{[0,\infty)} \int_{S^{n-1}} f(ry) r^{n-1}\,d\sigma(y)\,dr.$$

$\sigma(S^{n-1}) = \frac{2\pi^{n/2}}{\Gamma(n/2)}$ and $m(B^n) = \frac{1}{n}\sigma(S^{n-1}) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}$. For any $\epsilon > 0$, we have $S^{n-1} \subseteq B^n(0;1+\epsilon) - B^n(0;1)$

$$\begin{aligned} m(S^{n-1}) &\le m\big(B^n(0;1+\epsilon)\big) - m\big(B^n(0;1)\big) \\ &\le (1+\epsilon)^n m(B^n) - m(B^n). \end{aligned}$$

Take $\epsilon \to 0^+$, it is easy to see that $m(S^{n-1}) = 0$. surface area

Chapter 4 Structure of measures and integrals

4.A Hahn–Jordan decomposition of signed measures

Previously we generalized integrals of nonnegative function to integrals of general signed functions and complex functions. We can make a similar generalization of positive measures to **R** and **C**-valued measures. One of the key goals of this chapter is to explore the intrinsic relationships between measures, functions, and integrals.

4.1 Definition. Given a measurable space $(X, \mathcal{A})$, a *signed/real measure* (resp. *complex measure*) on the space is a function $\mu : \mathcal{A} \to \mathbf{R}$ (resp. $\mu : \mathcal{A} \to \mathbf{C}$) such that

(a) $\mu(\emptyset) = 0$;

(b) μ is σ-additive, i.e., $\mu(E) = \sum_{n=1}^{\infty} \mu(E_n)$ for all measurable partitions $\{E_n\}$ of E.

Note condition (b) implicitly requires the series $\sum \mu(E_n)$ to be absolutely convergent. An important result that says a series is absolutely convergent if and only if any rearrangement of terms in a series yields the same limiting sum; see [Rud76, Theorems 3.54 and 3.55]. Also note that condition (b) implies condition (a), but we have stated it for clarity.

Many textbooks define the codomain of a signed measure to include one of $+\infty$ or $-\infty$. We do not adopt this convention because it is hardly used in applications, and many complications are avoided. Furthermore, restricting the codomain to the reals allows us to discuss signed and complex measures simultaneously.

In this section, we will state all our proofs for signed measures, which can all be easily extended to complex measures. To distinguish signed/complex measures from the measures we have been discussing previously, we call measures that take nonnegative values *positive measures*.

Continuity from above and below still holds for signed and complex measures. The proof here is the same as the one for positive measures.

4.2 Exercise. Let μ be a signed/complex measure. If $E_n \uparrow E$ or $E_n \downarrow E$ in $\mathcal{A}$, then $\mu(E) = \lim_n(E_n)$.

Also the inclusion-exclusion formula still holds by countable additivity. However monotonicity no longer holds for signed/complex measures, but we may make the following definitions for a signed measure.

4.3 Definition. For a signed measure μ, a measurable set A is a *positive (negative, or null) set* if for every measurable subset B of A, $\mu(B) \geq 0$ (≤ 0, or $= 0$). Equivalently, the measurable set A is positive (negative, or null) if for all $E \in \mathcal{A}$, $\mu(E \cap A) \geq 0$ (≤ 0, or $= 0$).

4.4 Hahn decomposition. Let μ be a signed measure on $(X, \mathcal{A})$. Then X has a partition into P and N such that P is a positive set and N is a negative set.

Furthermore, if P' and N' is another such partition, then $P \triangle P' = N \triangle N'$ is null. This means that the Hahn decomposition is *essentially unique*.

Proof. First we show the essential uniqueness. Consider a measurable set $E_1 \subseteq P - P'$. This E_1, as a subset of P, must have measure ≥ 0. Yet at the same time $E_1 \subseteq N' - N \subseteq N'$, which implies that $\mu(E_1) \leq 0$. Therefore $\mu(E_1) = 0$. By the same reasoning with P' switching P and N switching N', we should have $\mu(E_2) = 0$ for all measurable subsets E_2 of $P' - P$. Since $P \triangle P' = N \triangle N' = (P - P') \cup (P' - P)$, it is clear that this is a null set with respect to the signed measure μ.

Now we prove the existence. We follow the presentation in [Fal19], which avoids the axiom of dependent choice used in the proofs of most textbook authors.

To show the existence of the partition $X = P \cup N$, it suffices[1] to find some measurable N such that for all $E \in \mathcal{A}$, $\mu(E) \geq \mu(N)$. Now we prove this claim. By assumption we have $\mu(N) \leq \mu(\emptyset) = 0$. Now for any $A \in \mathcal{A}$, we have

$$\mu(N) + \mu(N \cap A) \leq \mu(N - A) + \mu(N \cap A) = \mu(N).$$

Therefore N is a negative set. For any $A \in \mathcal{A}$, we also have $P \cap A = A - N$ and

$$\mu(N) \leq \mu(A) = \mu(A - N) + \mu(N).$$

Therefore $\mu(P \cap A) \geq 0$, which means P is a positive set.

Now we find such an N with the smallest measure over all measurable sets. Let $L = \inf\{\mu(A) : A \in \mathcal{A}\}$, then we need to find $N \in \mathcal{A}$ such that $L = \mu(N)$. Since $\mathcal{A} \neq \emptyset$, by countable choice we can take a sequence $\{D_n\} \subseteq \mathcal{A}$ with $\mu(D_n) \to L$.

Let $\mathcal{A}_n$ be the algebra of subsets of $\bigcup_{n=1}^{\infty} D_n$ generated by $\{D_k\}_{k=1}^{n}$, which is a finite collection[2]. Therefore $\mu_n := \mu|_{\mathcal{A}_n}$ achieves its minimum on the collection $\mathcal{A}_n$, say at E_n. Note the same argument that proved the sufficient condition for finding a Hahn decomposition clearly works for the premeasure $\mu|_{\mathcal{A}_n}$ on the algebra $\mathcal{A}_n$: we have E_n is a μ_n-negative set and E_n^{c} is a μ_n-positive set on $\mathcal{A}_n$.

We claim that the desired $N = \liminf_m E_m$. First let $A_m^n = \bigcap_{k=m}^{n} E_m$ and let $A_m = \bigcap_{k \geq m} E_m$. Then

$$\mu(A_m^n) \to \mu(A_m)$$

as $n \to \infty$. Furthermore the limit above is a decreasing one: note

$$\begin{aligned} \mu(A_m^{n-1}) &= \mu(A_m^n) + \mu(A_m^{n-1} - E_k) \\ &= \mu(A_m^n) + \mu(A_m^{n-1} \cap E_n^{\mathrm{c}}) \\ &\geq \mu(A_m^n), \end{aligned}$$

where the last inequality follows from the observation that E_n^{c} is μ_n-positive set on $\mathcal{A}_n$ and $A_m^{n-1} \in \mathcal{A}_n$.

Now by our choice of E_m, we have

$$\mu(D_m) \geq \mu(E_m) = \mu(A_m^m) \geq \mu(A_m^{m+1}) \geq \cdots.$$

Therefore

$$\mu(D_m) \geq \mu(A_m) \geq L,$$

and taking $m \to \infty$ gives us $\mu(A_m) \to L$ as $m \to \infty$. Now the magic takes place. We know $A_m \uparrow \liminf_m E_m$, and thus $\mu(\liminf E_m) = \lim \mu(A_m)$. The two limits must agree, and hence $L = \mu(\liminf E_m)$. This finishes the proof. □

[1] This is also a necessary condition.

[2] As an exercise, show that the (σ-)algebra generated by a collection of n sets can have at most 2^{2^n} sets. (Generating a σ-algebra from a finite collection is the same as generating a topology from subbasic open sets.)

In the proof above, our negative set N attains $\inf\{\mu(A) : A \in \mathcal{A}\}$, and by symmetry our positive set P attains $\sup\{\mu(A) : A \in \mathcal{A}\}$. This implies the boundedness of μ from both above and below.

We define the *total variation* of the signed/complex measure μ to be a function $|\mu| \colon \mathcal{A} \to [0, \infty]$ given by

$$|\mu|(E) = \sup\left\{\sum_{n=1}^{\infty} |\mu(E_n)| : \{E_n\} \text{ is a measurable partition of } E\right\}, \tag{4.5}$$

the maximized "variation" over all partitions of a given set in $\mathcal{A}$.

The definition in (4.5) can be significantly simplified. Because the summands are nonnegative, we can break it into two sums:

$$\begin{aligned}\sum_{n=1}^{\infty} |\mu(E_n)| &= \sum_{j:\mu(E_j)\geq 0} |\mu(E_j)| + \sum_{k:\mu(E_k)<0} |\mu(E_k)| \\ &= \left|\sum_{j:\mu(E_j)\geq 0} \mu(E_j)\right| + \left|\sum_{k:\mu(E_k)<0} \mu(E_k)\right| \\ &= |\mu(\widehat{E})| + |\mu(\widetilde{E})|,\end{aligned}$$

where $\widehat{E} = \bigcup\{E_j : \mu(E_j) \geq 0\}$ and $\widetilde{E} = \bigcup\{E_k : \mu(E_k) < 0\}$. Therefore

$$|\mu|(E) = \sup\{|\mu(E_1)| + |\mu(E_2)| : E_1 \text{ and } E_2 \text{ are measurable and partition } E\}. \tag{4.6}$$

It is clear that we may also take finite partitions here. We may also take the partition to a measurable partition of any measurable subsets of E instead.

By the equivalent definition in (4.6), since μ is a bounded function on $\mathcal{A}$, $|\mu|$ is also bounded. This is in fact the hardest part[3] of establishing the following fact.

4.7 Theorem. The total variation $|\mu|$ of a signed/complex measure μ is a finite positive measure on $(X, \mathcal{A})$.

Proof. Obviously $|\mu|(\emptyset) = 0$. It remains to check countable additivity. □

4.8 Definition. Let the space of signed (resp. complex) measure on $(X, \mathcal{A})$ be denoted by $\mathcal{M}(X)$. The *total variation norm* is defined to be the function $\|\cdot\| \colon \mathcal{M}(X) \to \mathbf{R}$ (resp. $\mathbf{C}$) given by

$$\|\mu\| = |\mu|(X).$$

Let us first show that this $\|\cdot\|$ is indeed a norm on $\mathcal{M}(X)$.

4.9 Theorem. The space of signed/complex measures $\mathcal{M}(X)$ with the total variation norm is a Banach space.

Proof. □

The most important implication of Hahn decomposition is a *unique* decomposition of a signed measure μ into a positive and negative part, known as the *Jordan decomposition*. As we will see soon, the Jordan decomposition offers another characterization of the total variation measure we have just discussed.

Before we start, we need an additional definition.

[3] There is a very interesting direct argument that proves the finiteness of $|\mu|$ using the axiom of dependent choice; see [Rud87; ADM11; Axl20].

4.10 Definition. Let μ and ν be two positive/signed/complex measures on $(X, \mathcal{A})$. We say μ and ν are *mutually singular*, denoted by $\mu \perp \nu$, if X can be partitioned into two measurable subsets A and B, such that

$$\mu(B) = 0 \quad \text{and} \quad \nu(A) = 0,$$

or equivalently, for all $E \in \mathcal{A}$,

$$\mu(E) = \mu(E \cap A) \quad \text{and} \quad \nu(E) = \nu(E \cap B).$$

4.11 Jordan decomposition. Let μ be a signed measure on $(X, \mathcal{A})$. Then there exist unique two finite positive measures μ^+ and μ^- on $(X, \mathcal{A})$ such that

$$\mu = \mu^+ - \mu^- \quad \text{and} \quad \mu^+ \perp \mu^-.$$

4.12 Definition. Let μ be a positive measure and ν be a positive/signed/complex measure on $(X, \mathcal{A})$. We say ν is *absolutely continuous* with respect to μ, or ν is *dominated by* μ, denoted by $\nu \ll \mu$, if for all $E \in \mathcal{A}$,

$$\mu(E) = 0 \implies \nu(E) = 0. \tag{4.13}$$

More generally, to define absolute continuity $\nu \ll \mu$ for signed/complex μ, we change (4.13) to

$$|\mu|(E) = 0 \implies \nu(E) = 0. \tag{4.14}$$

This is a definition not used much in practice.

One should check that $\nu \ll \mu$ if and only if $|\nu| \ll \mu$ if and only if $\nu^+ \ll \mu$ and $\nu^- \ll \mu$. Also check that ν and ν are *equivalent measures*, in the sense that

$$\nu \ll |\nu| \ll \nu.$$

4.B Radon–Nikodym theorem and Lebesgue decomposition

Depending on what kind of measures we are looking at, there exists multiple versions of the Radon–Nikodym theorem. The following version is the most basic one in practice. It considers a pair of σ-finite and finite measures.

4.15 Radon–Nikodym theorem. Let μ be a σ-finite measure and ν be a finite measure on $(X, \mathcal{A})$, where $\nu \ll \mu$. Then there exists an $\mathcal{A}$-measurable function f such that

$$\nu(E) = \int_E f \, d\mu \quad \text{for all } E \in \mathcal{A}.$$

Furthermore this f is nonnegative and unique in $L^1(X, \mathcal{A}, \mu)$.

If the ν above is given as a signed/complex measure instead, then the same conclusions still hold after dropping f is nonnegative. If ν is given as a σ-finite measure instead, the function f becomes nonnegative real-valued[4], and is unique a.e.

Our f here is called the *Radon–Nikodym derivative/density* of ν with respect to μ, denoted by $d\nu/d\mu$.

We summarize two standard proofs of this theorem. The first of which uses results from Hilbert spaces, while the second one is based on variational principles.

[4] i.e., f takes values in $[0, \infty)$.

Proof 1, using Hilbert spaces. □

Proof 2, using variational principles. □

4.16 Lebesgue decomposition. Let μ be a positive measure and ν be a signed/complex measure on $(X, \mathcal{A})$. Then

(a) there exist two unique signed/complex measures ν_a and ν_s on $(X, \mathcal{A})$ such that

$$\nu = \nu_a + \nu_s, \text{ where } \nu_a \ll \mu \text{ and } \nu_s \perp \mu;$$

(b)

We briefly discuss Lebesgue decomposition for other types of measures below.

- If ν is given as a positive/finite/σ-finite measure instead, then "positive" becomes "positive"/"finite"/"σ-finite" in conclusion (a).
- If ν is given as a σ-finite measure instead, then in conclusion (a) ν_a and ν_s become σ-finite.
- Conclusion (a) continues to hold if μ and ν are both signed or complex. Recall the definition of absolute continuity in this case from (4.14).
- The theorems can be generalized to the case when μ has no assumption while ν is an *s-finite measure*, which is a sum of countably many finite measures. See [Fal19].

4.17 Remark. If ν is given as a signed measure instead, then write $\nu = \nu^+ - \nu^-$, and then use the above version of Lebesgue decomposition to write

For each $n \in \mathbf{N}$, set $\nu_n(E) = \nu(E \cap X_n)$ for all $E \in \mathcal{A}$ and get a finite measure ν_n. Now apply Lebesgue decomposition for finite ν above

Radon–Nikodym derivative with respect to counting measure
Lebesgue decomposition of a monotonic function (p344 345 Bogachev)

4.C Differentiation

4.18 Vitali covering lemma.

Besicovitch covering theorem
There is a class of function, slightly weaker than the usual integrable L^1 functions, that is used frequently in some advanced analysis (e.g., distribution and PDE theory). Let the underlying space be $(\mathbf{R}^d, \mathcal{B}, m)$. The class of *locally integrable function*, denoted by L^1_{loc}, consists of (the equivalence class of) all measurable functions f satisfying $\int_K f(x)\,dx < \infty$ for all compact subsets of $\mathbf{R}^d$. (Since we are in $\mathbf{R}^d$, compact subsets may be replaced by bounded subsets.) The main difference between L^1 functions and L^1_{loc} functions is that the tail convergence behavior of L^1_{loc} functions is not controlled.

4.19 Definition. For $f \in L^1_{\text{loc}}$, its *Hardy–Littlewood maximal function* Mf is defined by

$$Mf(x) = \sup_{r>0} \frac{1}{m\big(B(x;r)\big)} \int_{B(x;r)} |f(y)|\,dy.$$

4.20 Lebesgue differentiation theorem.

density point

4.D Bounded variations and absolutely continuity

It is well-known that there are continuous yet nowhere differentiable functions, such as the famous Weierstrass function.

4.21 Definition. Let $J \subseteq \mathbf{R}$ be any interval between a and b (possibly unbounded). A function $F\colon J \to \mathbf{R}$

(a) has *bounded variation* if

$$V(F,J) := \sup \sum_{j=1}^{n} \big|F(t_j) - F(t_{j-1})\big| < \infty,$$

where the supremum is taken over all n and $t_0 < t_1 < \cdots < t_n$ contained in the interval J.

(b) is *absolutely continuous* if for all $\epsilon > 0$, there exists $\delta > 0$ such that

$$\sum_{j=1}^{n} (b_j - a_j) < \delta \implies \sum_{j=1}^{n} \big|F(b_j) - F(a_j)\big| < \epsilon$$

holds for any finite family of pairwise disjoint open intervals $\{(a_j, b_j)\}_{j=1}^{n}$ contained in J.

We will write $V_a^b(F)$ for $V(F,[a,b])$, and define a function $T_F\colon [-\infty,\infty] \to [0,\infty]$ by

$$T_F(x) = \begin{cases} 0 & \text{if } x = -\infty, \\ V(F,(-\infty,x]) & \text{if } x \in \mathbf{R}, \\ \lim_{x\to\infty} V(F,(-\infty,x]) & \text{if } x = \infty. \end{cases}$$

The limit as $x \to \infty$ in the last line make sense because T_F is an increasing function on $\mathbf{R}$.

4.22 Theorem. A function $F \in \mathrm{BV}[a,b]$ is differentiable a.e., with F' being integrable.

Recall we defined the distribution function F_μ of a positive measure μ by $F_\mu(x) = \mu(-\infty, x]$. We carry this definition to signed and complex measures.

4.23 Theorem. If μ is a signed/complex Borel measure on $\mathbf{R}$, then F_μ is BV, right-continuous, with $F_\mu(-\infty) = 0$.

Conversely, if F is BV, right-continuous, with $F(-\infty) = 0$, then there exists a unique signed/complex Borel measure μ on $\mathbf{R}$ such that $F = F_\mu$.

We have hence established a one-to-one correspondence between μ and right-continuous F with $F(-\infty) = 0$. Also, $|\mu| = \mu_{T_{F_\mu}}$.

4.E Fundamental theorem of calculus

4.24 Fundamental theorem of calculus. For $f\colon [a,b] \to \mathbf{R}$, the following are equivalent:

(a) f is absolutely continuous;

(b) there exists a Lebesgue integrable function g on $[a,b]$ such that

$$f(x) = f(a) + \int_a^x g(t)\,dt$$

for all $x \in [a,b]$.

(c) f has derivative f' almost everywhere, and f' is Lebesgue integrable with

$$f(x) = f(a) + \int_a^x f'(t)\,dt$$

for all $x \in [a, b]$.

Bogachev 5.4.5 4.7.60

4.25 Integration by parts. For absolutely continuous functions f and g on $[a, b]$, we have

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

For completeness state the change of variables formula on the real line. Note the distinction between $\int_{[\varphi(a),\varphi(b)]}$ and $\int_{\varphi(a)}^{\varphi(b)}$.

4.26 Substitution method. Let $\varphi\colon [a, b] \to \mathbf{R}$ be monotonic and absolutely continuous, and let J be the closed interval between $\varphi(a)$ and $\varphi(b)$. If $f \in L^1(J)$, then $f(\varphi)\varphi' \in L^1[a, b]$, with

$$\int_{\varphi(a)}^{\varphi(b)} f(x)\,dx = \int_a^b f\big(\varphi(t)\big)\varphi'(t)\,dt. \tag{4.27}$$

The interval $[a, b]$ above can in fact be any intervals, including unbounded ones.

If we drop the monotonicity of φ above, but instead impose that $f(\varphi)\varphi' \in L^1$, then (4.27) remains true.

Chapter 5 Measures and function spaces

5.A L^p when $1 \leq p < \infty$

Let $(X, \mathcal{A}, \mu)$ be the underlying measure space and $0 < p < \infty$. We define the *p-norm* of a measurable function f by

$$\|f\|_p = \left(\int_X |f| \, d\mu\right)^{1/p} \in [0, \infty].$$

The $\mathcal{L}^p$ *space* is the space of measurable functions with finite p-norms.

The space $\mathcal{L}^p$ is not quite a normed space under $\|\cdot\|_p$. We will soon see that only when $1 \leq p < \infty$, $\|\cdot\|_p$ will become a seminorm on $\mathcal{L}^p$. Hence if we consider the equivalence classes of functions in $\mathcal{L}^p$ that are a.e. the same, then $\|\cdot\|_p$ becomes a norm. The set of equivalence classes we described here is called the L^p *space*. We make the appearance of equivalence classes in the definition of L^p spaces implicit in our exposition, as long as it does not need to confusion; for example, we always write a function $f \in L^p$ instead of $f \in \mathcal{L}^p$.

The L^1 space of integrable functions have been the sole focus in the previous chapters. In this chapter we will look at the functional analytic structure of the L^p spaces, and touch on their connections to the theory of Fourier analysis.

5.1 Hölder's inequality. Let $\frac{1}{p} + \frac{1}{q} = 1$, then

$$\|fg\|_1 \leq \|f\|_p \|f\|_q.$$

(p and q satisfying $\frac{1}{p} + \frac{1}{q} = 1$ are called conjugate exponents.)

5.2 Minkowski's inequality. For $1 \leq p < \infty$, we have $\|f + g\|_p \leq \|f\|_p + \|g\|_p$.

5.3 Theorem. L^p is complete.

5.4 Proposition. The equivalence class simple functions are dense in L^p hence $L^q \cap L^p$ is dense in L^p (6.7)

5.5 Proposition. For any finite measure μ on a metric space, we have $C_b(X)$ is dense in $L^p(\mu)$. [ADM11, Proposition 3.16]

A separable metric space with Borel σ-algebra is countably generated. To see this, one can take open balls centered at a countable dense subset with rational radius.

5.6 Theorem [Coh13, Proposition 4.3.5]. If $\mathcal{A}$ is countably generated and μ is σ-finite, then $L^p(X)$ is separable for $1 \leq p < \infty$.

5.B L^p when $p = \infty$

5.7 Theorem. L^∞ is complete.

For any Borel measure that assigns positive values to all open sets (e.g., the Lebesgue measure on $\mathbf{R}^d$), we have $\|f\|_\infty = \|f\|_u$ when f is continuous, since $\{x : |f(x)| > t\}$ is open. Notice that the equivalence class of $(C_b(X), \|\cdot\|_u)$ may be regarded as a closed subspace of $(L^\infty(X), \|\cdot\|_\infty)$, since $(C_b(X), \|\cdot\|_u)$ is complete. It is clear that we do not have the density of C_b in L^∞ in general.

5.8 Proposition [Fol99, Proposition 6.10]. For $1 \le p < q < r \le \infty$, then $L^p \cap L^r \subseteq L^q$, and

$$\|f\|_q \le \|f\|_p^\lambda \|f\|_r^{1-\lambda},$$

where

$$\lambda = \frac{q^{-1} - r^{-1}}{p^{-1} - r^{-1}} \in (0, 1).$$

5.9 Proposition [Fol99, Exercise 6.7]. If $f \in L^p \cap L^\infty$ for some $p < \infty$, then

$$\|f\|_\infty = \lim_{q \to \infty} \|f\|_q.$$

The condition $f \in L^p \cap L^\infty$ enforces $f \in L^q$ for all $q > p$. One might ask if $f \in L^q$ for all $q \ge p$, then $f \in L^\infty$ automatically. This is unfortunately wrong: on the unit interval endowed with the Lebesgue measure, the function $\log(x)$ has finite L^p norm $\Gamma(p+1)^{1/p}$ for all $p < \infty$ (verify this!), and is close to p/e for large p. However, the logarithm function is not bounded a.e.

5.C Hilbert spaces and L^2

There is a canonical choice of orthonormal basis on the Hilbert space $L^2[0, 1]$.

Haar function

Every Hilbert space with orthonormal basis $\{e_\alpha\}_{\alpha \in A}$ is unitarily isomorphic to $\ell^2(A)$.

Every infinite-dimensional separable Hilbert space is unitarily isomorphic to $L^2[0, 1]$.

5.D Duality of L^p

5.10 Riesz representation theorem (L^p spaces). Let $(X, \mathcal{A}, \mu)$ be a σ-finite measure space. and let $1 \le p < \infty$. For every $\Phi \in (L^p)^*$, there is a unique $f \in L^q$ such that for all $g \in L^p$, we have

$$\Phi(g) = \int fg \, d\mu.$$

Meanwhile $\|\Phi\| = \|f\|$, which means $(L^p)^*$ is isometrically isomorphic to L^q.

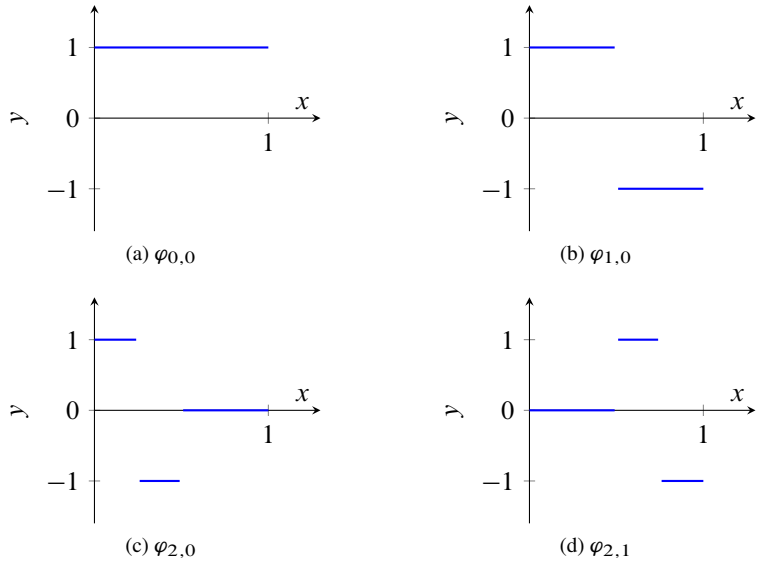
The statement above remains true if μ is not σ-finite, as long as $1 < p < \infty$.

5.E Convolutions and smooth approximation

Let f and g be measurable, the *convolution* of f and g is the function

$$f * g(x) = \int f(x - y) g(y) \, d\mu(y)$$

for all x such that the integral exists.



(a) $\varphi_{0,0}$ (b) $\varphi_{1,0}$

(c) $\varphi_{2,0}$ (d) $\varphi_{2,1}$

Figure 5.1: Haar basis functions

5.11 Young's inequality. For $1 \leq p, q, r \leq \infty$ and $\frac{1}{p} + \frac{1}{q} = \frac{1}{r} + 1$, then if $f \in L^p$ and $g \in L^q$, then $f * g$ is defined a.e. (if $r = \infty$ then everywhere) and is in L^r, with

$$\|f * g\|_r \leq \|f\|_p \|g\|_q.$$

It is clear that we can

5.F Riesz' theorems and convergence of measures

We will let X always be a locally compact metric space in this section. All results in this section can be generalized to locally compact Hausdorff spaces with more sophisticated topological tools and arguments.

5.F.1 The topology of locally compact spaces

For a Radon measure μ on a locally compact metric space X, we have $C_c(X)$ is dense in $L^p(\mu)$. (Folland 7.9)

5.12 Urysohn's lemma. (locally compact spaces) Let X be a locally compact metric space, and let K be compact and U be open in X such that $K \subseteq U$.

(a) We can construct a precompact open set G in X such that $K \subseteq G \subseteq \overline{G} \subseteq U$.

(b) It follows that we can construct a continuous function $f : X \to [0, 1]$ with $f = 1$ on K and supp f is compact and is contained in U.

Proof.

(a) First consider the case when $K = \{x\}$. We know there is an open set V containing x with compact closure $\overline{V}$. Taking U to be the smaller set $U \cap V$ if necessary, we may always assume that U has compact closure in X.

Now take $G = B(x;r) \subseteq \overline{B}(x;r) \subseteq U$ for some $r > 0$. We then have $\overline{G} \subseteq \overline{B}(x;r) \subseteq U$.[1] Since U has compact closure, $\overline{G}$ is compact. Hence we have found our desired G.

Now consider the general case for any compact set K. Let G_x be the precompact open set containing x discussed above, which satisfies $G_x \subseteq \overline{G_x} \subseteq U$. The collection $\{G_x\}_{x\in K}$ has a finite subcovering $\{G_1, \ldots, G_m\}$ that covers K. Set $G = \bigcup_{k=1}^m G_k$, which is open. Since $\overline{G} = \bigcup_{k=1}^m \overline{G_k}$ is compact and is contained in U, the proof is complete.

(b) We have two closed sets K and $X - G$, where G is given in part (a). Now by Urysohn's lemma (ordinary version), there is a continuous function $f : X \to [0,1]$ such that $f(K) = \{1\}$ and $f(X - G) = \{0\}$, which means that $\operatorname{supp} f \subseteq \overline{G} \subseteq U$. Since $\overline{G}$ is compact, $\operatorname{supp} f$ must be compact. □

To state the result above in fancy topology terms, we have proved exactly that locally compact metric spaces are *completely regular*.

Now we give a finite partition of unity over a compact set in a locally compact space.

5.13 Partition of unity. In a locally compact metric space X, let K be a compact subset, and $\{U_j\}_{j=1}^n$ be a finite open cover of K. Then there exists a collection of $\{\psi_j\}_{j=1}^n \subseteq C_c(X, [0,1])$ such that $\operatorname{supp} \psi_j \subseteq U_j$ and $\sum_{j=1}^n \psi_j(x) = 1$ for all $x \in K$.

Proof. We know each $x \in K$ is contained in some U_j, and therefore it has an open set G_x satisfying $x \in G_x \subseteq \overline{G_x} \subseteq U_j$. As in the proof of Urysohn's lemma part (a), we have a finite open cover $\{G_{x_k}\}_{k=1}^m$ of K such that $\bigcup_{k=1}^m \overline{G_{x_k}}$ is compact. For each $j \in [n]$ now define

$$F_j = \bigcup \{\overline{G_{x_k}} : \overline{G_{x_k}} \subseteq U_j\},$$

which as a compact subset of U_j allows us to define $g_j = 1$ on F_j and $\operatorname{supp} g_j \subseteq U_j$. Note also $\{F_j\}_{j=1}^n$ covers K by construction.

Now we have $\sum_{j=1}^n g_j \geq 1$ for all points on K. We want to normalize over K but still get a continuous function over the entire space. Here we use Urysohn's lemma to create a function $f \in C_c(X, [0,1])$ with $f = 1$ on K and $\operatorname{supp} f \subseteq \{x : \sum_{j=1}^n g_j(x) > 0\}$. Therefore $g_0 := 1 - f$ is a continuous function that is 0 on K but 1 on $\{x : \sum_{j=1}^n g_j(x) > 0\}$. Now $\sum_{j=0}^n g_j > 0$ on the entire X, so we can safely normalize and define

$$\psi_j = \frac{g_j}{\sum_{j=0}^n g_j}$$

for all $j \in [n]$. Clearly $\operatorname{supp} \psi_j = \operatorname{supp} g_j \subseteq U_j$, and so we are done with our construction. □

For $f \in C_c(X)$, we will use the notation $f \prec U$ to mean $0 \leq f \leq 1$ while $\operatorname{supp} f \subseteq U$.

On a locally compact metric space X, we have just seen that Urysohn's lemma guarantees the existence of a function $f \in C_c(X, [0,1])$ that equals 1 On a compact subset. Now we look at a particular example of such a bump function on $\mathbf{R}^n$, which is in fact smooth. (This allows to prove a smooth partition of unity on $\mathbf{R}^n$, which is crucial in theory of smooth manifolds.) It is noteworthy that a global partition of unity subordinate to a countably infinite open cover presents much more difficulty than a local partition of unity subordinate to a finite open cover of a compact set. In particular, we need to make sense of summing over a countably infinite number of functions.

[1] Please see Exercise B.2 for a relevant exercise.

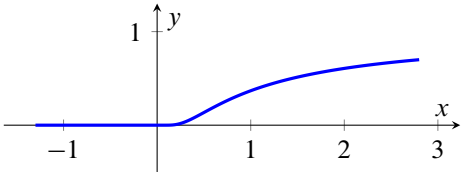


Figure 5.2: the function defined in (5.14), which is smooth at 0

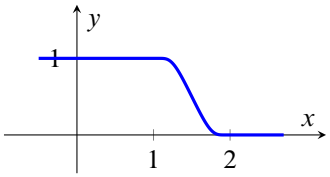


Figure 5.3: the transition function defined in (5.15), when $a = 1$ and $b = 2$

Recall that the function

$$f(x) = \begin{cases} \exp(-1/x) & \text{if } x > 0, \\ 0 & \text{if } x \le 0. \end{cases} \tag{5.14}$$

is a smooth function from $\mathbf{R}$ to $[0, 1)$; see Fig. 5.2. Now for any $a < b$, consider

$$g(x) = \frac{f(b-x)}{f(b-x) + f(x-a)}. \tag{5.15}$$

It is clear that g is smooth (the denominator is nowhere zero) and increasing on $\mathbf{R}$, with $g(x) = 1$ when $x \le a$ and $g(x) = 0$ when $x \ge b$. Such a function g is usually called a *transition function*, for the obvious reason.

Let $0 \le a < b$, then the function $h\colon x \mapsto g(\|x\|_2)$ is a smooth function that is 1 on $\overline{B}(0;a)$ and 0 outside $B(0;b)$. Alternatively if we define h using the max norm instead of the Euclidean norm, then the closed balls should be replaced by closed cubes.

The transition functions and the bump functions are very important approximants to other functions, e.g., indicator functions. The following smooth bump function also appears frequently in the literature as an approximant, because of its straightforward formula. Compose the function $x \mapsto 1 - \|x\|^2$ with the function f defined in (5.14), we get a new smooth function

$$\hat{f}(x) = \begin{cases} \exp\left(\frac{1}{\|x\|^2 - 1}\right) & \text{if } \|x\| < 1, \\ 0 & \text{if } \|x\| \ge 1, \end{cases}$$

which has a closed ball/cube as its support, depending on the norm used.

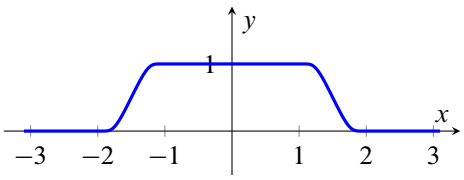


Figure 5.4: the bump function defined by h in dimension 1

We remark that $C_c^\infty(\mathbf{C})$ functions must be the zero function due to Liouville's theorem, in stark contrast to the real case.

5.16 Proposition [Fol99, Proposition 8.6].

(a) $f * g = g * f$.

(b) $f * (g * h) = (f * g) * h$.

(c) $\tau_z(f * g) = (\tau_z f) * g = f * (\tau_z g)$.

(d) $\operatorname{supp} f * g \subseteq \overline{\operatorname{supp} f + \operatorname{supp} g}$.

In PDE theory, we are interested in functions defined on an open subset U of $\mathbf{R}^n$; however, it can be hard to prove results directly about functions defined such U's. We establish tools for functions defined on the entire $\mathbf{R}^n$, and then use extension and restriction arguments to specialize down to functions defined on U.

[Bre11, Proposition 4.20] proves the second part directly

5.17 Proposition [Fol99, Proposition 8.10, Exercise 8.7]. For $f \in L^1(\mathbf{R}^n)$ and $g \in C_b^k(\mathbf{R}^n)$ (i.e., $\partial^\alpha g$ is bounded for all $|\alpha| \le k$), we have $f * g \in C^k(\mathbf{R}^n)$ with

$$\partial^\alpha(f * g) = f * (\partial^\alpha g) \text{ for all } |\alpha| \le k.$$

If $f \in L^1_{\text{loc}}(\mathbf{R}^n)$ and $g \in C_c^\infty(\mathbf{R}^n)$, then the same claim still holds.

Proof. The first part of theorem is really just differentiation under the integral sign of parameterized functions, Corollary 2.41.

The second part of the theorem □

[Jos05, Lemma 19.22] For $f \in L^1(U)$, for any open set $V \subseteq \overline{V} \subseteq U$ and $\epsilon < \operatorname{dist}(V, \partial U)$, we have $f^\epsilon \in C^\infty(V)$.

Extend f to $\mathbf{R}^n$ by defining it to be 0 outside U. Then we may see $f \in L^1(\mathbf{R}^n)$

We now introduce the notion of a mollifier. It can smooth out functions, and allows us to approximate a given function by its smoothed-out versions.

Dividing $\hat{f}$ by a constant, we obtain a function $\eta \in C_c^\infty$ with $\int \eta = 1$. Define $\eta_\epsilon(x) = \frac{1}{\epsilon^n}\eta(x/\epsilon)$ for all $\epsilon > 0$. Then η_ϵ continues to be smooth, $\int \eta_\epsilon = 1$, and now with support contained in $\overline{B}(0;\epsilon)$. The collection $\{\eta_\epsilon\}_{\epsilon \ge 0}$ is an *approximation to the identity*. (converges to the Dirac delta function in the Schwartz sense) The function η is called the *standard mollifier*.

Given $f \in L^1_{\text{loc}}(U)$, define $f^\epsilon = f * \eta_\epsilon$ in the open set $U_\epsilon = \{x \in \mathbf{R}^d : \operatorname{dist}(x, \partial U) > \epsilon\}$.

$f^\epsilon \in C^\infty(U_\epsilon)$ $f^\epsilon \to f$ a.s. as $\epsilon \to 0$ uniformly on compact subsets of U

For any open set $U \subseteq \mathbf{R}^n$, $C_c^\infty(U)$ is dense in $C_c(U)$, and hence dense in $C_0(U)$.

extend $C_c^\infty(U)$ to $C_c^\infty(\mathbf{R}^n)$,

This tells us that the Riesz–Markov–Kakutani theorem for finite measures holds for C_c^∞ test functions when the space X considered is $\mathbf{R}^n$.

given $f \in C_c$, consider f^ϵ

C_c^∞ dense in L^p for $1 \le p < \infty$

Fundamental theorem of calculus of variation

5.18 Theorem. For $f \in L^1_{\text{loc}}(U)$, if

$$\int fg = 0$$

for all $g \in C_c^\infty(U)$, then $f \equiv 0$ on U.

