From Measure to Probability

A probabilist's survey of measure-theoretic results

Feng Cheng*

Draft as of March 20, 2026

*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.

Remarks on Notation. In Part I we use X to denote a nonempty set, but in Part II we use X instead to denote a random variable. As a replacement a nonempty set is denoted by S. Oftentimes S is a metric space with metric ρ, and we recommend to assume S to be separable at all times.

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

fecheng@math.washington.edu.

Literature Review

The most commonly used textbook for measure theory and functional analysis these days in the US is [Fol99], but is a very hard book at first exposure based on our personal experience. On the other hand the book is too analysis for a probabilist. It would serve as a good preparation for people interested in Fourier analysis, PDEs, and mathematical physics, but may not be the best choice for people ultimately interested in probability theory.

The nice little treatise [ADM11] is based on the undergraduate classes at Scuola Normale Superiore in Italy, and is very accessible for students with concrete background in elementary real analysis. The exposition is very different from Folland, and various materials that would be more of interest to probabilists/optimal transport theorists. One could say that the measure theory part of the book is a combination of the Folland perspective and the Ambrosio–Prato–Mennucci perspective. We also mention [Coh13] and [Tay06] contain additional materials that are not usually covered in a first class in measure theory. These materials in particular include Polish spaces, integration on differential forms, and Haar and Hausdorff measures.

Functional analysis is not the focus, but sufficient understanding of the basics of Hilbert spaces, Banach spaces, weak topology, and semigroups play a crucial role throughout the text. [Bre11], [BS20], and [BS18] are our personal favorite. Both books are very detailed and well-structured. The book [Sch17] also heavily inspired to treat Riesz' theorem and Fourier transform of measures in the correct way suitable for probabilistic applications.

For probability from measure theory all the way up to stochastic calculus, the best recommendation is go with either Durrett [Dur19][Dur96] or Le Gall [LeG22][LeG16]. They are both very good books with very different perspectives, and one should probably be familiar with the content of both books.

We have to mention the encyclopedia [Kal21], which is really a reference book but contains literally 99% of the content[1] one will encounter as background knowledge in probability theory. Some of the results are a bit too general, but the presentation is usually concise and optimal.

stochastic calculus [KS91] [RY99]

The blog-style lecture notes by George Lowther and Fabrice Baudoin

[DaP14][DaP06] has a somewhat different yet valuable perspective for a thorough course on stochastic calculus. By this I mean it does not get deep into the convoluted continuous filtration and martingale theory. He opted C_b for uniformly continuous and bounded functions to develop many results, which may be nonstandard.

[Bas11] contains various good content, but has way too many typos for the readers

convergence of measures [Dud02][Bil99][Bog18]

[ABS24][San15][Vil15]

[1] exaggeration, perhaps

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,

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

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.

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

We know that the preimage of a function $f\colon X \to Y$ is a mapping $f^{-1}\colon \wp(Y) \to \wp(X)$ 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.

1.7 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

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

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

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

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

1.12 Exercise. Define $\sum_{n\in E} a_n = \sup\{\sum_{n\in I} a_n : I \text{ is a finite subset of } E\}$, where E is an uncountable index set and all $a_n \geq 0$. If the sum is finite, then only countably many a_n's are nonzero. (Hint: consider the set $E_k = \{n : a_n \geq 1/k\}$, which should be countable.)

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.

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

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

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

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

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.

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

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

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

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

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

(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 $\delta_X\colon \mathcal{A} \to \{0, 1\}$ 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 $1/0$.

[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.

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 $\mu\colon \mathcal{A} \to [0,\infty]$ 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 say μ is a *continuous measure* if it assigned zero measure to all singletons. By comparison, this should be naturally connected to continuous distributions in probability.

We now introduce two additional elementary results about measures, which are simple consequences from Proposition 1.15. These two results are important in probability theory, but both are indeed purely measure-theoretic.

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

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

If in addition μ is finite, then

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

1.17 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.

1.18 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. □

Let μ be a σ-finite measure. The set $A \in \mathcal{A}$ is called an *atom* of the measure μ if the set has measure $\mu(A) > 0$ (including $+\infty$), but all its measurable subsets must be either of measure 0 or of measure $\mu(A)$. The measure is *atomless* if there are no atoms. The 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., $\mu(X - \bigcup_n A_n) = 0$.

Atomic measures are essentially just discrete measures.

1.19 Proposition. Let X be a separable metric space. For any σ-finite measure on the Borel σ-algebra of X, an atomic measure is precisely the discrete measure.

Proof. Let X be concentrated on $\bigcup_{j=1}^{\infty} A_j$. The rest of proof should resemble that of Proposition 7.19, which has all the details written out. For each A_j, there should be a countable subcover of open balls $\{B(x; \frac{1}{n}) : x \in A\}$. Out of this subcover there exists one ball $B_{1/n}$ centered at some x such that $\mu(B_{1/n} \cap A_j) = \mu(A_j)$. Notice that $\mu\big((B_{1/m} - B_{1/n}) \cap A_j\big) = 0$ for any pair of n, m, and therefore

$$\mu(B_1 \cap A_j) = \mu(B_{1/2} \cap B_1 \cap A_j) = \cdots .$$

This implies that

$$\mu\left(\bigcap_{n=1}^{\infty} B_{1/n} \cap A_j\right) = \mu(A_j),$$

and since $\bigcap_{n=1}^{\infty} B_{1/n}$ is a singleton, we have "reduced" A_j to a single point. □

1.B Two tools from set theory

1.20 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)

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

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

1.23 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 $\mathcal{A}_0$.