5.19 Definition.

5.F.2 Spaces of test functions

Let X be any (Hausdorff) topological space. We define $C_c(X)$ to be the space of continuous functions on X with compact support. We also define $C_0(X)$ to be the space of continuous functions on X such that $\{x : |f(x)| \geq \epsilon\}$ is compact[2] for all $\epsilon > 0$.

Since C_c and C_0 are both subsets of C_b, it is natural to endow the uniform norm $\|\cdot\|_u$ on the two new spaces, and we will always do this henceforth. It turns out quite naturally that $C_0(X)$, as a closed normed subspace of $C_b(X)$, is complete.

In analysis one is often interested in test function classes C_0 or C_c. Below is one reason why the choice of C_b is desirable to probabilists. Suppose the sequence $\{\mu_n\} \subseteq \mathcal{P}(S)$ converges weakly to $\mu \in \mathcal{M}(S)$. If we take $f = 1$ on the entire S in (5.21), then we have $\mu(S) = \lim_n \mu_n(S) = 1$, thus proving that the weak limit μ is a Borel probability measure as well. Hence no "mass" is lost in this convergence, in contrast to ...

We use $\mathcal{M}(S)$ for the space of finite signed/complex Borel measures on S, $\mathcal{M}_+(S)$ for the space of finite positive Borel measures on S

5.20 Definition. A sequence $\{\mu_n\}$ is said to

(a) *converge weakly* to μ if for all $f \in C_b(S)$, we have

$$\int_S f \, d\mu_n \to \int_S f \, d\mu, \tag{5.21}$$

which we denote by $\mu_n \Rightarrow \mu$;

(b) *converge vaguely* to μ if for all $f \in C_c(S)$, we have

$$\int_S f \, d\mu_n \to \int_S f \, d\mu. \tag{5.22}$$

It is common to see the notation μf in place of $\int_S f \, d\mu$, because we may see μ as a linear operator acting on the space $C_b(S)$ with the topology of uniform convergence. The bounded continuous functions in the definition are called *test functions*, because this mode of convergence is tested with respect to $C_b(S)$.

Weak convergence is a subject of greater importance to probability compared to general measure theory and analysis. This has led to our choice (and many authors' choice) to present weak convergence solely in the context of probability.

If X is second countable topological space, then $C_c(X)$ is separable. It follows that its uniform closure $C_0(X)$ is also separable.

Let X be locally compact. If $\sup_n \|\mu_n\| < \infty$, then it is equivalent to say $\int_S f \, d\mu_n \to \int_S f \, d\mu$ either over all $f \in C_c(X)$ or over all $f \in C_0(X)$, as a consequence of Proposition D.6. In particular this applies to the space of subprobability measures.

loss of mass at infinity

5.23 Definition. Let X be an LCH space. A positive *Radon measure* μ on X is a Borel measure that is locally finite, outer regular on all Borel sets, and compact inner regular on all open sets.

Also may known as *Borel regular measure* or *Riesz measure*

5.24 Proposition. Every Radon measure is compact inner regular on σ-finite Borel sets. In particular, every σ-finite Radon measure is compact inner regular on all Borel sets.

[2] The compactness stated here should be viewed as a generalization of boundedness when the space concerned does not have any metric.

Since we are in a Hausdorff space, a compact inner regular measure is also closed inner regular, and because the measure is finite, it is also outer regular. Bogachev [Bog07; Bog18] defines a signed measure to be inner regular

Indeed, for finite measures, Radon measures are the same as compact inner regular measures.

5.25 Proposition. Every locally finite Borel measure on an lcscH space is compact inner regular on all Borel sets, and is hence a Radon measure.

On an lcscH spcae, finite/signed/complex Borel measures are always Radon.

Hence we may identify $\mathcal{M}(X)$ with the space of linear functionals on $C_c(X)$. This allows us to define the weak-star topology on $\mathcal{M}(X)$ by defining the convergence $\mu_n \to \mu$ in $\mathcal{M}(X)$ if

$$\int f \, d\mu_n \to \int f \, d\mu \quad \text{for all } f \in C_c(X).$$

In notation we can write $\sigma\big(\mathcal{M}(X), C_c(X)\big)$

In fact, this allows us to define a weak-star topology on $\mathcal{M}(X)$. It is clear that $\mu_n \to \mu$ in the weak-star sense if and only if

5.26 Riesz–Markov–Kakutani theorem for positive measures.

5.27 Riesz–Markov–Kakutani theorem for finite measures. Let X be a locally compact metric space, then the dual space $C_c(X)^*$ is isometrically isomorphic to $\mathcal{M}_{\mathrm{R}}(X)$, i.e., for all linear functionals $L \in C_c(X)^*$, there is a unique signed/complex inner regular Borel measure μ such that

$$L(f) = \int_X f \, d\mu \quad \text{for all } f \in C_c(X);$$

meanwhile $\|L\| = \|\mu\|$.

In particular, if X is separable, then in the above statement $\mathcal{M}_{\mathrm{R}}(X) = \mathcal{M}(X)$; if furthermore X is compact,[3] then $C_c(X) = C(X)$.

Every instance of $C_c(X)$ above can be replaced by its uniform closure $C_0(X)$.

vague limit of Radon measures is Radon, and when μ_n are all positive measures, then the vague limit μ is positive. To see this, take f to be any nonnegative C_b function. Then $\int f \, d\mu_n \to \int f \, d\mu \geq 0$, enforcing $\mu(X) \geq 0$.

5.28 Corollary. $\mathcal{M}_{\mathrm{R}}(X)$ is a closed subspace of $\mathcal{M}(X)$ and hence a Banach space.

5.29 Proposition. In a locally compact metric space X with $\mu_n \in \mathcal{M}_r^+(X)$ such that $\mu_n \to \mu$, one has $\mu(X) \leq \liminf_n \mu_n(X)$. If μ_n is allowed to be signed, then $|\mu|(X) \leq \liminf_n |\mu_n|(X)$.

If one is familiar with Banach space theory, this is an immediate consequence of the uniform boundedness principle.

5.G Fourier series

5.H Fourier transform of functions and measures

5.I Laplace transform

[3] Compact metric spaces are obviously Lindelöf, and hence separable; see Proposition A.17.

Chapter 6 Elements of Polish spaces

To use the word of , two spaces really stands out when studying measure and integration on topological spaces, one being locally compact space and the other being Polish spaces. In fact, an LCH space X is second countable if and only if it is Polish. The reverse direction is immediate by Proposition A.17. For the forward direction, the usual proof is to consider the one-point compactification $\widetilde{X}$ of X. Our new $\widetilde{X}$ is second countable and compact Hausdorff, and hence it is metrizable. Any such metric must be complete by compactness, and therefore $\widetilde{X}$ is Polish. Now our X, as an open subset of $\widetilde{X}$, must be Polish.

For example, we immediately get a version of RMK theorem for second countable LCH space, and hence an integration theory on manifolds (which are second countable LCH) can be obtained.

6.1 Definition. A *Polish space* is a separable topological space that admits a complete metrization.
A *standard Borel space* is a measurable space isomorphic to a Borel subset of a Polish space.

countable product of Polish spaces is Polish
subspace of Polish space is Polish if and only if it is a G_δ set

6.2 Proposition. Any Polish space is homeomorphic to a G_δ set of $[0,1]^{\mathbf{N}}$.

6.3 Ulam's theorem. In a Polish space X, every Borel measure is tight.

Proof. Let $S = \{x_j\}_{j=1}^{\infty}$ be a countable dense subset of X. Since any point X must be arbitrarily close to some point in S, the collection $\{\overline{B}(x_j; 1/n)\}_{j=1}^{\infty}$ covers X for any $n \in \mathbf{N}$. It follows that for any $\epsilon > 0$, there is some M_n such that

$$\mu\left(X - \bigcup_{j=1}^{M_n} \overline{B}(x_j; 1/n)\right) < 2^{-n}\epsilon.$$

To make the approximation set independent of n, consider

$$K = \bigcap_{n=1}^{\infty} \bigcup_{j=1}^{M_n} \overline{B}(x_j; 1/n).$$

It follows that $\mu(X - K) \leq \lim_{n\to\infty} 2^{-n}\epsilon = \epsilon$.

Since X is complete and K is closed, K is complete. In addition, K has finite $\frac{1}{n}$-net for each n, and it follows by the Theorem A.20 that K is compact. This proves the claim. □

6.4 Theorem. Two standard Borel spaces are Borel isomorphic if and only if they have the same cardinality, which must be finite, countably infinite, or that of the continuum.

universal measurability

Q is not Polish, Baire category theorem

For two separable metrizable spaces X and Y, if $f : X \to Y$ is Borel measurable, then its graph is a Borel subset of $X \times Y$.

limit of measurable function is measurable

6.5 Corollary. A standard Borel space A is isomorphic to a Borel subset B of the real line with the Borel σ-algebra. If A is an infinite set, then we can take B to be the entire real line.

Interlude: Between Measure and Probability

Part II

Probability

Chapter 7 Interpreting probability using measure theory

7.A Distributions

From now on (σ-)algebras will be called (σ-)fields. The measure space $(X, \mathcal{A}, \mu)$ will be replaced by $(\Omega, \mathcal{F}, P)$ with $P(\Omega) = 1$, which we call a *probability space*. In the probability triplet Ω is called the *sample space*, and $\mathcal{F}$ is called the *event space*, which contains all the possible *events*. If Ω is a countable set and $\mathcal{F} = \wp(\Omega)$, then the probability space is *discrete*.

Given an underlying measurable spaces $(\Omega, \mathcal{F})$, a measurable function $X\colon (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ is called a *random variable*. If $(\Omega, \mathcal{F}, P)$ is discrete, then the image of any function X is forced to be countable. We may then let $S = X(\Omega)$ and $\mathcal{S} = \wp(S)$, and X is obviously measurable. The random variable defined on a discrete space is called a *discrete random variable*, and its distribution is also *discrete*. If $(S, \mathcal{S})$ is a measurable subspace of $(\mathbf{R}, \mathcal{B})$, we call the random variable *real-valued*. In general when $(S, \mathcal{S})$ is a measurable subspace of $(\mathbf{R}^d, \mathcal{B}^d)$, then X may be called a *real random vector*. The preference of Borel σ-field over the Lebesgue σ-field has been discussed in Section 2.A.

Given a random variable X, following Section 2.I we may define a probability measure μ on $(S, \mathcal{S})$ given by

$$\mu(A) = P\big(X^{-1}(A)\big) = P(X \in A) \text{ for all } A \in \mathcal{S}. \tag{7.1}$$

We call this the *probability distribution/law*[1] of X, denoted by $X \sim \mu$. It characterizes how probability of (the image of) X is distributed across the target space $(S, \mathcal{S})$[2]. The $X \in A$ above is a shorthand for $\{\omega \in \Omega : X(\omega) \in A\}$, and this convention[3] is widely adopted throughout probability, as long as the context is clear. It also corresponds to the intuitive understanding of a random variable X as a "variable" taking random values by ignoring the underlying ω, but we must not take this formally. When two $(S, \mathcal{S})$-valued random variables X and Y (on possibly different underlying spaces) have the same distribution μ, we write $X \overset{\mathrm{D}}{=} Y$.

It is clear that a measure μ on a measurable subspace of $(\mathbf{R}, \mathcal{B})$ can be naturally extended to a measure on $(\mathbf{R}, \mathcal{B})$ (by setting all the new sets to measure 0). Therefore it always makes sense to regard the distribution of any real-valued random variable as a Borel measure on $\mathbf{R}$.

7.2 Remark. Another perspective we can take is to always let real-valued random variables take $(S, \mathcal{S})$ to be exactly $(\mathbf{R}, \mathcal{B})$. In this setup μ will always be a Borel measure. When X is a random variable with $S \coloneqq X(\Omega) \subsetneq \mathbf{R}$, we can always consider the restriction of the distribution μ_X to $(S, \mathcal{B}|_S)$ to obtain our adopted definition of probability distribution in (7.1). This alternative perspective is suitable for discussing distribution functions, while our previous perspective is suitable for discussing density functions, as we will see.

7.3 Remark. Throughout the notes, random variables are *almost always* taken to be real-valued[4]. The

[1] Another common notation is $\mathcal{L}$ that stands for "law".

[2] In comparison, P characterizes the *underlying* space $(\Omega, \mathcal{F})$.

[3] In fact we have used this shorthand before, when discussing uniform integrability.

[4] We have only discussed the integration of real/complex-valued functions. Some generalizations can definitely be made (to

exceptions should be noted by the readers on their own.

The *(cumulative) distribution function* (c.d.f.) of a real-valued random variable X is defined to be a function $F : \mathbf{R} \to [0, 1]$ given by

$$F(x) = P(X \leq x) = \mu(-\infty, x].$$

We now slightly modify Theorem 1.31(a)(b) to suit our purpose. Note now we instead start with the original part (b).

7.4 Theorem. Let X be a real-valued random variable with distribution μ on $(\mathbf{R}, \mathcal{B})$, then its distribution function F has the following properties:

- $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$;
- it is increasing and right-continuous;
- it has left limits in the sense that

$$F(x-) = \lim_{y\to x^-} F(y) = \mu(-\infty, x),$$

  which also implies $\mu\{x\} = F(x) - F(x-)$.

Since μ is now a probability measure, the first bullet point follows directly. The rest has been proved already before. We remark also that every distribution function has countably many discontinuities (by Proposition A.5), and is hence continuous a.e.

Recall Theorem 1.31(a). We can slightly modify its statement and proof to get the version for obtaining a unique Borel probability measure.

7.5 Theorem. Conversely, let $F : \mathbf{R} \to [0, 1]$ be an increasing, right-continuous function with

$$\lim_{x\to-\infty} F(x) = 0 \quad \text{and} \quad \lim_{x\to\infty} F(x) = 1,$$

then there is a unique probability measure μ on $(\mathbf{R}, \mathcal{B})$ such that

$$\mu(-\infty, x] = F(x) \quad \text{for all } x \in \mathbf{R}.$$

Theorem 7.5 tells us that as long as we have the distribution function of a random variable X, which increases from 0 to 1 and is right-continuous, then the distribution function determines the distribution of the random variable. Formally we are now ready to state

7.6 Corollary. For two real-valued random variables X and Y, we have $F_X = F_Y$ if and only if $\mu_X = \mu_Y$, i.e., a one-to-one correspondence between distribution functions and distributions.

This observation is very fundamental because it tells us we can see the distribution of a real random variables from two distinct perspectives. The corollary further suggests that given a random variable, we may specify its distribution solely in terms of a function $F : \mathbf{R} \to [0, 1]$ that is increasing, right-continuous, with

$$\lim_{x\to-\infty} F(x) = 0 \quad \text{and} \quad \lim_{x\to\infty} F(x) = 1.$$

We call such a function F a *(cumulative) distribution function* on its own. And we write $X \sim F$ if $X \sim \mu_F$, the unique probability measure associated to the distribution function F.

for example, Banach-valued functions/random variables), but it is beyond the scope of this survey.

7.7 Theorem. Indeed any distribution function $F : \mathbf{R} \to [0, 1]$ can be realized as the distribution function of some real random variable X on some probability space $(\Omega, \mathcal{F}, P)$. In particular we can take the probability space to be $([0, 1], \mathcal{B}_{[0,1]}, m)$, and realize $X \sim F$ from a Uniform$[0, 1]$ random variable on this probability space.

First construction. By Theorem 7.5, we know every distribution function F gives rise to a unique probability measure μ on $(\mathbf{R}, \mathcal{B})$. Now let $(\Omega, \mathcal{F}, P) = (\mathbf{R}, \mathcal{B}, \mu)$ and let X be the identity map on $\mathbf{R}$. □

As long as one knows Theorem 7.5, this first proof is indeed a very trivial construction. The second proof, independent of Theorem 7.5, is more interesting and certainly of significance to us.

Second construction. Let $(\Omega, \mathcal{F}, P) = ([0, 1], \mathcal{B}_{[0,1]}, m)$, and we define

$$X(\omega) = \inf\{y : F(y) \geq \omega\} := F^{-1}(\omega). \tag{7.8}$$

It is clear to see that $X(\omega) \leq y$ if and only if $\omega \leq F(y)$. Therefore for all $y \in \mathbf{R}$,

$$P(X \leq y) = P\big(\omega \leq F(y)\big) = F(y).$$

The construction still works out perfectly if one replaces ω by $U(\omega)$, where $U \sim$ Uniform$[0, 1]$. This is because the identity map is the special case of a Uniform$[0, 1]$ random variable. We conclude that we can use any Uniform$[0, 1]$ random variable U to generate a μ-distributed random variable on the probability space $([0, 1], \mathcal{B}_{[0,1]}, m)$, via the recipe $F_\mu^{-1}(U)$. □

The realization of $X \sim F$ described above will play a pivotal role later in Section 10.A.

There are several things we need to mind here. Firstly, one can show that $\inf\{y : F(y) \geq \omega\} = \sup\{y : F(y) < \omega\}$. The "$\geq$" direction is obvious. To see the "$\leq$" direction, consider any $x > \sup\{y : F(y) < \omega\}$. It is clear that $F(x) \geq \omega$, and thus by right-continuity we have $F(\sup\{y : F(y) < \omega\}) \geq \omega$. Note that we have also just proved that the infimum in (7.9) can be attained.

Secondly, the X defined here in (7.8) is sometimes called the *generalized inverse/quantile function* of the distribution function F, denoted by F^{-1}. Distributions functions are not in general invertible, but this almost invertibility between $\mathbf{R}$ and $[0, 1]$ motivates our definition.

We now show $X(\omega)$ is continuous from the left, i.e., for all $a \in (0, 1]$,

$$\lim_{\omega \to a^-} X(\omega) = X(a). \tag{7.9}$$

Since F is increasing, the limit exists and the "$\leq$" direction follows. Now suppose we have the strict inequality "$<$". This implies $F\big(\lim_{\omega \to a^-} X(\omega)\big) < a$. Since $F\big(X(\omega)\big) \geq \omega$, we get a contradiction. Hence we have the equality in (7.9).

Thirdly, we remark that $\overline{X}(\omega) = \sup\{y : F(y) \leq \omega\} = \inf\{y : F(y) > \omega\}$ has the same distribution as our X defined in (7.8). In fact X and $\overline{X}$ differ at countably many points; $X(\omega) \neq \overline{X}(\omega)$ if and only if $X([0, \omega]) - X\big([0, \omega)\big)$, i.e., there is a jump for X at ω. For distinct $\omega \in [0, 1]$ these intervals have to be disjoint, and hence there are only countably many such intervals. The proof of this final step is included in Proposition A.5. We leave it as an exercise to reader to show that this $\overline{X}$ is right-continuous. (This will be used in the proof of Lemma 8.14, when constructing a right-continuous candidate for a distribution function.)

We will generalize this result later. Generalization of Theorem 7.7

7.10 Theorem [Coh13, Exercise 8.3.4].

Let $X\colon (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ have distribution μ, and the codomain $(S, \mathcal{S})$ has a natural underlying measure ρ with $\mu \ll \rho$. The *(probability) density function* (p.d.f.)[5] of the random variable X is Radon–Nikodym derivative $d\mu/d\rho$ of the probability distribution with respect to this underlying measure for the image space.

Specifically, when X is a discrete random variable, then the counting measure is a natural measure for $(S, \mathcal{S})$, and obviously $\mu \ll$ count. Hence $d\mu/d(\text{count})\colon x \mapsto \mu\{x\}$ is the density function, which is also called the *probability mass function* (p.m.f.)[6].

On the other hand, recall Fact 1.13. Given a random variable X, if the codomain S is a Borel subset of $\mathbf{R}^d$ and $\mathcal{S} = \mathcal{B}^d|_S$, and in addition $\mu \ll m|_S$, then $d\mu/d(m|_S)$ is the density function. We call such X a *continuous random variable*[7]. Note in this continuous case the density function is a.e. defined, but in the discrete case the density (p.m.f.) is exact. Later on when discussing continuous random variables, we usually only write out the case $(S, \mathcal{S}) = (\mathbf{R}, \mathcal{B})$ for brevity, since the density function $d\mu/d(m|_S)$ defined on S can be naturally extended to the entire $\mathbf{R}$.

The definition of density function for a continuous random vector is the same as above, with the Lebesgue measure replaced by the product Lebesgue measure. Also notice that the product of counting measures on marginal spaces is the counting measure on the product space, so we do not need to make a separate note for p.m.f. when $(S, \mathcal{S})$ is a product of discrete spaces. In contrast to distribution functions which are only nice to work with in dimension 1, density functions is defined for general random vectors in $\mathbf{R}^d$, as long as $\mu \ll m$.

We can define the class of distributions with densities solely in terms of their density functions. When the desired distribution of X is discrete, it is clear that we can specify this distribution using a *probability mass function* (on its own), i.e., a function $p\colon X(\Omega) \to [0, 1]$ such that

$$\sum_{x \in X(\Omega)} p(x) = 1.$$

When the desired distribution of X is continuous, then a nonnegative Borel measurable function f satisfying

$$\int_{\mathbf{R}} f(x)\, dx = 1,$$

called a *(probability) density function* (on its own) will specify the distribution. In summary, probability mass and density functions let us generate discrete and continuous random variables.

7.B Moments, independence, and joint distributions

7.B.1 Expectations as integrals

The average value of function

Following the theory of Lebesgue integration we have developed,

7.11 Definition. Let X be a nonnegative random variable, its *expectation/expected value* is given by

$$\mathrm{E}X = \int_\Omega X\, dP.$$

[5] or *frequency function*

[6] to emphasize we are in the discrete setting

[7] The term "continuous" here refers to the absolute continuity of the distribution function, and does not require that the density function must be continuous.

If X is a signed real-valued random variable, with one of $\mathrm{E}X^+$ and $\mathrm{E}X^-$ being finite, then we can define the *expectation* of X to be

$$\mathrm{E}X = \int_\Omega X\,dP = \mathrm{E}X^+ - \mathrm{E}X^-.$$

In particular, when $\mathrm{E}|X| < \infty$[8], $\mathrm{E}X$ always exists. This is the case we are interested in mostly.

Since the distribution μ on $(S, \mathcal{S})$ is given as the image measure $P \circ X^{-1}$, by Proposition 2.42 we have for $g\colon (S, \mathcal{S}) \to (\mathbf{R}, \mathcal{B})$, if $g \geq 0$ or $g \circ X \in L^1(\Omega)$, then

$$\mathrm{E}g(X) = \int_\Omega g\big(X(\omega)\big)\,dP(\omega) = \int_S g(x)\,d\mu(x).$$

In particular, if X is real-valued, then

$$\mathrm{E}X = \int_\Omega X(\omega)\,dP(\omega) = \int_S x\,d\mu(x).$$

Furthermore, if X is discrete, then

$$\mathrm{E}X = \sum_{x\in S} x\mu\{x\};$$

and if X is continuous with density f, then

$$\mathrm{E}X = \int xf(x)\,dx$$

It should be clear that $X =_d Y$ (on possibly different probability spaces), then $\mathrm{E}X = \mathrm{E}Y$.

7.12 Cauchy–Schwarz inequality. For any random variables X and Y,

$$\mathrm{E}|XY| \leq \big(\mathrm{E}X^2\big)^{1/2}\big(\mathrm{E}Y^2\big)^{1/2}$$

7.13 Jensen's inequality. Let $\mathrm{E}|X| < \infty$. Suppose I is an interval containing the range of X, and we have a convex function $\varphi\colon I \to \mathbf{R}$. Then

$$\varphi(\mathrm{E}X) \leq \mathrm{E}\varphi(X).$$

7.14 Lyapunov's inequality. For $1 \leq p \leq q < \infty$, we have $(\mathrm{E}|X|^p)^{1/p} \leq (\mathrm{E}|X|^q)^{1/q}$. It follows that

$$L^1 \supseteq L^2 \supseteq \cdots \supseteq L^\infty.$$

However, $L^\infty \neq \bigcap_{p=1}^\infty L^p$. The Gaussian measure is the counterexample.

7.B.2 Independence, a new measure-theoretic notion

7.15 Definition. We say events $A_1, \ldots, A_n \in \mathcal{F}$ are *independent* if for every subcollection $J \subseteq [n]$,

$$P\left(\bigcap_{j\in J} A_j\right) = \prod_{j\in J} P(A_j).$$

[8] One often prefers to write $\mathrm{E}|X| < \infty$ for integrability of X in probability. However, when we are dealing integration with respect to two different measures, then the L^1 notation should again be helpful.

Collections of events $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n$ are *independent* if for every subcollection $J \subseteq [n]$,

$$P\left(\bigcap_{j\in J} A_j\right) = \prod_{j\in J} P(A_j)$$

for all possible $A_j \in \mathcal{A}_j$ ($j \in J$). Random variables $X_1, X_2, \dots, X_n$ are *independent* if $\sigma(X_1), \dots, \sigma(X_n)$ are independent collections of events.

When the number of events/collection of events/random variables are infinite, then events/collection of events/random variables are said to be *independent* if every finite subcollection of these events/collection of events/random variables satisfies their independence definitions given above.

We will be concerned mostly with the finite collection in this section. Their extension to be infinite case should be easy.

7.16 Proposition. The following statements are equivalent.

(a) $A_1, A_2, \dots, A_n$ are independent;

(b) $A_1^c, A_2, \dots, A_n$ are independent;

(c) $\mathbf{1}_{A_1}, \mathbf{1}_{A_2}, \dots, \mathbf{1}_{A_n}$ are independent.

Given $(\Omega, \mathcal{F}, P)$, and let X and Y be two random variables taking values on $(S_1, \mathcal{S}_1)$ and on $(S_2, \mathcal{S}_2)$ respectively, with distributions μ_X and μ_Y. The *joint distribution* $\mu_{X,Y}$ of the pair (X, Y) is given by

$$\mu_{X,Y}(A) = P \times P\big((X, Y) \in A\big) \quad \text{for all } A \in \mathcal{S}_1 \otimes \mathcal{S}_2.$$

The $P \times P$ here is a product probability measure on $(\Omega \times \Omega, \mathcal{F} \otimes \mathcal{F})$.

The definition of joint distributions can obviously be generalized to any finite and countably infinite number of random variables, by our previous discussions on product measure spaces.

7.17 Theorem (independence characterizations). For two random variables X and Y taking values in $(S_1, \mathcal{S}_1)$ and $(S_2, \mathcal{S}_2)$ respectively, the following are equivalent characterization that X and Y are independent (which we sometimes denote by $X \perp Y$).

(a) $P(X \in A_1)P(Y \in A_2) = P(X \in A_1, Y \in A_2)$ for all $B_1 \in \mathcal{S}_1$ and $B_2 \in \mathcal{S}_2$;

(b) $\mu_X \times \mu_Y = \mu_{X\times Y}$;

(c) $P(X \in A_1)P(Y \in A_2) = P(X \in A_1, Y \in A_2)$ for all $A_1 \in \mathcal{K}_1$ and $A_2 \in \mathcal{K}_2$, where $\mathcal{K}_1$ and $\mathcal{K}_2$ are two π-systems such that $\mathcal{S}_1 = \sigma(\mathcal{K}_1)$ and $\mathcal{S}_2 = \sigma(\mathcal{K}_2)$;

(d) for all $f(X), g(Y) \in L^2$,

$$\mathrm{E}[f(X)g(Y)] = \mathrm{E}f(X)\,\mathrm{E}g(Y).$$

Here the L^2 requirement is a sufficient condition for us to assert the integrability of $f(X)g(Y)$, by Cauchy–Schwarz inequality.

Proof. Recall that the product measure is the unique extension of the product of marginal measures on measurable rectangles. □

7.18 Proposition. A real-valued random variable X independent of itself must take a constant value a.s.

If we know $X \in L^2$, then the result is immediate: by part (d) above we have $\mathrm{E}X^2 = (\mathrm{E}X)^2$, which implies $\mathrm{Var}(X) = 0$, i.e, $X = \mathrm{E}X$ a.s. But there is no need to make the L^2 assumption.

Proof. For any $A \in \mathcal{B}$, we have

$$P(X \in B)P(X \in B) = P(X \in B),$$

which implies $P(X \in B) = 0$ or 1.

We now prove a more general claim that directly implies the proposition:

> a $\{0, 1\}$-valued Borel probability measure μ on a separable metric space S must be a point mass.[9]

We know every open cover has a countable subcover in S (this is Proposition A.17). Fix $\epsilon > 0$ and consider the ϵ-balls $B(x; \epsilon)$ around each $x \in S$. Now we can choose a countable subcollection $\{B(x_j; \epsilon)\}_{j=1}^{\infty}$ that covers S, and this implies there exists one unique $j \in \mathbf{N}$ such that $\mu\big(B(x_j; \epsilon)\big) = 1$. We call this ball B_ϵ.

The intersection of any two such balls $B_{\epsilon_1} \cap B_{\epsilon_2}$ must have measure 1. This is because if it has measure 0, then $B_{\epsilon_1} - B_{\epsilon_2}$ and $B_{\epsilon_2} - B_{\epsilon_1}$ both have measure 1 despite being disjoint. Let $\epsilon_n = 1/n$, and it follows that

$$\mu\left(\bigcap_{n=1}^{\infty} B_{1/n}\right) = \lim_{k\to\infty} \mu\left(\bigcap_{n=1}^{k} B_{1/n}\right) = 1.$$

Since $B := \bigcap_n B_{1/n}$ has diameter 0, B is a singleton set of measure 1.

One has to be amazed that for any choice of countable subcover of open balls above, the end product is always *the* unique singleton set. (When $S = \mathbf{R}^d$ we can let these balls be 2^{-n}-cubes, whose countable disjoint union is the entire space.) □

As a consequence of Fubini–Tonelli, for Borel measurable $g\colon S_1 \times S_2 \to \mathbf{R}$ such that $g \geq 0$ or $\mathrm{E}|g(X, Y)| < \infty$, we have

$$\begin{aligned} \mathrm{E}g(X, Y) &= \int_{\mathbf{R}^2} g(x, y)\, d(\mu_X \times \mu_Y) \\ &= \int_{\mathbf{R}} \int_{\mathbf{R}} g(x, y)\, d\mu_X\, d\mu_Y. \end{aligned}$$

marginal density

7.19 Proposition (Factorization). Let X and Y be two discrete/continuous random variables. Then X and Y are independent if and only if for all $x, y \in \mathbf{R}$,

(a) $f_{X,Y}(x, y) = f_X(x) f_Y(y)$, where the f's are density functions;

(b) $f_{X,Y}(x, y) = g(x)h(y)$ for some functions g and h.

To be precise the equalities above are up to measure zero.

Proof. We show the case when X and Y are continuous random variables on $\mathbf{R}$. For all $A_1, A_2 \in \mathcal{B}$, we have

$$\begin{aligned} \mu_{X,Y}(A_1 \times A_2) &= \int_{A_1 \times A_2} f_{X,Y}(x, y)\, dx\, dy, \\ \mu_X(A_1) \times \mu_Y(A_2) &= \int_{A_1} f_X(x)\, dx \int_{A_2} f_Y(y)\, dy \\ &= \int_{A_1} \int_{A_2} f_X(x) f_Y(y)\, dx\, dy. \end{aligned}$$

[9] Hence the "real-valued random variable X" in the proposition statement may be replaced by "random variable X taking values in a separable metric space".

Part (a) now follows easily. To see the "if" direction of part (b), integrate both sides of $f_{X,Y}(x, y) = g(x)h(y)$ over $A_1 \times A_2$, we have

$$\mu_{X,Y}(A_1 \times A_2) = \int_{A_1} g(x)\,dx \int_{A_2} h(y)\,dy.$$

Consider $C = \int_{\mathbf{R}} h(y)\,dy$. We may divide h by this constant C and multiply g by this C, and assume without loss of generality that

$$\mu_X(A_1) = \mu_{X,Y}(A_1 \times \mathbf{R}) = \int_{A_1} g(x)\,dx,$$
$$\mu_Y(A_2) = \mu_{X,Y}(\mathbf{R} \times A_2) = \int_{A_2} h(y)\,dy.$$

This completes the proof. □

7.20 Definition. The *variance* of an L^2 random variable X is defined by

$$\begin{aligned}\operatorname{Var}(X) &= \mathrm{E}(X - \mathrm{E}X)^2 \\ &= \mathrm{E}(X^2) - 2\,\mathrm{E}X \cdot \mathrm{E}X + (\mathrm{E}X)^2 \\ &= \mathrm{E}(X^2) - (\mathrm{E}X)^2.\end{aligned}$$

Given two L^2 random variables X and Y, their *covariance* is defined by

$$\begin{aligned}\operatorname{Cov}(X, Y) &= \mathrm{E}\big((X - \mathrm{E}X)(Y - \mathrm{E}Y)\big) \\ &= \mathrm{E}(XY) - \mathrm{E}X \cdot \mathrm{E}Y;\end{aligned}$$

they are said to be *uncorrelated* if $\operatorname{Cov}(X, Y) = 0$, i.e.,

$$\mathrm{E}X \cdot \mathrm{E}Y = \mathrm{E}(XY);$$

and their *correlation* is defined by

$$\operatorname{Corr}(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(y)}}.$$

As mentioned perviously, the L^2 requirement is a sufficient, but not necessary condition for covariance to always exist. This is similar to the L^1 requirement sufficient for the expectation of a random variable to always exist.

Let A and B be two events and consider two indicators $\mathbf{1}_A$ and $\mathbf{1}_B$. Notice

$$\operatorname{Cov}(\mathbf{1}_A, \mathbf{1}_B) = \mathrm{E}(\mathbf{1}_{A\cap B}) - \mathrm{E}\mathbf{1}_A\,\mathrm{E}\mathbf{1}_B = P(A \cap B) - P(A)P(B).$$

We say A and B are *positively correlated* if the covariance above is ≥ 0, i.e., $P(A \cap B) \geq P(A)P(B)$, or equivalently $P(A \mid B) \geq P(A)$. We say A and B are *negatively correlated* if the $\geq$'s are replaced by $\leq$'s. Note that the covariance and correlation are symmetric.

7.B.3 Sum of independent random variables

Fourier transform

The *tail σ-field* of a sequence of random variables $X_1, X_2, \dots$ to be

$$\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, \dots).$$

Let $\pi : \mathbf{N} \to \mathbf{N}$ be a map such that $\pi(n) = n$ for all n larger than some finite M, which means that π only permutes finitely many indices. We call such a map a finite permutation of $\mathbf{N}$.

$\omega = (\omega_1, \omega_2, \dots)$, $\omega_j = X_j(\omega)$, the random variables X_j works as projection maps, and we have identified random variables with the coordinates of the samples (in our constructed product space).

An event is *permutable* if $\pi^{-1}A = \pi\{\omega : \pi\omega \in A\}$ for all finite permutations, which means exactly that an event remains invariant when we exchange the order of finitely many random variables.

7.21 Kolmogorov zero–one law. Let $X_1, X_2, \dots$ be a sequence of independent random variables, then any event in its tail σ-field $\mathcal{T}$ has probability 0 or 1.

7.22 Hewitt–Savage zero–one law. Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables, then any event in its exchangeable σ-field $\mathcal{E}$ has probability 0 or 1.

exchangeable family of random variables

7.23 Proposition.

7.C Basic concentration and deviation inequalities

We begin with the vanilla Markov's inequality that imposes minimal assumptions on the distribution of the random variable X considered.

7.24 Markov's inequality. Let $0 < p < \infty$. For any $a > 0$, we have

$$P(|X| \geq a) \leq \frac{1}{a^p} \mathrm{E}(|X|^p).$$

In particular, for nonnegative X, we have

$$P(X \geq a) \leq \frac{\mathrm{E}X}{a}.$$

These are just special cases of the following specialized version.

7.25 Generalized Markov's inequality. Let $\varphi\colon \mathbf{R} \to [0, \infty)$ be increasing. Then for any random variable X, and any $a \in \mathbf{R}$ with $\varphi(a) \neq 0$, we have

$$P(X \geq a) \leq \frac{1}{\varphi(a)} \mathrm{E}\varphi(X).$$

In probability theory it is often useful to take this φ to be an exponential function. If we assume $\mathrm{E}\exp(X) < \infty$, we get tail probabilities that are exponentially decreasing in a. In fact, the derivation

of many concentration inequalities depends in general on a technique called *Chernoff method*, where you set $\varphi(x) = \exp(\lambda x)$, and in the end you aim to minimize

$$\frac{1}{\exp(\lambda a)} \operatorname{E} \exp(\lambda X)$$

over all $\lambda \in \mathbf{R}^{>0}$ (so that φ is increasing).

Unfortunately $\operatorname{E} \exp(\lambda X)$ can be infinite, in particular for X with heavy-tailed distributions. For random variables with tails thinner than exponential or Gaussian random variables, the Chernoff method provides us valuable insights. Such random variables are known as *subexponential* and *sugaussian random variables*, and with information about its tail behavior, or equivalently, $\operatorname{E} \exp(\lambda X)$, we can derive much better concentration bounds than the vanilla Markov's inequality (and its consequences). See [Ver18] and [Han14] for the study of these concentration results, and their applications.

The exponential and Gaussian random variables represent the two canonical tail behavior of a random variable. To get this idea, we will let the reader verify that $\operatorname{E} \exp(\lambda X) = \frac{\rho}{\rho - \lambda}$ for $X \sim \text{Exponential}(\rho)$ and $\lambda < \rho$; and also $\operatorname{E} \exp(\lambda Y) = \exp(\lambda^2/2)$ for $Y \sim N(0, 1)$.

The transform $\operatorname{E} \exp(\lambda X)$ of the random variable X is called the *moment generating function* of X, denoted by $M_X(\lambda)$. Apart from its significance in proving concentration bounds, it also recovers the distribution of X, which we will study later. We will also see that under suitable conditions for λ, if $M_X(\lambda) < \infty$, then $X \in L^p$ for all $p \in [1, \infty)$. This is expected by considering the Taylor expansion of the exponential function.