[9] see Definition 1.9

We deferred the proofs of both theorems to Appendix G; 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]

1.24 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 $\mathcal{A}$. (X is σ-finite in this case.)

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 $\mu_1 = \mu_2$. □

1.C Extension theorems

1.25 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 $D_1, D_2, \dots$ 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 $E \in \wp(X)$

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

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 $f\colon \mathbf{R} \to \mathbf{R}$ 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.

1.26 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 $\mu^*(E) < \infty$,

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

1.27 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

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

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

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

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

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

$$\mu^*(E) \le \mu^*(E \cap N) + \mu^*(E \cap N^{\mathrm{c}}) \le \mu^*(E \cap N^{\mathrm{c}}) \le \mu(E). \qquad \square$$

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

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

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 $\mu^*(E) \ge \mu_0(E)$ 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 $\sum_{n=1}^{\infty} \mu_0(A_n) \ge \sum_{n=1}^{\infty} \mu_0(B_n) = \mu_0(E)$. 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

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

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

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

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

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.

1.29 Proposition [Coh13, Lemma 3.4.6, 3.4.7]. Let $\mathcal{A}$ be a σ-algebra generated from the algebra $\mathcal{A}_0$.

(a) Suppose μ is a finite measure on $\mathcal{A}$. Then for any $A \in \mathcal{A}$ and $\epsilon > 0$, there exists some $B \in \mathcal{A}_0$ such that $\mu(A \triangle B) < \epsilon$.

(b) Now μ is allowed to be any measure on $\mathcal{A}$, but instead suppose $X = \bigcup_{n=1}^{\infty} C_n$, where $C_n \in \mathcal{A}_0$ and $\mu(C_n) < \infty$. Then for any $A \in \mathcal{A}$ with finite μ-measure any any $\epsilon > 0$, there exists some $B \in \mathcal{A}_0$ such that $\mu(A \triangle B) < \epsilon$.

Proof. In part (a) we want to show that the collection of $A \in \mathcal{A}$ that can be approximated by sets in the algebra $\mathcal{A}_0$ forms a σ-algebra. Since the collection contains $\mathcal{A}_0$, it must be $\mathcal{A}$. The proof is standard and hence omitted.

Moving on to part (b), we may first assume that C_n is an increasing sequence of sets. Since $\mu(A \cap C_n) \to \mu(A)$, for any given $\epsilon > 0$ there is some N such that

$$\mu(A \cap C_n) \geq \mu(A) - \epsilon/2.$$

This allows us to apply the first part to $A \cap C_n$: there exists some $A_0 \in \mathcal{A}_0$ such that

$$\mu\big((A \cap C_N) \triangle A_0\big) < \epsilon/2.$$

We claim that $B = A_0 \cap C_N$ is what we are looking for, which follows by using the triangular equality for symmetric difference of measures, stated below. □

$\mu(A \triangle B) \leq \mu(A \triangle C) + \mu(C \triangle B)$. Up to the equivalence relation $A \sim B$ if $\mu(A \triangle B) = 0$, we have a metric on the σ-algebra of measurable sets.

measure approximation in symmetric difference

1.30 Theorem.

1.D The Lebesgue measure

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

1.32 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.

1.33 Theorem. For an increasing right-continuous function $F \colon \mathbf{R} \to \mathbf{R}$, the function $\mu_0 \colon \mathcal{A}_0 \to [0, \infty]$ 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 $\mathcal{A}_0$.

1.34 Theorem [Fol99, Theorem 1.16].

(a) Let $F \colon \mathbf{R} \to \mathbf{R}$ 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 $F\colon \mathbf{R} \to \mathbf{R}$ 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,

$$F(x-) = \mu(-\infty, x). \tag{1.35}$$

The function F is known as the *(cumulative) distribution function* of μ.

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 (1.35), 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 1.33, 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 (1.35) 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.

The reason why we look at half intervals $(a, b]$ instead of $[a, b)$ in measure and probability theory is largely conventional. If we work with $[a, b)$, then the distribution function of μ would be increasing and left-continuous instead. Nothing changes essentially.

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 $\mathcal{L}$.

1.36 Exercise. For any bounded and Borel measurable U in $\mathbf{R}^n$, given any vector, there exists a hyperplane orthogonal to this vector such that U is bisected in Lebesgue measure. If U is further assumed to be open and connected, then the hyperplane is in fact unique.

Combine this with the famous Borsuk–Ulam theorem from topology, one can prove the Ham Sandwich Theorem, which says that for any n bounded measurable sets in $\mathbf{R}^n$, there exists an $(n-1)$-dimensional hyperplane that simultaneously bisects all n sets.

1.37 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.

1.E Regularity of measures

We study regularity of measures on a topological space X, which is almost always given the Borel σ-algebra $\mathcal{B}$ by default. If X is given the Borel σ-algebra, then the *support* of a Borel measure μ is given

$$\operatorname{supp} \mu = \bigcap \{F \text{ closed in } X : \mu(F^{\mathrm{c}}) = 0\},$$

which is a closed set and hence Borel measurable.