Going back to vanilla Markov's inequality, we have the moment bound

$$P(|X| \geq a) \leq \inf_{p \in \mathbf{N}} \frac{1}{a^p} \operatorname{E}(|X|^p),$$

which is in fact always as least as good as the Chernoff bound. However, the optimization over $p \in \mathbf{N}$ (or $p \in \mathbf{R}$) is hard to materialize.

7.26 Chebyshev's inequality. For X with $\operatorname{E} X^2 < \infty$, we have for all $t > 0$ that

$$P(|X - \operatorname{E} X| \geq t) \leq \frac{\operatorname{Var}(X)}{t^2}.$$

Markov's inequality gives an upper bound on the tail probability, with the first moment $\operatorname{E} X$. A lower bound can also be obtained, with in addition the second moment $\operatorname{E} X^2$.

7.27 Paley–Zygmund inequality. Let $X \geq 0$ with $\operatorname{E} X^2 < \infty$. For any $0 \leq \theta \leq 1$, we have

$$P(X > \theta \operatorname{E} X) \geq (1 - \theta)^2 \frac{(\operatorname{E} X)^2}{\operatorname{E} X^2}.$$

Proof. The case for $\theta = 1$ is trivial. We will fix $0 < \theta < 1$ first.

The key is to use Cauchy–Schwarz inequality:

$$\begin{aligned} \operatorname{E} X &= \operatorname{E}(X \mathbf{1}\{X \leq \theta \operatorname{E} X\}) + \operatorname{E}(X \mathbf{1}\{X > \theta \operatorname{E} X\}) \\ &= \theta \operatorname{E} X + \sqrt{\operatorname{E} X^2 P(X > \theta \operatorname{E} X)}, \end{aligned}$$

and then rearrange to get the desired expression.

Now let $\theta_n = 1/n$ and take $n \to \infty$ to get the case for $\theta = 0$. □

We remark Markov's inequality and Paley–Zygmund inequality are related respectively to the *first* and the *second moment method* in probabilistic combinatorics; see [Roc24, Chapter 2].

7.28 Hoeffding's inequality. Suppose $X_1, \dots, X_n$ are independent, where X_k is almost surely contained in $[a_k, b_k]$ with means μ_k for all $k \in [n]$. Then for any $t \geq 0$, we have

$$P\left(\sum_{k=1}^{n}(X_k - \mu_k) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\right).$$

Judging by the look of the above inequality, it is clear that we need the Chernoff method for a proof. An additional ingredient is the following well-known lemma, which is surprisingly hard to establish.

7.29 Hoeffding's lemma. For a mean zero random variable Y that is a.s. bounded within $[a, b]$, we have

$$M_Y(\lambda) = \mathrm{E}\exp(\lambda Y) = \exp\left(\frac{\lambda^2(b-a)^2}{8}\right) \tag{7.30}$$

Proof. Since $e^{\lambda x}$ is convex, we have for $x \in [a, b]$

$$e^{\lambda x} \leq \frac{b-x}{b-a}e^{\lambda a} + \frac{x-a}{b-a}e^{\lambda b}.$$

Taking expectation on both sides, and we have

$$\mathrm{E}\exp(\lambda Y) \leq \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} = e^{L(\lambda(b-a))}, \tag{7.31}$$

where for all $h \in \mathbf{R}$, $L(h) = -\gamma h + \log(1 - \gamma + \gamma e^h)$, with $\gamma = -\frac{a}{b-a} > 0$ (so that the log is well-defined).

Notice that $L(0) = 0$, and $L'(0) = -\gamma + \frac{\gamma e^h}{1-\gamma+\gamma e^h}\big|_{h=0} = 0$.

$$\begin{aligned} L''(h) &= \left(\frac{\gamma e^h}{1-\gamma+\gamma e^h}\right)' \\ &= \frac{\gamma e^h}{1-\gamma+\gamma e^h} - \frac{\gamma^2 e^{2h}}{(1-\gamma+\gamma e^h)^2} \\ &= t - t^2 \leq 1/4 \quad \text{for any } t \in \mathbf{R}, \end{aligned}$$

where we let $t = \frac{\gamma e^h}{1-\gamma+\gamma e^h}$. Now we appeal to Taylor's theorem: for $h \neq 0$, there exists a ξ between h and 0 such that

$$L(h) = 0 + 0 \cdot h + \frac{L''(\xi)}{2} \cdot h^2 \leq \frac{1}{8}h^2.$$

Now let $h = \lambda(b-a)$ and plug it back into (7.31), and we have shown (7.30). □

At the moment, proving Hoeffding's inequality rigorously is left as an exercise to the reader. In fact, later when studying martingales, we will prove a renowned generalization known as Azuma–Hoeffding inequality, and the above inequality will become a trivial special case.[10]

tight consider any binary random variables, with probabilities $1/2$

[10] Hence one may extract a proof of the above inequality from there.

7.D Miscellaneous but crucial facts and tools

7.32 Definition. Fix the dimension d. The *standard Gaussian measure* on $\mathbf{R}^d$ is the measure $\gamma\colon \mathcal{B}(\mathbf{R}^d) \to [0,\infty]$ given by

$$\gamma(A) = \frac{1}{\left(\sqrt{2\pi}\right)^d} \int_A \exp\left(\|x\|_2^2/2\right) dx.$$

It is quite clear that m and γ are equivalent measures, since $\exp(\cdot)$ is nonnegative. In fact, this crucial fact allows us to prove some deterministic facts in analysis, e.g., the space of all n-by-n matrix, when embedded into $\mathbf{R}^{n^2}$, is a.e. invertible.

7.33 Fact. For $Z \sim N(0,1)$, we have

$$\mathrm{E}[Zf(Z)] = \mathrm{E}f'(Z),$$

when the expectations on the two sides are defined.

7.34 Proposition [Ver18, Proposition 2.1.2] [MP10, Lemma 12.9]. For $Z \sim N(0,1)$, we have the following tail estimate: for any $t > 0$, it holds that

$$\frac{t}{t^2+1}\frac{1}{\sqrt{2\pi}}\exp(-t^2/2) \le P(X \ge t) \le \frac{1}{t}\frac{1}{\sqrt{2\pi}}\exp(-t^2/2).$$

Let $Z \sim N(0,1)$, then in general for all $k \in \mathbf{N}$,

$$\mathrm{E}Z^{2k-1} = 0 \quad \text{and} \quad \mathrm{E}Z^{2k} = \frac{(2k)!}{2^k k!}.$$

7.35 Proposition. For a sequence of normal random variables $Z_n \sim N(\mu_n, \sigma_n^2)$. Suppose $Z_n \Rightarrow Z$, then $Z \sim N(\lim_n \mu_n, \lim_n \sigma_n^2)$.

If $Z_n \to Z$ in probability (so they live in the same probability space), then $Z_n \to Z$ in L^p for all p.

Proof. □

Bogachev Theorem 1.4.3.

The coordinates of a normal random vector are independent if and only if they are uncorrelated.

Gaussian measures are orthogonally invariant. However, it is not translation invariant.

We now restate Borel–Cantelli lemma I.

7.36 Borel–Cantelli lemma I. For events $A_1, A_2, \ldots$, if $\sum_n P(A_n) < \infty$, then

$$P(A_n \text{ i.o.}) = 0.$$

In probability this theorem is typically applied to show the a.s. convergence of random variables. We may rewrite

$$\{\omega : X_n(\omega) \to X(\omega)\} = \bigcap_{\epsilon>0}\{\omega \in \Omega : |X_n(\omega) - X(\omega)| < \epsilon \text{ ev.}\}.$$

Therefore $X_n \to X$ a.s. is equivalent to

$$\forall\, \epsilon > 0,\ P\left(|X_n(\omega) - X(\omega)| \ge \epsilon \text{ i.o.}\right) = 0.$$

(This is true for infinite measure space as well, and hence provides a characterization of a.e. convergence.) Equivalently, since we are in a probability space, $X_n \to X$ a.s. is the same as saying

$$\forall\, \epsilon > 0,\, P\big(|X_n(\omega) - X(\omega)| < \epsilon \text{ ev.}\big) = 1.$$

7.37 Borel–Cantelli lemma II. For pairwise independent events $A_1, A_2, \dots$, if $\sum_n P(A_n) = \infty$, then

$$P(A_n \text{ i.o.}) = 1.$$

The proof is much easier if we assume that the events are independent.

Proof. □

non-measurable set of the coin-tossing space

uniform measure on the sphere

The following elementary inequality is widely useful in research, but not often discussed in the textbooks. The proof uses the very important technique of introducing an independent copy of a given random variable. It is truly magical that an exogenous random variable that does not appear in the problem statement itself can make such a difference to a problem.

7.38 Harris' inequality. [11] Given a random variable X taking values on some totally ordered set S, and increasing functions f and g such that $f(X)$ and $g(X)$ are L^2 (or nonnegative), we have

$$\mathrm{E} f(X) \cdot \mathrm{E} g(X) \le \mathrm{E}[f(X)g(X)].$$

More generally, the above inequality still holds if $X = (X_1, \dots, X_n)$ takes value on a product space $S_1 \times \cdots \times S_n$ and has independent components, and f and g are increasing in each component.

Proof. Let Y be an independent copy of X. Consider the expectation

$$\begin{aligned} &\mathrm{E}\big\{[f(X) - f(Y)] \cdot [g(X) - g(Y)]\big\} \\ &= \mathrm{E}[f(X)g(X)] - \mathrm{E}[f(Y)g(X)] - \mathrm{E}[f(X)g(Y)] + \mathrm{E}[f(Y)g(Y)] \\ &= 2\,\mathrm{E}[f(X)g(X)] - 2\,\mathrm{E} f(X) \cdot \mathrm{E} g(Y), \end{aligned} \tag{7.39}$$

where we have used $f(X) \perp g(Y)$ and $f(Y) \perp g(X)$.

For any outcome $\omega \in \Omega$, if $X(\omega) \ge Y(\omega)$, then by monotonicity of f and g we have

$$f(X) - f(Y) \ge 0 \quad \text{and} \quad g(X) - g(Y) \ge 0,$$

which implies that

$$[f(X) - f(Y)] \cdot [g(X) - g(Y)] \ge 0.$$

The above inequality also holds when $X(\omega) < Y(\omega)$. Therefore

$$\mathrm{E}\big\{[f(X) - f(Y)] \cdot [g(X) - g(Y)]\big\} \ge 0.$$

The desired inequality then follows by using (7.39).

If suffices to only consider the case where $X = (X_1, X_2)$, since the rest can be done by induction.

Say X_1 takes value in S_1 with distribution μ_1. Define $f_1(x_1) = \mathrm{E} f(x_1, X_2)$ and $g_1(x_1) = \mathrm{E} g(x_1, X_2)$. It is clear that f_1 and g_1 should be increasing. Note that by the Fubini–Tonelli theorem, we have

$$\mathrm{E} f(X) \cdot \mathrm{E} g(X) = \mathrm{E} f_1(X_1) \cdot \mathrm{E} g_1(X_1)$$

[11] also known as Fortuin–Kasteleyn–Ginibre (FKG) inequality

and

$$\mathrm{E}[f(X)g(X)] = \int_{S_1} \mathrm{E}[f(x_1, X_2)g(x_1, X_2)]\, d\mu_1(x_1).$$

By the 1-dimensional case, since f is increasing in the second coordinate, we know

$$\begin{aligned}\mathrm{E}[f(x_1, X_2)g(x_1, X_2)] &\geq \mathrm{E}f(x_1, X_2) \cdot \mathrm{E}g(x_1, X_2) \\ &= f_1(x_1)g_1(x_1),\end{aligned}$$

and therefore

$$\begin{aligned}\mathrm{E}[f(X)g(X)] &\geq \int_{S_1} f_1(x_1)g_1(x_1)\, d\mu_1(x_1) \\ &= \mathrm{E}[f_1(X_1)g_1(X_1)] \\ &\geq \mathrm{E}f_1(X_1) \cdot \mathrm{E}g_1(X_1) \\ &= \mathrm{E}[f(X)g(X)].\end{aligned}$$

Here we used the 1-dimensional case again in the second-to-last line. □

symmetrization technique
replace X by $X - X'$
replace X by εX

Chapter 8 Modes of convergence in probability

8.A Statistical distances

Important disclaimer. This section deals purely with comparisons of probability measures μ and ν on a given measurable space $(S, \mathcal{S})$, and has nothing to do with random variables. In practice we may want to see μ and ν indeed as probability distributions of random variables on the codomain space $(S, \mathcal{S})$. Please be very careful about this distinction.

Given two probability measure μ and ν on $(S, \mathcal{S})$, we have the signed measure $\mu - \nu \colon \mathcal{S} \to [-1, 1]$. Its total variation norm

$$\begin{aligned} \|\mu - \nu\| &= |\mu - \nu|(S) \\ &= \sup_{A \in \mathcal{S}} |(\mu - \nu)(A)| + |(\mu - \nu)(S - A)| \\ &= \sup_{A \in \mathcal{S}} |\mu(A) - \nu(A)| + |1 - \mu(A) - 1 + \nu(A)| \\ &= 2 \sup_{A \in \mathcal{S}} |\mu(A) - \nu(A)|. \end{aligned}$$

The factor 2 above is usually dropped in probabilistic applications. We define the *total variation distance* between μ and ν to be

$$d_{\mathrm{TV}}(\mu, \nu) = \frac{1}{2} \|\mu - \nu\| = \sup_{A \in \mathcal{S}} |\mu(A) - \nu(A)|.$$

It should be clear that the absolute value sign can be dropped in the definition above, since

$$\mu(A) - \nu(A) = \nu(A^{\mathrm{c}}) - \mu(A^{\mathrm{c}}).$$

8.1 Definition. On a given measurable space $(S, \mathcal{S})$, we say a sequence of probability measure $\{\mu_n\}$ *converges* to a probability measure μ *in total variation* if

$$d_{\mathrm{TV}}(\mu_n, \mu) \to 0. \tag{8.2}$$

Note that if μ_n are probability measures and (8.2) holds, then the TV-limit μ must be a probability measure. This is because

$$0 = \lim_n d_{\mathrm{TV}}(\mu_n, \mu) = \lim_n \sup_{A \in \mathcal{S}} |\mu_n(A) - \mu(A)|,$$

which in particular implies $\mu_n(S) - \mu(S) \to 0$. We remark that convergence in total variation may be understood as *uniform* setwise convergence. *Setwise convergence*, by its name, means that

$$\mu_n(S) - \mu(S) \to 0 \quad \text{for all } S \in \mathcal{S}.$$

The total variation convergence given above can of course be defined for general finite/signed/complex measures, by using the distance induced from the total variation norm $\|\cdot\|$ in place of d_{TV}. (We know $d_{\mathrm{TV}}(\mu_n, \mu)$ and $\|\mu_n - \mu\|$ differ by a constant factor of 2, which leads to the same definition of convergence.) We do not discuss this convergence in the general setting.

The following result is a restatement of something we have proved in Section 4.A.

8.3 Fact. If μ and ν have a common dominating measure ρ, then

$$d_{\mathrm{TV}}(\mu, \nu) = \frac{1}{2} \int_S \left| \frac{d\mu}{d\rho}(\omega) - \frac{d\nu}{d\rho}(\omega) \right| d\rho. \tag{8.4}$$

In particular, if $(S, \mathcal{S})$ is a discrete space, then

$$d_{\mathrm{TV}}(\mu, \nu) = \frac{1}{2} \sum_{x \in S} |\mu\{x\} - \nu\{x\}|.$$

And if $(S, \mathcal{S}) = (\mathbf{R}, \mathcal{B})$, with ρ being the Lebesgue measure, then

$$d_{\mathrm{TV}}(\mu, \nu) = \frac{1}{2} \int_{\mathbf{R}} |f(x) - g(x)| \, dx,$$

where $f = \frac{d\mu}{d\rho}$ and $g = \frac{d\nu}{d\rho}$ are the two probability densities[1]. In short, the total variation distance between two probability measures is half the L^1 distance between their densities.

The *Kullback–Leibler divergence/relative entropy* of μ with respect to ν is given by

$$D_{\mathrm{KL}}(\mu \| \nu) = \begin{cases} \int_S \log \frac{d\mu}{d\nu} \, d\mu & \text{if } \mu \ll \nu, \\ +\infty & \text{otherwise.} \end{cases}$$

Let $f = \frac{d\mu}{d\nu} \in L^1(\nu)$. It is very important to note that

$$\int_S \log \frac{d\mu}{d\nu} \, d\mu = \int_S f \log f \, d\nu$$

can be infinite, since $f \log f$ might not be integrable with respect to ν. Sometimes we just

8.5 Fact. If $\mu \ll \nu \ll \rho$, then

$$D_{\mathrm{KL}}(\mu \| \nu) = \int_S \left(\frac{d\mu}{d\rho} \right) \log \left(\frac{d\mu/d\rho}{d\nu/d\rho} \right) d\rho.$$

Therefore if the space is discrete, then we take ρ to be the counting measure and get

$$D_{\mathrm{KL}}(\mu \| \nu) = \sum_{x \in S} \mu\{x\} \log \frac{\mu\{x\}}{\nu\{x\}}.$$

And if $(S, \mathcal{S}) = (\mathbf{R}, \mathcal{B})$, with ρ being the Lebesgue measure, then

$$D_{\mathrm{KL}}(\mu \| \nu) = \int_{\mathbf{R}} f(x) \log \frac{f(x)}{g(x)} \, dx,$$

where $f = \frac{d\mu}{d\rho}$ and $g = \frac{d\nu}{d\rho}$ are the two probability densities. In this latter case we might as well write $D_{\mathrm{KL}}(f \| g)$.

[1] Of course we may consider μ and ν on some restricted subspace of $(\mathbf{R}, \mathcal{B})$, but as mentioned before we drop such consideration for brevity.

Given a probability measure ν, for a nonnegative $f \in L^1(\nu)$ such that $f \log f$ is also ν-integrable, we define its *entropy functional* to be

$$\operatorname{Ent}_\nu f = \mathrm{E}_\nu(f \log f) - (\mathrm{E}_\nu f)(\log \mathrm{E}_\nu f),$$

which should be compared with the variance functional

$$\operatorname{Var}_\nu f = \mathrm{E}_\nu f^2 - (\mathrm{E}_\nu f)^2.$$

But keep in mind the entropy functional can only be applied to (ν-a.e.) nonnegative[2] functions because of the logarithm in the definition. Also note that the entropy functional is homogeneous: we have

$$\operatorname{Ent} cf = c \operatorname{Ent} f \quad \text{for } c \geq 0,$$

which is "better" than

$$\operatorname{Var} cf = c^2 \operatorname{Var} f \quad \text{for } c \in \mathbf{R}$$

in some applications.[3]

If we have another probability measure μ with $\mu \ll \nu$, then

$$\operatorname{Ent}_\nu \frac{d\mu}{d\nu} = D_{\mathrm{KL}}(\mu \| \nu).$$

If $d\mu/d\nu$ can be explicitly expressed by some function h (as discussed in Fact 8.5), then the equation above gives a simple expression for the KL divergence.

Fisher information 5.1.2 Markov diffusion operators LSI

8.6 Pinsker's inequality. $d_{\mathrm{TV}}(\mu, \nu) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mu \| \nu)}$.

Say μ and ν have a common dominating measure ρ, with

$$\frac{d\mu}{d\rho} = f \quad \text{and} \quad \frac{d\nu}{d\rho} = g,$$

then the *Hellinger distance* between μ and ν is defined by

$$H(\mu, \nu) = \left(\frac{1}{2} \int_S \left[\sqrt{f(x)} - \sqrt{g(x)} \right]^2 d\rho(x) \right)^{1/2}.$$

The Hellinger distance always exists, since we may take $\rho = \mu + \nu$. The distance is well-defined, in the sense that it is independent of the choice of such ρ. (Clearly this is a straightforward exercise using the chain rule for Radon–Nikodym derivatives.) One can obviously write down the expression when ρ is the counting measure or the Lebesgue measure, which we omit here.

When the Hellinger distance exists, the following holds:

$$H^2(\mu, \nu) \leq d_{\mathrm{TV}}(\mu, \nu) \leq \sqrt{2} H(\mu, \nu).$$

This follows from a straightforward comparison with (8.4).

probability metric

[2] If $f = 0$ ν-a.e., since $0 \log 0$ is taken to be 0, we would have no problem.

[3] Some authors define *φ-entropy* for a convex function to mean $\mathrm{E}\varphi(X) - \varphi(\mathrm{E}X)$, which puts the "Ent" and "Var" under the same umbrella.

The *integral probability metric* (IPM) uses a class of test functions $\mathcal{F}$ to determine the distance between μ and ν:

$$d_{\mathcal{F}}(\mu,\nu) = \sup_{f\in\mathcal{F}}\left|\int_S f\,d\mu - \int_S f\,d\nu\right|.$$

To be precise $d_{\mathcal{F}}$ is in fact a pseudometric, and it is a metric if and only if there exists $f \in \mathcal{F}$ such that $\int_S f\,d\mu \neq \int_S f\,d\nu$.

If we take $\mathcal{F}$ to be the collection of all indicator functions, then $d_{\mathcal{F}} = d_{\mathrm{TV}}$.

The *Kolmogorov uniform metric* is defined by

$$d_{\mathrm{K}}(\mu,\nu) = \sup_{x\in\mathbf{R}}|F_\mu(x) - F_\nu(x)| = \sup_{x\in\mathbf{R}}\big|\mu(-\infty,x] - \nu(-\infty,x]\big|,$$

which is the an IPM $d_{\mathcal{F}}$ with $\mathcal{F} = \{\mathbf{1}_{(-\infty,x]} : x \in \mathbf{R}\}$.

Let (S,ρ) be a separable metric space, and $1 \le p < \infty$, the *Wasserstein distance* of order p is defined by

$$W_p(\mu,\nu) = \left[\inf_{\pi\in\Pi(\mu,\nu)}\int_{S\times S}\rho(x,y)^p\,d\pi(x,y)\right]^{1/p}.$$

$\pi = P\circ(X,Y)^{-1}$ on measurable rectangles, extend to the entire $(S\times S, \mathcal{S}\otimes\mathcal{S})$

Alternatively, one has the probabilistic interpretation

$$W_p(\mu,\nu) = \inf\big\{\mathrm{E}\big[\rho(X,Y)^p\big]^{1/p} : X\sim\mu, Y\sim\nu\big\},$$

and understand it as an L^p distance between two probability measures.

The separability of (S,ρ) ensures $\rho\colon S\times S\to[0,\infty)$ to be a measurable function with respect to the product σ-field $\mathcal{B}(S)\otimes\mathcal{B}(S)$, which we discussed in Remark 3.7.

Restricting μ and ν to be measures on the *Wasserstein space* enforces W_p to be finite, and henceforth a metric, as we will see. The *Wasserstein space* of order p is defined by

$$\mathcal{P}_p(S) = \left\{\mu\in\mathcal{P}(S) : \int_S \rho(x_0,x)^p\,d\mu(x) < \infty \text{ for all } x_0\in S\right\}.$$

8.7 Dual representation of W_1. For $\mu,\nu\in\mathcal{P}_p(S)$, we have

$$W_1(\mu,\nu) = \sup\left\{\left|\int f\,d\mu - \int f\,d\nu\right| : f \text{ is 1-Lipschitz}\right\}.$$

Thus W_1 is an IPM.

8.B Weak convergence of probability measures

Let (S,ρ) be a metric space. We use $\mathcal{P}(S)$ for the space of Borel probability measures. A *subprobability measure* μ is a measure with $\mu(S)\le 1$, and we denote the space of all Borel subprobability measures by $\mathcal{M}^{\le 1}(S)$.

Our attention will be restricted to the case when μ_n is a sequence of Borel probability measures.

The current section aims to present the tip of the iceberg of the theory of weak convergence. For the thorough treatment of weak convergence of Borel probability measures on metric spaces, see the classical [Bil99] and [Par67].

8.8 Definition. A sequence $\{\mu_n\}$ of Borel probability measures *converges weakly* to a Borel probability measure μ if for all $f \in C_b(S)$, we have

$$\int_S f \, d\mu_n \to \int_S f \, d\mu,$$

which we denote by $\mu_n \Rightarrow \mu$.

If each μ_n and μ represents the distribution of some $(S, \mathcal{B}_S)$-valued random variables X_n and X, then we usually say X_n *converges to* X *in distribution*, denoted by[4] $X_n \Rightarrow X$. Because of Corollary 7.6, when $S = \mathbf{R}$ we also write $F_{X_n} \Rightarrow F_X$.

Recall that vague convergence

8.9 Proposition. Weak convergence of integer-valued measures is equivalent to pointwise convergence.

8.10 Alexandroff portmanteau theorem. [5] The following statements are equivalent characterizations of the weak convergence of Borel probability measures on a metric space (S, ρ).

(a) $\int f \, d\mu_n \to \int f \, d\mu$ for all bounded Lipschitz functions f on S;

(b) $\int f \, d\mu_n \to \int f \, d\mu$ for all bounded uniformly continuous functions f on S;

(c) $\limsup_n \int f \, d\mu_n \le \int f \, d\mu$ for all USC functions bounded from above;

(d) $\liminf_n \int f \, d\mu_n \ge \int f \, d\mu$ for all LSC functions bounded from below;

(e) $\limsup_n \mu_n(F) \le \mu(F)$ for all closed sets F;

(f) $\liminf_n \mu_n(G) \ge \mu(G)$ for all open sets G;

(g) $\lim_n \mu_n(A) = \mu(A)$ for all *continuity sets* A with respect to μ, i.e., Borel sets A with $\mu(\partial A) = 0$.

The same convergence remains in force if we have $\lim_n \mu_n(S) = \mu(S)$ for $\mu_n, \mu \in \mathcal{M}^+(S)$.

8.11 Theorem. When $S = \mathbf{R}$, the weak convergence of probability measures $\mu_n \Rightarrow \mu$ is equivalent to $F_n(x) \to F(x)$ at every continuity point x of F, where F_n and F are the distribution functions of μ_n and μ, respectively.

Proof. Characterization (g) immediately tells us the direction that weak convergence implies convergence at all continuity points of the limiting distribution function. For the reverse direction, □

Compare with the Arzelà–Ascoli theorem for space of continuous functions.

8.12 Prohorov's theorem. Suppose a collection of random variables $\mathcal{K} \subseteq \mathcal{P}(S)$ is tight, then $\mathcal{K}$ is precompact in the topology of weak convergence on $\mathcal{P}(S)$.

The converse is true when S is Polish.

Proof. We follow [Kal21, Theorem 23.2]. □

8.13 Corollary. [ABS24, Corollary 2.9] Let S and T be Polish, then the space $\Pi(\mu, \nu)$ of couplings between $\mu \in \mathcal{P}(S)$ and $\nu \in \mathcal{P}(T)$ is a compact subspace of $\mathcal{P}(S \times T)$.

[4] sometimes even mix up and write $X_n \Rightarrow \mu$

[5] As Bogachev [Bog18] points out, "I do not know who invented such a nonsensical name for Alexandroff's theorem."

Proof. First we show $\Pi(\mu,\nu)$ is closed in $\mathcal{P}(S\times T)$. We know each $\pi\in\mathcal{P}(\mu,\nu)$ is characterized by

$$\int_{S\times T}\varphi\times \mathrm{Id}_Y\, d\pi=\int_S \varphi\, d\mu \quad \text{for all } \varphi\in C_b(S),$$

and symmetrically with respect to the marginal ν.

By Prohorov's theorem, it now suffices to show that $\Pi(\mu,\nu)$ is a tight family. By Ulam's theorem, for any $\epsilon>0$, there exists $K_1\subseteq S$ and $K_2\subseteq T$ such that

$$\mu(S-K_1)<\epsilon/2 \quad \text{and} \quad \mu(T-K_2)<\epsilon/2.$$

It follows that for any $\pi\in\Pi(\mu,\nu)$

$$\begin{aligned}\pi(S\times T-K_1\times K_2)&\le \pi\big((S-K_1)\times T\big)+\pi\big(S\times(T-K_2)\big)\\&=\mu(S-K_1)+\nu(T-K-2)<\epsilon,\end{aligned}$$

proving tightness. □

The proof of the following result (and its generalizations) resembles that of the classical Arzelà–Ascoli theorem on $\mathbf{R}^d$. The shared proof idea is to construct a desired subsequence (that converges pointwise on all rationals) by the so-called diagonal argument. The construct can be made very explicit, as we will show below.

8.14 Lemma. Let $\{F_n\}$ be a sequence of distribution functions, then there is a subsequence $\{F_{n_k}\}$ and a right-continuous increasing function F such that

$$\lim_{k\to\infty} F_{n_k}(x)=F(x)$$

for all continuity points x of F.

Proof. Let $q_1,q_2,\dots$ be an enumeration of $\mathbf{Q}$. First $\{F_n(q_1)\}$ is a sequence in $[0,1]$, a bounded interval, and therefore there is a subsequence that converges to $s_1:=\liminf_n F_n(q_1)$. We can construct such a subsequence by defining $F_{\nu(n)}(q_1)$ inductively for all n:

$$\nu(n)=\min\{m>\nu(n-1):|F_m(q_1)-s_1|<1/n\}.$$

Let $\{F_n^1\}$ be the new sequence $\{F_{\nu(n)}\}$, and in the same way one can construct its subsequence $\{F_n^2\}$ satisfying $\lim_n F_n^2(q_2)=s_2:=\liminf_n F_n^1(q_2)$. Proceeding in this fashion, we get

Subsequences listed in rows

$$\begin{array}{ccccc}F_1^1 & F_2^1 & F_3^1 & F_4^1 & \cdots\\ F_1^2 & F_2^2 & F_3^2 & F_4^2 & \cdots\\ F_1^3 & F_2^3 & F_3^3 & F_4^3 & \cdots\\ F_1^4 & F_2^4 & F_3^4 & F_4^4 & \cdots\\ \vdots & \vdots & \vdots & \vdots & \ddots\end{array}$$

Take the diagonal sequence $F_1^1,F_2^2,\dots$, which we call F_{n_k}. If we ignore the first $j-1$ terms of the diagonal sequence, this new F_{n_k} is a subsequence of $\{F_n^j\}_{n=1}^\infty$. Therefore this subsequence converges at all rational points. Let $G\colon\mathbf{Q}\to[0,1]$ be its pointwise limit:

$$G(q)=\lim_k F_{n_k}(q).$$

Take its increasing, right-continuous inverse F given by

$$F(x) = \inf\{G(q) : q \in \mathbf{Q}, q > x\}.$$

Notice for a rational q strictly between two reals x_1 and x_2, we have

$$F(x_1) \leq G(q) \leq F(x_2), \tag{8.15}$$

where the second inequality follows from the fact that G is increasing on $\mathbf{Q}$.

Now let x be a continuity point of F, then for any $\epsilon > 0$, we can find two reals $\tilde{r}$ and $\hat{r}$ with $\tilde{r} < x < \hat{r}$, such that

$$F(x) - \epsilon < F(\tilde{r}) \leq F(x) \leq F(\hat{r}) < F(x) + \epsilon.$$

Then we can find two rationals $\tilde{q}$ and $\hat{q}$ satisfying $\tilde{r} < \tilde{q} < x < \hat{q} < \hat{r}$, such that

$$F(x) - \epsilon < G(\tilde{q}) \leq F(x) \leq G(\hat{q}) < F(x) + \epsilon,$$

by (8.15). Since G is the pointwise limit of F_{n_k} on the rationals, for all k large enough we will have

$$F(x) - \epsilon < F_{n_k}(\tilde{q}) \leq F_{n_k}(x) \leq F_{n_k}(\hat{q}) < F(x) + \epsilon.$$

It follows that F_{n_k} converges to F at all continuity points x. □

Recall that the subsequential limit F constructed above associates to a Borel measure μ_F, by Theorem 1.31(a). This μ_F is a subprobability measure on $(\mathbf{R}, \mathcal{B})$, since for all continuity points x,

$$\mu_F(-\infty, x] = F(x) \leq 1.$$

Since the increasing function F has at most countably many discontinuities, we can construct a sequence of continuity points approaching ∞, and conclude $\mu_F(\mathbf{R}) \leq 1$.

See [Sch17, Theorem 21.18, Corollary 21.19] for a direct proof.

8.16 Helly selection theorem. Let S be a locally compact separable metric space. For any sequence $\{\mu_n\} \subseteq \mathcal{M}^{\leq 1}(S)$, it has a vague subsequential limit in $\mathcal{M}^{\leq 1}(S)$. This means exactly that $\mathcal{M}^{\leq 1}(S)$ is sequentially compact in the vague topology.

Proof. This follows by combining three results. We know $(C_c(S), \|\cdot\|_u)$ is a separable normed space, and by the sequential Banach–Alaoglu theorem, $C_c(S)^*$ must be weak-star sequentially compact. By the Riesz–Markov–Kakutani theorem for finite measures, the sequence $\{\mu_n\} \subseteq \mathcal{P}(S)$ is norm bounded in $\mathcal{M}(S) \cong C_c(S)^*$. Hence μ_n must have a subsequential vague limit μ that satisfies $\|\mu\| \leq 1$. Since μ_n are all positive measures, μ must also be positive measure, and hence a subprobability measure. □

8.17 Corollary. When S is a compact metric space, weak and vague convergence coincides. Hence for any sequence $\{\mu_n\} \subseteq \mathcal{P}(S)$, it has a weak subsequential limit in $\mathcal{P}(S)$. This shows that $\mathcal{P}(S)$ is sequentially compact in the vague/weak topology.

A finite-dimensional normed space $(S, \|\cdot\|)$ must be locally compact and separable.

The μ_F above is not in general a probability measure, for example, consider the sequence of distribution functions F_n of the uniform distributions over $[-n, n]$. The sequence $\{F_n\}$ *itself* (and hence all of its subsequences) converges vaguely to the 0 function. To ensure that the subsequential F constructed in Lemma 8.14 is indeed a distribution function, we require tightness over *the entire*

sequence of measures in addition. We say a family of measures Γ is *tight* if for each $\epsilon > 0$, there exists some compact set K_ϵ such that

$$\sup_{\gamma \in \Gamma} \mu_\gamma(S - K_\epsilon) < \epsilon.$$

Indeed this is just the generalization of tightness of one measure we have discussed previously. (Some authors use the term "uniformly tight" or "equi-tight" to stress the difference.)

When S is compact, we know weak and vague convergence for a sequence of measures are the same. To upgrade vague convergence to weak convergence in the general case of a locally compact separable metric space S, it seems natural to control the proximity-in-measure of S to a compact metric space.

When $S = \mathbf{R}^d$, vague convergence may be further defined by $\int f \, d\mu \to \int f \, d\mu$ for $f \in C_c^\infty(\mathbf{R}^d)$

8.18 Theorem [Sch17, Theorem 21.17]. Let S be locally compact and separable[6], and $\{\mu_n\} \subseteq \mathcal{P}(S)$, then the following are equivalent.

(a) $\mu_n \Rightarrow \mu$;

(b) $\mu_n \to \mu$ vaguely, with $\mu \in \mathcal{P}(S)$;

(c) $\mu_n \to \mu$ vaguely, with $\{\mu_n\}$ being a tight sequence of measures.

We are now ready to generalize Corollary 8.17 from compact to locally compact separable metric spaces. It also provides an accurate characterization for tightness.

8.19 Proposition. For a sequence of Borel probability measures in a locally compact separable metric space, every vague subsequential limit (which always exists by THEOREM 8.16) is a probability measure if and only if the sequence is tight.

Proof. One direction is already contained in the previous theorem. For the other direction, suppose every vague subsequential limit of $\{\mu_n\}$ is a probability measure, but the sequence is not tight. By Helly selection theorem, we may assume in addition that μ_{n_j} is vaguely convergent sequence, with limit as a probability measure by assumption. This contradicts Theorem 8.18. □

8.20 Helly selection theorem. If we assume that in Lemma 8.14 $\{F_n\}$ is a tight sequence of distribution functions, then the vague subsequential limit F constructed there is a distribution function.

generalization subprobability measure (21.16 17 18)

8.21 Skorohod representation theorem (Polish space). Let (S, ρ) be Polish. Suppose $\mu_n \Rightarrow \mu$, then there exist X_n and X defined on a common probability space $(\Omega, \mathcal{F}, P) = ([0,1], \mathcal{B}, m)$, such that $X_n \sim \mu_n$, $X \sim \mu$, and $X_n \to X$ pointwise everywhere on Ω.

Redefine X_n by X outside the set of convergence Weak compactness
Prohorov metric for $S = \mathbf{Z}$

8.B.1 The topology and metric of weak convergence

8.22 Theorem [Bog18, Theorem 3.1.2]. The weak topology on $\mathcal{M}^+(S)$

Prohorov metric On a Polish space

[6] Of course we can state this result in general for lc(sc)H spaces, but we chose not to due to our focus on metric spaces.

8.B.2 Problem of measurability

When S is infinite, $\mathcal{B}(S)$ is not separable.

8.C Comparisons between modes of convergence

8.23 Theorem. If $\mu_n \to \mu$ in total variation, $\mu_n \to \mu$ setwise, which implies that $\mu_n \Rightarrow \mu$.

Proof. The first part has already been discussed. For the second part, we know setwise convergence means that for all $A \in \mathcal{S}$,

$$\int \mathbf{1}_A \, d\mu_n \to \int \mathbf{1}_A \, d\mu.$$

The convergence then extends to all bounded measurable functions, which of course include $C_b(S)$.

Alternatively this also follows from characterization (g) in Alexandroff portmanteau theorem. □

For this reason, $\mu_n \to \mu$ setwise is often referred to as *strong convergence of measures* as opposed to weak convergence of meausres.

8.24 Theorem. If $X_n \to X$ a.s., then $X_n \to X$ in probability, which further implies $X_n \Rightarrow X$ when S is a separable metric space.

Proof. The first part was done in Theorem 2.20. □

8.25 Theorem. If $X_n \Rightarrow c$ for some real constant c, then $X_n \to c$ in probability.

Notice that for $g \in C_b(\mathbf{R})$ and $f \in C_b(S)$, $g \circ f \in C_b(S)$.

8.26 Continuous mapping theorems. Let f be a continuous function. If $X_n \to X$ weakly/in probability/almost surely, we then have $f(X_n) \to f(X)$ weakly/in probability/almost surely, respectively.

8.27 Lemma. If $X_n \Rightarrow X$ and $Y_n \Rightarrow c$ for some real constant c, then

$$(X_n, Y_n) \Rightarrow (X, c).$$

Proof. □

Convergence of one sequence in distribution and another to a constant implies joint convergence in distribution

The following result is a direct corollary of

8.28 Slutsky's theorem. Suppose $X_n \Rightarrow X$ and $Y_n \Rightarrow c$ as real random variables, then

(a) $X_n + Y_n \Rightarrow X + c$;

(b) $Y_n X_n \Rightarrow cX$;

(c) $X_n / Y_n \Rightarrow X/c$, provided that c is invertible.

Holds for random matrices as well

8.D Laws of large numbers

8.29 L^2 weak law. Let $X_1, X_2, \ldots$ be uncorrelated L^2 random variables with equal mean μ and $\sup_j \operatorname{Var}(X_j) < \infty$. Then

$$\frac{X_1 + \cdots + X_n}{n} \to \mu$$

in L^2 (and hence in probability).

Proof. We have

$$\mathrm{E}\left(\frac{X_1 + \ldots + X_n}{n} - \mu\right)^2 = \operatorname{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right) \leq \frac{1}{n} \sup_j \operatorname{Var}(X_j).$$

Take $n \to \infty$ gives the result. □

8.30 L^1 weak law. Let $X_1, X_2, \ldots$ be i.i.d. and L^1 with mean μ. Then

$$\frac{X_1 + \ldots + X_n}{n} \to \mu$$

in probability.

8.31 L^1 strong law. Let $X_1, X_2, \ldots$ be pairwise independent, identically distributed L^1 random variables with mean μ. We have

$$\frac{X_1 + \ldots + X_n}{n} \to \mu \quad \text{a.s.}$$

Furthermore the above convergence also holds in L^1.

Proof. The a.s. part will follow the Etemadi's classical truncation proof.

It remains to show that $\{\overline{X}_n\}_{n\in\mathbf{N}} = \left\{\frac{X_1+\ldots+X_n}{n}\right\}_{n\in\mathbf{N}}$ is uniformly integrable. We know each X_j, as an L^1 random variable, must be uniformly integrable. In particular, for any $\epsilon > 0$, there is some $\delta > 0$ such that for all $n \in \mathbf{N}$,

$$\begin{aligned} P(A) < \delta &\implies \mathrm{E}\big(|X_j|; A\big) < \epsilon \quad \text{for all } j \in [n] \\ &\implies \mathrm{E}\left(\left|\frac{X_1 + \ldots + X_n}{n}\right|; A\right) < \epsilon. \end{aligned}$$

Meanwhile

$$\begin{aligned} \sup_n \mathrm{E}\left|\frac{X_1 + \ldots + X_n}{n}\right| &\leq \sup_n \frac{\mathrm{E}|X_1| + \ldots + \mathrm{E}|X_n|}{n} \\ &= \mathrm{E}|X_1| < \infty. \end{aligned}$$

Combining the above information gives uniformly integrability of $\{\overline{X}_n\}$. □

8.32 L^4 strong law.

8.33 L^2 strong law.

Let $X_1, X_2, \dots$ follow a common distribution μ, or alternatively a common distribution function F. The *empirical distribution* of the first n random variables is defined to

$$\mu_n = \frac{1}{n}\sum_{k=1}^{n} \delta_{X_k},$$

which is the averaging of the first n observations. Notice that this is a random variable in terms of $X_1, \dots, X_n$. This gives us the *empirical distribution function*

$$F_n(x) = \mu(-\infty, x] = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}\{X_k \le x\}.$$

8.34 Glivenko–Cantelli theorem. As $n \to \infty$,

$$\sup_x |F_n(x) - F(x)| \to 0 \quad P\text{-a.s.}$$

11.4 Dudley
Kolmogorov–Smirnov statistics and test

8.35 Dvoretzky–Kiefer–Wolfowitz–Massart inequality. For every $\epsilon > 0$,

$$P\left(\sup_x |F_n(x) - F(x)| > \epsilon\right) \le 2\exp(-2n\epsilon^2).$$

8.E Moment generating functions and characteristic functions

Integral transform converts a given problem to one which is easier to solve, and then 'inverting' to solve the original problem

For a real random variable X, its *moment generating function* (m.g.f.) is a function $M_X : \mathbf{R} \to \mathbf{R}$ defined by $M_X(t) = \mathrm{E}\exp(itX)$, provided that $\exp(itX)$ is integrable. Its *characteristic function* (ch.f.) is a function $\varphi_X : \mathbf{R} \to \mathbf{C}$ defined by $\varphi_X(t) = \mathrm{E}\exp(itX)$. Notice that

$$\mathrm{E}\exp(itX) = \mathrm{E}\cos(tX) + i\,\mathrm{E}\sin(tX)$$

always exists, because the real and imaginary parts are both bounded by 1.

testing against coefficients give you enough information to recover information about the random variable

For a random vector $X \in \mathbf{R}^d$, we would define $M_X(t) = \mathrm{E}\exp(i\langle t, X\rangle)$ and $\varphi_X(t) = \mathrm{E}\exp(i\langle t, X\rangle)$ for $t \in \mathbf{R}^d$.

The *cumulant generating function* is defined to be the log moment generating function.

[Bog07, Theorem 7.13.1] Bochner

8.36 Example. For $Z \sim N(0,1)$ and $t \in \mathbf{R}$, we have the $M_Z(t)$ given by

$$\begin{aligned}\mathrm{E}\exp(tX) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}e^{tx}\,dx \\ &= e^{t^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\Big(-\frac{1}{2}(x-t)^2\Big)\,dx = \exp(t^2/2).\end{aligned}$$

It turns out that the $\varphi_Z(t)$ has almost the same expression (except for the sign):

$$\mathrm{E}\exp(itX) = \exp(-t^2/2)$$

for all $t \in \mathbf{R}$. It suffices to show that

$$\mathrm{E}\exp(tX) = \exp(t^2/2) \tag{8.37}$$

for all $t \in \mathbf{C}$.

We wish to use the uniqueness theorem given (8.37) already holds for $t \in \mathbf{R}$. The left-hand side is holomorphic:

$$\begin{aligned}\partial_t \mathrm{E}\exp(tX) &= \mathrm{E}\partial_t \exp(tX)\\ &= \mathrm{E}X\exp(tX);\end{aligned}$$

and the right-hand side is obviously holomorphic[7].

8.38 Example. The L^2 limit of a sequence of normal random variables must be normal.

8.39 Recovery theorem for m.g.f. Suppose $M(t)$ exists for t in some neighborhood $(-\delta, \delta)$ of 0, then

(a) $\mathrm{E}|X|^k < \infty$ for all $k \in \mathbf{N}_0$, with $\mathrm{E}X^k = M^{(k)}(0)$;

(b) we have the Taylor expansion $M(t) = \sum_{k=0}^{\infty} \frac{\mathrm{E}X^k}{k!} t^k$ in $(-\delta, \delta)$.

8.40 Recovery theorem for ch.f. When the high-order derivatives of φ is finite, they recover the high-order moments of X. More precisely,

(a) if $\varphi^{(2k)}(0)$ exists, then $\mathrm{E}X^{2k} < \infty$;

(b) if $\mathrm{E}|X|^k < \infty$, then we have the Taylor approximation

$$\varphi(t) = \sum_{j=0}^{k} \frac{\mathrm{E}(iX)^j}{j!} t^j + o(t^k),$$

and in particular $\varphi^{(k)}(t) = i^k \mathrm{E}X^k$.

8.41 Inversion formula on the real line. Let $X \sim F$ with ch.f. φ, and define $\overline{F}\colon \mathbf{R} \to [0, 1]$

$$\overline{F}(x) = \frac{1}{2}[F(x) + F(x-)].$$

We have for any $a < b$,

$$\overline{F}(b) - \overline{F}(a) = \lim_{T\to\infty} \int_{-T}^{T} \frac{\exp(-iat) - \exp(-ibt)}{2\pi it} \varphi(t)\, dt$$

Note $\overline{F}(b) - \overline{F}(a) = \mu(a, b) + \frac{1}{2}(\mu\{a\} + \mu\{b\})$. In particular, if a and b are not atoms of μ_F, then the expression is equal to $\mu(a, b]$.

8.42 Theorem (c.d.f. and ch.f. correspondence). For any real random vectors X and Y, $X =_d Y$ if and only if $\varphi_X = \varphi_Y$.

[7] Recall we defined complex exponentials as power series, and power series/polynomials are differentiable term-by-term.

Proof. A short proof can be given when X and Y are $\mathbf{R}$-valued. One direction is obvious. Now assume $\varphi_X = \varphi_Y$, which gives

$$\overline{F}_X(b) - \overline{F}_X(a) = \overline{F}_Y(b) - \overline{F}_Y(a)$$

for all real numbers $a \leq b$. Take $a \to -\infty$ gives us $\overline{F}_X(b) = \overline{F}_Y(b)$.

We show the agreement of $\overline{F}$ implies the agreement of F. Now take F in fact to be any distribution function. For any $x \in \mathbf{R}$, consider a sequence $b_n = x + \frac{1}{n}$. Now

$$\lim_n F(b_n) = F(x)$$

and $\lim_n \mu(-\infty, b_n) = \mu(-\infty, x]$, i.e.,

$$\lim_n F(b_n-) = F(x).$$

Hence

$$\lim_n \overline{F}(b_n) = \frac{1}{2} \lim_n [F(b_n) + F(b_n-)] = F(x).$$

The conclusion now follows. □

8.43 Corollary. Two $\mathbf{R}^d$-random variables X and Y are independent if and only if $\varphi_{X,Y} = \varphi_X \varphi_Y$.

8.44 Theorem.

(a) If $\mu_n \Rightarrow \mu$, then the corresponding ch.f.'s have $\varphi_n \to \varphi$ pointwise everywhere;

(b) if $\varphi_n \to \varphi$ pointwise, and φ is continuous at 0, then the measures μ_n associated to φ_n are tight and converges weakly to some measure μ whose characteristic function is φ.

8.45 Classical central limit theorem. Let $X_1, X_2, \ldots$ be i.i.d. L^2 random variables with variance $\sigma^2 \neq 0$, then we have

$$\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \Rightarrow N(0, 1)$$

8.46 Lindeberg–Feller condition. For each $n \in \mathbf{N}$, let $\{X_{n,m}\}_{m=1}^n$ be a sequence of L^2 random variables with zero mean. If

(a) $\sum_{m=1}^n \mathrm{E}(X_{n,m})^2 = \sigma_n^2 > 0$, and

(b) for all $\epsilon > 0$, we have

$$\frac{1}{\sigma_n^2} \sum_{m=1}^n \mathrm{E}\big(|X_{n,m}^2; X_{n,m}^2 > \epsilon\sigma_n|\big) \to 0,$$

then

$$\frac{X_1 + \ldots + X_n}{\sigma_n} \Rightarrow N(0, 1).$$

8.47 Lyapunov condition.

8.48 Bochner's theorem. A characteristic function $\varphi\colon \mathbf{R} \to \mathbf{C}$ is precisely characterized by the following three properties:

(a) $\varphi(0) = 1$, and $\sup_t |\varphi(t)| \leq 1$;

(b) φ is uniformly continuous on $\mathbf{R}$;

(c) φ is a positive semidefinite function, i.e., for any finite number of real numbers $t_1, \ldots, t_n$, the matrix $[\varphi(x_j - x_k)]_{j,k}$ is positive semidefinite.

moment problem

8.49 Berry–Essen bound. For $X_1, X_2, \ldots$ i.i.d. with $\mathrm{E}|X_1|^3 < \rho$, $\mathrm{E}X_1 = 0$, and $\mathrm{E}X_1^2 = \sigma^2$, let

$$G_n(x) = P\left(\frac{X_1 + \cdots + X_n}{\sigma\sqrt{n}} \le x\right)$$

be the empirical CLT-scaled distribution. We have

$$|G_n(x) - G(x)| \le \frac{C\rho}{\sigma^3\sqrt{n}}$$

for some absolute constant $C > 0$.

One can show that

$\sqrt{n}$ rate of convergence to the distribution function

de Moivre–Laplace central limit theorem for binomial distributions with fixed p's can be proved directly for example with the help of Stirling's formula

8.50 Theorem.

$$\frac{S_n - np}{\sqrt{npq}} \Rightarrow N(0,1)$$

$$\frac{S_n}{\sqrt{n}} \Rightarrow N(0,1)$$

[Dur19, Theorem 3.12] $2k/\sqrt{2n} \to x$

$$P(S_{2n} = 2k) \simeq \frac{\exp(-x^2/2)}{\sqrt{\pi n}}.$$

If p_n decreases inversely in n, then we have the following theorem.

8.51 Poisson limit theorem. For a sequence of $X_n \sim \text{Binomial}(n, p_n)$, where $np_n \to \lambda$ for some positive constant λ, we have

$$X_n \Rightarrow \text{Poisson}(\lambda).$$

Explicitly, this means given $Y \sim \text{Poisson}(\lambda)$, for all $k \in \mathbf{N}_0$, we have

$$P(X_n = k) \to P(Y = k) \quad \text{as } n \to \infty.$$

Chapter 9 Conditional expectations and discrete martingales

9.A Conditional expectations

9.1 Definition. Let $\mathrm{E}|X| < \infty$, and $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. Define the *conditional expectation* of X given $\mathcal{G}$ to be the random variable Y satisfying

(a) Y is $\mathcal{G}$-measurable;

(b) $\mathrm{E}(Y\mathbf{1}_G) = \mathrm{E}(X\mathbf{1}_G)$ for all $G \in \mathcal{G}$.

This Y is denoted by $\mathrm{E}(X \mid \mathcal{G})$.

We first show that the above definition makes sense from a purely measure-theoretic point of view, and is unique a.s. Notice that the function $\nu\colon \mathcal{G} \to \mathbf{R}$ given by

$$\nu(G) = \mathrm{E}(X\mathbf{1}_G) = \int_G X\, dP \tag{9.2}$$

is a signed measure, and $\nu \ll P|_{\mathcal{G}}$. Therefore by the Radon–Nikodym theorem for a signed measure and a finite positive measure, there exists a random variable Y, unique in $L^1(\Omega, \mathcal{G}, P|_{\mathcal{G}})$, such that

$$\nu(G) = \int_G Y\, dP = \mathrm{E}(Y\mathbf{1}_G)$$

for all $G \in \mathcal{G}$. *Be aware that conditional expectations are unique up to measure zero.*

9.3 Definition. Define the *conditional probability* of $A \in \mathcal{F}$ given a sub-σ-field $\mathcal{G}$ of $\mathcal{F}$ to be $\mathrm{E}(\mathbf{1}_A \mid \mathcal{G})$, which we denote by $P(A \mid \mathcal{G})$.

Our new definitions of conditional expectation and conditional probability are very abstract, and particularly distinct from the undergraduate version, and the following example is almost included in all textbooks, which explains how our new definitions generalizes the old definitions.

9.4 Example. Let $\Omega_1, \Omega_2, \ldots$ be a countable partition of the sample space Ω, where each Ω_n has strictly positive measure. In an undergraduate class we would define

$$\mathrm{E}(X \mid \Omega_n) = \frac{\mathrm{E}(X; \Omega_n)}{P(\Omega_n)}$$

for any n. Now define $\mathcal{G} = \sigma(\{\Omega_n\}_{n=1}^{\infty})$. It is easy to see that

$$\int_{\Omega_n} \frac{\mathrm{E}(X; \Omega_n)}{P(\Omega_n)}\, dP = \int_{\Omega_n} X\, dP. \tag{9.5}$$

We claim that $\mathrm{E}(X \mid \mathcal{G})$ is given by

$$Y = \frac{\mathrm{E}(X;\Omega_n)}{P(\Omega_n)} \text{ on each } \Omega_n,$$

and hence coincides with our undergraduate definition.

First the candidate Y is $\mathcal{G}$-measurable since it is a constant on each Ω_n. Also since $\{\Omega_n\}$ is a partition of Ω and generates $\mathcal{G}$, equation (9.5) immediately implies that

$$\int_G Y\,dP = \int_G X\,dP$$

for all $G \in \mathcal{G}$. This finishes the proof.

Now we look at condition probability. Set $X = \mathbf{1}_A$, and we have

$$\begin{aligned} P(A \mid \mathcal{G}) &= \mathrm{E}(\mathbf{1}_A \mid \mathcal{G}) \\ &= \frac{\mathrm{E}(\mathbf{1}_A \mathbf{1}_{\Omega_n})}{P(\Omega_n)} \text{ on each } \Omega_n \\ &= \frac{P(A \cap \Omega_n)}{P(\Omega_n)} \text{ on each } \Omega_n, \end{aligned}$$

which was our undergraduate definition of conditional probability $P(A \mid \Omega_n)$.

9.6 Fact (characteristic property). Let all $X \in \mathcal{F}$ and $Z \in \mathcal{G}$ satisfying $\mathrm{E}|X| < \infty$ and $\mathrm{E}|XZ| < \infty$, we have

$$\mathrm{E}(\mathrm{E}(X \mid \mathcal{G})Z) = \mathrm{E}(XZ).$$

This property characterizes the conditional expectation $\mathrm{E}(X \mid \mathcal{G})$.

Proof. Left as an exercise, using the standard limiting argument. □

9.7 Proposition. Let $X, Y \in L^1(\Omega, \mathcal{F}, P)$.

(a) For X that is $\mathcal{G}$-measurable, $\mathrm{E}(X \mid \mathcal{G}) = X$.

(b) For X and $\mathcal{G}$ that are independent, $\mathrm{E}(X \mid \mathcal{G}) = \mathrm{E}X$.

(c) Linearity: $\mathrm{E}(aX + Y \mid \mathcal{G}) = a\,\mathrm{E}(X \mid \mathcal{G}) + \mathrm{E}(Y \mid \mathcal{G})$.

(d) Monotonicity: if $X \geq Y$ a.s., then $\mathrm{E}(X \mid \mathcal{G}) \geq \mathrm{E}(Y \mid \mathcal{G})$.

(e) Contractivity (in L^1): $\big|\mathrm{E}(X \mid \mathcal{G})\big| \leq \mathrm{E}\big(|X| \;\big|\; \mathcal{G}\big)$, and taking expectation on both sides gives $\mathrm{E}\big(\big|\mathrm{E}(X \mid \mathcal{G})\big|\big) \leq \mathrm{E}|X|$.

9.8 Conditional Jensen's inequality. Let $\varphi\colon \mathbf{R} \to \mathbf{R}$ be convex, and X and $\varphi(X)$ be both integrable, then

$$\varphi\big(\mathrm{E}(X \mid \mathcal{G})\big) \leq \mathrm{E}(\varphi(X) \mid \mathcal{G}).$$

9.9 Corollary (Contraction property). The conditional expectation $\mathrm{E}(\cdot \mid \mathcal{G})$ is a 1-Lipschitz linear operator on any L^p ($1 \leq p < \infty$): for $X \in L^p(\Omega, \mathcal{F}, P)$, $\big|\mathrm{E}(X^p \mid \mathcal{G})\big| \leq \mathrm{E}\big(|X|^p \;\big|\; \mathcal{G}\big)$, and taking expectations on both sides gives

$$\mathrm{E}\big(\big|\mathrm{E}(X \mid \mathcal{G})\big|^p\big) \leq \mathrm{E}|X|^p.$$

In particular, this implies that if $X_n \to X$ in L^p, then $\mathrm{E}(X_n \mid \mathcal{G}) \to \mathrm{E}(X \mid \mathcal{G})$ in L^p.

9.10 Theorem (alternative Hilbert space definition). Let $X \in L^2(\mathcal{F})$, which is a Hilbert space. Then $\mathrm{E}(X \mid \mathcal{G})$ is exactly the projection to the closed subspace $L^2(\mathcal{G})$. Furthermore, this projection linear operator $\pi\colon L^2(\mathcal{F}) \to L^2(\mathcal{G})$ can be uniquely extended to a bounded linear operator $\Pi\colon L^1(\mathcal{F}) \to L^1(\mathcal{G})$, which is exactly the conditional expectation defined by Radon–Nikodym in Definition 9.1.

Proof. The projection theorem says that it suffices to show that for all $Y \in L^2(\mathcal{G})$,

$$\mathrm{E}(\mathrm{E}(X \mid \mathcal{G})Y) = \mathrm{E}(XY).$$

This is true by Fact 9.6 and $\mathrm{E}(X \mid \mathcal{G}) \in L^2(\mathcal{G})$, which follows from Corollary 9.9.

To extend the linear operator π to a larger domain $L^1(\mathcal{F})$, recall that $L^2(\mathcal{F})$ is dense when considered as a metric subspace of $L^1(\mathcal{F})$, and $L^1(\mathcal{G})$ is complete. Now consider π as a function from $(L^2(\mathcal{F}), \|\cdot\|_1)$ to $(L^1(\mathcal{G}), \|\cdot\|_1)$. We claim this π is bounded, in particular 1-Lipschitz. To see this, it suffices to verify that

$$\mathrm{E}\big|\mathrm{E}(X \mid \mathcal{G})\big| \le \mathrm{E}|X|$$

for all $X \in L^2(\mathcal{F})$. Now let $A = \mathrm{E}(X \mid \mathcal{G}) \ge 0$, then

$$\begin{aligned}\mathrm{E}\big|\mathrm{E}(X \mid \mathcal{G})\big| &= \mathrm{E}\big(\mathrm{E}(X \mid \mathcal{G})\mathbf{1}_A\big) - \mathrm{E}\big(\mathrm{E}(X \mid \mathcal{G})\mathbf{1}_{A^c}\big)\\ &= \mathrm{E}(X\mathbf{1}_A) - \mathrm{E}(X\mathbf{1}_{A^c}) \le \mathrm{E}|X|.\end{aligned}$$

With all these information, by Theorem A.22 we have a continuous linear operator $\Pi\colon L^1(\mathcal{F}) \to L^1(\mathcal{G})$, and by the uniqueness of the extension, Π should exactly be the conditional expectation $\mathrm{E}(\cdot \mid \mathcal{G})$ defined previously. □

L^2-contractivity follows from Hilbert subspace projection reduces norm

9.11 Tower property. For $\mathcal{G}_1 \subseteq \mathcal{G}_2$, we have

$$\mathrm{E}(\mathrm{E}(X \mid \mathcal{G}_1) \mid \mathcal{G}_2) = \mathrm{E}(X \mid \mathcal{G}_1) = \mathrm{E}(\mathrm{E}(X \mid \mathcal{G}_2) \mid \mathcal{G}_1).$$

This means that the iterated conditioning is ultimately conditioning on the smallest σ-field. Note in particular, we have

$$\mathrm{E}(\mathrm{E}(X \mid \mathcal{G})) = \mathrm{E}X.$$

9.12 Proposition. For $Y \in \mathcal{G}$, we have $\mathrm{E}(XY \mid \mathcal{G}) = Y\,\mathrm{E}(X \mid \mathcal{G})$.

9.13 Proposition. For two sub-σ-fields $\mathcal{G}_1$ and $\mathcal{G}_2$ of $\mathcal{F}$, then the following three are equivalent:

(a) $\mathcal{G}_1$ and $\mathcal{G}_2$ are independent;

(b) $\mathrm{E}(X \mid \mathcal{G}_1) = \mathrm{E}X$ for every $X \in L^+(\mathcal{G}_2)$ or $L^1(\mathcal{G}_2)$;

(c) $\mathrm{E}(\mathbf{1}_{G_2} \mid \mathcal{G}_1) = P(G_2)$ for every $G_2 \in \mathcal{G}_2$.

In particular, let X and Y be two random variables. Consider $\mathcal{G}_1 = \sigma(X)$ and $\mathcal{G}_2 = \sigma(Y)$. Then X and Y are independent if and only if

$$\mathrm{E}[f(X) \mid Y] = \mathrm{E}f(X)$$

for all f such that $\mathrm{E}|f(X)| < \infty$.

9.14 Proposition. Let $X\colon (\Omega,\mathcal{F})\to(T,\mathcal{T})$ and $Y\colon(\Omega,\mathcal{F})\to(S,\mathcal{S})$, and say $\mathcal{G}$ is a sub-σ-field of $\mathcal{F}$. If X is $\mathcal{G}$-measurable and Y is independent of $\mathcal{G}$, then for any $f\colon(T\times S,\mathcal{T}\otimes\mathcal{S})\to(\mathbf{R},\mathcal{B})$ such that $\mathrm{E}|f(X,Y)|<\infty$, we have

$$\mathrm{E}[f(X,Y)\mid\mathcal{G}]=h(X),\ \text{where } h(x)=\mathrm{E}f(x,Y).$$

In particular, when X and Y are independent, we have

$$\mathrm{E}[f(X,Y)\mid X]=h(X).$$

9.15 Definition. Let X be nonnegative $\mathcal{F}$-measurable, then we define its *conditional expectation* given $\mathcal{G}$ to be

$$\mathrm{E}(X\mid\mathcal{G})=\lim_{n\to\infty}\mathrm{E}(X\wedge n\mid\mathcal{G}).$$

Radon–Nikodym theorem fails to help us

9.16 de Finetti's theorem. For a sequence of exchangeable random variables, conditioning on the exchangeabale σ-field $\mathcal{E}$, $X_1,X_2,\dots$ are i.i.d. More precisely, we can show that for any bounded measurable functions f_j's, it holds that

$$\mathrm{E}\Big(\prod_{j=1}^{n}X_j\;\Big|\;\mathcal{E}\Big)=\prod_{j=1}^{n}\mathrm{E}(X_j\mid\mathcal{E}).$$

9.B Conditional distributions and transition kernels

9.17 Definition. Let $(\Omega,\mathcal{F})$ and $(S,\mathcal{S})$ be two measurable space. A *random probability measure* is $\nu\colon\Omega\times\mathcal{S}\to[0,1]$ such that

(a) for (P-a.e.) $\omega\in\Omega$, the function $\nu(\omega,\cdot)$ is a probability measure on $(S,\mathcal{S})$;

(b) for each $A\in\mathcal{S}$, the function $\omega\mapsto\nu(\omega,A)$ is $\mathcal{F}$-measurable.

Let $Y\colon(\Omega,\mathcal{F})\to(S,\mathcal{S})$, and $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. A *regular conditional distribution* of Y given $\mathcal{G}$ is a random probability measure $\nu\colon\Omega\times\mathcal{S}\to[0,1]$ such that satisfies:

(c) for each $A\in\mathcal{S}$, the function $\omega\mapsto\nu(\omega,A)$ is a version of $P(Y\in A\mid\mathcal{G})$.

Recall that the function $P(Y\in A\mid\mathcal{G})$ is unique only P-a.e., hence a version of $P(Y\in A\mid\mathcal{G})$ means a function defined at each $\omega\in\Omega$. Note that condition (c) may be replaced by the following: for each f such that $\mathrm{E}|f(Y)|<\infty$, we have for P-a.e. ω,

$$\mathrm{E}[f(Y)\mid\mathcal{G}]=\int f(y)\,\nu(\omega,dy).$$

This follows by observing that

$$\mathrm{E}\big(\mathbf{1}_{\{Y\in A\}}\;\big|\;\mathcal{G}\big)(\omega)=P(Y\in A\mid\mathcal{G})(\omega)=\int\mathbf{1}_A(y)\,\nu(\omega,dy),$$

and then employing the standard approximation argument.

Most often we take $\mathcal{G}=\sigma(X)$, and then our ν defined above is the *regular conditional distribution* of Y given X. In this case however, our notation above turns out to be awkward, since we want to make the role of ω implicit, but the role of $x=X(\omega)$ explicit. To accomplish this the following general definition is introduced.

9.18 Definition. Given two measurable spaces $(T, \mathcal{T})$ and $(S, \mathcal{S})$, the *stochastic/transition kernel* from T to S is a function $\kappa : T \times \mathcal{S} \to [0, 1]$ that satisfies:

(a) for each $x \in T$, the function $\kappa(x, \cdot)$ is a probability measure on $(S, \mathcal{S})$;

(b) for each $A \in \mathcal{S}$, the function $x \mapsto \kappa(x, A)$ is $\mathcal{T}$-measurable.

Note that a transition kernel ν from Ω to S is just a random measure. Now rephrasing Definition 9.17, the regular conditional distribution of Y given $\mathcal{G}$ is a transition kernel ν from Ω to S such that

$$\text{the function } \omega \mapsto \nu(\omega, A) \text{ is a version of } P(Y \in A \mid \mathcal{G}),$$

for each $A \in \mathcal{S}$.

If we consider two random variables $X \colon (\Omega, \mathcal{F}) \to (T, \mathcal{T})$ and $Y \colon (\Omega, \mathcal{F}) \to (S, \mathcal{S})$, respectively, then the regular conditional distribution of Y given X is $\kappa \circ X$, where $\kappa \colon S \times \mathcal{T} \to [0, 1]$ is a transition kernel such that

$$\kappa(X(\omega), A) \text{ is a version of } P(Y \in A \mid X)(\omega). \tag{9.19}$$

for each $A \in \mathcal{S}$.

It turns out quite surprising that the notion of a transition kernel fully generalizes Definition 9.17. Indeed it is clear that random measures are just transition kernels. To recover the definition of regular condition probability in Definition 9.17, we may set X to be the identity map from $(\Omega, \mathcal{G})$ to itself.

These are all formal definitions. When doing concrete computations,

It turns out that regular condition distributions do not always exist.

However, if we assume that Y takes value in a standard Borel space $(S, \mathcal{S})$, then the regular conditional probability of Y given X exists.

9.20 Theorem. Let ρ be a probability measure on the product space $(T \times S, \mathcal{T} \otimes \mathcal{S})$, where $(S, \mathcal{S})$ is a standard Borel space. Then $\rho = \mu \otimes \kappa$, where $\mu = \rho(\cdot \times S)$ and κ is a transition kernel from T to S.

(a) For two discrete random variables X and Y, we want

$$\kappa(x, A) = \begin{cases} P(Y \in A \mid X = x) & \text{if } P(X = x) > 0, \\ \delta_{y_0}(A) & \text{if } P(X = x) = 0. \end{cases}$$

(b) For two continuous random variables $X \in \mathbf{R}^m$ and $Y \in \mathbf{R}^n$, with joint density $f(x, y)$. We know the marginal density of X is given by

(c) Gaussian

[Kal21, Theorem 8.5]

9.21 Disintegration of random variables. Consider two random variables $X \colon (\Omega, \mathcal{G}) \to (T, \mathcal{T})$ and $Y \colon (\Omega, \mathcal{F}) \to (S, \mathcal{S})$, where $(S, \mathcal{S})$ is a standard Borel space. Then the joint distribution ρ of (X, Y) is equal to the product measure $\mu \times \kappa$, where μ is the marginal distribution of X and κ is a transition kernel that satisfies (9.19).

It follows that for any measurable $f \geq 0$ or $\mathrm{E}|f(X, Y)| < \infty$, we have

$$\mathrm{E}[f(X, Y) \mid X] = \int f(X, y)\, \kappa(X, dy).$$

$$\mathrm{E}[f(X, Y) \mid \mathcal{G}] = \int f(X, y)\, \nu(dy).$$

$$\mathrm{E} f(X, Y) = \mathrm{E} \int f(X, y)\, \kappa(X, dy).$$

9.C Stopping times

A *discrete filtration* on a given a probability space $(\Omega, \mathcal{F}, P)$ is an expanding sequence of sub-σ-fields $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$ of $\mathcal{F}$. Given a sequence of random variables $X_0, X_1, \ldots$ we define its *natural filtration* by setting $\mathcal{F}_n = \sigma(X_0, X_1, \ldots)$ for all $n \in \mathbf{N}_0$.

9.22 Exercise. Let S and T be two stopping times. Prove the following claims.

(a) $S \wedge T$ and $S \vee T$ are both stopping times.

(b) If $S \leq T$, then $\mathcal{F}_S \subseteq \mathcal{F}_T$.

(c) $\mathcal{F}_{S \wedge T} = \mathcal{F}_S \cap \mathcal{F}_T$.

9.23 Wald's equations.

(a) Let $X_1, X_2, \ldots$ be i.i.d. L^1 random variables. If T is a stopping time with $\mathrm{E}T < \infty$, then $\mathrm{E}(X_1 + \ldots + X_T) = \mathrm{E}X_1 \, \mathrm{E}T$.

(b) Let $X_1, X_2, \ldots$ be i.i.d. mean zero L^2 random variables. If T is a stopping time with $\mathrm{E}T < \infty$, then $\mathrm{E}(X_1 + \ldots + X_T)^2 = \mathrm{E}(X_1)^2 \, \mathrm{E}T$.

9.D Martingales in discrete time

Given a filtration $\{\mathcal{F}_n\}$, a *discrete martingale* is a sequence of L^1 random variables X_n, adapted to $\mathcal{F}_n$, such that

$$\mathrm{E}(X_{n+1} \mid \mathcal{F}_n) = X_n.$$

9.24 Example. Here are some of the most important examples of martingales coming up in applications. Let $\{X_j\}_j$ be a sequence of random variables, and $S_n = \sum_{j=1}^n X_j$. Let $\mathcal{F}_n$ be the natural filtration with respect to X_n.

(a) Linear martingales: if the sequence $\mathrm{E}X_j = 0$ for all j, then S_n is a martingale.

(b) Quadratic martingales: if $\mathrm{E}X_j = 0$ and $\mathrm{E}X_j^2 = \sigma^2$ for all j, we have $S_n^2 - n\sigma^2$ as a martingale.

(c) Exponential martingales: say an unrelated sequence Y_j's are nonnegative, i.i.d., with $\mathrm{E}Y_j = 1$, then $M_n = \prod_{j=1}^n Y_j$ is a martingale.

It is clear that $\frac{\exp(tX_j)}{M_{X_j}(t)}$ is a candidate for our Y_j.

9.25 Exercise. Let $\{X_n\}$ be a martingale (resp. supermartingale, submartingale). For every $0 \leq n \leq m$, $\mathrm{E}(X_m \mid \mathcal{F}_n) = X_n$ (resp. $\leq$, $\geq$).

9.26 Proposition (Martingale transformations under convex functions). Let $\{X_n\}$ be adapted. For a convex $\varphi\colon \mathbf{R} \to \mathbf{R}$ such that $\mathrm{E}|\varphi(X_n)| < \infty$, we have

(a) if $\{X_n\}$ is a martingale, then $\{\varphi(X_n)\}$ becomes a submartingale.

(b) if $\{X_n\}$ is a submartingale (resp. supermartingale), and φ is in addition increasing (resp. decreasing), then $\{\varphi(X_n)\}$ remains a submartingale (resp. supermartingale)

9.27 Definition. A sequence of random variables $\{H_n\}$ is a predictable sequence if the sequence is bounded and each H_{n+1} is $\mathcal{F}_n$ measurable.

The *discrete stochastic integral* from time 0 to $n \in \mathbf{N}_0$ is defined by

$$(H \cdot X)_n = H_1(X_1 - X_0) + H_2(X_2 - X_1) + \ldots + H_n(X_n - X_{n-1}),$$

for $n \geq 1$, and $(H \cdot X)_0 = 0$.

It is useful to see that $(H \cdot -X)_n = -(H \cdot X)_n$.

9.28 Proposition.

(a) If $\{X_n\}$ is a martingale, then $\{(H \cdot X)_n\}$ is a martingale.

(b) If $\{X_n\}_n$ is a submartingale (resp. supermartingale), and $H_n \geq 0$ for all n, then $\{(H \cdot X)_n\}$ is a submartingale (resp. supermartingale).

9.29 Optional stopping theorem, basic version. Let $\{X_n\}$ be a martingale (resp. supermartingale), and T be a stopping time, both with respect to $\{\mathcal{F}_n\}$, then

(a) the stopped process $\{X_{n\wedge T}\}$ remains a martingale (resp. supermartingale);

(b) moreover, if $T \leq M$ a.s. for some $M < \infty$ (bounded stopping time), then $\mathrm{E}X_T = \mathrm{E}X_0$ (resp. $\leq \mathrm{E}X_0$).

9.30 Doob's decomposition. Any adapted integrable process $\{X_n\}$ can be uniquely decomposed by $X_n = M_n + A_n$, where $\{M_n\}$ is a martingale and $\{A_n\}$ is a predictable sequence starting from $A_0 = 0$. This is known as the *Doob decomposition*.

Furthermore, an adapted integrable process $\{X_n\}$ is a submartingale (resp. supermartingale) if and only if it has a Doob decomposition with an increasing (resp. decreasing) predictable sequence.

Proof. The proof is quite elementary and may be left as an exercise. We want $X_n = M_n + A_n$, and conditioning both sides on $\mathcal{F}_{n-1}$ gives

$$\mathrm{E}(X_n \mid \mathcal{F}_{n-1}) = M_{n-1} + A_n = X_{n-1} - A_{n-1} + A_n.$$

This gives for all $n \in \mathbf{N}$,

$$A_n - A_{n-1} = \mathrm{E}(X_n \mid \mathcal{F}_{n-1}) - X_{n-1}. \tag{9.31}$$

Set $A_0 = 0$, and by repeatedly applying the above identity we get

$$A_n = \sum_{k=1}^{n} \mathrm{E}(X_k - X_{k-1} \mid \mathcal{F}_{k-1})$$

that is $\mathcal{F}_{k-1}$ measurable. We have shown that the decomposition, if exists, must be unique.

It remains to check that M_n is indeed a martingale:

$$\begin{aligned} \mathrm{E}(M_n \mid \mathcal{F}_{n-1}) &= \mathrm{E}(X_n \mid \mathcal{F}_{n-1}) - A_n \\ &= X_{n-1} - A_{n-1} = M_n, \end{aligned}$$

where we used (9.31).

The furthermore part follows immediately from (9.31). □

9.32 Martingale convergence theorem. Say $\{X_n\}$ is a submartingale bounded in L^1, then the sequence X_n converges a.s. to some $X_\infty \in L^1$.

supermartingale/martingale

It is clear that the limit X_∞ cannot be explicitly computed, and we have to employ some clever trick to prove the existence of the limit.

9.33 Lemma. A sequence of real numbers $x = \{x_n\}$ converges if and only for any two rationals $a < b$, we have $U_\infty([a,b], x) < \infty$.

9.34 Doob's upcrossing inequality. Let $X = \{X_n\}$ be a submartingale. Then for every $a < b$ and every $n \in \mathbf{N}$

$$(b-a)\,\mathrm{E}U_n([a,b], X) \le \mathrm{E}(X_n - a)^+ - \mathrm{E}(X_0 - a)^+.$$

Proof of the martingale convergence theorem. □

Optional stopping theorem, basic version may fail when the stopping time T is unbounded.