1.38 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

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

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

A measure is locally finite if it is finite on all compact subsets of X. We say a finite measure μ is *tight* if

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

We remark that in the literature people both inner regularities are common. Compact inner regularity is sometimes restricted to only open sets, which is a natural choice for Section 5.F.

1.39 Proposition. Every finite measure on a topological space with the Borel σ-algebra 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.

1.40 Theorem [Bil99, Theorem 1.1] [Sch17, Theorem H.2]. 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.

Furthermore, if X is σ-compact (in particular, when X is locally compact and separable), then X is compact inner regular.

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 $E_1, E_2, \dots$ 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

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

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.

Now assume in addition that X is σ-compact. Let $\{L_m\}$ be a sequence of compact sets that increases to X, and we define $K_{n,m} = F_n \cap L_m$, which is compact. It should be clear to see that $K_{n,m} \cap E$ approximates E from the inside:

$$\begin{aligned} \mu(E - K_{n,m}) &\le \mu(E - F_n) + \mu(F_n - K_{n,m}) \\ &\to \mu(E - F_n) \quad \text{as } m \to \infty \text{ since } K_{n,,m} \uparrow F_n, \\ &\to \mu(E) \quad \text{as } n \to \infty. \end{aligned}$$

This completes the proof. □

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

The inner regularity can be easily passed to *locally* finite measures.

1.41 Corollary [Sch17, Theorem H.3]. A locally finite measure on a σ-compact metric space with the Borel σ-algebra is compact inner regular.

Proof. We now have a sequence of compact sets L_m, each of finite measure, that increases to X. For each $E \in \mathcal{B}$, we have

$$\mu(E) = \sup_m \mu(E \cap L_m) = \sup_m \mu_m(E),$$

where μ_m is the restriction of μ to L_m (recall Fact 1.13), now a finite measure. Hence

$$\mu(E) = \sup_m \sup_K \mu_m(K) = \sup_K \sup_m \mu_m(K) = \sup_K \mu(K),$$

where the supremum over K is over all compact subsets of E. □

In fact, Schilling [Sch17, Theorem H.3] also proves that for a locally finite measure μ on any metric space X, if there exists a sequence of open sets G_n such that $\mu(G_n) < \infty$ and $G_n \uparrow X$, then μ is out regular. Because this will not be used in the text, we omit its proof.

In elementary analysis, recall that a set E is of Lebesgue measure zero if for any $\epsilon > 0$, there exists a countable cover of open intervals $\bigcup_{j=1}^{\infty}(a_j, b_j) \supseteq E$ such that $\sum_{j=1}^{\infty}|b_j - a_j| < \epsilon$. We now justify this, which requires little modification from the formula for Lebesgue–Stieltjes measures

$$\mu(E) = \inf\left\{\sum_{j=1}^{\infty}\mu(a_j, b_j] : \bigcup_{j=1}^{\infty}(a_j, b_j] \supseteq E\right\} \tag{1.42}$$

1.43 Proposition. For any Lebesgue–Stieltjes measure μ, we have

$$\mu(E) = \inf\left\{\sum_{j=1}^{\infty}\mu(a_j, b_j) : \bigcup_{j=1}^{\infty}(a_j, b_j) \supseteq E\right\} \tag{1.44}$$

for any $E \in \mathcal{B}(\mathbf{R})$.[13]

Proof. We will give a sketch and let the reader fill out the detail. Each open interval (a_j, b_j) can be written as $\bigcup_{k=1}^{\infty}(c_j^k, c_j^{k+1})$, where

$$c_j^1 = a_j,\ c_j^2 = \tfrac{b_j + c_j^1}{2},\ c_j^3 = \tfrac{b_j + c_j^2}{2}, \dots .$$

This allows to reduce open intervals to half intervals, and hence the right hand side of (1.42) is at most (1.44).

For the reverse inequality, we may use the right-continuity of F_μ at each b_j. Consider the larger open interval $(a_j, b_j + \delta_j)$, where $F(b_j + \delta_j) - F(b_j) < \epsilon/2^j$. □

1.45 Corollary. Lebesgue–Stieltjes measures on $\mathcal{B}(\mathbf{R})$[14] are compact (and closed) inner regular and outer regular.

Proof. The compact inner regular part follows directly from Corollary 1.41. The outer regular part follows from the previous proposition. □

[13] Or even $E \in \mathcal{L}$; but as mentioned before, the Lebesgue σ-algebra will be of little interest to us.
[14] again, or $\mathcal{L}$

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

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

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

1.47 Proposition. For any $0 < \epsilon < 1$, there is some interval $I \subseteq [0,1)$ with $0 < m(I) \leq \epsilon$ such that

$$m(A \cap I) \geq (1-\epsilon)m(I) \quad \text{for all } A \in \mathcal{B}[0,1).$$

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 $f\colon X \to Y$ 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 $f\colon X \to Y$ 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 $f_n\colon X \to \mathbf{R}$, 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. Say $f\colon (a, b) \to \mathbf{R}$ is differentiable, then f' is measurable.

Write f' as the limit of a sequence of measurable functions.

2.5 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 $f\colon X \to Y$, 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 $\sigma(f)$.

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

More generally, consider a collection of measurable spaces $(Y_\alpha, \mathcal{N}_\alpha)$ over all $\alpha \in I$. Suppose we are given $f_\alpha \colon X \to Y_\alpha$ 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.)