The same example also show that the martingale convergence theorem does not hold in the L^1 sense.

9.E Uniformly integrable martingales

9.35 Proposition. The collection $\{\mathrm{E}(X \mid \mathcal{G}) : \mathcal{G} \text{ is a sub-}\sigma\text{-field of } \mathcal{F}\}$ is uniformly integrable.

9.36 Theorem (characterizations of uniformly integrable martingales). For an $\mathcal{F}_n$-adpated martingale X_n, the following are equivalent.

(a) $\{X_n\}$ is uniformly integrable;

(b) X_n converges a.s. and in L^1;

(c) X_n converges in L^1;

(d) there exists an integrable X such that $X_n = \mathrm{E}(X \mid \mathcal{F}_n)$.

Proof. (d) $\Longrightarrow$ (a) follows from Proposition 9.35. (b) $\Longrightarrow$ (c) is trivial. (a) $\Longrightarrow$ (b) is true because $\{X_n\}$ is bounded, and hence we may apply the martingale convergence theorem.

(c) $\Longrightarrow$ (d) is also not difficult. Let X be the L^1 limit of X_n. Then for any $m > n$, we have $\mathrm{E}(X_m \mid \mathcal{F}_n) = X_n$. If we can show that $\mathrm{E}(X_m \mid \mathcal{F}_n) \to \mathrm{E}(X \mid \mathcal{F}_n)$ in L^1, then the proof is complete.

$$\mathrm{E}\big|\mathrm{E}(X_m \mid \mathcal{F}_n) - \mathrm{E}(X \mid \mathcal{F}_n)\big| \le \mathrm{E}\big[\mathrm{E}\big(|X_m - X| \;\big|\; \mathcal{F}_n\big)\big] \le \mathrm{E}|X_m - X|,$$

which goes to 0 as $m \to \infty$, as desired. □

9.37 Levy's zero–one law. Let $\{\mathcal{F}_n\}$ is a filtration with $\mathcal{F}_\infty = \sigma(\cup_n \mathcal{F}_n)$, which we write as $\mathcal{F}_n \uparrow \mathcal{F}_\infty$. Suppose $\mathrm{E}|X| < \infty$, then

$$\mathrm{E}(X \mid \mathcal{F}_n) \to \mathrm{E}(X \mid \mathcal{F}_\infty) \quad \text{a.s. and in } L^1.$$

In particular, for $A \in \mathcal{F}_\infty$, we have

$$\mathrm{E}(\mathbf{1}_A \mid \mathcal{F}_n) \to \mathbf{1}_A \quad \text{a.s. and in } L^1.$$

9.38 DCT for conditional expectations. Suppose $X_n \to X$ a.s. and for all $n \in \mathbf{N}_0$, $|X_n| \le Y$ for some $Y \in L^1$. Given that the σ-fields $\mathcal{F}_n \uparrow \mathcal{F}_\infty$, then

$$\mathrm{E}(X_n \mid \mathcal{F}_n) \to \mathrm{E}(X \mid \mathcal{F}_\infty).$$

9.F Backward martingales and their applications

9.39 Definition. A *backward filtration* is a $\mathbf{Z}^{\leq 0}$-indexed filtration, i.e., a sequence of sub-σ-fields of $\mathcal{F}$

$$\cdots \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_0.$$

Let $\mathcal{F}_\infty = \bigcap_{n=-\infty}^{0} \mathcal{F}_n$, which we know is again a sub-σ-field of $\mathcal{F}$.

9.40 Backward martingale convergence theorem. The sequence $X_n \to X_{-\infty}$ a.s. and in L^1.

9.41 Example (another proof of the L^1 strong law).

9.42 Example (another proof of the Hewitt–Savage zero–one law).

9.G L^p convergence of martingales

9.43 Doob's maximal inequality. Let $\{X_n\}$ be a submartingale, then for every $a > 0$, we have

$$a P\big(\max_{0\leq k\leq n} X_k \geq a\big) \leq \mathrm{E}\Big(X_n \mathbf{1}\big\{\max_{0\leq k\leq n} X_n \geq a\big\}\Big) \leq \mathrm{E}X_n^+.$$

If $\{Y_n\}$ is a supermartingale, then for every $a > 0$, we have

$$a P\big(\max_{0\leq k\leq n} Y_k \geq a\big) \leq \mathrm{E}Y_0 + \mathrm{E}Y_n^-.$$

Combining the two cases above, we get for a submartingale or supermartingale $\{X_k\}$, it holds that

$$a P\big(\max_{0\leq k\leq n} |X_k| \geq a\big) \leq \mathrm{E}|X_0| + 2\,\mathrm{E}|X_n|.$$

The technique of introducing an appropriate stopping time

Integrating the two sides of the first inequality, we can obtain an L^p moment bound on $\mathrm{E}(\max X_k)$ for $1 < p < \infty$.

9.44 Doob's L^p inequality. Let $1 < p < \infty$ and $\{X_n\}$ be a nonnegative submartingale. For each $n \in \mathbf{N}_0$, we have

$$a^p P\big(\max_{0\leq k\leq n} X_k \geq a\big) \leq \mathrm{E}\big(\max_{0\leq k\leq n} X_k\big)^p \leq \left(\frac{p}{p-1}\right)^p \mathrm{E}(X_n)^p.$$

Therefore if $\{Z_n\}$ is a martingale, then $\{|Z_n|\}$ is a nonnegative submartingale. Therefore we have

$$a^p P\big(\max_{0\leq k\leq n} |Z_k| \geq a\big) \leq \mathrm{E}\big(\max_{0\leq k\leq n} |Z_k|\big)^p \leq \left(\frac{p}{p-1}\right)^p \mathrm{E}|Z_n|^p.$$

We say $\{X_n\}$ is a *square integrable martingale* if $\{X_n\}$ is a martingale, and each $X_n \in L^2(P)$. When $\{X_n\} \subseteq L^2$ is a martingale, then we have by Doob's L^p inequality ($p = 2$) that

$$\mathrm{E}\max_{0\leq k\leq n} X_k^2 \leq 4\,\mathrm{E}X_n^2. \tag{9.45}$$

We now show that a uniform control on $\{X_n\}_{n\in\mathbf{N}}$ can be obtained. Doob's decomposition tells us that we can decompose the submartingale X_n^2 into $M_n + A_n$, where $\{M_n\}$ is a martingale, and $\{A_n\}$ is an increasing predictable sequence given by

$$\begin{aligned} A_n &= \sum_{k=1}^{n} \mathrm{E}(X_k^2 - X_{k-1}^2 \mid \mathcal{F}_{k-1}) \\ &= \sum_{k=1}^{n} \mathrm{E}(X_k^2 - 2X_k X_{k-1} + X_{k-1}^2 \mid \mathcal{F}_{k-1}) \\ &= \sum_{k=1}^{n} \mathrm{E}[(X_k - X_{k-1})^2 \mid \mathcal{F}_{k-1}], \end{aligned}$$

where we have used $\mathrm{E}(X_k \mid \mathcal{F}_{k-1}) = X_{k-1}$ in the second equality. This increasing sequence has a special name, called the *quadratic variation* of the square integrable martingale $\{X_n\}$, which we denote by $\{\langle X\rangle_n\}$.

Note $\mathrm{E}X_n^2 = \mathrm{E}M_n + \mathrm{E}\langle X\rangle_n = \mathrm{E}X_0^2 + \mathrm{E}\langle X\rangle_n$, and we may plug this into (9.45). By the monotone convergence theorem, we can therefore conclude

9.46 Proposition. For martingale $\{X_n\} \subseteq L^2$, we have

$$\mathrm{E}\sup_n X_n^2 \le 4\,\mathrm{E}\langle X\rangle_\infty + 4\,\mathrm{E}X_0^2,$$

where $\langle X\rangle_\infty = \lim_n \langle X\rangle_n$, which is possibly infinite.

9.47 L^p convergence theorem for martingales. Let $1 < p < \infty$, and $\{X_n\}$ be a uniformly L^p-bounded martingale. Then X_n converges a.s. and in L^p to some X_∞ satisfying

$$\mathrm{E}|X_\infty|^p = \sup_n \mathrm{E}|X_n|^p.$$

Meanwhile

$$\mathrm{E}\big(\sup_n |X_n|\big)^p \le \left(\frac{p}{p-1}\right)^p \mathrm{E}|X_\infty|^p.$$

9.48 Theorem (convergence of L^2 summable random series). Let $\{X_n\}$ be a sequence of independent mean zero L^2 random variables, then the following are equivalent:

(a) $\sum_{n=1}^{\infty} \mathrm{E}X_n^2 < \infty$;
(b) $\sum_{n=1}^{\infty} X_n^2$ converges a.s. and in L^2;
(c) $\sum_{n=1}^{\infty} X_n^2$ converges in L^2.

9.H Martingales of bounded increments

9.49 Theorem (convergence behavior). For a martingale $\{X_n\}$ with $\sup_n |X_{n+1} - X_n| < \infty$, we have almost surely either $\lim_n X_n$ exists and is finite, or $\limsup_n X_n = +\infty$ and $\liminf_n X_n = -\infty$.

9.50 Fact. For a martingale $\{X_n\}$, we have for any Borel measurable function f that

$$\mathrm{E}\big[(X_{n+1} - X_n) f(X_0, X_1, \ldots, X_n)\big] = 0$$

by the tower property. In particular, we have

$$\mathrm{E}\big[(X_{n+1} - X_n)(X_{m+1} - X_m)\big] = 0$$

for any $n > m$, i.e., martingale differences are uncorrelated.

9.51 Azuma–Hoeffding inequality. Let $\{X_n\}$ be a supermartingale, and $\{A_n\}$ and $\{B_n\}$ are predictable with respect to the filtration $\{\mathcal{F}_n\}$, such that

$$A_n \le X_n - X_{n-1} \le B_n.$$

If for all A_n and B_n we have some positive constant c_n such that $B_n - A_n \le c_n$, then we have

$$P(X_n - X_0 \ge t) \le \exp\left(-\frac{2t^2}{\sum_{k=1}^n c_k^2}\right). \tag{9.52}$$

If the $\{X_n\}$ above is a submartingale instead, then we get

$$P(X_0 - X_n \ge t) \le \exp\left(-\frac{2t^2}{\sum_{k=1}^n c_k^2}\right).$$

Hence by a simple union bound, we get for a martingale $\{X_n\}$ with the described conditions, it holds that

$$P(|X_n - X_0| \ge t) \le 2\exp\left(-\frac{2t^2}{\sum_{k=1}^n c_k^2}\right).$$

Proof. We first show (9.52) when $\{X_n\}$ is only a martingale, and at the end we extend the inequality to the supermartingale case by invoking Doob's decomposition.

The Chernoff method is expected, just by observation of the inequality. For any $\lambda \in \mathbf{R}$, we have

$$P(X_n - X_0 \ge t) \le \frac{\mathrm{E}\exp\big(\lambda(X_n - X_0)\big)}{e^{\lambda t}}. \tag{9.53}$$

Now focus on $\mathrm{E}\exp\big(\lambda(X_n - X_0)\big)$, which is equal to

$$\begin{aligned}
\mathrm{E}\exp\left(\lambda\sum_{k=1}^n X_k - X_{k-1}\right) &= \mathrm{E}\left[\exp\lambda(X_n - X_{n-1}) \cdot \exp\left(\lambda\sum_{k=1}^{n-1} X_k - X_{k-1}\right)\right] \\
&= \mathrm{E}\left[\mathrm{E}\left(\exp\lambda(X_n - X_{n-1}) \cdot \exp\left(\lambda\sum_{k=1}^{n-1} X_k - X_{k-1}\right) \,\middle|\, \mathcal{F}_{n-1}\right)\right] \\
&= \mathrm{E}\left[\exp\left(\lambda\sum_{k=1}^{n-1} X_k - X_{k-1}\right)\mathrm{E}\left(\exp\big(\lambda(X_n - X_{n-1})\big) \,\middle|\, \mathcal{F}_{n-1}\right)\right].
\end{aligned} \tag{9.54}$$

Since $\{X_n\}$ is an $\{\mathcal{F}_n\}$-adapted martingale, for the difference sequence $Y_n = X_n - X_{n-1}$, we should have

$$\mathrm{E}(Y_n \mid \mathcal{F}_{n-1}) = 0.$$

Also by assumption for constant $c_n > 0$ and random variable A_n that is $\mathcal{F}_{n-1}$-measurable, we have

$$A_n \le Y_n \le A_n + c_n,$$

therefore by the conditional version of Hoeffding's lemma, line (9.54) is

$$\leq \mathrm{E}\Big[\exp\Big(\lambda \sum_{k=1}^{n-1} X_k - X_{k-1}\Big)\Big] \exp\Big(\frac{\lambda^2 c_n^2}{8}\Big).$$

We may repeat the procedure above by first conditioning and then applying Hoeffding's lemma, and obtain in the end that

$$\begin{aligned}\mathrm{E}\exp\Big(\lambda \sum_{k=1}^{n} X_k - X_{k-1}\Big) &\leq \prod_{k=1}^{n} \exp\Big(\frac{\lambda^2 c_k^2}{8}\Big) \\ &= \exp\Big(\frac{\lambda^2 \sum_{k=1}^n c_k^2}{8}\Big).\end{aligned}$$

Going back to (9.53), we have

$$P(X_n - X_0 \geq t) \leq \exp\Big(\frac{\lambda^2 \sum_{k=1}^n c_k^2}{8} - \lambda t\Big).$$

The quadratic expression $\frac{\sum_{k=1}^n c_k^2}{8}\lambda^2 - t\lambda$ in λ has minimum value

$$-\frac{t^2}{4 \cdot \frac{\sum_{k=1}^n c_k^2}{8}} = -\frac{2t^2}{\sum_{k=1}^n c_k^2}.$$

when $\lambda = -\frac{-t}{2 \cdot \frac{\sum_{k=1}^n c_k^2}{8}}$. Therefore

$$P(X_n - X_0 \geq t) \leq \exp\Big(-\frac{2t^2}{\sum_{k=1}^n c_k^2}\Big),$$

finishing the proof for the martingale case.

For a supermartingale $\widehat{X}_n$, we know by Doob's decomposition that there is a (unique) decomposition $\widehat{X}_n = X_n + D_n$, where X_n is a martingale and D_n is a decreasing sequence. This implies that

$$\begin{aligned}P(\widehat{X}_n - \widehat{X}_0 \geq t) &= P(X_n - X_0 + D_n - D_0 \geq t) \\ &\leq P(X_n - X_0 \geq t),\end{aligned}$$

as $D_n - D_0 \leq 0$. The proof is now complete. □

It is clear that Hoeffding's inequality is simply a special case of the above martingale inequality. Another applicable consequence of Azuma–Hoeffding inequality is the following result about concentration of functions taking vector inputs of independent components.

9.55 McDiarmid's bounded difference inequality. Let a measurable function $g\colon \prod_{k=1}^n S_k \to \mathbf{R}$ satisfy the bounded difference property with constants $c_1, \ldots, c_n$. This means for each $k \in [n]$, we have

$$\sup_{\substack{x_1,\ldots,x_n \\ x_k' \in S_k}} |g(x_1, \ldots x_n) - g(x_1, \ldots x_{k-1}, x_k', x_{k+1}, \ldots, x_n)| \leq c_k.$$

Let $\{X_k\}_{k=1}^n$ be independent S_k-valued random variables, then

$$P\big(g(X_1,\dots,X_n) - \mathrm{E}g(X_1,\dots,X_n) \geq t\big) \leq \exp\left(\frac{-2t^2}{\sum_j c_j^2}\right).$$

(Note that by the triangle inequality, $g(\mathbf{x}) - g(\mathbf{x}') \leq \sum_{k=1}^n c_k$ for any $\mathbf{x}, \mathbf{x}' \in \prod_k S_k$, and hence the function g is bounded.)

Proof. Define the martingale $Y_k = \mathrm{E}[g(X_1,\dots,X_n) \mid \mathcal{F}_k]$ for $0 \leq k \leq n$, where $\mathcal{F}_k$ is the natural filtration with respect to $\{X_k\}$. It suffices to check the conditions of Azuma–Hoeffding inequality. Note $Y_k - Y_{k-1}$ can be expanded into

$$\mathrm{E}[g(X_1,\dots,X_k,\xi_{k+1},\dots,\xi_n) - g(X_1,\dots,X_{k-1},\xi_k,\dots,\xi_n) \mid \mathcal{F}_k], \tag{9.56}$$

where $\xi_k,\dots,\xi_n$ are copies of $X_k,\dots,X_n$ that are independent of everything else.

Now we define our natural candidate for A_k by

$$\inf_{z\in S_k} \mathrm{E}[g(X_1,\dots,X_{k-1},z,\xi_{k+1}\dots,\xi_n) - g(X_1,\dots,X_{k-1},\xi_k,\xi_{k+1}\dots,\xi_n) \mid \mathcal{F}_k],$$

and let B_k be the corresponding supremum; both are $\mathcal{F}_{k-1}$-measurable. The bounded difference condition gives that

$$\begin{aligned} &B_k - A_k \\ &= \sup_{z,w\in S_k} \mathrm{E}[g(X_1,\dots,X_{k-1},z,\xi_{k+1}\dots,\xi_n) - g(X_1,\dots,X_{k-1},w,\xi_{k+1}\dots,\xi_n) \mid \mathcal{F}_k] \\ &\leq c_k, \end{aligned}$$

and hence we have verified all the conditions for invoking Azuma–Hoeffding. □

9.I Gamblers' ruin and random walks

For a random process $\{X_t\}$ that starts $X_0 = x$, we often use $\mathbf{P}_x$ instead of P as the notation for the underlying probability measure. *There is no absolutely difference between the two in the context of this section.* The subscript x here is merely used to emphasize where the random process starts. However, it deserves attention that $\mathbf{P}_x$ is a distinct probability measure living on a different space that is induced from the usual P on $(\Omega, \mathcal{F})$. We will discuss this at a detailed level in the upcoming chapter, when discussing the canonical probability space for a Markov chain.

9.57 Theorem. Let S_n be the symmetric random walk on $\mathbf{Z}$ that starts at 0, and define $T = \min\{n : S_n \notin (-a, b)\}$, where $-a < 0 \leq b$ are integers. We have $T < \infty$ a.s., and

$$\mathbf{P}_0(S_T = -a) = \frac{b}{a+b},\ \mathbf{P}_0(S_T = b) = \frac{a}{a+b},\ \text{and}\ \mathbf{E}_0 T = ab.$$

Take $-a = -1$ and $b = N - 1 \geq 0$. For any $y \in \mathbf{N}$, define the hitting time $T_y = \min\{n : S_n = y\}$, then

$$\mathbf{P}_0(T_{-1} < T_{N-1}) = \frac{N-1}{N} \text{ and } \mathbf{P}_0(T_{-1} > T_{N-1}) = \frac{1}{N}.$$

Therefore

$$\mathbf{P}_0(T_{-1} < \infty) = \mathbf{P}_0\left(\bigcup_{N=1}^{\infty} \{T_{-1} < T_{N-1}\} \right) = 1.$$

However, $\mathbf{E}_0 T_{-1} = \infty$.

by monotone convergence

9.58 Theorem. Let $S_n = \sum_{j=1}^{\infty} \xi_j$ be the asymmetric random walk that starts from 0, where each ξ_j is i.i.d., with $P(\xi_j = 1) = p > 1/2$ and $P(\xi_j = -1) = q = 1 - p < 1/2$.

(a) First, one can verify that $\{(q/p)^{S_n}\}_n$ is a martingale.

(b) Therefore let $f(y) = (q/p)^y$. Again define the hitting time $T_z = \min\{n : S_n = z\}$. For $-a < 0 < b$, we have

$$\mathbf{P}_0(T_{-a} < T_b) = \frac{\varphi(b) - \varphi(0)}{\varphi(b) - \varphi(-a)} \text{ and } \mathbf{P}_0(T_{-a} > T_b) = \frac{\varphi(0) - \varphi(-a)}{\varphi(b) - \varphi(-a)}.$$

(c) Now we look at the two hitting times individually:

$$\mathbf{P}_0(\min_n S_n \le -a) = \mathbf{P}_0(T_{-a} < \infty) = \left(\frac{1-p}{p} \right)^a.$$

$$\mathbf{P}_0(\max_n S_n \ge b) = \mathbf{P}_0(T_b < \infty) = 1 \quad \mathbf{E}_0 T_b = \frac{b}{2p-1}.$$

This stopping time conversion is very standard

Chapter 10 Construction of random processes

10.A Independent sequences

A major theme of the previous chapters is about the asymptotic behavior of a sequence of independent random variables. However, it is not immediate that there is an appropriate probability space for us to construct such an *independent* sequence.

If the sequence or random variables is assumed to be $\mathbf{R}$-valued but not necessarily independent, then the sequence always exists on the common probability space $([0,1], \mathcal{B}_{[0,1]}, m)$, since we can use Theorem 7.7 to realize the distribution μ_n for each X_n. If we have a finite list of independent random variables $X_1, \dots, X_n$ with distributions $\mu_1, \dots, \mu_n$, then we can always take $(\mathbf{R}^n, \mathcal{B}, \mu_1 \times \cdots \times \mu_n)$ to be the probability space.

It is in fact possible to either continue with the space $([0,1], \mathcal{B}_{[0,1]}, m)$ and define an independent sequence, or prove a theorem on the existence of countable product of probability measures. We will go with the first approach, and leave the existence theorem to Appendix K.

We closely follow [LeG22] below, and summarize the main idea first.[1] The binary digits of a single Uniform$[0,1)$ is an i.i.d. sequence of Bernoulli$(1/2)$ random variables. By a clever expansion of indices, the sequence (X_n) can now be used to generate a *sequence* of i.i.d. Uniform$[0,1]$ random variables. An application of Theorem 7.7 gives us the desired construction.

On the probability space $\Omega = [0,1), \mathcal{F} = \mathcal{B}_{[0.1)}, P = m$, define

$$X_n(\omega) = \lfloor 2^n \omega \rfloor - 2\lfloor 2^{n-1} \omega \rfloor,$$

the proper binary expansion of $\omega \in [0,1)$. We claim that the $X_n(\omega)$'s form an i.i.d. Bernoulli$(1/2)$ sequence in our probability space. This is easy because for any finite subcollection,

$$P(X_1 = b_1, \dots, X_n = b_n) = 2^{-n} = \prod_{k=1}^{n} P(X_k = b_k).$$

Let $\varphi \colon \mathbf{N} \times \mathbf{N} \to \mathbf{N}$ be a fixed one-to-one and onto map, and define

$$Y_{(i,j)} = X_{\varphi(i,j)}.$$

Further define $U_i = \sum_{j=1}^{\infty} Y_{(i,j)} 2^{-j}$, which forms an i.i.d. sequence of Uniform$[0,1]$ random variables.

We may now invoke Theorem 7.7 and conclude that

$$F_{\mu_i}^{-1}(U_i)$$

produces an independent sequence of μ_i-distributed real random variables.

[1] See also [Kal21, Theorem 4.19] for a quick introduction.

Given a sequence of i.i.d. random variables

$$\xi_n = \begin{cases} 1 & \text{with probability } p, \\ -1 & \text{with probability } q, \end{cases}$$

we define $S_n = \xi_1 + \cdots + \xi_n$.

Nearest neighborhood rw lazy

There are two different perspectives on may look at random walks.

()

10.B Consistent family of probability measures

Let $\{\mu_n\}_{n=1}^{\infty}$ be a sequence of measures each defined on S. We say the sequence is a *consistent family of probability measures* if

$$\mu_n(A_1 \times \cdots \times A_n) = \mu_{n+1}(A_1 \times \cdots \times A_n \times S)$$

for any $A_1, \ldots, A_k \in \mathcal{B}(S)$. The μ_n's are called *finite-dimensional distributions*.

10.1 Daniell–Kolmogorov existence theorem. For a consistent family of distributions $\{\mu_n\}$ on $\mathbf{R}$, then there exists some probability space $(\Omega, \mathcal{F}, P)$ on which we can define a stochastic process $\{X_n\}_{n\in\mathbf{N}}$ with $\{\mu_n\}$ as its finite-dimensional distributions.

We follow the second proof in [Bil95, Section 36]. See also [Kal02, Chapter 8].

For uncountable indices, we have to adjust our definition.

generalize to Polish spaces

10.C Poisson processes

10.2 Proposition. For two independent random variables $X \sim \text{Exponential}(\lambda)$ and $Y \sim \text{Exponential}(\mu)$, we have

(a) $\min\{X, Y\} \sim \text{Exponential}(\lambda + \mu)$;

(b) $P(X \le Y) = \frac{\lambda}{\lambda+\mu}$;

(c) $\min\{X, Y\}$ and $\{X \le Y\}$ are independent.

Proof.

(a) $P(X > t, Y > t) = P(X > t)P(Y > t) = e^{-(\lambda+\mu)t}$.

(b)

$$\begin{aligned} P(X - Y \le 0) &= \int_0^\infty P(X \le y) f_Y(y)\, dy \\ &= \int_0^\infty (1 - e^{-\lambda y}) \mu e^{-\mu y}\, dy \\ &= \int_0^\infty \mu e^{-\mu y} - \int_0^\infty \mu e^{-(\lambda+\mu)y}\, dy \\ &= \frac{\lambda}{\lambda + \mu}. \end{aligned}$$

(c)

$$\begin{aligned} P(X > t, Y > t, X \le Y) &= P(X > t, X \le Y) \\ &= \int_{x=t}^{\infty} \int_{y=x}^{\infty} \lambda e^{-\lambda x} \mu e^{-\mu y} \, dy \, dx \\ &= \int_{x=t}^{\infty} \lambda e^{-\lambda x} e^{-\mu x} \, dx \\ &= \frac{\lambda}{\lambda+\mu} e^{(-\lambda+\mu)t} \\ &= P(X > t, Y > t) P(X \le Y). \end{aligned}$$

□

The above claim can be similarly stated for the smallest of n independent exponential random variables, with the exact same proof; see [Dur19, Exercise 3.7.3] for the statement.

However, the converse is not true. Consider a random variable $Z \sim \text{Exponential}(\lambda + \mu)$, and an independent random variable $\xi \sim \text{Bernoulli}\left(\frac{\lambda}{\lambda+\mu}\right)$. Let

$$(X, Y) = \begin{cases} (Z, Z+1) & \text{if } \xi = 1, \\ (Z+1, Z) & \text{if } \xi = 0. \end{cases}$$

Then clearly all conditions (a)(b)(c) are met, but X and Y are clearly not the desired Exponential random variables.

This might be a little disappointing. But with some deliberation, we notice that the three conditions only capture the distribution of the smaller of X and Y, and the probability that which of two is smaller one. In other words, we have no information about the larger of the two random variables after the "time" $\min\{X, Y\}$ is reached. (Recall an exponential random variable is usually interpreted as the random time at which a light bulb went off.)

Poisson distributions and exponential distributions are duals to each other. Because exponential distributions are memoryless, they form the basis of all continuous-time processes. (for a clock to ring)

A *counting process* $\{N(t)\}_{t\ge0} = \{N_t(\omega)\}_{t\ge0}$ is a continuous-time stochastic process with these properties:

(i) for each t, $N(t) \in \mathbf{N}_0$;

(ii) $N(t)$ is increasing;

(iii) $N(t)$ is right-continuous for almost every ω.

Given a stochastic process X_t, for each sample ω we have the so-called *sample path* $X_t(\omega)$. It is customary to assume càdlàg sample paths for continuous-time stochastic processes. This is mostly an assumption for theoretic purposes on the regularity of the process. For a counting process this assumption is also somewhat natural, since once an increment in N is supposed to take place, it should take place at precisely this time instant t, and not at $t+$.

A *Poisson (point) process* with arrival rate λ is the particular counting process that "follows" the Poisson distribution. It satisfies

(a) $N(0) = 0$;

(b) $N(t)$ has right-continuous sample paths;

(c) for $(s_1, t_1] \cap (s_2, t_2] = \emptyset$, the increments $N(t_1) - N(s_1)$ and $N(t_2) - N(s_2)$ are independent random variables; (independent increments)

(d) the number of events in any interval of length t follows Poisson(λt). (stationary increments depending solely on t)

The independent and stationary increments assumption can be expressed explicitly as follows: for any arbitrary t, h, and $k \geq 0$, we have

$$P(N(t+h) - N(t) = k) = e^{-\lambda h} \frac{(\lambda h)^k}{k!}.$$

We have the equivalent infinitesimal definition that, for any $t \geq 0$, $N(t)$ follows the equation that for very small positive h,

$$P(N(t+h) - N(t) = 1) = \lambda h + o(h) \quad \text{and} \quad P(N(t+h) - N(t) = 0) = 1 - \lambda h + o(h).$$

To illustrate why the two definitions are the same, recall how the Poisson distribution may be interpreted as infinite coin flips. Thus, the coin with success probability λh to increase N by 1 in all intervals with very small length h is what approximates the Poisson process. Note that when $h \to 0$, the $o(h)$ in the two expressions above will vanish. The rigorous proof of the equivalence between the two definitions requires differential equations and is omitted here.

Poisson thinning and superposition

10.3 Theorem. Given a Poisson process $\{N(t)\}$ with rate λ, and an independent sequence of i.i.d. random variables $\xi_j \sim$ Bernoulli(p), then the process $\{N_1(t)\}$, given by

$$N_1(t) = \sum_{j=1}^{N(t)} \xi_j,$$

is a Poisson process with rate $p\lambda$.

10.4 Theorem. Given two independent Poisson processes $N_1(t)$ and $N_2(t)$ with rate λ and μ respectively, the process $N(t) := N_1(t) + N_2(t)$ is a Poisson process with rate $\lambda + \mu$.

Again both claims

Compound Poisson process

10.5 Poisson limit theorem.

10.D Explicit construction of discrete Markov chains

Let S be a finite or countably infinite set, implicitly with the σ-field $\mathcal{P}(S)$. A (row) *stochastic matrix* is a countable-dimensional real matrix $\{Q(x, y) : x, y \in S\}$ satisfying

(a) each value takes value in $[0, 1]$: for every $x, y \in S$, $0 \leq Q(x, y) \leq 1$;

(b) each row sums to 1: for each $x \in S$, $\sum_{y \in S} Q(x, y) = 1$.

10.6 Theorem. Let Q be a stochastic matrix on S, we can find a probability space $(\Omega, \mathcal{F}, P)$, on which we can construct a Markov chain $\{X_n\}$ started at any initial distribution for X_0.

The construction of such a probability space is insufficient for our theory. By definition a Markov chain forgets its past. At a future state $X_n = x_n$, we can pretend that moving forward $\{X_k : k \geq n\}$ is a new Markov chain started at x_n, as if the past has never happened. (This is known as the *Markov*

property, and will be introduced in Section 12.A.) To formalize this notion, we are forced to consider a probability space on which the entire Markov chain $\{X_n\}_{n\geq 0}$ can be shifted into the future.

Now we describe the canonical probability space for a Markov chain. Let $\mathbf{\text{-}} = S^{\mathbf{N}}$, on which we define functions $\mathbf{X}_0, \mathbf{X}_1, \dots$ to be the sequence of coordinate projections. This means that for $\omega = (\omega_0, \omega_1, \dots)$, we define

$$\mathbf{X}_n(\omega) = \omega_n.$$

Let $\mathfrak{F} = \sigma(X_0, X_1, \dots)$. Under this setup, we show that the probability space $(\Omega, \mathcal{F}, P)$ in the previous theorem can be pushed to another probability space $(\mathbf{\text{-}}, \mathfrak{F}, \mathbf{P})$, the canonical one.

Recall Exercise 3.6, which tells us exactly that f is $/\mathfrak{F}$

10.7 Theorem. Let Q be a stochastic matrix on S. For any distribution μ on S, there exists a unique probability measure $\mathbf{P}_\mu$ on $(\mathbf{\text{-}}, \mathfrak{F})$ such that under $\mathbf{P}_\mu$, the sequence of coordinate projections $\{\mathbf{X}_n\}$ becomes a Markov chain with initial distribution μ and transition matrix Q.

10.E Lévy's construction of Brownian motions

Let $(\Omega, \mathcal{F}, P)$ be the underlying space. An $\mathbf{R}^d$-valued stochastic process $\{B_t\}_{t\geq 0}$ is called a d-dimensional *Brownian motion* started from x if it satisfies the following three conditions:

(a) (independent increments) $B_0 = x$, and for any $n \in \mathbf{N}$ and possible $0 = t_0 < t_1 < \cdots < t_n$, the increments

$$B_{t_1} - B_{t_0}, \dots, B_{t_n} - B_{t_{n-1}}$$

are all independent;

(b) (stationary increments) for any $t, s \geq 0$, $B_{t+s} - B_t \overset{\mathrm{D}}{=} B_s - B_0$;

(c) (Gaussian increments) $B_t - B_0 \sim N(0, tI_d)$;

(d) (continuous sample paths) the $t \mapsto B_t(\omega)$ is continuous P-a.s.

The sample path can be made surely continuous.
When $x = 0$, $\{B_t\}_{t\geq 0}$ is called a *standard Brownian motion*.

10.8 Theorem [Bil95, Theorem 36.3]. For a family of functions $X_t : \Omega \to \mathbf{R}$ over $t \in T$,

(a) if $A \in \sigma(X_t : t \in T)$ and $\omega \in A$, if $X_t(\omega) = X_t(\omega')$ for all $t \in T$, then $\omega' \in A$;

(b) if $A \in \sigma(X_t : t \in T)$, then $A \in \sigma(X_t : t \in S)$ for some countable $S \subseteq T$.

$\mathcal{C}\big([0,\infty), \mathbf{R}\big) \subsetneq \mathcal{B}\big(\mathbf{R}^{[0,\infty)}\big)$
a direct proof using a special complete orthonormal system
It is possible to endow two different metrics on $C[0,\infty)$.

10.9 Proposition [Coh13, Exercise 8.1.6]. We can define a metric $d(\cdot,\cdot)$ on $C[0,\infty)$ given by the recipe

$$d(f,g) = \sup\big\{1 \wedge |f(t) - g(t)| : t \in [0,\infty)\big\}.$$

The metric characterizes uniform convergence of continuous functions on $[0,\infty)$:

$$f_n \to f \text{ uniformly on } [0,\infty) \iff d(f_n, f) \to 0.$$

However, the topology on $C[0,\infty)$ induced from this metric is not separable.

10.10 Proposition [Coh13, Exercise 8.1.7]. We can define another metric $d(\cdot, \cdot)$ on $C[0, \infty)$ given by

$$d(f, g) = \sum_{n=1}^{\infty} \frac{1}{2^n} \max\{1 \wedge |f(t) - g(t)| : t \in [0, n]\}.$$

The metric characterizes uniform convergence on compact subsets of $[0, \infty)$:

$$f_n \to f \text{ uniformly on } [0, N] \text{ for all } N \in \mathbf{N} \iff d(f_n, f) \to 0.$$

Under this metric, $C[0, \infty)$ is in fact complete and separable. (This shows the topology of uniform convergence on compact sets is in fact Polish.)

10.F Other constructions of Brownian motions

modify its path so that it becomes continuous
- global Hölder
- local Hölder means on any compact subsets
- Hölder at a point

10.11 Kolmogorov–Chenstov continuity lemma. Given a complete separable metric space (S, ρ), let $X : [0, \infty) \times \Omega \to S$ be a stochastic process. Suppose we have positive constant α, β, C such that

$$\mathrm{E}\rho(X_s, X_t)^\alpha \le C|s - t|^{1+\beta} \tag{10.12}$$

for $x, y \in [0, \infty)$, then we have a continuous modification $\widetilde{X}$ of X whose sample paths are locally Hölder-γ continuous for all $\gamma \in (0, \beta/\alpha)$.

Again the separability of S is assumed to ensure the measurability of $\rho(X_s, X_t)$, as discussed in Remark 3.7.

We may generalize the time index set $[0, \infty)$ to be any subset of $\mathbf{R}^d$, while changing the exponent of $|s - t|$ in (10.12) from $1 + \beta$ to $d + \beta$.

Chapter 11 Ergodic theory and stationary processes

11.A Elementary notions

Given a probability space $(S, \mathcal{S}, \mu)$, a *measure-preserving transformation* (MPT) T is a measurable function from $(S, \mathcal{S})$ to itself such that

$$\mu(T^{-1}A) = \mu(A) \text{ for all } A \in \mathcal{S}.$$

In this circumstance we would also say the measure μ is T*-invariant*. The resulting quartet $(S, \mathcal{S}, \mu, T)$ is called a *measure-preserving dynamical system* (MPDS). If T is invertible, and T^{-1} is measurable, then it is equivalent to say T is measure-preserving if

$$\mu(TA) = \mu(A) \text{ for all } A \in \mathcal{S}.$$

We say a measurable function f is *(almost) invariant* (resp. strictly invariant) with respect to T is $f \circ T = f$ a.s. (resp. x-pointwise).

Recall $T_*\mu$ is the image measure from Section 2.I. Note T is measure-preserving precisely means $T_*\mu = \mu$. Therefore by Proposition 2.42, we have T is measure-preserving if[1] and only if

$$\int f \, d\mu = \int f \circ T \, d\mu$$

for any $f \in L^+$. This is also true $f \in L^1(\mu)$: breaking $f = f^+ - f^-$, it is clear that $f \circ T \in L^1(\mu)$ is guaranteed.

An MPT T is said to be μ*-ergodic* (or the measure μ is said to be T*-ergodic*) if for all $A \in \mathcal{S}$, we have

$$\mu(A \triangle T^{-1}A) = 0 \implies \mu(A) = 0 \text{ or } 1.$$

A set $A \in \mathcal{S}$ satisfying $\mu(A \triangle T^{-1}A) = 0$ is called *(almost) invariant*. If instead we have $T^{-1}A = A$, then T is *strictly invariant*. The ergodicity of T can be equivalently defined by

$$T^{-1}A = A \implies \mu(A) = 0 \text{ or } 1,$$

that is, we only need to check strictly invariant sets must be of measure 0 or 1.

One direction is obvious. For the other direction, one can check that for any set $A \in \mathcal{S}$, the set $B = \limsup_n T^{-n}A$ is always going to be strictly invariant.