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

2.7 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 càdlàg

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

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

2.9 Proposition. A measure μ is an order-preserving positive linear functional on $L^+(\mu)$. (Strictly speaking it is a linear functional allowed to take value $+\infty$.) Let the integration all be with respect to μ below.

(a) $\int f + g = \int f + \int g$ for $f, g \in L^+(\mu)$;

(b) $c\int f\,d\mu = \int cf\,d\mu$ for $\lambda \geq 0$;

(c) $\int f\,d\mu \leq \int g\,d\mu$ if $f \leq g$.

With this view in mind, sometimes it is preferable to write μf in place of $\int f\,d\mu$.

Furthermore for $f \in L^+(\mu)$, $A \mapsto \int_A f\,d\mu$ defines a (positive) measure on $(X, \mathcal{A})$.

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

$$\int \big(\liminf_n f_n\big) \leq \liminf_n \int f_n$$

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

Extending Proposition 2.9, one can easily see that $L^1(\mu)$ is a vector space, and μ is again a positive linear functional over $L^1(\mu)$, sending $f \geq 0$ to $\int f\,d\mu \geq 0$. In Section 4.A we will define real and complex-valued μ, and in those cases μ will become a general linear functional over $\mathbf{R}$ and $\mathbf{C}$.

2.11 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

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

2.12 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.

2.13 Markov's inequality. Let $f\colon X \to \mathbf{R}$ be measurable and $\varphi\colon \mathbf{R} \to [0,\infty)$ 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 $f\colon X \to \mathbf{R}$, then we get for any $a > 0$,

$$\mu\{x : |f| \geq a\} \leq \frac{1}{a^p} \int |f|^p\,d\mu. \tag{2.14}$$

2.15 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 $\varphi\colon I \to \mathbf{R}$. Then

$$\varphi\left(\int f\,d\mu\right) \leq \int \varphi \circ f\,d\mu. \tag{2.16}$$

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

We use $\int_a^b f(x)\,dx$ for Riemann integrals, and $\int_{[a,b]} f(x)\,dm(x)$ for Lebesgue integrals.

2.17 Theorem. For a Riemann integrable function f on a bounded interval $[a,b]$, we have

$$\int_a^b f\,dx = \int_{[a,b]} f\,dm$$

The following result was proved by [Lew86] using an elementary method without the Lebesgue theory.

2.18 Bounded convergence theorem (Riemann integration). For a sequence of Riemann integrable functions $\{f_n\}$ on $[a,b]$, suppose its pointwise limit f is also Riemann integrable on $[a,b]$, and $\sup_n \|f_n\|_u \le K$. Then

$$\int_a^b f_n\,dx \to \int_a^b f\,dx.$$

In essence, the Lebesgue theory was introduced to handle the limits of integrals nicely. We will also see that the Riemann theory is easier to develop in higher dimensions.

2.19 Proposition. An improper Riemann integral over an unbounded interval is Lebesgue integrable if it is absolutely convergent. Furthermore, the improper Riemann integral and the Riemann integral would coincide.

In case the reader has forgot how improper Riemann integrals are defined, we have provided the full proof below. However, this should be left as an exercise to the reader.

Proof. We consider the interval $[a,\infty)$ for $a > -\infty$. The proof can be adapted easily to other types of improper integral.

Assume $\int_a^\infty |f|\,dx < \infty$, we have

$$\lim_{K\to\infty} \int_a^K |f|\,dx = \lim_{K\to\infty} \int |f|\mathbf{1}_{[a,K)}\,dm,$$

which by MCT converges to

$$\lim_{K\to\infty} \int |f|\mathbf{1}_{[a,\infty)}\,dm,$$

This shows that $f\mathbf{1}_{[a,\infty)} \in L^1(m)$, and then by repeating the above argument again with f in place of $|f|$ we conclude that the

$$\int_a^\infty f\,dx = \int_{[a,\infty)} f\,dm. \qquad \square$$

The improper Riemann integral that is not Lebesgue integrable is the Dirichlet integral: $\frac{\sin x}{x}\mathbf{1}_{[0,\infty)}$ is not Lebesgue integrable, but is improperly integrable. Although it is not Lebesgue integrable, we will see in that the Lebesgue theory will still be tremendously helpful in justifying

$$\lim_{K\to\infty} \int_0^K \frac{\sin x}{x}\,dx = \frac{\pi}{2}.$$

2.E Modes of convergence

2.20 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 $\epsilon > 0$,
$$\lim_n \mu\{x : |f_n(x) - f(x)| > \epsilon\} = 0. \tag{2.21}$$

We say $\{f_n\}$ is

- *Cauchy/fundamental in measure* if for any $\epsilon > 0$, there exists $N \in \mathbf{N}$ such that for all $m > n \ge N$,
$$\mu\{x : |f_n(x) - f_m(x)| > \epsilon\} < \epsilon \tag{2.22}$$

Note that the "$>$" in both (2.21) and (2.22) can be replaced by "$\ge$", obviously. It suffices to use only one ϵ in (2.22) because we can always choose the smaller of two distinct ϵ's.