It is easy to see that the strictly invariant σ-field $\mathcal{I}$ and the almost invariant σ-field must *almost* be the same:

11.1 Fact [Kal21, Lemma 25.4]. The almost invariant σ-field is precisely generated by $\mathcal{I}$ and the μ-null sets in $\mathcal{S}$.

[1] follows by just taking f to be any indicator functions

11.2 Definition. An MPDS $(S, \mathcal{S}, \mu, T)$ is said to be *strong mixing* if for all $A, B \in \mathcal{S}$,

$$\lim_n \mu(A \cap T^{-n} B) = \mu(A)\mu(B); \tag{11.3}$$

it is said to be *weak mixing* if for all $A, B \in \mathcal{S}$,

$$\lim_n \frac{1}{n} \sum_{k=0}^{n-1} \left|\mu(A \cap T^{-k} B) - \mu(A)\mu(B)\right| = 0, \tag{11.4}$$

i.e., $\left|\mu(A \cap T^{-k} B) - \mu(A)\mu(B)\right|$ converges to 0 in the Cesàro sense.

Hence strong mixing implies weak mixing. In fact weak mixing further implies the system is ergodic. Let $A = B \in \mathcal{S}$ be strictly invariant, then we may replace $T^{-k} B$ by B in (11.4) and get $\mu(B) = \mu(B)^2$.

Notice that the above argument remains true if we remove the $|\cdot|$ in the definition (11.4) of weak mixing. It turns out that

$$\lim_n \frac{1}{n} \sum_{k=0}^{n-1} \mu(A \cap T^{-k} B) = \mu(A)\mu(B) \quad \text{for all } A, B \in \mathcal{S}$$

is in fact equivalent to the saying that the system is ergodic. But the converse requires the Birkhoff pointwise ergodic theorem, which is the most important result of ergodic theory.

Most ergodic dynamical systems of interest to probabilists turns out to be strong mixing. Indeed, one may interpret it as eventual independence.

Dyadic transformation

strongly ergodic completely positive entropy isomorphic to Bernoulli shift

occurrence time recurrence time sojourn time

11.5 Poincaré recurrence theorem. $\mu(\{x \in A : T^n x \in A \text{ i.o.}\}) = \mu(A)$.

Proof. Consider the set

$$\begin{aligned} B := \{x \in A : T^n x \notin A \text{ ev.}\} &= \bigcup_{n=1}^{\infty} \bigcap_{m \geq n} \{x \in A : T^m x \notin A\}, \\ &= \bigcap_{n=1}^{\infty} \{x \in A : T^n x \notin A\} \\ &= A \cap \left(\bigcap_{n=1}^{\infty} T^{-n}(X - A) \right). \end{aligned}$$

which we want to show is of measure 0.

Notice that for any $j < k$ in $\mathbf{N}$, $T^{-j} B$ and $T^{-k} B$ are disjoint because $T^{-j} B \subseteq T^{-k}(X - A)$ while $T^{-k} B \subseteq T^{-k} A$. Therefore

$$\mu\left(\bigcup_{k=1}^{\infty} T^{-k} B \right) = \sum_{k=1}^{\infty} \mu(B) \leq 1,$$

and this forces $\mu(B) = 0$. □

11.6 Lemma. Given a π-system $\mathcal{K}$ that generates $\mathcal{S}$, if for each $A \in \mathcal{K}$ we have $T^{-1}A \in \mathcal{S}$ and $\mu(T^{-1}A) = \mu(A)$, then μ is measure-preserving.

11.7 Lemma. Given a π-system $\mathcal{K}$ that generates $\mathcal{S}$, if (11.3) holds for all $A, B \in \mathcal{K}$, then the system is strong mixing.

Proof. Apply the π-λ theorem twice □

more general [Bil95, Lemma 24.2]
may also use caratheodory as appropriate

11.8 Example. (a) Bernoulli shift or i.i.d. sequence strong mixing Kolmogorov zero–one law

(b) Rotation on a circle

(c) Markov shift (delayed)

from [Bil95, Theorem 36.5]

Proof of Hewitt–Savage zero–one law. □

11.B The ergodic theorems

11.9 von Neumann mean ergodic theorem. Let U be a contraction operator on a Hilbert space H, and let Π be the projection onto the closed subspace $\operatorname{null}(I - U)$. We then have

$$\frac{1}{n}\sum_{k=0}^{n-1} U^k \to \Pi$$

in the strong operator topology, i.e., pointwise on H.

Proof. easy for unitary

Let $N = \operatorname{null}(I - U)$, $R = \operatorname{range}(I - U)$, and $A_n = \frac{1}{n}\sum_{k=0}^{n-1} U^k$.

If $x \in N$, then $Ux = x$ and $\Pi x = x$, so the convergence holds. If $x \in R$, which means $x = (I - U)v$ for some $v \in H$, then

$$\begin{aligned}\|A_n x\| &= \frac{1}{n}\|v - U^n v\| \\ &\le \frac{1}{n}\|I - U^n\|\|v\| \\ &\le \frac{1}{n}(1 + 1^n)\|v\| \to 0.\end{aligned}$$

We can extend the convergence above to all $x \in \overline{R}$, essentially because $\|A_n\| \le 1$. Take a sequence $\{x_j\} \subseteq R$ converging to x. We have for any n and j that

$$\begin{aligned}\|A_n x\| &\le \|A_n x_j\| + \|A_n\|\|x_j - x\| \\ &\le \|A_n x_j\| + \|x_j - x\|.\end{aligned}$$

Taking $n \to \infty$ first and $j \to \infty$ next, we have shown $A_n x \to 0$ for all $x \in \overline{R}$.

The desired claim now follows from the orthogonal decomposition $H = N \oplus \overline{R}$. □

some tricks is needed We follow [Tay06, Lemma 14.1].

11.10 Birkhoff pointwise ergodic theorem. For f nonnegative measurable or in $L^1(\mu)$, it holds that

$$\frac{1}{n}\sum_{k=0}^{n-1} f(T^k x) \to \mathrm{E}_\mu(f \mid \mathcal{I}) \quad \text{a.s.}$$

If $f \in L^p$, then the convergence also holds in L^p.

Consider the context when T is the shift operator on $(S^{\mathbf{Z}}, \otimes^{\mathbf{Z}}\mathcal{S}, \mu)$, then for μ-a.e. $x \in S^{\mathbf{Z}}$ that

$$\frac{1}{n}\sum_{k=0}^{n-1} f(T^k x) \to \hat{f}(x),$$

where $\hat{f}$ is $\mathcal{I}$-measurable and

$$\int_G \hat{f}\, d\mu = \int_G f\, d\mu \quad \text{for all } G \in \mathcal{I}.$$

Say X is a random variable on $(\Omega, \mathcal{F}, P)$ that is $(S^{\mathbf{Z}}, \otimes^{\mathbf{Z}}\mathcal{S})$-valued with distribution μ. Pushing forward, we get for P-a.e. ω that

$$\frac{1}{n}\sum_{k=0}^{n-1} f\big(T^k X(\omega)\big) \to \hat{f}\big(X(\omega)\big),$$

where $\hat{f} \circ X$ is $X^{-1}\mathcal{I}$-measurable, and

$$\int_A \hat{f}(X)\, dP = \int_A f(X)\, dP \quad \text{for all } A \in X^{-1}\mathcal{I}.$$

This confirms our suspicion that when f is nonnegative measurable or $\mathrm{E}|f(X)| \geq 0$,

$$\frac{1}{n}\sum_{k=0}^{n-1} f(T^k X) \to \mathrm{E}_P[f(X) \mid X^{-1}\mathcal{I}] \quad \text{a.s.}$$

Clearly L^p convergence should hold as well.

11.11 Maximal ergodic theorem. For a T-invariant function $g\colon S \to S$ with $g^+ \in L^1(\mu)$, we have

$$\mathrm{E}\big(f - g;\, f^* - g > 0\big) \geq 0.$$

11.12 Subadditive ergodic theorem.

Consider a one-parameter semigroup $\{T_t\}_{t\geq 0}$ such that

$$(x, t) \mapsto T_t x$$

is $\mathcal{S} \otimes \mathcal{B}[0,\infty)/\mathcal{S}$-measurable. This is called a *measurable flow*. We say the flow is *measure-preserving* if $\mu(T_t^{-1}A) = \mu(A)$ for all $t \geq 0$. Now we naturally define the invariant σ-field $\mathcal{I}$ to be the collection

$$\{I \in \mathcal{S} : T_t^{-1} I = I\}.$$

11.13 Continuous-time von Neumann theorem. Say there is a one-parameter semigroup $\{U_t\}_{t\geq 0}$ of contraction linear operators on a Hilbert space H, and let Π be the projection onto the closed subspace $\{x \in H : U_t x = x \text{ for all } t \geq 0\}$ (invariant under U_t). We have as the time $N \to \infty$,

$$\frac{1}{N}\int_0^N U_t \, dt \to \Pi$$

in the strong operator topology, i.e., pointwise on H.

11.14 Continuous-time Birkhoff's theorem. For f is nonnegative measurable or in $L^1(\mu)$, it holds that as the time $N \to \infty$, we have

$$\frac{1}{N}\int_0^N f(T_t x)\, dt \to \mathrm{E}_\mu(f \mid \mathcal{I}) \quad \text{a.s.}$$

If $f \in L^p$, then the convergence also holds in L^p.

Again we should have for f nonnegative measurable or $\mathrm{E}|f(X)| < \infty$, that

$$\frac{1}{N}\int_0^N f(Q_t X)\, dt \to \mathrm{E}_P[f(X) \mid X^{-1}\mathcal{I}] \quad \text{a.s.},$$

and also L^p convergence.

11.15 Shannon–McMillan–Breiman theorem. Let H be the entropy rate of a given discrete-time finite-state stationary ergodic process $\{X_n\}$, then almost surely

$$-\frac{1}{n}\log p(X_0, X_1, \dots) \to H.$$

generalization to countable-state and densities
induced transformation

11.C Invariant measures, ergodicity, and weak convergence

Throughout this section we may assume S to be a locally compact and separable metric space, and let $T : S \to S$ be a measurable mapping.

Occasionally we also assume that S is just compact, so that vague convergence are automatically weak convergence. Recall in this case the space of Borel subprobability measures $\mathcal{M}^{\leq 1}(S)$, as the closed unit ball in $C^*(S)$, is a sequentially compact space in the topology of weak convergence. Also $\mathcal{P}(S)$, as a weakly closed subset of $\mathcal{M}^{\leq 1}(S)$ (since mass is preserved), is also a sequentially compact space.

It turns out a nonempty set of invariant measures arises naturally from continuous maps on a compact metric space S.

Denote the space of invariant measures by $\mathcal{P}^T(S)$. It is a closed and convex subset of $\mathcal{P}(S)$.

11.16 Krylov–Bogoliubov theorem. Let S be compact and T be continuous. Given any measure $\nu \in \mathcal{P}(S)$, we may define a sequence $\mu_n = \frac{1}{n}\sum_{k=0}^{n-1} T_*^k \nu$ of Cesàro sums of image measures.

Any subsequential limit of $\{\mu_n\}$ in the topology of weak convergence is an invariant probability measure. Since $\mathcal{P}(S)$ is sequentially compact (see Corollary 8.17), $\mathcal{P}^T(S)$ must be nonempty.

(If S is in general locally compact and separable, then if $\{\mu_n\}$ is tight, it has a subsequential limit that is an invariant probability measure, by Proposition 8.19.)

To make things slightly more general, we may also replace ν be a sequence of measures $\{\nu_n\} \subseteq \mathcal{P}(S)$, and define $\mu_n = \frac{1}{n}\sum_{k=0}^{n-1} T_*^k \nu_n$. The same proof below carries over.

Proof. Let $\{\mu_{n_j}\}$ be a subsequence converging weakly to μ. To check μ is T-invariant, it suffices to show that as $j \to \infty$,

$$\int f \circ T - f \, d\mu_{n_j} \to 0$$

for all $f \in C(S)$. Expanding the left-hand side, we get

$$\begin{aligned} &\frac{1}{n_j} \int \sum_{k=1}^{n_j} f \circ T^k - f \circ T^{k-1} \, d\nu \\ \leq &\frac{1}{n_j} \int |f \circ T^{n_j} - f| \, d\nu \\ \leq &\frac{2}{n_j} \|f\|_u \to 0, \end{aligned}$$

finishing the proof. □

11.17 Continuous Krylov–Bogoliubov theorem. Let S be locally compact and separable. Given a measure ν and a continuous measurable flow $\{T_t\}_{t\geq 0}$, define for each $N > 0$ and any $x \in S$

$$\mu_{N,x}(A) = \frac{1}{N} \int_{t=0}^{N} (T_t)_* \nu(A) \, dt.$$

If the family of probability measures $\{\mu_{N,x}\}_{N>0}$ is tight for some $x \in S$, then there is an invariant measure μ with respect to $\{T_t\}_{t\geq 0}$.

if and only if

The following result characterizes the ergodic measures among the invariant measures.

11.18 Theorem. Let T be measurable, then the ergodic measures in $\mathcal{P}^T(S)$ are precisely the extreme points of $\mathcal{P}^T(S)$.

11.19 Proposition. Two distinct T-invariant measures μ and ν must be mutually singular.

Proof. This is a simple application of the Birkhoff pointwise ergodic theorem. Pick some $B \in \mathcal{S}$ such that $\mu(B) \neq \nu(B)$. Therefore

$$\frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_B(T^k x) \to \mu(B) \quad \mu\text{-a.s.}, \tag{11.20}$$

and

$$\frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_B(T^k x) \to \nu(B) \quad \nu\text{-a.s.}$$

Say (11.20) holds on the set A with $\mu(A) = 1$. It is immediate that $\nu(A) = 0 = \mu(A^c)$, which proves mutual singularity. □

Chapter 12 Markov chains

12.A Markov properties

12.1 Simple Markov property. Let $G : \text{-} \to \mathbf{R}$ be a nonnegative or bounded Borel measurable function, then

$$\mathbf{E}_\mu(G \circ \theta_n \mid \mathfrak{F}_n) = \mathbf{E}_{X_n} G \quad \text{for all } n \in \mathbf{N}.$$

Note that G is just a random variable

12.2 Chapman–Kolmogorov equation. $\mathbf{P}_\mu(X_{n+m} = x) = \sum_{y \in S} \mathbf{P}_\mu(X_n = y)\mathbf{P}_y(X_m = x)$. In particular we have

$$\mathbf{P}_\mu(X_{n+1} = x) = \sum_{y \in S} \mathbf{P}_\mu(X_n = y)\mathbf{P}_y(X_1 = x) = \sum_{y \in S} \mathbf{P}_\mu(X_n = y) Q(y, x).$$

By an inductive argument one gets $\mathbf{P}_\mu(X_n = x) = \sum_y \mu(y) Q^n(y, x)$, and in particular

$$\mathbf{P}_y(X_n = x) = Q^n(y, x).$$

12.3 Strong Markov property. Let $G_n : \text{-} \to \mathbf{R}$ be a sequence of Borel measurable functions bounded by M for all $n \in \mathbf{N}$, then

$$\mathbf{E}_\mu\big(\mathbf{1}_{\{T<\infty\}} G_T \circ \theta_T \;\big|\; \mathfrak{F}_T\big) = \mathbf{1}_{\{T<\infty\}}\, \mathbf{E}_{X_T} G_T.$$

12.B Recurrence and transience

Let the *hitting time* to y be $T_y = \inf\{n \geq 1 : X_n = y\}$, then the expected hitting time to y starting from x is $\mathbf{E}_x N_y$. Let the number of visits to y be $N_y = \sum_{n=1}^{\infty} \mathbf{1}\{X_n = y\}$. The goal of this section is to establish the connection between the three quantities.

12.4 Fact. $\mathbf{E}_x N_y = \sum_{n=1}^{\infty} \mathbf{P}_x(X_n = y) = \sum_{n=1}^{\infty} Q^n(x, y)$.

12.5 Theorem. For any $x \in S$, then there are only two possibilities for a state:

(a) *recurrent*, i.e., $\mathbf{P}_x(T_x < \infty) = 1$. In this case $\mathbf{P}_x(N_x = \infty) = 1$ and hence $\mathbf{E}_x N_x = \sum_n Q^n(x, x) = \infty$.

(b) *transient*, i.e., $\mathbf{P}_x(T_x < \infty) < 1$. In this case $\mathbf{P}_x(N_x < \infty) = 1$, and furthermore $\mathbf{E}_x N_x = \sum_n Q^n(x, x) < \infty$.

positive recurrent
a finite mc has at least one recurrent state

12.6 Recurrence as an equivalence relation.

12.C Stationary distributions

12.7 Definition. Given a nonzero measure π such that $\pi(x) < \infty$ for all $x \in S$, we say

(a) π is a *stationary/invariant measure* with respect to Q if for all $y \in S$,

$$\pi(y) = \sum_{x \in S} \pi(x) Q(x, y); \tag{12.8}$$

(b) π is a *reversible measure* with respect to Q if for all $x, y \in S$, we have

$$\pi(x) Q(x, y) = \pi(y) Q(y, x). \tag{12.9}$$

A stationary measure can be easily interpreted in the matrix notation. If we write π as a row vector indexed by S, then (12.8) is equivalent to $\pi = \pi Q$. Furthermore, this gives $\pi = \pi Q^n$ for any n.

12.10 Fact. A reversible measure is a stationary measure, which is clear by doing a summation over x on both sides of (12.9).

12.11 Kolmogorov cycle condition for stationarity. Suppose Q is irreducible. A necessary and sufficient condition for Q to have a reversible measure is

(a) $Q(x, y) > 0 \implies Q(y, x) > 0$;

(b) for any cycle $x_0, x_1, \ldots, x_n = x_0$,

$$\prod_{j=1}^{n} Q(x_{j-1}, x_j) = \prod_{j=1}^{n} Q(x_j, x_{j-1}).$$

time reversal

12.12 Proposition. Let

12.D Convergence to stationarity

12.13 Proposition.

(a) In an irreducible and aperiodic chain, there exists an N such that $Q^n(x, x) > 0$ for all $n \geq N$;

(b) if the chain is furthermore finite, then there exists M such that $Q^n(x, y) > 0$ for all $n \geq M$.

12.14 Convergence in total variation. Let Q be irreducible and aperiodic for $\{X_n\}$, and let π be its stationary distribution. We have

$$\max_{x \in S} d_{\mathrm{TV}}\big(Q^n(x, \cdot), \pi\big) \to 0.$$

If the state space is finite, then we have an exponential convergence rate: for all time n, there exists some rate $r < 1$ such that

$$\max_{x \in S} d_{\mathrm{TV}}\big(Q^n(x, \cdot), \pi\big) \leq C e^{rn}$$

for some absolute constant $C > 0$.

12.E Ergodicity of Markov chains

12.15 Ergodic theorem for Markov chains. Let Q be irreducible with stationary distribution π, and let $f \in L^1(S, \pi)$. For any initial distribution μ, we have

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_n) \to \sum_y f(y)\pi(y) \quad \mathbf{P}_\mu\text{-a.s.}$$

Proof. □

12.F Harmonic Markov chains

12.G Random walks as Markov chains

12.16 Reflection principle. For any real number $a \geq 0$, we have

$$\mathbf{P}_0\Big(\max_{m\leq n} S_m \geq a\Big) = \mathbf{P}_0(S_n \geq a) + \mathbf{P}_0(S_n \geq a+1)$$

The key to many results about Markov processes is to write the desired event in terms of stopping times, which we have control over by the strong Markov property.

12.H Major examples

Ehrenfest urn model Pólya's urn

12.I Continuous-time Poisson jump Markov chains

12.J The general continuous-time theory

Given a collection $\{Q_t\}$ of transition kernels, if we have in addition

(a) for every $x \in S$, $Q_0(x, \cdot) = \delta_x$;

(b) the Chapman–Kolmogorov equation holds: for every $A \in \mathcal{S}$

$$Q_{s+t}(x, A) = \int_{y\in S} Q_t(y, A)\, Q_s(x, dy);$$

(c) for each fixed $A \in \mathcal{S}$, the mapping $(x, t) \mapsto Q_t(x, A)$ is a $\mathcal{S} \otimes \mathcal{B}_{[0,\infty)}$-measurable function.

Notice that $\|Q_t f\| \leq \|f\|$
$Q_0 = I$

Also for any $x \in S$ and $A \in \mathcal{S}$,

$$\begin{aligned} Q_s Q_t \mathbf{1}_A(x) &= Q_s \int_{y \in S} \mathbf{1}_A(y)\, Q_t(x, dy) \\ &= \int_{z \in S} \int_{y \in S} \mathbf{1}_A(y)\, Q_t(z, dy)\, Q_s(x, dz) \\ &= \int_{z \in S} Q_t(z, A)\, Q_s(x, dz) \\ &= Q_{s+t}(x, A) = Q_{s+t} \mathbf{1}_A(x), \end{aligned}$$

and the property $Q_s Q_t = Q_{s+t}$ follows.

Therefore a Markov semigroup is structurally a one-parameter strongly continuous contractive semigroups on the space of bounded Borel measurable functions. Very often we are interested in other function spaces, in particular C_0 and L^p spaces.

A transition semigroup $\{Q_t\}$ is called a *Feller semigroup* if

(a) Q_t takes $f \in C_0(S)$ to $Q_t f \in C_0(S)$;

(b) for any $f \in C_0(S)$, we have the pointwise convergence $Q_t f(x) \to f(x)$ at all $x \in S$.

It turns out that condition implies the strong convergence $\| Q_t f - f \| \to 0$.

We have a

Hille–Yosida theorem

Hunter Evans Taylor

12.17 Simple Markov property. Let $\Phi\colon D(S) \to \mathbf{R}$ be nonnegative/bounded measurable, and assume that $\{X_t\}$ has càdlàg sample paths. Then

$$\mathbf{E}_\mu[\Phi(X_{s+t}) \mid \mathfrak{F}_s] = \mathbf{E}_{X_s} \Phi(X_t) \quad \text{for all } n \in \mathbf{N}.$$

Carré du champ operator

measures leftover from being a product rule

$$\Gamma(f, g) = \tfrac{1}{2}\big[L(fg) - (Lf)g - f(Lg)\big].$$

12.K Harris chains

Chapter 13 Brownian motions

13.A Some sample path properties

13.1 Theorem. Almost surely, the sample paths of Brownian motions are locally α-Hölder continuous for $\alpha < 1/2$, but at no points for $\alpha > 1/2$.

13.2 Lévy's modulus of continuity. Almost surely

$$\limsup_{\delta \to 0^+} \sup_{0 \le t \le 1-\delta} \frac{|B_{t+\delta} - B_t|}{\sqrt{2\delta \log(1/\delta)}} = 1.$$

13.3 Theorem. For a standard Brownian motion $\{B_t\}$,

(a) under an orthogonal transformation U, $\{UB_t\}_t$ is still a standard Brownian motion;

(b) for any $\gamma > 0$, the scaled $\{\frac{1}{\gamma} B_{\gamma^2 t}\}_t$ is still a standard Brownian motion;

(c) the process

$$W_t := \begin{cases} tB_{1/t} & \text{when } t > 0; \\ 0 & \text{when } t = 0. \end{cases}$$

is also a standard Brownian motion, called the *time inversion* of B_t.

$\mathcal{F}_{0+} = \bigcap_{t>0} \mathcal{F}_t$
Let $\mathcal{T} = \bigcap_{s \ge 0} \sigma(B_t : t \ge s)$, the tail σ-field of the Brownian motion $\{B_t\}$.

13.4 Blumenthal's zero–one law. For any $A \in \mathcal{F}_{0+}$, we have $\mathbf{P}_x(A) = 0$ or 1.

If we complete the natural filtration of the Brownian motion, then the filtration $\widetilde{\mathcal{F}}_t$ becomes right-continuous, and hence $\widetilde{\mathcal{F}}_{0+} = \widetilde{\mathcal{F}}_0$.

13.5 Theorem. For any $A \in \mathcal{T}$, we have $\mathbf{P}_x(A) = 0$ or 1.

13.6 Theorem.

In any small interval $[0, \epsilon)$ right of 0, the Brownian motion is almost surely positive, negative, and zero at some time instant.

The zero set of a Brownian motion is an closed set without isolated points. Therefore it is also uncountable.

Almost surely $t \mapsto B_t$ is not monotone on any nondegenerate intervals

Almost surely $t \mapsto B_t$ is of unbounded variation on any nondegenerate intervals.

13.B Markov properties

13.7 Simple Markov properties. For every fixed time $s \geq 0$, the process $B_{s+t} - B_s$ is a Brownian motion that is independent of $\mathcal{F}_{s+}$.

with transition density given by

$$p_t(x, y) = \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{|x-y|^2}{2t}\right).$$

13.8 Strong Markov properties. Given a stopping time T such that $P(T < \infty) > 0$. Then under the conditional probability measure $P(\cdot \mid T < \infty)$, we have

$$\mathbf{1}\{T < \infty\}(B_{T+t} - B_T) \quad \text{is a Brownian motion independent of} \quad \mathcal{F}_{T+}.$$

infinitesimal peak into the future

13.9 Proposition (reflected Brownian motion). If T is a stopping time, then

$$B_t \mathbf{1}\{t \leq T\} + (2B_T - B_t)\mathbf{1}_{t>T}$$

is also a standard Brownian motion indexed by t.

divergence behavior

13.10 Theorem. For a one-dimensional Brownian motion starting from any x, $\limsup_t \frac{B_t}{\sqrt{t}} = +\infty$ and $\liminf_t \frac{B_t}{\sqrt{t}} = -\infty$ $\mathbf{P}_x$-a.s.

As in the case for random walks, this can be proven using the . However,

13.11 Reflection principle. For any $a \geq 0$, we have

$$\mathbf{P}_0\Big(\max_{s \leq t} B_s \geq a\Big) = 2\mathbf{P}_0(B_t \geq a) = \mathbf{P}_0(|B_t| \geq a)$$

If we let the running maximum of Brownian motion $\max_{s \leq t} B_s$ be S_t, then $S_t \overset{\mathrm{D}}{=} |B_t|$.

More generally, we have for any $a \geq 0$ and $b \leq a$ that

$$\mathbf{P}_0\Big(\max_{s \leq t} B_s \geq a, B_t \leq b\Big) = \mathbf{P}_0(B_t \geq 2a - b).$$

Clearly

$$\mathbf{P}_0\Big(\max_{s \leq t} B_s \geq a, B_t \geq 2a - b\Big) = \mathbf{P}_0(B_t \geq 2a - b).$$

13.C A third return to random walks

13.12 Law of iterated logarithms.

$$\limsup_t \frac{B_t}{\sqrt{2t \log\log t}} = 1 \quad \text{a.s.}$$

Since $B_t \overset{\mathrm{D}}{=} -B_t$, we also have a.s. $\liminf_t \frac{B_t}{\sqrt{2t\log\log t}} = -1$. Therefore some authors would write

$$\limsup_t \frac{|B_t|}{\sqrt{2t\log\log t}} = 1 \quad \text{a.s.}$$

martingales
linear martingales B_t quadratic martingales $B_t^2 - t$ exponential martingales
We write $T_a = \inf\{t \geq 0 : B_t = a\}$

13.13 Theorem. For $-a < 0 < b$, let $T = \inf\{t \geq 0 : B_t \notin [-a, b]\}$, which we usually call the *exit time* to the interval $[-a, b]$. Then

$$\mathbf{P}_0(B_T = -a) = \frac{b}{a+b},\ \mathbf{P}_0(B_T = b) = \frac{a}{a+b}, \text{ and } \mathbf{E}_0 T = ab.$$

13.14 Skorohod representation theorem. For $X \in L^2$ with $\mathrm{E}X = 0$, we have a stopping time T with respect to the natural filtration of the Brownian motion, such that

$$B_T \overset{\mathrm{D}}{=} X \quad \text{and} \quad \mathrm{E}T = \mathrm{E}X^2.$$

We can embed the symmetric random walk into a Brownian motion.

13.15 Corollary. For i.i.d. real-valued $X_1, \ldots, X_n$ with mean 0 and variance 1. Define $S_n = \sum_{j=1}^n X_j$. We can find a sequence of stopping times $\{T_k\}_{k=0}^\infty$ with $T_0 = 0$, such that

$$\text{each } S_n = B_{T_n} \quad \text{and} \quad T_n - T_{n-1} \text{ are i.i.d.}$$

As usual, we use $S_n = \xi_1 + \cdots + \xi_n$, where $\xi_1 \ldots, \xi_n$ are independent with zero mean and unit variance. Define the function

$$S(t) = S_{\lfloor t \rfloor}(1 + \lfloor t \rfloor - t) + S_{\lfloor t \rfloor + 1}(t - \lfloor t \rfloor),$$

which extends S_n by linearly interpolates between the points (n, S_n) on the graph.

The following result tells us the symmetric random walks $S(t)$, when scaled, becomes a Brownian motion in the weak limit.

13.16 Donsker's invariance principle. On the space $C[0, 1]$ with the Borel σ-field, we have

$$\frac{S(n\,\cdot)}{\sqrt{n}} \Rightarrow B(\,\cdot\,)$$

It is important to be clear about what the weak convergence here actually means.

13.D Introduction to Gaussian processes

fractional Brownian motion

Given a parameter $0 < H < 1$, we may define a standard Gaussian process $\{B_H(t)\}$ with zero mean and covariance function

$$\mathrm{E}\big[B_H(s)B_H(t)\big] = \frac{1}{2}\big(s^{2H} + t^{2H} - |t-s|^{2H}\big)$$

for any s, t. This process is known as the standard one-dimensional *fractional Brownian motion* with *Hurst parameter* H.

non independent increments, but still remains stationary (fractional Gaussian noise)

When $H = 1/2$, it is clear that we recover the standard Brownian motion. When $H < 1/2$ ($> 1/2$), $\mathrm{E}\big[B_H(s)B_H(t)\big] < 0$ (> 0), negatively (positively correlated) correlation function.

locally Hölder-α continuous for any $\alpha < H$

13.E Processes induced from Brownian motions

Throughout this section, the time index of processes will be in the parentheses instead of the subscripts.

The stopping time process $\{T_b : b \geq 0\}$ is an increasing homogeneous Markov process. Its transition density is given by

$$p_a(s,t) = \frac{a}{\sqrt{2\pi(t-s)^3}} \exp\left(-\frac{a^2}{2(t-s)}\right),$$

for $s < t$.

The statement is indeed confusing. The stopping time process is indexed by the states b, while s and t are candidates for stopping times. The transition density $p_a(s,t)$ describes the density of the conditional distribution

$$P(T_{b_2} = t \mid T_{b_1} = s),$$

where $a = b_2 - b_1$.

A (standard) reflected Brownian motion is given by $\{|B_t|\}$, where B_t is a standard Brownian motion. The name comes from the observation that once B_t hits zero in its sample, it must "reflect" to stay nonnegative.

A (standard) Brownian bridge process $W^0(t)$ is defined in distribution by $\{B(t) - tB(1)\}_{0\leq t\leq 1}$.

13.17 Proposition. A Brownian bridge has the distribution of a Brownian motion conditioning on hitting 0 at time 0. To be precise,

$$\{B(t) - tB(1)\}_{0\leq t\leq 1} \overset{\mathrm{D}}{=} \{B(t) \mid B(1) = 0\}_{0\leq t\leq 1}.$$

Recall that the conditional distribution of the right-hand side is given by the f.d.d.

$$P\big(B(t_1) = x_1, \ldots . B(t_n) = x_n \;\big|\; B(1) = 0\big) = \frac{1}{p_1(0,0)} \prod_{j=0}^{n} p_{t_{j+1}-t_j}(x_j, x_{j+1}).$$

for $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} = 1$.

13.18 Proposition. The Brownian bridge is a continuous Gaussian process with zero mean and covariance function

$$\mathrm{E}(X_s X_t) = s(1-t) \quad \text{for all } 0 \leq s \leq t \leq 1.$$

13.19 Vervaat transform. Let $\tau_m = \arg\min_t W^0(t)$, which is a.s. unique. It turns out that

$$W^{\oplus}(\cdot) \overset{\mathrm{D}}{=} W^0(\tau_m + \cdot) - W^0(\tau_m).$$

Fix time $T > 0$. we define the *last passage time* at level 0 before time T by

$$\sigma = \sigma_T = \sup\{s \leq T : B_s = 0\},$$

and the *first passage time* after time T to be

$$\tau = \tau_T = \inf\{s \geq T : B_s = 0\}.$$

Be aware that only τ is a stopping time with respect to the natural filtration of the Brownian motion. The random time σ, being the *last* passage time, is dependent on the future up to the fixed time T. However, we may show that it is a stopping time under time-reversal.

A (standard) d-dimensional squared Bessel process

13.F Generalization of Brownian motions

13.20 Definition. A (standard) Lévy process $X_0 = 0$ Independent and stationary increments as time $t \to 0^+$, we have $X_t \to 0$ in probability continuity in probability: at all times $t \geq 0$, for any $\epsilon > 0$, we have

$$\lim_{h \to 0} P(|X_{t+h} - X_t| > \epsilon) = 0.$$

A version that is càdlàg

Brownian motion and Poisson process now falls under the same umbrella

Lévy–Khintchine formula

Chapter 14 Stochastic calculus

14.A Continuous filtration and martingales

If X_t is progressively measurable, then X_t is adapted to $\mathcal{F}_t$.

A continuous adapted process $\{M_t\}$ with $M_0 = 0$ is called a *continuous local martingale* if there exists an increasing sequence of stopping times $\{T_n\}$ such that $T_n(\omega) \to +\infty$ a.s., while for each n, the stopped process $\{M_{t\wedge T_n}\}_t$ is a uniformly integrable martingale.

Suppose $\{M_t\}$ is a continuous local martingale started at 0 and also a finite-variation process. then $M_t(\omega) = 0$ for a.e. ω over $t \geq 0$.

For a continuous local martingale M started from 0, we have $M = 0$ if and only if $\langle M \rangle = 0$.

14.1 Proposition Proposition 3.4. A left-continuous/right-continuous adapted process is progressively measurable.

14.2 Doob's maximal inequality.

14.3 Doob's L^p inequality.

A process $\{X_t\}$ is said to be of class D if $\{X_\tau : \tau$ is a finite stopping time$\}$ is uniformly integrable.

14.4 Doob–Meyer decomposition. The process $\{X_t\}$ is a submartingale of class D if and only if

$$X_t = M_t + A_t,$$

where M is a uniformly integrable martingale, and A is an increasing predictable process such that $\mathrm{E}A_\infty < \infty$. The decomposition is unique.

(If the process X is a supermartingale then we have $X_t = M_t - A_t$ instead.)

14.5 Undefined Theorem Name.

finite variation process

quadratic variation

Let $\{\mathcal{F}_t\}$ be a right-continuous and complete filtration, and $\{X_t\}$ be an adapted submartingale (or supermartingale) such that $t \mapsto \mathrm{E}X_t$ is right-continuous (which is clearly satisfied when $\{X_t\}$ is just a martingale). Then $\{X_t\}$ has a càdlàg modification $\{\widetilde{X}_t\}$ that remains a submartingale (or supermartingale).

A local martingale is a martingale if and only if it is uniformly integrable.

A continuous martingale must be a continuous local martingale, but the converse is false in general.

$B_t^2 - t$ is a continuous martingale

and

Fix any $t > 0$, and let $p = \{t_0, \ldots, t_{n(p)}\}$ be any partition of the time interval $[0, t]$, where

$$0 = t_0 < t_1 < \cdots < t_{n(p)} = t.$$

If we have a sequence of partitions p_m of $[0,t]$ such that the mesh $\|p_m\| \to 0$, then

$$\sum_{j=1}^{n(p)} (M_{t_j} - M_{t_{j-1}})^2 \to \langle M \rangle_t \text{ in } L^2,$$

and hence in probability.

14.6 Theorem. For a continuous local martingale $\{M_t\}$, there exists an increasing process $\{\langle M \rangle_t\}$ unique up to distinguishability, called the *quadratic variation* of M_t, such that

$$M_t^2 - \langle M \rangle_t$$

gives a new continuous local martingale.

The name comes from the following result. Fix any $t > 0$, and let $p = \{t_0, \ldots, t_{n(p)}\}$ be any partition of the time interval $[0,t]$, where

$$0 = t_0 < t_1 < \cdots < t_{n(p)} = t.$$

We define the QV of the continuous local martingale M with respect to a partition of $[0,t]$ by

$$\mathrm{QV}(M,p) = \sum_{j=1}^{n(p)} (M_{t_j} - M_{t_{j-1}})^2$$

If we have a sequence of partitions p_m of $[0,t]$ such that the mesh $\|p_m\| \to 0$, then

$$\mathrm{QV}(M,p_m) \to \langle M \rangle_t \text{ in probability.}$$

Given two continuous local martingales M_t and N_t, we define their *covariation process* by

$$\langle M,N \rangle_t = \frac{1}{2}\big(\langle M+N \rangle_t - \langle M \rangle_t - \langle N \rangle_t\big)$$

symmetric bilinear form

A process X_t is a *continuous semimartingale* the sum of a continuous local martingale X_t and a finite variation process A_t.

Let two stochastic processes $\{X_t\}$ and $\{\widetilde{X}_t\}$ be indexed by a common set T. The two processes are *indistinguishable* if there exists a null set $N \subseteq \Omega$ such that for all $\omega \in \Omega - N$, it holds that

$$\widetilde{X}_t(\omega) = X_t(\omega) \quad \text{for all } t \in T.$$

The process $\widetilde{X}_t$ is said to be a *modification* of X_t if for each $t \in T$, it holds that

$$P(\omega : \widetilde{X}_t = X_t) = 1.$$

Modification means that we are modifying at each time instant, but indistinguishable means that the entire sample paths are indistinguishable with respect to the samples.

14.7 Kunita–Watanabe inequality. For two continuous local martingales M and N, and two measurable processes H and K, we have

$$\int_0^\infty |H_s|\,|K_s|\,|d\langle M,N \rangle_s| \le \left(\int_0^\infty H_s^2\, d\langle M \rangle_s\right)^{1/2} \left(\int_0^\infty K_s^2\, d\langle N \rangle_s\right)^{1/2}$$

almost surely.

A continuous local supermartingale that is bounded below is a true supermartingale.