2.23 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

$$|f(x) - g(x)| > \epsilon \implies |f(x) - h(x)| > \epsilon/2 \text{ or } |h(x) - g(x)| > \epsilon/2 \tag{2.24}$$

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 (2.24). 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

$$\mu\{x : |f_n(x) - f_m(x)| > 2^{-k}\} < 2^{-k} \quad \text{for all } m \geq n \geq n_k. \tag{2.25}$$

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

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

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 $n \geq k$,

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

[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.

This yields $g_k \to f$ in measure.

The final step is to use this to show $f_n \to f$ in measure. We again resort to the containment relation (2.24):

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

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

(c) Contained in the previous part.

(d) This is clearly a consequence of (2.14).

(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 $y_n = \mu\{x : |f_n(x) - f(x)| > \epsilon\}$. □

2.26 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.

2.27 Exercise. Show in one line that if the measure space is finite, then uniform convergence implies convergence in L^1. (In fact in L^p, as we will see later.)

2.28 Exercise. Give a proof of Theorem 2.23(e) using the bounded convergence theorem.

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

2.30 Corollary. Let μ be finite, and $f_n \to f$ and $g_n \to g$ in measure. Say $\Psi : \mathbf{R}^2 \to \mathbf{R}$ 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

$$\mu\big\{x : \big|\Psi\big(f_{n_k}(x), g_{n_k}(x)\big) - \Psi\big(f(x), g(x)\big)\big| > \epsilon\big\} \ge \epsilon. \tag{2.31}$$

Recall the construction of the subsequence in Theorem 2.23(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 (2.31). □

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.

2.32 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.

2.33 Exercise. Use Theorem 2.23(c) to prove the monotone convergence theorem and Fatou's lemma with convergence in measure.

2.F Littlewood's second and third principles

2.34 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

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

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.

2.35 Classical Luzin's theorem. Let $f\colon [a,b] \to \mathbf{R}$ (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 $f\colon A \to Y$ restricted to a closed/compact subset, while the *strong* Luzin's theorem also considers whether we may find a continuous function $g\colon A \to Y$ 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]:

2.36 Theorem. Let $A \in \mathcal{L}(\mathbf{R})$, and let $f\colon A \to \mathbf{R}$ be Borel measurable. For any $\epsilon > 0$, there is a continuous function $g\colon \mathbf{R} \to \mathbf{R}$ 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.

2.37 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

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

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.

2.38 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 2.37.

2.39 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.

2.40 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.

2.41 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

2.42 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. □

2.43 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

2.44 Corollary [Jos05, Theorem 16.10]. Let A be a metric space[3], and $f\colon X \times A \to \mathbf{R}$. 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

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

is continuous on A.

We need to check if $\{y_n\} \subseteq A$ converges to y, 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

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

2.45 Corollary [Jos05, Theorem 16.11]. Let $I \subseteq \mathbf{R}$ be an open interval, and $f\colon X \times I \to \mathbf{R}$. 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 $t \in I$,

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

We may then conclude that the function

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

is differentiable in 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 a measurable function $\varphi\colon (X,\mathcal{M}) \to (Y,\mathcal{N})$, 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 be a measure on $(Y,\mathcal{N})$, and we call this the *image/pushforward measure* of μ by φ, denoted by $\varphi_*\mu$ or $\varphi_\#\mu$, since are pushing a measure from the domain X forward to its range Y.

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.

2.46 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. □

There are some properties of φ_* that will be useful. First, the mass is preserved under φ_*:

$$\varphi_*\mu(Y) = \mu(\varphi^{-1}Y) = \mu(X).$$

In particular, if μ is a probability measure, then $\varphi_*\mu$ remains a probability measure. Second, the map φ_* is additive and positively homogeneous on the space of all measures on X: for two measures μ, ν on X, we have

$$\varphi_*(\mu + \nu) = \varphi_*\mu + \varphi_*\nu \quad \text{and} \quad (\lambda\varphi)_*\mu = \lambda(\varphi_*\mu) \text{ for all } \lambda \geq 0.$$

In particular, this implies that the map φ_* is affine:

$$\varphi_*\big((1-\lambda)\mu + \lambda\nu\big) = (1-\lambda)\varphi_*\mu + \lambda\varphi_*\nu \quad \text{for } 0 < \lambda < 1.$$

We can further show that φ_* is continuous, when the domain and range are endowed with either the norm or weak topologies. These will all come in Section 4.A and Section 8.C.

Also note that $\varphi_*\delta_x = \delta_{\varphi(x)}$.

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

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

[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.

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. □

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.

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

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 $f\colon X \to Y$ is $\mathcal{M}/\mathcal{N}$-measurable if and only if each $f_\alpha = \pi_\alpha \circ f$ is $\mathcal{M}/\mathcal{N}_\alpha$-measurable.

3.7 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 $f\colon X \times Y \to \mathbf{R}$ (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 H. 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.

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

complex integration, Laplace transform, and Feynmann's trick often requires justification including uniform convergence

We now generalize Theorem 1.37

3.11 Theorem.

3.C Change of variables

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

3.13 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 $\mathbf{R}^m$.

for injective C^1 functions

3.14 Change of variables. 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

Convex sets in $\mathbf{R}^n$ are not necessarily Borel measurable. Of course in $\mathbf{R}^1$, convex sets are just intervals, so they are always Borel measurable. In the general case, we can consider the union of $B(0;1)$ and a non-Borel subset of $S(0;1)$.[4] However, convex sets in $\mathbf{R}^n$ are always Lebesgue measurable. See [Lan86].

3.15 Brunn–Minkowski inequality. For two compact/open bounded 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.) We remark that if we let A and B be n-by-n matrices and replace m by det, than the same inequalities are true as well. This is a useful fact, for example, used in optimal transport.

3.16 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 $a = b$.

3.17 Prékopa–Leindler inequality. Let $0 < \lambda < 1$ and $f, g, h \colon \mathbf{R}^n \to [0, \infty)$ be measurable. Suppose

$$f(x)^{1-\lambda} g(y)^{\lambda} \le h\big((1-\lambda)x + \lambda y\big)$$

for all $x, y \in \mathbf{R}^n$, then

$$\|f\|_1^{1-\lambda} \|g\|_1^{\lambda} \le \|h\|_1.$$

Take $f = \mathbf{1}_A$, $g = \mathbf{1}_B$, and $h = \mathbf{1}_{(1-\lambda)A+\lambda B}$, we recover Brunn–Minkowski inequality.
log-concave measure
Gaussian measure

3.18 Exercise. For A compact (resp. open bounded), the spherical cap $A_\epsilon = A + \epsilon\overline{B}$ (resp. ϵB) is minimized when A is the closed (resp. open) ball with volume $m(A)$.

By taking limit as $\epsilon \to 0$ one may obtain the Euclidean isoperimetric inequality, which states that the boundary "content" of a Borel set is minimized when the set is the Euclidean ball.

As similar result exists for Gaussian measures
intimate connection with L^p space, Hölder's inequality, and the *reverse Young's inequality*.

3.E The Gamma function and polar coordinates

There is the polar identification between $\mathbf{R}^n - \{0\}$ and $(0,\infty) \times S^{n-1}$. Let $\Phi \colon x \to |x| \times \frac{x}{|x|}$ be this map. The point at zero can be safely ignored throughout the section, since it does not contribute anything to the integral.

[4] The convexity of the set is obvious. Stereographic projection tells us that S^n minus one point and $\mathbf{R}^{n-1}$ are homeomorphic, and hence the non Borel subset of $S(0;1)$ exists.

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

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

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

3.20 Example (Gaussian integral). Let us show $\int_{\mathbf{R}} \exp(-x^2)\,dx = \int_{-\infty}^{\infty} \exp(-x^2)\,dx = \sqrt{\pi}$. It is well-known that one can consider the double integral

$$\int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} \exp(-x^2-y^2)\,dx\,dy = \left(\int_{x=-\infty}^{\infty} \exp(-x^2)\,dx\right)^2.$$

By Tonelli's theorem, the above integral is equal to

$$\int_{\mathbf{R}^2} \exp(-x^2-y^2)\,dm(x,y),$$

and by polar change of coordinates we get

$$\int_{\theta=0}^{2\pi} \int_{r=0}^{\infty} e^{-r^2} r\,dr\,d\theta,$$

which after computation turns out to be π.

3.21 Corollary. Say we have a function $f \in L^+(\mathbf{R}^n)$ or $L^1(\mathbf{R}^n)$ that is radial, i.e., there exists some $f(x) = g(|x|)$ for $x \in [0,\infty)$. Then

$$\int_{\mathbf{R}^n} f\,dx = \sigma(S^{n-1}) \int_0^{\infty} g(r)\, r^{n-1}\,dr.$$

Note that 2π is by definition the proportion of the circumference of S^1.
$\sigma(S^{n-1}) = \frac{2\pi^{n/2}}{\Gamma(n/2)}$ and
By definition $m(B^n) = \rho \times \sigma(\Phi^{-1}B^n) = \int_0^1 r^{n-1}\,dr = \frac{1}{n}$, and therefore

$$\alpha_n \coloneqq m(B^n) = \frac{1}{n}\sigma(S^{n-1}) = \frac{\pi^{n/2}}{\Gamma\big(\frac{n}{2}+1\big)}.$$

We will write $\sigma(S^{n-1}) = n\alpha_n$.
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 $\mathbf{R}$ and $\mathbf{C}$-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 $\mu\colon \mathcal{A} \to \mathbf{R}$ (resp. $\mu\colon \mathcal{A} \to \mathbf{C}$) 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 $\mu_n \coloneqq \mu|_{\mathcal{A}_n}$ 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

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

[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.)

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. □

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 any finite partitions. We may also take the partition to be a measurable partition of any measurable subsets of E instead, not necessarily the entire E.

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.

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

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.

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

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

For signed ν, define $L^1(\nu) = L^1(\nu^+) \cap L^1(\nu^-)$, and for $f \in L^1(\nu)$, define

$$\int f \, d\nu = \int f \, d\nu^+ - \int f \, d\nu^-.$$

Observe for $f \in L^1(|\nu|)$, we require

$$\int |f| \, d|\nu| = \int |f| \, d\nu^+ + \int |f| \, d\nu^- < \infty.$$

This shows that $L^1(|\nu|) = L^1(\nu)$

$|\int f \, d\nu| \leq \int |f| \, d|\nu|$.

4.15 Proposition. $|\nu|(A)$ is the "operator norm" of the linear functional $f \mapsto \int_A f\,d\nu$. By this we mean

$$|\nu|(A) = \sup_{|f|\le 1} \left| \int_A f\,d\mu \right|,$$

where the supremum is taken over all pointwise bounded functions.