Domination property: a continuous local martingale M such that $\sup_t |M_t| < Y$ for some $Y \in L^1$ is a uniformly integrable (true) martingale.

14.8 Burkholder–Davis–Gundy inequality. For $0 \le p < \infty$, there exists two absolute constants c_p and C_p such that for any continuous local martingale M, it holds that

$$c_p \, \mathrm{E}\big(\sup_t |M_t|\big)^p \le \mathrm{E}\langle M\rangle_\infty^{p/2} \le C_p \, \mathrm{E}\big(\sup_t |M_t|\big)^p.$$

In particular, this means that for a continuous local martingale M such that $\mathrm{E}\langle M\rangle_\infty^{1/2} < \infty$, $\mathrm{E}\sup_t |M_t| < \infty$, which implies that M is in fact a uniformly integrable martingale.

14.B Construction of stochastic integrals

14.B.1 The Brownian case

agrees with the Wiener integral

14.B.2 The L^2 martingale case

Itô's isometry

$$\mathrm{E}\bigg(\int H_s \, dM_s\bigg)^2 = \mathrm{E}\int H_s^2 \, d\langle M\rangle_s$$

14.9 Itô–Döblin formula. For n continuous semimartingales $X_1, \dots, X_n$ and $F \in C^2(\mathbf{R}^n)$, we have for all $t \ge 0$, it holds that

$$\begin{aligned} F(X_t^1, \dots, X_t^n) = F(X_0^1, \dots, X_0^n) &+ \sum_{j=1}^n \int_0^t \frac{\partial F}{\partial x^j}(X_s^1, \dots, X_s^n) \, dX_s^j \\ &+ \frac{1}{2} \sum_{j,k=1}^n \int_0^t \frac{\partial F}{\partial x^j \partial x^k}(X_s^1, \dots, X_s^n) \, d\langle X^j, X^k\rangle_s. \end{aligned}$$

Note that the Itô's formula is usually proved for functions F defined globally on $\mathbf{R}^n$. To make this in general applicable to an open subset U of $\mathbf{R}^n$ (which will occur in the context of PDEs), we need to introduce a continuous bump function and

14.10 Corollary. If we take $n = 2$ and $F(x, y) = xy$, then we have for two continuous semimartingales X and Y that

$$X_t Y_t = X_0 Y_0 + \int_0^t X_s \, dY_s + \int_0^t Y_s \, dX_s + \langle X, Y\rangle_t.$$

If we $\{X_t\}$ is a continuous local martingale, then the continuous martingale

$$X_t^2 - \langle X\rangle_t = 2\int_0^t X_s \, dX_s + \langle X\rangle_t.$$

Product rule

$$d(X_t Y_t) = X_t \, dY_t + Y_t \, dX_t + d\langle X, Y\rangle_t$$

$dX_t = \sigma(t, X_t) \, dB_t + b(t, X_t) \, dt$

14.11 Martingale representation theorem. Let $\mathcal{F}_t$ be the minimal completed filtration of a standard Brownian motion. For any random variable $Y \in L^2(\Omega, \mathcal{F}_\infty, P)$, there exists a unique progressive process $H \in L^2(B)$ such that

$$Y = \mathrm{E}Y + \int_0^\infty H_s \, dB_s.$$

Replacing Y by a not necessarily continuous true martingale bounded in L^2, then there exists a unique progressive process $H \in L^2(B)$ and some constant $C > 0$ such that

$$M_t = C + \int_0^t H_s \, dB_s.$$

The same claim still holds if Y is a continuous local martingale that is $L^2_{\mathrm{loc}}(B)$.

Given a Brownian motion B_t, its completed filtration $\mathcal{F}_t$ is automatically right-continuous.
All martingales with respect to $\mathcal{F}_t$ has not only a continuous modification, not just càdlàg.
change of measures right-continuous and complete filtration

14.12 Girsanov's theorem.

Novikov's condition. $\mathrm{E}\exp\left(\frac{1}{2}\langle L\rangle_\infty\right) < \infty$
Kazamaki's condition. L is a uniformly integrable martingale, and $\mathrm{E}\exp\left(\frac{1}{2}L_\infty\right) < \infty$
$\mathcal{E}(L)$ is a uniformly integrable martingale
For a continuous local martingale M and any real/complex number λ, we define the *stochastic exponential* (also known as *Doléans-Dade exponential*) of λM by

$$\mathcal{E}(\lambda M)_t = \exp\left(\lambda M_t - \frac{\lambda^2}{2}\langle M\rangle_t\right).$$

Note that since $\mathcal{E}(\lambda M)$ is bounded below, it is a continuous supermartingale, is a martingale if and only if $\mathrm{E}\mathcal{E}(\lambda M)_t = 1$. It is unique solution to the SDE

$$dZ_t = \lambda Z_t \, dM_t, \text{ where } X_0 = 0.$$

$$\mathcal{E}(X)_t \mathcal{E}(Y)_t = \mathcal{E}(X + Y + \langle X, Y\rangle)_t$$

For two continuous local martingales M and N, $\mathcal{E}(M)\mathcal{E}(N)$ is a continuous local martingale if $\langle M, N\rangle = 0$.

14.13 Theorem. For a continuous local martingale D that take strictly positive values, it has a *stochastic logarithm* L, in the sense that

$$D_t = \mathcal{E}(L)_t = \exp\left(L_t - \frac{1}{2}\langle L\rangle_t\right).$$

An explicit formula for L is given by

$$L_t = \log D_0 + \int_0^t \frac{1}{D_s} \, ds.$$

14.14 Lipschitz existence and uniqueness. Let $b\colon [0,\infty)\times\mathbf{R}^n \to \mathbf{R}^n$ and $\sigma\colon [0,\infty)\times\mathbf{R}^n \to \mathbf{R}^{n\times m}$ satisfy the global Lipschitz condition in the space parameter: for each $t \in [0,\infty)$, it holds that

$$\|b(t,x)-b(t,y)\|_2 + \|\sigma(t,x)-\sigma(t,y)\|_{\mathrm{F}} \le C\,\|x-y\|_2$$

for all $x, y \in \mathbf{R}^n$. (The matrix norm $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm, which is simply the 2-norm of the vector associated to the matrix.) Let B_t be a Brownian motion, and let $\mathcal{F}_t$ be its completed filtration. Assume ξ is an L^2 random variable independent of $\mathcal{F}_\infty$, then we have a unique pathwise solution to

$$\begin{cases} dX_t = \sigma(t,X_t)\,dB_t + b(t,X_t)\,dt, \\ X_0 = \xi. \end{cases}$$

linear growth condition is automatic when we have the global Lipschitz condition
strong Markov property of solution to SDE

14.15 Lévy's martingale characterization of Brownian motions. For a continuous process adapted to $\mathcal{F}_t$, the process X_t is a d-dimensional standard Brownian motion if and only if it is a continuous local martingale with

$$\langle X^j, X^k\rangle_t = \delta_{jk} t \quad \text{for all components } j \text{ and } k.$$

14.16 Yamada–Watanabe.

Geometric Brownian motions

$$dX_t = \sigma X_t\,dB_t + \mu X_t\,dt$$

Assume $X_0 > 0$. We differentiate $d\log X_t$ using the Itô's formula:

$$\begin{aligned} d\log X_t &= \frac{1}{X_t}\,dX_t - \frac{1}{2X_t^2}\,d\langle X\rangle_t \\ &= \sigma\,dB_t + \mu\,dt - \frac{1}{2X_t^2}\sigma^2 X_t^2\,dt \\ &= \sigma\,dB_t + \left(\mu - \frac{\sigma^2}{2}\right)dt \end{aligned}$$

Therefore

$$\log X_t - \log X_0 = \sigma B_t + \left(\mu - \frac{\sigma^2}{2}\right)t,$$

which gives

$$X_t = X_0 \cdot \exp\left(\sigma B_t + \left(\mu - \frac{\sigma^2}{2}\right)t\right).$$

There is one subtle issue with this approach. We need to show that $X_t > 0$ must holds for all $t \ge 0$, so that $\log X_t$ makes sense for all $t \ge 0$. localization trick

Assume $X_0 = 0$

Ornstein–Uhlenbeck process is the solution to the classical Langevin's equation

$$dX_t = \sigma dB_t - \lambda X_t\,dt \tag{14.17}$$

The explicit solution can be easily computed $d(e^{\lambda t} X_t)$ using the product rule:

$$X_t = X_0 e^{-\lambda t} + \sigma e^{-\lambda t} \int_0^t e^{\lambda s}\, dB_s.$$

Note that the second term is distributed as a Brownian motion indexed by t. It also has the name of stochastic convolution.

Let $\lambda = 1$ and $\sigma = \sqrt{2}$, we get

$$\begin{aligned} X_t &= e^{-t} X_0 + \sqrt{2} e^{-t} \int_0^t e^s\, dB_s \\ &\overset{\mathrm{D}}{=} e^{-t}\big(X_0 + \beta_{e^{2t}-1}\big), \end{aligned}$$

where β is a standard Brownian motion independent of X_0.

Now we perturb (14.17) and consider the overdamped Langevin's equation

$$dX_t = \sigma\, dB_t - [\lambda X_t + \nabla U(X_t)]dt$$

Linear equations

$$dX_t = C X_t\, dB_t + D X_t\, dt$$

n-dimensional squared Bessel processes

$$dX_t = 2\sqrt{X_t}\, dB_t + n\, dt$$

Exterior cone condition

a convex domain

Chapter 15 Special Topics

15.A Random matrices

15.A.1 The moment problem

15.A.2 Stieltjes transform

15.A.3 Ensembles

A Gaussian orthogonal ensemble (GOE)

A Gaussian unitary ensemble (GUE)

15.A.4 Asymptotic laws on the spectrum of random matrices

15.1 Semicircle law.

15.2 Marchenko–Pastur law.

$$d\mu_{\mathrm{MP}} = \frac{\sqrt{}}{}$$

15.3 Tracy–Widom law.

15.B Determinantal point processes

15.C Large deviation theory

We

15.D Mixing times of Markov chains

Let the state space S be finite.

Define the worst scenario distance between t-step and the stationary distribution by

$$\begin{aligned} d(t) &= \max_{\mu \in \mathcal{P}(S)} d_{\mathrm{TV}}(\mu Q_t, \pi) \\ &= \max_{x \in S} d_{\mathrm{TV}}\big(Q_t(x, \cdot), \pi\big). \end{aligned}$$

We define the *ϵ-mixing time* to be

$$t_{\mathrm{mix}}(\epsilon) = \inf\{t \geq 0 : d(t) \leq \epsilon\}$$

relaxation time
$d_{\mathrm{TV}}(\mu,\nu) \le d_{\mathrm{TV}}(\mu,\rho) + d_{\mathrm{TV}}(\nu,\rho)$
submultiplicativity of coupling distance

15.4 Theorem. $d(s+t) \le 2d(s)d(t)$

15.5 Fekete's lemma. For a subadditive function $f : [0,\infty) \to \mathbf{R}$, i.e.,

$$f(s+t) \le f(s) + f(t) \quad \text{for all } s,t > 0.$$

We have

$$\lim_t \frac{f(t)}{t} = \inf_{t>0} \frac{f(t)}{t} \in [-\infty,\infty).$$

The same result holds for a subadditive sequence of real numbers. See [Kal21, Lemma 25.19]

$$t_{\mathrm{mix}}(\epsilon) \ge t_{\mathrm{rel}} \log\left(\frac{1}{2\epsilon}\right)$$

The mixing time for symmetric random walks on the n-cycle $\mathbf{Z}_n$ is n^2.
The mixing time for random walks[1] on the boolean hypercubes $\{0,1\}^n$ is $n \log n$.

15.E Percolation

15.E.1 Bernoulli bond percolation

Define

$$\theta(p) = P_p(\text{the origin is contained in an infinite open cluster}),$$

and

$$p_c(d) = \sup\{p : \theta(p) = 0\}.$$

$\theta(p)$ is an increasing function in p
$p_c(d+1) \le p_c(d)$, and in fact the strict inequality holds for $d \ge 1$
$0 < p_c(p) < 1$ for $d \ge 2$
For any increasing $L^2(P_p)$ function, we have
Burton–Keane trifurcation argument

15.F Optimal transport

Proof of the dual representation of W_1. □

Monge's problem

15.G Local times

[1] If discrete-time, we can let the chain to be $1/2$-lazy

Epilogue

Appendices

A Helpful results from analysis and topology

A.1 Fact. The metric function $\rho\colon X \times X \to [0, \infty)$ on the space X is continuous.

A.2 Proposition. In a given (Hausdorff)[2] topological space X, a sequence $\{x_n\}$ converges to x if and only if every subsequence of x_n has a further subsequence that converges to x.

Proof. The only if direction is obvious. To prove the if direction, suppose $x_n \not\to x$ under the assumption. Let $n_0 = 1$. There is some (open) neighborhood U of x such that for every $k \in \mathbf{N}$, we can find a smallest $n_k \geq n_{k-1}$ such that $x_{n_k} \notin U$. However, this implies that the subsequence $\{x_{n_k}\}$ of $\{x_n\}$ does not have a subsequence that converges to x, which contradicts the assumption. □

A.3 Theorem [Mun00, Theorem 30.1]. Let X be a topological space, and A be a subset. If some $\{x_n\} \subseteq A$ converges to $x \in X$, then $x \in \overline{A}$. The converse is true when X is first countable.

Now let $f\colon X \to Y$. If function f is continuous, then for all sequences $x_n \to x$, we have $f(x_n) \to f(x)$. The converse is true when X is first countable.

A.4 Fact. Every real sequence has a monotonic subsequence.

A.5 Proposition. For an increasing function $f : \mathbf{R} \to \mathbf{R}$, the set of discontinuities is countable.

A.6 Proposition. Given a set A in a metric space (X, d), the function $d(\cdot, A)\colon X \to [0, \infty)$ given by

$$d(x, A) = \inf\{d(x, y) : y \in A\}$$

is a continuous function. Also $d(x, A) = 0$ if and only if $x \in \overline{A}$.

A.7 Proposition.

(a) Closed subspace of a complete metric space is complete.

(b) Complete subspace of any metric space must be closed.

A.8 Abel's theorem. Assume $S(x) = \sum_{n=0}^{\infty} a_n x^n$ converges, and let R be the radius of convergence

$$\frac{1}{\limsup_n |a_n|^{1/n}}.$$

If the series converges at $x = R > 0$, then the series converges uniformly over $[0, R]$. In particular this implies that $S(x)$ is continuous at R^-.

A.9 Proposition. Infinite subset of a compact set has a limit point.

[2] to ensure that the sequential limit must be unique; actually not necessary for this proposition

A.10 Proposition. Intersection of a closed set and a compact set is compact.

A.11 Proposition. Compact subsets of a Hausdorff space are closed.

A.12 Proposition. For $A \subseteq B \subseteq X$, where A and B are given the subspace topology of X. Then A is dense in X if and only if A is dense in X.

Note that A is dense in B means that $\overline{A} \supseteq B$.

A.13 Urysohn's lemma. Let X be normal. If A and B are two disjoint closed sets in X, then there exists a continuous function $f : X \to [0, 1]$ such that $f(B) = \{1\}$ and $f(A) = \{0\}$.

If X is a metric space (which is necessarily normal), then this is easy. We may just take

$$f(x) = \frac{d(x, A)}{d(x, A) + d(x, B)}.$$

Here is a sketch of the standard proof of this important result in topology. Based on normality, we may inductively dyadically choose (i.e., using DC) an increasing sequence of sets $U_{j/2^n}$ that "lie between" A and B:

$$A \subseteq U_{1/2^n}, \quad \ldots \quad , \overline{U}_{(j-1)/2^n} \subseteq U_{j/2^n}, \quad \ldots \quad , \overline{U}_{(2^n-1)/2^n} \cap B = \emptyset.$$

One can show that the function $f : X \to [0, 1]$ given by

$$f(x) = \begin{cases} \inf\{r : x \in U_r\} & \text{if the set is nonempty,} \\ 1 & \text{otherwise} \end{cases}$$

is continuous.

The use of DC can be avoided when X is second countable and regular, by the constructive proof of the following proposition.

A.14 Proposition. Every second countable regular space is normal.

A.15 Urysohn metrization theorem. Every second countable regular space is metrizable.

In particular, every lcscH space is metrizable.
[Fol99, Theorem 4/16, Corollary 4.17]

A.16 Tietze extension theorem. Let X be normal and $A \subseteq X$ be closed. For $f \in C(A)$, we can extend it to $F \in C(X)$ with $F|_A = f$.

application of Urysohn's lemma
in the case $X = \mathbf{R}$, a particular simple proof can be obtained as follows. The complement of A is a countable union of open intervals, and by continuously connecting all the endpoints of these intervals we may extend f to a continuous function on the real line.

A.17 Proposition.

(a) A second countable space is separable; the converse is also true when we are in a metric space.

(b) A second countable space is Lindelöf, the converse is also true when we are in a metric space.

A subspace of a Lindelöf space is not necessarily Lindelöf. Therefore it is sometimes useful to introduce the definition of a *hereditary Lindelöf* space, whose subspaces are all Lindelöf.

A.18 Fact. A second countable space is hereditary Lindelöf, since any subspace of a second countable space is second countable.

A.19 Fact. Closure of separable space is separable.

A.20 Theorem (Characterization of compactness in metric spaces). A subset of a metric space is compact if and only if it is sequentially compact if and only if it is totally bounded and complete.[3]

A.21 Proposition. Let $f, g\colon X \to Y$ be two continuous functions, where X is a topological space and Y is Hausdorff. If f and g agree on a dense subset of X, then $f = g$ on X.

A.22 Theorem. Let X and Y be metric spaces, with Y being complete. Let D be a dense subspace of X, and $f : D \to Y$ be a uniformly continuous function. Then there is a unique extension of f to $F\colon X \to Y$, such that F is still uniformly continuous.

Proof. Any $x \in X$ can be written as the limit of a sequence $\{x_n\} \subseteq D$. For each such sequence $\{x_n\}$, by uniform continuity it holds that for all $\epsilon > 0$, for all $m, n \in \mathbf{N}$ there exists $\delta > 0$ such that

$$|x_n - x_m| < \delta \implies |f(x_n) - f(x_m)| < \epsilon.$$

Since $\{x_n\}$ is a convergent sequence it also holds that there is some $N_\delta \in \mathbf{N}$ such that for all $m > n \geq N_\delta$, it holds that $|x_n - x_m| < \delta$. With these information combined, we get $\{f(x_n)\}$ is a Cauchy sequence in Y, which is complete. Therefore $\lim_n f(x_n)$ exists.

Now let us show that $\lim_n f(x_n) = \lim_n f(w_n)$ is the same for any $\{x_n\}$ and $\{w_n\}$ that approach x. We know $x_n - w_n \to 0$, and hence (using the same reasoning as above) $f(x_n) - f(w_n) \to 0$.

Now define $F(x) = \lim_n f(x_n)$ for any $\{x_n\}$. The function F is (sequentially) continuous everywhere. It is clear $F|_D = f$, and such an extension must be unique by Proposition A.21.

It remains to show that F is uniformly continuous. Consider $a, b \in X$, which are respectively limits of some $\{a_n\}$ and $\{b_n\}$ in D. We want to show that for any $\epsilon > 0$, for all $a, b \in X$, there exists $\delta > 0$ such that

$$|a - b| < \delta \implies |F(a) - F(b)| < \epsilon.$$

We leave it to the reader to use the uniform continuity of $F|_D$, $F(a) = \lim_n F(a_n)$, and the triangular inequality to meet the above inequality. □

This result is frequently used as one way to extend linear functionals $f \in D^*$ on the dense subspace D to the entire normed space X. Notice that linearity on the dense subspace carries easily over to the whole space, and if $\|f\| \leq C$, then $\|F\| \leq C$, by the continuity of F.

We emphasize X and D here have the same metric structure. Compare this result with the upcoming Hahn–Banach theorem.

A.23 Uniqueness theorem. Let G be a region (i.e., nonempty open connected subset of $\mathbf{C}$). If f and g are both holomorphic in G, and f and g agree on some $S \subseteq G$ that has a limit point in G, then f and g agrees everywhere on G.

A.24 Mean value inequality for $\mathbf{R}^d$-valued functions [Rud76, Theorem 5.19]. Let $f : [a, b] \to \mathbf{R}^d$ be continuous, and f be differentiable in (a, b), then there exists $x \in (a, b)$ such that

$$|f(b) - f(a)| \leq (b - a) \sup_{a<x<b} |f'(x)|.$$

[3] In the usual proof of totally bounded and complete $\implies$ compact, DC is used. The alternative proof that only requires CC is given in [Her06, Proposition 3.26].

Proof. Apply the ordinary mean-value theorem to the continuous $\varphi\colon [a,b] \to \mathbf{R}$ defined by

$$\varphi(t) = \langle f(b) - f(a), f(t) \rangle,$$

and use the Cauchy–Schwarz inequality. □

A.25 Mean value inequality for **C**-valued functions. Let f be defined on an open set containing the segment γ^* between z and z_0, and f be differentiable everywhere on γ^*. Then

$$\frac{|f(z) - f(z_0)|}{|z - z_0|} \leq \sup_{w \in \gamma^*} |f'(w)|.$$

Proof. This follows from the Fundamental theorem of calculus for parameterized paths and the Estimation lemma:

$$\begin{aligned} |f(z) - f(z_0)| &= \left| \int_\gamma f'(w)\, dw \right| \\ &\leq \sup_{w \in \gamma^*} |f'(w)| \cdot \operatorname{length}(\gamma) \\ &= \sup_{w \in \gamma^*} |f'(w)| \cdot |z - z_0|. \end{aligned}$$

□

A.26 Uniform convergence of derivatives [Rud76, Theorem 7.17][4]. Let $f_n\colon (a,b) \to \mathbf{R}$ be a sequence of differentiable functions that converges pointwise to f. If f_n' converges uniformly to some function g, then $f_n \to f$ uniformly and also $f' = g$.

The key part of the proof is the use of the mean value theorem on $f_n' - f_m'$.

A.27 Tychonoff's theorem. Arbitrary product of compact topological spaces is compact.

A.28 Theorem (Tychonoff's theorem for countable product). Countable product of compact topological spaces is compact.

Tychonoff's theorem is equivalent to the axiom of choice.
See discussion in [Her06, Section 4.8].
for countable product of compact metric space, only CC is needed
If the product is finite, then no choice is needed.

A.29 Exercise. Give a direct proof of Tychonoff's theorem for the countable product of compact metric spaces, using metrization.

A.30 Theorem. The countable product of sequentially compact spaces is sequentially compact.

B Normed spaces

Let X and Y be normed spaces in this section.

We use $\mathcal{L}(X,Y)$ for the space of linear maps between normed spaces X and Y, and we denote $\mathcal{L}(X,\mathbf{F})$ by X^*, called the dual space of X.

[4] Also see Theorem 8.15 and Remark 8.16 in [Kra22].

B.1 Fact. Let $(X, \|\cdot\|)$ be a normed vector space. Then vector addition $X \times X \to X$ and scalar multiplication $\mathbf{F} \times X \to X$ are both continuous. Also by the reverse triangular inequality,

$$\big|\|x\| - \|y\|\big| \leq \|x - y\|,$$

the norm function $\|\cdot\|$ is continuous with respect the topology generated by it.

B.2 Exercise. For a general metric space, one has $\overline{B(x;r)} \subseteq \overline{B}(x;r)$. Provide a example that shows that equality may not hold. (Hint: discrete metric.) Show that in addition that when the space is a normed vector space, then $\overline{B(x;r)} = \overline{B}(x;r)$.

B.3 Proposition. For $T \in \mathcal{L}(X, Y)$, then T is bounded if and only if Lipschitz continuous if and only if it is continuous if and only if it is continuous at any point of X.

Proposition A.6 When X is a normed space and A is a subspace, then $d(\cdot, A)$ is furthermore linear. Hence it is a continuous linear functional on X with kernel A.

B.4 Proposition. A normed space X is Banach if and only if for every sequence $\{x_n\} \subseteq X$ satisfying

$$\sum_n \|x_n\| < \infty,$$

the series $\sum_n x_n$ converges to some element of X in norm. (Every absolutely convergent series converges in the norm topology of X.)

This alternative criterion for completeness can be useful at times.

B.5 Proposition.

(a) For a normed space X and its closed proper subspace V, we can define a norm on the quotient space X/V by

$$\|[x]\|_{X/V} = \inf\{\|x - v\| : v \in V\},$$

where $[x]$ is the coset $x + V$. If X is Banach, then V is Banach as well.

(b) The topology induced by the quotient norm $\|\cdot\|_{X/V}$ is the same as the quotient topology on X/V.

(c) (Riesz' lemma) For any $\epsilon > 0$, there is some $x \in X$ with $\|x\| = 1$ satisfying

$$\|[x]\|_{X/V} \geq 1 - \epsilon.$$

B.6 Proposition. The closed unit ball is compact if and only if the Banach space is infinite-dimensional.

Therefore a Banach space is locally compact if and only if it is infinite-dimensional. Hence an infinite-dimensional separable Banach space is a Polish space that is not locally compact.

B.7 Theorem. The closed unit ball is compact in a normed space if and only if the normed space is finite-dimensional.

B.8 Fact. Let Y be a dense subspace of a normed space X, then Y^* and X^* can be isometrically identified in a natural way since a continuous function[5] is uniquely determined by its value on a dense subset.

[5] consider $f \in X^*$ and $\frac{|f(x)|}{\|x\|}$ in our context

B.9 Proposition. If Y is complete, then $\mathcal{L}(X, Y)$ is complete. In particular the dual space of any normed space is complete.

B.10 Hahn–Banach theorem. Let X be a real vector space, and p be a sublinear functional on X. Say E is a vector subspace of X, on which we have a linear functional $f \in E^*$. If $f(x) \leq p(x)$ for all $x \in E$ (f is dominated by p on the subspace), then we can extend f to a linear functional F defined on the entire space X, such that $F(x) \leq p(x)$ now holds for all $x \in X$.

Let X be a complex vector space, and p be a seminorm[6] on X. Say E is a vector subspace of X, on which we have a linear functional $f \in E^*$. If $|f(x)| \leq p(x)$ for all $x \in E$, then we can extend f to a linear functional F defined on the entire space X, such that $|F(x)| \leq p(x)$ now holds for all $x \in X$.

Let X be a real separable topological vector space, and p be a continuous sublinear functional, then the Hahn–Banach theorem can be proved in ZF without any choice. The term *topological vector space* will be clarified in Appendix D, but one can probably guess what it means.

In many applications, our p is automatically continuous (e.g., bounded linear functionals when X is a normed space). Also note that when p is a linear functional, then $|p|$ is a seminorm, and since p is continuous, $|p|$ must also be continuous. Hence with the separability topological assumption on X, most consequences of Hahn–Banach are retained.

The most significant consequence of the Hahn–Banach theorem is the existence of nontrivial linear functionals that satisfy certain properties.

B.11 Corollary. Let X be a normed space.

(a) Let V be a closed proper subspace of X. Take any $x \in X - V$, then there exists $f \in X^*$ such that $f(x) = \inf_{v \in V} \|x - v\| \neq 0$, $f|_V \equiv 0$, and $\|f\| = 1$.

(b) For $x \neq 0_X$, there exists $f \in X^*$ such that and $f(x) = \|x\|$ and $\|f\| = 1$.

(c) For any $f \in X^*$, there exists $x, y \in X$ such that $f(x) \neq f(y)$.

The hat map $\hat{}: X \to X^{**}$ such that $\hat{x}(f) = f(x)$ is an isometric injection. When the hat map is also surjective, the normed space X is called *reflexive*, which means exactly that we can always identify X with X^{**} as the same. Notice in particular that a reflexive space must be Banach because X^{**}, as a dual space, is complete under its norm.

For $A \subseteq X$, the *Minkowski functional/gauge* of A is defined by

$$p_A(x) = \inf\{r \in \mathbf{R} : r > 0 \text{ and } x \in rA\}$$

for all $x \in A$, where we take $\inf \emptyset = +\infty$ as usual.

We claim that p_A is continuous if and only if $0 \in \operatorname{Int} A$. If in addition A is convex, then p_A is a sublinear functional.

B.12 Uniform boundedness principle. [7] Let X be Banach and Y only be normed. For $\{T_\alpha\}_{\alpha \in A} \subseteq \mathcal{L}(X, Y)$, suppose $\sup_\alpha \|T_\alpha x\| < \infty$ for all $x \in X$, then $\sup_\alpha \|T_\alpha\| < \infty$.

B.13 Open mapping theorem. For two Banach spaces X and Y, if $T \in \mathcal{L}(X, Y)$ is surjective, then the map is open.

B.14 Corollary. For two Banach spaces X and Y, if $T \in \mathcal{L}(X, Y)$ is bijective, then the inverse T^{-1} is also a bounded linear map.

[6] Note that seminorms are always nonnegative, in contrast to sublinear functionals. The absolute value signs that pop up later are expected.

[7] or the Banach–Steinhaus theorem

B.15 Closed graph theorem. For two Banach spaces X and Y, if $T \in \mathcal{L}(X, Y)$ is closed, then the operator is bounded.

B.16 Baire category theorem. Every complete (pseudo)metric space is a Baire space, i.e., a space where a countable intersection of nowhere dense sets is nowhere dense. This implies that a complete metric space is not the countable union of nowhere dense sets.

The above result also holds for all locally compact regular spaces, which includes locally compact Hausdorff spaces.

It is a well-known fact that Baire category theorem for complete metric space is equivalent to DC. However, a Polish space is Baire can be proven in ZF; see [Her06, Theorem 4.102]. Also, it is shown in [Fel17] that only CC is needed to establish the uniform boundedness principle.

B.17 Proposition. A closed and countable nonempty subset of a complete metric space X must have an isolated point.

Proof. If X have no isolated point, then every singleton $\{x\} \subseteq X$ is nowhere dense, which implies that X is a countable union of nowhere dense set. □

C Hilbert spaces

A *Hilbert space* is an inner space with a complete metric induced from the inner product. We assume the underlying field is $\mathbf{C}$ for this section.

C.1 Proposition. An inner product space (resp. Hilbert space) is a normed space (resp. Banach space) with the *parallelogram law/polarization identity*:

$$\|x - y\|^2 + \|x + y\|^2 = 2\|x\|^2 + 2\|y\|^2 \quad \text{for all } x \text{ and } y.$$

C.2 Cauchy–Schwarz inequality. On an inner product space V, we have

$$|\langle u, v\rangle| \leq \|u\|\|v\|,$$

with equality if and only if one is a scalar multiple of the other.

Proof. Expand the nonnegative expression $f(\lambda) := \|u + \lambda v\|^2$ for all $\lambda \in \mathbf{R}$, which contains the desired real part of the inner product and has discriminant ≤ 0. After getting

$$|\mathrm{Re}\langle u, v\rangle| \leq \|u\|\|v\|,$$

replace u by $\frac{|\langle u,v\rangle|}{\langle u,v\rangle}$[8]. □

With the additional topological assumption that Hilbert spaces have complete metric, most of the results for finite-dimensional inner product spaces carry over to infinite dimensional Hilbert spaces. To motivate the upcoming results, it is recommended to review their finite-dimensional analogs, and understand why these results should be true.

C.3 Projection theorem. Given a Hilbert space H and a closed convex subset Y,

[8] This change-of-direction trick is a prevalent trick to extend results proved over real vector spaces to over complex vector spaces

(a) for each $x \in H$ there exists a unique

$$y = \arg\min_{z \in Y} \|x - z\|,$$

which we call the *projection* of x to Y, denoted by $\pi_Y(x)$.

Moreover, the projection $y = \pi_Y(x)$ is characterized by the property

$$\operatorname{Re}\langle x - y, z - y\rangle \le 0 \quad \text{for all } z \in Y. \tag{C.4}$$

(b) if Y is furthermore a closed subspace of H, then the characterization above for $\pi_Y(x)$ may be further replaced by

$$\langle x - y, z\rangle = 0 \quad \text{for all } z \in Y. \tag{C.5}$$

Proof.

(a) Let $D = \inf_{z \in Y} \|x - z\|$, and since Y is close, we may choose a sequence $\{y_n\}$ such that $\|x - y_n\| \to D$ from above. Our goal is to show that it is a Cauchy sequence, and hence converges.

For $n > m \ge 1$, by the parallelogram law we have

$$\|y_n - y_m\|^2 = 2\|x - y_n\|^2 + 2\|x - y_m\|^2 - 4\left\|x - \frac{y_n + y_m}{2}\right\|^2.$$

Since $\frac{y_n + y_m}{2} \in Y$ by convexity, we have

$$\|y_n - y_m\|^2 \le 2\|x - y_n\|^2 + 2\|x - y_m\|^2 - 4D^2.$$

It follows that as $n, m \to \infty$, $\|y_n - y_m\| \to 0$, as desired. Since closed subset of a complete metric space is complete, y_n should converges to some $y \in Y$. By $\|x - y_n\| \to \|x - y\|$ we conclude that $\|x - y\| = D$.

To show the uniqueness of y: for two y and y' that attains the infimum D, use the parallelogram law again we have

$$\begin{aligned}\|y - y'\|^2 &= 2\|x - y\|^2 + 2\|x - y'\|^2 - 4\left\|x - \frac{y + y'}{2}\right\|^2 \\ &\le 2D^2 + 2D^2 - 4D^2 = 0.\end{aligned}$$

Now we want to show this y satisfies (C.4). Let $z \in Y$ be arbitrary. To get (the real part of) the inner product[9] we consider the expression

$$f(\lambda) := \|\lambda(z - y) - (x - y)\|^2 = \|y + \lambda(z - y) - x\|^2.$$

For all $\lambda \in [0, 1]$, by convexity $y + \lambda(z - y) \in Y$, and hence $f(\lambda) \ge \|x - y\|^2$. Now expanding $f(\lambda)$ gives us

$$\lambda^2\|z - y\|^2 - 2\lambda \operatorname{Re}\langle x - y, z - y\rangle \ge 0.$$

Hence

$$\lambda\|z - y\|^2 \ge 2\operatorname{Re}\langle x - y, z - y\rangle \quad \text{for all } \lambda \in [0, 1],$$

[9] like in the proof of Cauchy–Schwarz

and take $\lambda \to 0^+$ gives us (C.4).

For the converse, now suppose (C.4) holds for some $y \in Y$, and we want to show

$$\|x - y\| \le \|x - z\| \quad \text{for all } z \in Y.$$

We trace our steps back: first,

$$2\,\mathrm{Re}\langle x - y, z - y\rangle \le 0 \le \|z - y\|^2.$$

It follows that

$$\|x - y\|^2 \le \|(z - y) - (x - y)\|^2,$$

as desired.

(b) To show the second part, it suffice to prove that (C.4) and (C.5) are equivalent. Because Y is now a subspace of H, equation (C.4) is equivalent to

$$\mathrm{Re}\langle x - y, z\rangle = 0 \quad \text{for all } z \in Y.$$

Notice that

$$\mathrm{Im}\langle x - y, z\rangle = \mathrm{Re}\,{-i}\langle x - y, z\rangle = \mathrm{Re}\langle x - y, iz\rangle,$$

which completes the proof. □

C.6 Proposition. For H and its closed subspace Y, π_Y has the following properties:

(a) $\pi_Y \in \mathcal{L}(H)$;
(b) $\pi_Y^2 = \pi_Y$;
(c) $\mathrm{range}\,\pi_Y = Y$ and $\mathrm{null}\,\pi = Y^\perp$;
(d) $\|\pi_Y(x)\| \le \|x\|$ for all $x \in H$.

C.7 Riesz representation theorem (Hilbert space). For each linear functional $f \in H^*$, there exist a unique $v \in H$ such that

$$f(x) = \langle x, v\rangle \quad \text{for all } x \in H.$$

Moreover $\|f\| = \|v\|$, and hence we have a isometric isomorphism between H^* and H.

An *orthonormal system* $\{e_\alpha\}_{\alpha\in A}$ is a possibly infinite collection of vectors such that

$$\langle e_\alpha, e_\beta\rangle = \begin{cases} 1 & \alpha = \beta, \\ 0 & \alpha \ne \beta. \end{cases}$$

The order of α does not matter when A is countable.

C.8 Proposition. Suppose we have a finite orthonormal system $\{e_j\}_{j=1}^n$ that spans Y. If $Y \subseteq H$. Then the projection of any $x \in H$ is explicitly $\pi_Y(x) = \sum_{j=1}^n \langle x, e_j\rangle e_j$.

C.9 Proposition. $\sum_{\alpha\in A}\langle x, e_\alpha\rangle e_\alpha$

C.10 Theorem. Let $\{e_\alpha\}_{\alpha\in A}$ be an orthonormal system, then

(a) $\sum_{\alpha\in A}\langle x, e_\alpha\rangle^2 \le \|x\|^2$, which is known as *Bessel's inequality*;
(b) the equality above holds if and only if the series $x = \sum_{\alpha\in A}\langle x, e_\alpha\rangle e_\alpha$ in H.

Orthonormal decomposition
Parseval's identity
Gram–Schmidt process

C.11 Theorem (complete orthonormal system). $\{e_\alpha\}_{\alpha\in A}$ is an orthonormal basis of H if and only if $\operatorname{span}\{e_\alpha\}$ is dense in H.

C.12 Theorem. H has a countable orthonormal basis if and only if H is separable. Additionally in this case, all bases have the same cardinality.

D Weak topologies and topological vector spaces

Some motivation is needed before we start the main material of this section.

$f: X \to Y$ is continuous if and only if for every $x_\alpha \to x$, we have $f(x_\alpha) \to f(x)$.

A related results $x_\alpha \to x$ in the initial topology on X generated by $\mathcal{F} = \{f_\beta : X \to Y_\beta\}_{\beta\in B}$ if and only if $f(x_\alpha) \to f(x)$ for all $f \in \mathcal{F}$. This is true for both nets and sequences.

convergence in product spaces

If the target spaces Y_β's are all Hausdorff, then X is Hausdorff if and only if the collection $\mathcal{F}$ separates points in X.

Whenever you see "separates points" below, it ensures the imposed topology on to be Hausdorff.
Sequential limits are unique

The subbasis of $\mathcal{F}$ can be specified by $f_\beta^{-1}(V)$, where V ranges over any open sets of Y_β, for any $\beta \in B$. One may take Y to be any basic or subbasic open set as well, by the property of the preimage. If $\mathcal{F}$ consists of only one function f, then the preimage f^{-1} takes (subbasic/basic) open sets in Y precisely to (subbasic/basic) open sets in X.