If f is in some continuous function space *with the uniform/supremum norm* ($C_c/C_0/C_b$, which we will see later), the operator norm becomes a genuine one. In this case a signed measure can be viewed as a bounded linear functional on this continuous function space with the uniform norm.

In particular, if we view ν as the linear operator mapping $f \mapsto \int_X f\,d\nu$, then the "operator norm" of ν is just the total variation norm.

Note that within the vector spaces of signed measures, the space of positive measures is a convex cone. Living in the vector space of signed/complex measure, we can upgrade the property of pushforward maps mentioned in Section 2.I.

4.16 Proposition. Given a measurable function $\varphi\colon (X,\mathcal{A}) \to (Y,\mathcal{E})$, the pushforward $\varphi_*\colon \mathcal{M}(X) \to \mathcal{M}(Y)$ is a linear map that is also a contraction.

Proof. Linearity is clear. We need to check $\|\varphi_*\mu\| \le \|\mu\|$. By definition

$$\begin{aligned} \|\varphi_*\mu\| &= \sup_{E\in\mathcal{E}} \big\{ |\varphi_*\mu(E)| + |\varphi_*\mu(E^c)| \big\} \\ &= \sup_{E\in\mathcal{E}} \big\{ |\mu(\varphi^{-1}E)| + |\mu(\varphi^{-1}E^c)| \big\} \end{aligned}$$

Notice that $\varphi^{-1}E$ and $\varphi^{-1}E^c$ partitions E, and therefore the previous line is

$$\le \sup_{A\in\mathcal{A}} \big\{ |\mu(A)| + |\mu(A^c)| \big\} = \|A\| = \|\mu\|. \qquad \square$$

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.

4.17 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 an arbitrary 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. We follow [Roy88]. Suppose μ is finite. For each $q \in \mathbf{Q}$, let P_q and N_q be the positive and negative parts of $\nu - q\mu$. Define $\sup_q q\mathbf{1}_{P_q}$

For each $q \in \mathbf{Q}$, if for some measurable function f we have $\{f > q\}, \{f \leq q\}$ as a Hahn decomposition of $\nu - q\mu$, then f the function desired in the Radon–Nikodym theorem.[5] □

4.18 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].

4.19 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

All measures are absolutely continuous with respect to the counting measure, and the Radon–Nikodym derivative of μ with respect to counting measure is just $x \mapsto \mu\{x\}$, the evaluation of at the singleton.

For any finite/signed/complex μ on $(\mathbf{R}^n, \mathcal{B}^n)$, we have an important decomposition. First, consider any measure space $(X, \mathcal{A})$. We can write μ as the sum of a discrete measure (atomic part) and a continuous measure. Similar to the discrete finite measure, a *discrete signed/complex measure* μ can be expressed as

$$\sum_{y \in Y} c(y)\delta_y,$$

where Y is a countable subset of X and the function $c\colon Y \to \mathcal{F}$ is absolutely summable:

$$\sum_{y \in Y} |c(y)| < \infty.$$

(The definition for a continuous measure remains unchanged as in Section 1.A: since $|\mu|$ and μ are equivalent measures, they should simultaneously assign zero measure to the same

[5] The converse is true as well. It is interesting that both the existence of the Hahn decomposition also had such equivalent criterion which helped our proof.

point.) Define $D = \{x \in \mathbf{R}^n \colon |\mu\{x\}| > 0\}$. It turns out that $\mu_d = \mu|_D$ and $\mu_c = \mu|_{D^c}$. This follows directly from Exercise 1.12.

Now let $(X, \mathcal{A}) = (\mathbf{R}^n, \mathcal{B}^n)$. Notice $\mu_d \perp m$ and $\mu_c \ll m$. We can then break μ_c into two μ_{ac} and μ_{sc}, where $\mu_{ac} \ll m$ and $\mu_{sc} \ll m$. (The subscript "sc" stands for singular continuity.) Therefore

$$\mu = \mu_d + \mu_{ac} + \mu_{sc}.$$

In the case of a Borel measure on the real line, the above result has a very nice interpretation. First, recall any finite Borel measure μ can be identified as a right-continuous function F increasing from 0 to $\mu(\mathbf{R})$ on the real line.[6] The previous decomposition can then be written as

$$F = F_d + F_{ac} + F_{sc},$$

where F_d is a piecewise constant function, F_{ac} is an absolutely continuous function, and F_{sc} is a continuous but not absolutely continuous function. Of course, all of them are distribution functions, so they are increasing and right-continuous.

Lebesgue decomposition of a monotonic function (p344 345 Bogachev)

4.20 Chain rule.

4.21 Proposition [Axl20, 9.10]. If for a positive measure μ and a function $f \in L^1(\mu)$ we have $\frac{d\nu}{d\mu} = f$, then

$$\frac{d|\nu|}{d\mu} = |f|.$$

product of signed/complex measure

$$d(\mu \times \nu) = \frac{d\mu}{d|\mu|}\frac{d\nu}{d|\nu|} d(|\mu| \times |\nu|)$$

4.C Differentiation

In elementary analysis we learned that

$$\frac{1}{r}\int_x^{x+r} f(y)\,dy = f(x) = \frac{1}{2r}\int_{x-r}^{x+r} f(y)\,dy.$$

for any $r \neq 0$ and x at which f is continuous. A result like this was virtually impossible to generalize before to $\mathbf{R}^n$, but now with the machinery developed, this is possible. Furthermore, the continuity assumption is not longer needed, at the cost of being an almost everywhere result.