Suppose we have two vector spaces X and Y. We say X and Y are in duality if there is a bilinear pairing $\langle\cdot,\cdot\rangle: X \times Y \to \mathbf{F}$. Assume also that Y separates points in X, which means that for each $x \neq 0_X$, there exists some $y \in Y$ such that $\langle x, y\rangle \neq 0$, since we are in the setting of vector spaces. We assign a topology $\sigma(X, Y)$ to X, known as the *weak topology*, the weakest topology that makes the collection of mappings

$$\{x \mapsto \langle x, y\rangle : y \in Y\}$$

continuous. If X also separates points in Y, then the pairing $(X, Y, \langle\cdot,\cdot\rangle)$ is called a dual pairing.

Bogachev 1.6.5 6 8

We need a new type of convergence on vector spaces

$x_n \to x$ weakly (i.e., converges in the weak topology) if and only if for all $f \in X^*$, $f(x_n) \to f(x)$

$f_n \to f$ weakly (i.e., converges in the weak-star topology) if and only if for all $x \in X$, $\hat{x}(f_n) = f_n(x) \to \hat{x}(f) = f(x)$

The basis for $\sigma(X, X^*)$ is usually expressed in the following explicit way.

For any $x_0 \in X$, a neighborhood basis for x_0 is given by

$$\bigcap_{j=1}^{n} f_j^{-1}\big(f_j(x_0) - \epsilon, f_j(x_0) + \epsilon\big),$$

or equivalently,

$$\big\{x \in X : |f_j(x - x_0)| < \epsilon \text{ for all } j \in [n]\big\},$$

for any finite number of f_j's and $\epsilon > 0$.

You push x_0 to the target field $\mathbf{F}$, vary $f_j(x_0)$ in a small neighborhood in $\mathcal{F}$, and then push back to X to get a neighborhood for x_0.

The weak and weak-star topology can alternatively be seen as *seminorm topologies*, which we discuss here. Say X is a vector space, on which we have $\{p_\alpha\}_{\alpha\in A}$ as a family of seminorms that separates points in X. The *topology on X generated by* $\{p_\alpha\}$ is the initial topology with respect to the family of functions

$$\{x \mapsto p_\alpha(x - x_0) : x_0 \in X, \alpha \in A\}.$$

The seminorms we used to define the weak topology on X are $\{|f_\alpha| : f_\alpha \in X^*\}$.

Be very careful that this is *not* the initial topology that makes all $p_\alpha(\cdot)$ continuous. Rather, due to the vector space structure of X, the translation by y in the functions $x \mapsto p_\alpha(x - y)$ is an important requirement, such that $(x, y) \mapsto x + y$ and $(\lambda, x) \mapsto \lambda x$ are continuous. A vector space with a Hausdorff topology that makes vector addition and scalar multiplication continuous is called a *topological vector space*, which we have mentioned earlier.

A topological vector space X is *locally convex* if every neighborhood of 0 contains a convex neighborhood of 0. The topology on a vector space induced from seminorms is locally convex because the neighborhood basis is made of locally convex sets

$$\{x \in X : p_j(x) < \epsilon \text{ for all } j \in [n]\}$$

for any finite number of p_j's and $\epsilon > 0$. In fact more surprisingly, all locally convex topology can be generated by a family of seminorms, using the Minkowski functional. For details of the two equivalent characterizations of locally convex spaces, see [BS20, Section 8.1].[10] Hence we have generalized a very wide class of topological vector spaces from weak and weak-star topologies on normed spaces.

If the number $|A|$ of seminorms p_α used to generated the locally convex topology on X is countable, then the topology on X is metrizable with

$$d(x, y) := \sum_{j=1}^{\infty} 2^{-j} \frac{p_j(x - y)}{1 + p_j(x - y)}.$$

The converse of this statement is also true. The proof of this equivalence again can be found in [BS20, Proposition 8.6.1] Note that if (X, d) is complete, the locally convex space is called a *Fréchet space*. The *Schwartz space* of rapidly decreasing functions $\mathcal{S}(\mathbf{R}^n)$ useful in Fourier analysis is the primary example.

Given two normed spaces X and Y, we are already familiar that we can assign a norm topology to the vector space $\mathcal{L}(X, Y)$. With all our previous discussions, it is possible to assign two other topologies to $\mathcal{L}(X, Y)$.

First, we have the *strong operator topology* generated by the seminorms

$$T \mapsto \|Tx\| \text{ over } x \in X.$$

Hence $T_n \to T$ in the strong operator topology if and only if $T_n x \to Tx$ in Y-norm for all $x \in X$. Clearly the limit T is unique since Tx is uniquely determined for all x.

Second,we have the *weak operator topology* on $\mathcal{L}(X, Y)$ generated by the seminorms

$$T \mapsto f(Tx) \text{ over } x \in X, f \in Y^*.$$

[10] Some authors ask the convex neighborhoods to be *balanced*, i.e., $\alpha U \subseteq U$ for any $|\alpha| \leq 1$ in the definition. One may safely drop this assumption, which is also discussed in the reference.

Therefore $T_n \to T$ in the weak operator topology if and only if for all $x \in X$ and $f \in Y^*$, $f(T_n x) \to f(Tx)$, which is equivalent to saying that $T_n x \to Tx$ weakly in Y for all $x \in X$. Since the weak limit in Y is unique, T is unique.

The norm topology on $\mathcal{L}(X, Y)$ is stronger than strong operator topology, which is again stronger than the weak operator topology.

D.1 Proposition. Weak and weak-star topologies are Hausdorff (for different reasons). In fact, one can further show that weak-star topologies are completely regular.

Proof. The weak topology is Hausdorff because continuous linear functionals separates points. □

There is only one topology that one can assign to a finite-dimensional vector space such that vector addition and scalar multiplications become continuous.

D.2 Proposition [Rud91, Theorem 1.21]. A real/complex topological vector space X of finite dimension n is homeomorphic to $\mathbf{R}^n/\mathbf{C}^n$ with the Euclidean topology.

Proof. Consider the real case. We have a linear isomorphism T from $\mathbf{R}^n$ to X by identifying the standard basis elements $e_1, \ldots, e_n$ of $\mathbf{R}^n$ with a basis $x_1, \ldots, x_n$ of X. For $a = (a_1, \ldots, a_n) \in \mathbf{R}^n$, we have

$$T(a) = a_1 x_1 + \ldots + a_n x_n.$$

The coordinate projections $a \mapsto a_j$ are of course continuous, and since addition and scalar multiplications are both continuous, T is continuous.

Showing that T^{-1} is continuous requires more work. □

.

D.3 Proposition [Rud91, Theorem 1.22]. A topological vector space is locally compact if and only if it is finite-dimensional.

D.4 Proposition. For a normed vector space, weak topology is always weaker than the norm topology. Furthermore, the weak topology is strictly weaker than the norm topology if and only if the space is infinite-dimensional.

Proof. First, weak convergence is weaker than norm convergence, since

$$\|f(x_n) - f(x)\| \leq \|f\| \|x_n - x\|$$

for all $f \in X^*$. Therefore the weakest topology that makes all linear functionals continuous is weaker than the norm topology.

It suffices to show that all weakly open sets are norm-unbounded, which can be further reduced to showing that any neighborhood basis

$$U = \bigcap_{j=1}^{n} \{x : |f_j(x)| < \epsilon\}$$

around 0_X is unbounded in norm. Consider the linear map $F : X \to \mathbf{F}^n$ given by

$$F(x) = \big(f_1(x), \ldots, f_n(x)\big).$$

Note that $F^{-1}(\{0\})$ is a subspace of the considered neighborhood basis U. Hence if U is norm bounded then $F^{-1}(\{0\})$ must only contain 0. However, the injective linear map F cannot map an infinite-dimensional space X to a finite-dimensional one. □

D.5 Proposition. Suppose $x_n \to x$ weakly, then $\sup_n \|x_n\| < \infty$, and $\|x\| \le \liminf_n \|x_n\|$.

D.6 Proposition [Fol99, Proposition 5.17]. For $\{T_n\} \subseteq \mathcal{L}(X, Y)$ with $\sup_n \|T_n\| < \infty$. If for some $T \in \mathcal{L}(X, Y)$, we have $\|T_n x - Tx\| \to 0$ on for all $x \in D$ dense in X, then $T_n \to T$ in the strong operator topology.

D.7 Sequential Banach–Alaoglu theorem. For a separable normed vector space X, the closed unit ball in X^* is weak-star sequentially compact. This means precisely that for any normed bounded sequence in X^*, it has a subsequence that is weak-star convergent to some $F \in X^*$ with the same norm bound.

close connection to Helly selection theorem

Proof. Let $\{f_n\} \subseteq X^*$ be norm bounded by some positive constant C, and take a countable dense subset $\{x_j\}$ of X. We follow the diagonalization procedure. Since $\sup_n |f_n(x_1)| \le C\|x_1\|$, $\{f_n(x_1)\}$ lives in a bounded interval, there is a subsequence $\{f_{\nu(n)}(x_1)\}$ that converges.[11] Let $\{f_n^1\} = \{f_{\nu(n)}\}$, and we can now extract a further subsequence $\{f_n^2\}$ from $\{f_n^1\}$ such that f_n^2 converges on $\{x_1, x_2\}$. Proceeding inductively, we get the following table of subsequences listed in rows:

Table 1: subsequences listed in rows

$$\begin{array}{ccccc} f_1^1 & f_2^1 & f_3^1 & f_4^1 & \cdots \\ f_1^2 & f_2^2 & f_3^2 & f_4^2 & \cdots \\ f_1^3 & f_2^3 & f_3^3 & f_4^3 & \cdots \\ f_1^4 & f_2^4 & f_3^4 & f_4^4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

Take the diagonal sequence $f_1^1, f_2^2, \ldots$. If we ignore the first $j-1$ terms of the diagonal sequence, this new $\{f_n^n\}$ is a subsequence of $\{f_n^j\}_{n=1}^\infty$. Therefore f_n^n converges on the dense subset $\{x_j\}$ of X. We need to show that the convergence in fact holds on the entire space X.

(One may want to proceed using Theorem A.22, but unfortunately this does not work because the dense subset might not contain 0.) Take any $x \in X$, for any $\epsilon > 0$ there exists some x_j such that $\|x - x_j\| < \epsilon$, which implies that

$$|f_n^n(x) - f_n^n(x_j)| < C\epsilon \quad \text{for all } n.$$

Now

$$f_n^n(x_j) - C\epsilon \le f_n^n(x) \le f_n^n(x_j) + C\epsilon$$

Let f satisfy $f(x_j) = \lim_n f_n^n(x_j)$ for all j, then taking limits we have

$$f(x_j) - C\epsilon \le \liminf_n f_n^n(x) \le \limsup_n f_n^n(x) \le f(x_j) + C\epsilon.$$

It follows that

$$\limsup_n f_n^n(x) - \liminf_n f_n^n(x) \le 2C\epsilon,$$

[11] This can be done explicitly by letting

$$\nu(n) = \min\{m > \nu(n-1) : |f_m(x_1) - s_1| < 1/n\},$$

where $s_1 := \liminf_n f_n(x_1)$.

and since ϵ is arbitrary, $f(x) = \lim_n f_n^n(x)$ for all $x \in X$. We then know f should be linear, and also that

$$|f(x)| = \lim_n |f_n^n(x)| \leq C.$$

for $x \in X$ with unit norm, which shows that $f \in X^*$ with $\|f\| \leq C$, as desired. □

D.8 Banach–Alaoglu theorem. For a normed vector space X, every closed and bounded-in-norm subset of X^* is weak-star compact.

metrizability

D.9 Exercise. The converse of Banach–Alaoglu theorem is also correct when X is a Banach space. (Hint: use the uniform boundedness principle)

A set S in a vector space is called *balanced* if $\lambda S \subseteq S$ for all $|\lambda| \leq 1$.

E Some relevant operator theory

The current section only covers the mere basics of operator theory useful to the study of stochastic processes. In particular, we will discuss adjoint and unbounded operators on Banach and Hilbert spaces, but completely omit compact operators and spectral theory.

E.1 Theorem. For $T \in \mathcal{L}(H)$, there is a unique $T^* \in \mathcal{L}(H)$ such that $\langle Tx, y\rangle = \langle x, Ty\rangle$ for all $x, y \in H$. This T^* is known as the *adjoint* of T, which has the following properties:

(a) $\|T^*\| = \|T\|$, $\|T^*T\| = \|T\|^2$, $T^{**} = T$,

$$(\operatorname{range} T)^\perp = \operatorname{null} T^* \quad \text{and} \quad (\operatorname{null} T)^\perp = \overline{\operatorname{range} T^*}.$$

unitary operators
annihilators

E.2 Closed range theorem.

F Semigroups

G Convex geometry, optimization, and analysis

Let X be a nonempty vector or topological space, and let $f : X \to \overline{\mathbf{R}}$ throughout this section.

The *epigraph* of f, denoted by epi f, is the set

$$\{(x, y) \in X \times \mathbf{R} : y \geq f(x)\},$$

the set of all points lying on or above the graph of the function.

G.1 Fact. Let X be convex. The function f is convex if and only if its epigraph is convex.

A function $f : X \to (-\infty, +\infty]$ is *lower-semicontinuous* at $a \in X$ if for all $y < f(a)$, we have an open neighborhood U_a such that $y < f(x)$ for all $x \in U_a$. Equivalently this means

$$\liminf_{x \to a} f(x) \geq f(a).$$

Instead, the function $f : X \to [-\infty, +\infty)$ is *upper-semicontinuous* at $a \in X$ if for all $y > f(a)$, we have an open neighborhood U_a such that $y > f(x)$ for all $x \in U_a$. Equivalently this means

$$\limsup_{x \to a} f(x) \leq f(a).$$

We say the function f is *lower-semicontinuous* (LSC) or *upper-semicontinuous* (USC) if the function it is pointwise LSC/USC. Because of symmetry we will focus on LSC functions from now on.

A function is LSC if and only if

(a) $f^{-1}(-\infty, c]$ is closed for all $c \in \mathbf{R}$;

(b) $f^{-1}(c, +\infty]$ is open for all $c \in \mathbf{R}$;

(c) epi f is a closed in $X \times \mathbf{R}$.

geometric consequence of the Hahn–Banach theorem theorem. Let X be a real topological vector space, a hyperplane is a set

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

for some linear functional f and $t \in \mathbf{R}$. It is a codimension-1 affine subspace, and one can show that

G.2 Fact. A hyperplane is closed if and only if the f is a continuous linear functional.

A hyperplane $\{x \in X : f(x) = t\}$ separates two sets $A, B \subseteq X$ if

$$f(x) \leq t \text{ for all } x \in A \quad \text{and} \quad f(x) \geq t \text{ for all } x \in B.$$

The hyperplane strictly separates A and B if

$$f(x) \leq t - \epsilon \text{ for all } x \in A \quad \text{and} \quad f(x) \geq t + \epsilon \text{ for all } x \in B.$$

G.3 Hyperplane separation theorem. Let X be a finite-dimensional real vector space, and A and B be disjoint convex subsets. Then there is a hyperplane that separates A and B.

Notice that such a hyperplane must be closed because the algebraic dual and continuous dual space coincides in the finite-dimensional case.

G.4 Hyperplane separation theorem. Let X be an infinite-dimensional real topological vector space. For two disjoint convex sets A and B in X, if

(a) A is open, then there is a closed hyperplane that separates A and B.

(b) A is closed and B is compact, then there is a closed hyperplane that strictly separates A and B.

See [Bre11, Chapter 1] for details.

G.5 Fenchel–Rockafellar theorem.

We already know convex sets. A is an *affine set* if for all $\lambda \in \mathbf{R}$ and $x, y \in X$,

$$(1-\lambda)x + \lambda y \in X.$$

Different from a convex set, an affine set must contain each line through any two points within, not just the line segment. The vector subspaces of $\mathbf{R}^d$ are precisely the affine subspaces of $\mathbf{R}^d$ containing 0.

Given a vector space X and a subset A, a point $p \in A$ is called an *extreme point* of A if it is on any line connecting two distinct points. This means precisely there does not exist $x \neq y$ in A such that

$$p \neq (1-\lambda)x + \lambda y \quad \text{for any } 0 < \lambda < 1.$$

Given a set of points S in a vector space X, the *convex hull* conv S is the smallest set in X that contains S. Equivalently it can be explicitly written as all finite sums $\sum_{j=1}^n \lambda_j x_j$, where $x_j \in S$, $0 \leq \lambda_j \leq 1$, and $\sum_{j=1}^n \lambda_j = 1$. If X is a topological vector space, then the *closed convex hull* (resp. *open convex hull*) is the closure (resp. interior) of the conv S.

We can define *affine hull* similarly, without restricting λ_j to be nonnegative.

G.6 Krein–Milman theorem. A compact convex subset of a locally convex topological vector space is equal to the closed convex hall of its extreme point.

A set $C \subseteq X$ is a cone if $x \in C$ implies $\lambda x \in C$ for all $\lambda > 0$.

G.7 Radon's theorem. Any set of $d+2$ points in $\mathbf{R}^d$ can be partitioned into two subsets whose convex hulls intersect.

G.8 Carathéodory's theorem. Given some set $S \subseteq \mathbf{R}^d$, for any point in conv S, it is the convex combination of at most $d+1$ points of S.

G.9 Helly's theorem. Let $A_1, A_2, \ldots, A_n$ be convex subsets of $\mathbf{R}^d$, where $n \geq d+1$. If every $d+1$ number of A_γ's have nonempty intersection, then the intersection of the whole collection $\bigcap_\gamma A_\gamma \neq \emptyset$.

The result remains in force if we let $\{A_\gamma\}_{\gamma\in\Gamma}$ be an (infinite) indexed family of compact convex subsets of $\mathbf{R}^d$. This case follows by the finite intersection characterization of compactness. (Fix one A' in the collection, and replace each A_γ by $A_\gamma \cap A'$.)

G.10 Lemma. For $F\colon X \times Y \to [-\infty, \infty]$, we have

$$\sup_{x\in X} \inf_{y\in Y} F(x,y) \leq \inf_{x\in X} \sup_{y\in Y} F(x,y).$$

H Hausdorff measures

Let (X, ρ) be a metric space, and $E \subseteq X$. For every $\alpha \geq 0$ and $\epsilon > 0$, we define

$$H^\alpha_\epsilon(E) = \inf\left\{\sum_{j=1}^\infty (\operatorname{diam} A_j)^\alpha : \sup_j(\operatorname{diam} A_j) \leq \epsilon \text{ and } \{A_j\} \text{ covers } E\right\}.$$

When $\alpha = 0$, $H^0_\epsilon(E)$ is just the *covering number* of E, i.e., the smallest cardinality for an ϵ-net of E. We also consider the case where $\epsilon = \infty$, which means that there is no restriction on the diam A_j's.

Notice that H_ϵ^α is increasing as ϵ decreases to 0. We define the *α-Hausdorff measure* of E to be

$$H^\alpha(E) = \sup_{\epsilon>0} H_\epsilon^\alpha(E) = \lim_{\epsilon\to 0^+} H_\epsilon^\alpha(E).$$

Notice that we get a Carathéodory outer measure.

We define the Hausdorff dimension of E by

$$\dim_{\mathrm{H}} E = \inf\{\alpha : H^\alpha(E) = 0\} = \inf\{\alpha : H_\infty^\alpha(E) = 0\}.$$

The ternary Cantor set on $[0, 1]$ has Hausdorff dimension $\frac{\log 2}{\log 3}$.

When f is Lipschitz, $\operatorname{diam}_H f(E) \le \operatorname{diam} E$, since the diameter of a set is stretched by at most a constant under a Lipschitz map. More generally, one can show as an exercise that for f that is α-Hölder continuous with constant C, we have

$$H^\beta\big(f(E)\big) \le C^\beta H^{\alpha\beta}(E),$$

which gives

$$\operatorname{diam} f(E) \le \frac{1}{\alpha} \operatorname{diam} E.$$

(This is the reason why we chose α for the exponent in the definition of Hausdorff measures.)

Minkowski dimension

I Topological groups and Haar measures

compact groups are unimodular

left and right Haar measures agree precisely when the group is unimodular

J Proof of the two extension theorems

J.1 Dynkin's π-λ theorem. Within a nonempty set X, if $\mathcal{P}$ is a π-system that is contained in a λ-system $\mathcal{L}$, then $\sigma(\mathcal{P}) \subseteq \mathcal{L}$.

Proof. Let $\Gamma = \lambda(\mathcal{P})$, the λ-system that contains $\mathcal{P}$ (see Definition 1.9).

We then need to show Γ is a σ-algebra. Once this has been shown, we can claim that $\sigma(\mathcal{P}) \subseteq \Gamma \subseteq \mathcal{L}$, which finishes the proof. To prove Γ is a σ-algebra, we need the key fact that Γ is in fact a π-system, i.e., for $E \in \Gamma$ and $F \in \Gamma$, we wish to prove $E \cap F \in \Gamma$.

Here is the major trick. Define

$$\mathcal{K}_E = \{F \subseteq X : E \cap F \in \Gamma\} \tag{J.2}$$

for any $E \in \Gamma$. We show that $\mathcal{K}_E$ is a λ-system for any fixed $E \in \Gamma$.

First, $X \in \mathcal{K}_E$ since for $E \in \Gamma$, $E \cap X = E \in \Gamma$. Next for $A \subseteq B$ in $\mathcal{K}_E$, $E \cap A \subseteq E \cap B$ are both in Γ. Therefore

$$\begin{aligned} E \cap (B - A) &= E \cap (B \cap A^{\mathrm{c}}) \\ &= (E \cap B) \cap (E \cap A)^{\mathrm{c}} \\ &= E \cap B - E \cap A \in \Gamma, \end{aligned}$$

which proves that $F - E \in \mathcal{K}_E$. Finally for the ascending sequence of sets $A_1 \subseteq A_2 \subseteq \cdots$ in $\mathcal{K}_E$, we have

$$E \cap \left(\bigcup_{j=1}^{\infty} A_j \right) = \bigcup_{j=1}^{\infty} (E \cap A_j).$$

Since $E \cap A_j \in \Gamma$ for all $j \in \mathbf{N}$ and

$$E \cap A_j \uparrow \bigcup_{j=1}^{\infty} (E \cap A_j) \quad \text{as } j \to \infty,$$

we have $\bigcup_{j=1}^{\infty} A_j \in \mathcal{K}_E$. Hence we have proved that $\mathcal{K}_E$ is a λ-system for any $E \in \Gamma$.

Now we restrict our attention to $E \in \mathcal{P}$. Since $\mathcal{P}$ is closed under finite intersections, we have $\mathcal{P} \subseteq \mathcal{K}_E$, and therefore $\lambda(\mathcal{P}) = \Gamma \subseteq \mathcal{K}_E$. In summary, we have

$$E \in \mathcal{P} \text{ and } F \in \Gamma \Rightarrow E \cap F \in \Gamma.$$

Here is where the magic takes place. By symmetry we may switch E and F, and see that now given any $E \in \Gamma$, we have $F \in \mathcal{P} \Rightarrow E \cap F \in \Gamma$, i.e., $\mathcal{P} \subseteq \mathcal{K}_E$. Therefore for general $E \in \Gamma$, it holds that $\Gamma \subseteq \mathcal{K}_E$. More explicitly, this means

$$E \in \Gamma \text{ and } F \in \Gamma \Rightarrow E \cap F \in \Gamma,$$

i.e., Γ is closed under finite intersections.

It remains to show that Γ is a σ-algebra. We check the three axioms for a σ-algebra:

(i) $X \in \Gamma$; (by λ-system axiom 1)

(ii) for $A \in \Gamma$ with $A \subseteq X$, we have $X - A \in \Gamma$; (by λ-system axiom 2)

(iii) for $A_1, A_2 \in \Gamma$, $A_1 \cup A_2 = X - \big((X - A_1) \cap (X - A_2)\big)$. By (ii) above and Γ being a π-system it is clear to see $A_1 \cup A_2 \in \Gamma$. Therefore for $A_1, A_2, \ldots$ from Γ, we $\bigcup_{j=1}^{n} A_j \in \Gamma$. Now by axiom 3 of a λ-system,

$$\bigcup_{j=1}^{n} A_j \uparrow \bigcup_{j=1}^{\infty} A_j \quad \text{as } n \to \infty.$$

Thus $\bigcup_{j=1}^{\infty} A_j \in \Gamma$.

The proof is now complete. □

The key idea in these proofs is always to explore "the structure generated from $\mathcal{E}$ is the smallest containing $\mathcal{E}$." This is the reason we define collection $\mathcal{K}_E$ in (J.2), as our end goal is to show that for any $E \in \Gamma$, it holds that $E \cap F \in \Gamma$ for any $F \in \Gamma$, which is the λ-system generated by $\mathcal{P}$.

The reason why we can switch the role of E and F in the proof is the symmetry of "$\cap$" operation. It simplifies the proof, but there is nothing truly magical in the end.

The exact same idea (including this symmetry switch) can be applied to prove the monotone class theorem, which we will do now.

J.3 Monotone class theorem. Given an algebra $\mathcal{A}_0$ of sets, then the monotone class $\mathcal{M}$ generated by $\mathcal{A}_0$ coincides with the σ-algebra $\sigma(\mathcal{A}_0)$ generated by $\mathcal{A}_0$.

Proof. To prove $\mathcal{M} \supseteq \mathcal{A}_0$, it suffices to show that $\mathcal{M}$ is a σ-algebra.

First of all we note that every monotone class closed under finite unions must be closed under countable unions. Suppose $\mathcal{M}$ is closed under finite unions. Then if $A_j \in \mathcal{M}$ for all j, we have $B_n := \bigcup_{j=1}^n A_j \in \mathcal{M}$. Meanwhile $B_n \uparrow \bigcup_{j=1}^\infty A_j$ as $n \to \infty$, and therefore $\bigcup_{j=1}^\infty A_j \in \mathcal{M}$.

Since $\mathcal{M}$ contains $\emptyset$ and X, we only need to show $\mathcal{M}$ is closed under complements and closed under finite unions.

We first show $\mathcal{M}$ is closed under complements. If we can show that the collection

$$\mathcal{K} := \{A \subseteq X : A^c \in \mathcal{M}\}$$

is a monotone class, then since $\mathcal{K} \supseteq \mathcal{A}_0$, it follows that $\mathcal{K} \supseteq \mathcal{M}$, which proves our claim that $\mathcal{M}$ is closed under complements. To see why $\mathcal{K}$ is a monotone class, for an ascending sequence of sets $A_1 \subseteq A_2 \subseteq \cdots$ in $\mathcal{K}$,

$$\left(\bigcup_{j=1}^\infty A_j\right)^c = \bigcap_{j=1}^\infty A_j^c \in \mathcal{M}.$$

The same argument applies to any descending sequence of sets in $\mathcal{K}$.

It remains to prove that $\mathcal{M}$ is closed under finite unions. For any $E \in \mathcal{M}$, let us define

$$\mathcal{K}_E = \{F \subseteq X : E \cup F \in \mathcal{M}\}.$$

First we prove $\mathcal{K}_E$ is a monotone class. Consider an ascending sequence of sets $F_1 \subseteq F_2 \subseteq \cdots$ in $\mathcal{K}_E$. This gives an ascending sequence of sets

$$E \cup F_1 \subseteq E \cup F_2 \subseteq \cdots$$

in $\mathcal{M}$, which implies

$$\bigcup_{j=1}^\infty (E \cup F_j) = E \cup \left(\bigcup_{j=1}^\infty F_j\right) \in \mathcal{M}.$$

Therefore $\bigcup_{j=1}^\infty F_j \in \mathcal{K}_E$. A decreasing sequence of sets from $\mathcal{K}_E$ can be handled in the same way.

Just like in the proof of the π-λ theorem, we first fix $E \in \mathcal{A}_0$. Since for $F \in \mathcal{A}_0$, $E \cup F \in \mathcal{A}_0 \subseteq \mathcal{M}$, we have $\mathcal{K}_E \supseteq \mathcal{A}_0$. Therefore $\mathcal{K}_E \supseteq \mathcal{M}$, given that $\mathcal{K}_E$ is a monotone class. This shows that

$$E \in \mathcal{A}_0 \text{ and } F \in \mathcal{M} \Rightarrow E \cup F \in \mathcal{M}.$$

Now switch E and F to see that for any given $E \in \mathcal{M}$, if $F \in \mathcal{A}_0$, then $E \cup F \in \mathcal{M}$, i.e., $\mathcal{K}_E \supseteq \mathcal{A}_0$. Again we get $\mathcal{K}_E \supseteq \mathcal{M}$. This shows that for any $E \in \mathcal{M}$ and $F \in \mathcal{M}$, we have $E \cup F \in \mathcal{M}$, as desired. □

Given the resemblance of these two theorems, one might wonder if there is a shortcut to directly prove one from the other. Sadly the answer is no, in both direction.

A proof of Dynkin's theorem from the monotone class theorem is outlined in [Bil95, Exercise 3.12]. The idea is as follows: given $\mathcal{P} \subseteq \mathcal{L}$, we consider the algebra $\mathcal{A}_0$ generated by $\mathcal{P}$. By the monotone class theorem, we can conclude that $\sigma(\mathcal{P})$ is exactly the monotone class generated by $\mathcal{A}_0$. Since $\mathcal{L}$ by definition, if we can show $\mathcal{A}_0 \subseteq \mathcal{L}$, then it follows that $\sigma(\mathcal{P}) \subseteq \mathcal{L}$. Recall we have an explicit description of the sets in $\mathcal{A}_0$, which will help us here. However, the proof is by no means simple.

It is unlikely to prove the monotone class theorem directly from Dynkin's theorem. Since $\mathcal{A}_0$ is a π-system, if we can show that the monotone class $\mathcal{M}$ generated by $\mathcal{A}_0$ is a λ-system, then we

are done. The main difficulty is that we cannot show easily verify that $\mathcal{M}$ is closed under proper difference. We might want to define

$$\mathcal{Q}_A = \{B \subseteq X : B \supseteq A \text{ and } B - A \in \mathcal{M}\}$$

for $A \in \mathcal{M}$, but this does not really work out because of the constraint $B \supseteq A$.

K Existence theorems for probability measures on product spaces

It is noteworthy that all results here use the axiom of dependent choice in the proof.

K.1 Existence of product probability measures on infinite spaces. The probability premeasure μ_0 defined above is σ-additive, and hence by Carathéodory extension theorem, there is a unique extension of μ_0 to a probability measure on $\bigotimes_n \mathcal{F}_n$.

Proof. The tradition approach requires Tonelli's theorem on finite products, see for example [ADM11, Section 6.3]. We follow [Sae96], which proceeds from first principles and is much simpler. □

It is clear that this can also be proved as a consequence of the following Ionesco-Tulcea existence theorem. One has to extend from countable indices to arbitrary indices, but we have done this in the proof of Daniell–Kolmogorov existence theorem.

[Kal21, Theorem 8.24]

K.2 Ionesco-Tulcea existence theorem. For any sequence of measurable spaces $\{(S_n, \mathcal{S}_n)\}$ and kernels $\mu_n : S_1 \times \cdots \times S_{n-1} \to S_n$ for $n \geq 2$. Then there exists a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ each living in $\{S_n\}_{n=1}^{\infty}$, such that the f.d.d. is given by

$$(X_1, \ldots, X_n) \sim \mu_1 \times \cdots \times \mu_n.$$

K.3 Nelson extension theorem [Fol99, Theorem 10.18].

L Facts and tools in probability

$e^x \geq x + 1$ log sum inequality $\frac{x-1}{x} \leq \log x \leq x - 1$ for $x > 0$

$$\frac{1}{x} \leq \log\Big(\frac{x}{x-1}\Big) = \int_{x-1}^{x} \frac{1}{t}\, dt \leq \frac{1}{x-1}$$

Therefore for all n,

$$\sum_{x=2}^{n} \frac{1}{x} \leq \log n = \int_{1}^{n} \frac{1}{t}\, dt \leq \sum_{x=2}^{n} \frac{1}{x-1}$$

Hence

$$\log(n+1) \leq \sum_{x=1}^{n} \frac{1}{x} \leq \log(n) + 1$$

L.1 Coupon collector's problem.

L.2 Bernoulli bond percolation.

Bibliography

[ABS24] Luigi Ambrosio, Elia Brué, and Daniele Semola. *Lectures on Optimal Transport*. 2nd ed. Springer, Cham, 2024, pp. xi+260.

[ADM11] Luigi Ambrosio, Giuseppe Da Prato, and Andrea Mennucci. *Introduction to Measure Theory and Integration*. Edizioni della Normale, 2011.

[Axl20] Sheldon Axler. *Measure, Integration & Real Analysis*. Springer International Publishing, 2020.

[Bil95] Patrick Billingsley. *Probability and Measure*. 3rd ed. John Wiley & Sons, 1995.

[Bil99] Patrick Billingsley. *Convergence of Probability Measures*. 2nd ed. John Wiley & Sons, 1999.

[Bog07] Vladimir I. Bogachev. *Measure Theory*. Springer Berlin Heidelberg, 2007.

[Bog18] Vladimir I. Bogachev. *Weak Convergence of Measures*. American Mathematical Society, 2018.

[Bre11] Haïm Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer New York, 2011, pp. xiv+599.

[BS20] Vladimir I. Bogachev and Oleg G. Smolyanov. *Real and Functional Analysis*. Springer International Publishing, 2020.

[Coh13] Donald L. Cohn. *Measure Theory*. 2nd ed. Birkhäuser/Springer, New York, 2013.

[Dur19] Rick Durrett. *Probability: Theory and Examples*. 5th ed. Cambridge University Press, 2019.

[Fal19] Neil Falkner. "Hahn's Proof of the Hahn Decomposition Theorem, and Related Matters". *The American Mathematical Monthly* 3 (Mar. 2019), pp. 264–268.

[Fel17] Adrian F. D. Fellhauer. "On the relation of three theorems of analysis to the axiom of choice". *Journal of Logic and Analysis* (2017), Paper No. 1, 23.

[Fol99] Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications.* 2nd ed. John Wiley & Sons, 1999.

[Han14] Ramon van Handel. "APC550 Lecture Notes: Probability in High Dimension". 2014.

[Her06] Horst Herrlich. *Axiom of Choice*. Springer Berlin Heidelberg, 2006.

[Jos05] Jürgen Jost. *Postmodern Analysis*. 3rd ed. Springer-Verlag, Berlin, 2005, pp. xvi+371.

[Kal02] Olav Kallenberg. *Foundations of Modern Probability*. 2nd ed. Springer New York, 2002.

[Kal21] Olav Kallenberg. *Foundations of Modern Probability*. 3rd ed. Springer Switzerland, 2021.

[Kra22] Steven G. Krantz. *Real Analysis and Foundations*. 5th Ed. CRC Press, Boca Raton, FL, 2022.

[LeG22] Jean-François Le Gall. *Measure Theory, Probability, and Stochastic Processes*. Springer International Publishing, 2022.

[MP10] Peter Mörters and Yuval Peres. *Brownian Motion*. With an appendix by Oded Schramm and Wendelin Werner. Cambridge University Press, Cambridge, 2010, pp. xii+403.

[Mun00] James R. Munkres. *Topology*. 2nd Ed. Pearson, 2000, pp. xvi+537.

[Par67] K. R. Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press, 1967.

[RF23] Halsey Royden and Patrick M. Fitzpatrick. *Real Analysis*. 5th ed. Pearson, 2023.

[Roc24] Sébastien Roch. *Modern Discrete Probability: An Essential Toolkit*. Cambridge University Press, 2024.

[Rud76] Walter Rudin. *Principles of Mathematical Analysis.* 3rd ed. McGraw-Hill, 1976.

[Rud87] Walter Rudin. *Real and Complex Analysis*. 3rd ed. McGraw-Hill, 1987.

[Rud91] Walter Rudin. *Functional analysis*. 2nd Ed. McGraw-Hill, Inc., New York, 1991, pp. xviii+424.

[Sae96] Sadahiro Saeki. "A Proof of the Existence of Infinite Product Probability Measures". *The American Mathematical Monthly* 8 (Oct. 1996), pp. 682–683.

[San15] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Calculus of variations, PDEs, and modeling. Birkhäuser/Springer, Cham, 2015, pp. xxvii+353.

[Sch17] René L. Schilling. *Measures, Integrals and Martingales*. 2nd ed. Cambridge University Press, 2017.

[Tay06] Michael E. Taylor. *Measure Theory and Integration*. American Mathematical Society, Providence, RI, 2006, pp. xiv+319.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, 2018, pp. xiv+284.

Index of Notations

$\vee$ maximum of two

$\wedge$ minimum of two

$\wp$ power set

$B(x; r)$ the ball centered at x with radius r

$S(x; r)$ the sphere centered at x with radius r

Function spaces

$\langle \cdot, \cdot \rangle$ inner product, or dual pairing

$\| \cdot \|$ norm

$\| \cdot \|_u$ uniform/supremum norm

$\| \cdot \|_p$ ℓ^p/L^p norm

C_0 space of continuous functions that vanishes at infinity

C_b space of bounded continuous functions

C_c space of continuous functions with compact supports

L^p L^p space

General measure theory

$\mathcal{A}$ a general σ-algebra

$\mathcal{B}$ Borel σ-algebra

$\bigotimes$ product σ-algebra

$\mathcal{M}$ space of signed/complex Borel measures

$\mathcal{M}_r$ space of signed/complex Radon measures

$\mathcal{L}$ Lebesgue σ-algebra

μ a general measure, or a probability distribution

m Lebesgue measure on $\mathbf{R}^d$

Probability

$\mathbf{E}_\mu$ expectation on the canonical space for stochastic processes, with initial distribution μ

$\mathbf{P}_\mu$ probability measure on the canonical space for stochastic processes, with initial distribution μ

E expected value

$\mathcal{F}$ a general σ-field, or a collection of functions

$\mathcal{P}$ space of (Borel) probability measures

$\mathcal{P}_p$ Wasserstein p-space of probability measures

$d_{\mathrm{TV}}(\cdot, \cdot)$ total variation distance between two probability measures

P probability measure

List of Definitions