4.22 Wiener covering lemma. Given an arbitrary collection of open balls $\{B_\alpha\}_{\alpha \in A}$ in $\mathbf{R}^n$ whose union is the open set U, for any $c < m(U)$, there is finite subcollection $\{B_1, \ldots, B_K\}$ of disjoint open balls such that

$$3^n \sum_{k=1}^{K} m(B_k) > c.$$

[6] A similar characterization also exists for signed/complex measures; see Theorem 4.31.

Proof. We can find a compact set $K \subseteq U$ with $m(K) > c$. Then we have a subcollection $\mathcal{G} = \{G_1, G_2, \dots, G_N\}$ of open balls that covers K. Let B_1 be the G with the largest radius. Throw away any other G's that intersects B_1, and let B_2 be the remaining G with the largest radius. Repeating this, we will obtain $\{B_1, \dots, B_K\}$ ($K \leq N$), where each G intersects some B. If you draw a picture, it is clear to see that if we triple the radii of each B_k and obtain $\widetilde{B}_k$, then each G_k would be contained in $\widetilde{B}_k$, and hence

$$3^n \sum_{k=1}^{K} m(B_k) = \sum_{k=1}^{K} m(B_k) \geq \sum_{k=1}^{K} m(G_k) > c. \qquad \square$$

4.23 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 $(\mathbf{R}^n, \mathcal{B}, m)$. 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 $\mathbf{R}^n$. (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.

4.24 Definition. For $f \in L^1_{\text{loc}}$, we define the *averaging operator* $A_r f$ by

$$A_r f(x) = \frac{1}{m\big(B(x;r)\big)} \int_{B(x;r)} f(y)\,dy.$$

The *Hardy–Littlewood maximal operator* is defined by

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

4.25 Proposition. Fix the dimension n, there is a constant C such that for any $\lambda \geq 0$ and $f \in L^1(\mathbf{R}^n)$, we have

$$m\{x \in \mathbf{R}^n : Mf(x) \geq \lambda\} \leq \frac{C}{\lambda} \int |f|\,dm.$$

Notice the resemblance with Markov's inequality

$$m\{x : |f(x)| \geq \lambda\} \leq \frac{1}{\lambda} \int |f|\,dm.$$

The above proposition tells us replacing $|f(x)|$ by the average of $|f(x)|$ over balls, we have the same inequality up to a dimension-dependent constant.

4.26 Lebesgue differentiation theorem. For $f \in L^1_{\text{loc}}$, we have $A_r f(x) \to f(x)$ for a.e. $x \in \mathbf{R}^n$.

This is the most applicable result out of this section, but one may generalize it with little effort. First, the convergence in fact holds in the L^1 sense. Second, the averaging does not have to be over balls. We say a collection $\{E_r\}_{r>0}$ of sets shrinks nicely to $x \in \mathbf{R}^n$ if $E_r \subseteq B(x;r)$ for each r, and a constant $\alpha > 0$ such that

$$m(E_r) \geq \alpha\, m\big(B(x;r)\big)$$

holds uniformly for all r. (Or course $E_r = B(x;r)$ is the trivial yet most important example.)

4.27 Lebesgue differentiation theorem, generalized. Given $f \in L^1_{\text{loc}}$ and E_r that shrinks nicely to x, we have for a.e. $x \in \mathbf{R}^n$

$$\lim_{r\to 0} \frac{1}{m(E_r)} \int_{E_r} |f(x) - f(y)|\, dy = 0,$$

and hence

$$\lim_{r\to 0} \frac{1}{m(E_r)} \int_{E_r} f(x)\, dx = f(x_0)$$

In fact this holds for Radon measures (Besicovitch density 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.

4.28 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

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

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 say a function $f\colon J \to \mathbf{R}$ is locally Lipschitz/BV/AC on the interval J if it is Lipschitz/BV/AC on any compact subintervals.

4.29 Fact. Convex functions are locally Lipschitz. Lipschitz functions are absolutely continuous.

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 $\mathbf{R}$.

4.30 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. It is not hard to imagine that we may generalize Theorem 1.34.

4.31 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

4.32 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 (measure theory)
[BS20, Corollary 4.3.8] For $f \in \mathrm{AC}[a,b]$, we have

$$V(f,[a,b]) = \int_a^b |f'(x)|\,dx.$$

Let $\gamma\colon [0,1] \to \mathbf{R}$ be a continuous curve on $\mathbf{R}$, it makes sense to define the length of the curve γ by

$$\mathrm{length}(\gamma) = V(\gamma,[0,1]).$$

When γ is absolutely continuous, we then recover our familiar definition

$$\mathrm{length}(\gamma) = \int_a^b |\gamma'(x)|\,dx.$$

This example may seem naïve, but the principle generalizes to continuous curves on $\mathbf{R}^n$ and even general metric spaces, by appropriately generalizing absolutely continuity, and defining the metric derivative by

$$|\gamma'(x)| = \lim_{h\to 0} \frac{\rho\big(\gamma(t+h),\gamma(t)\big)}{|h|}.$$

See [San15, Box 5.1 & 5.2] for a brief account, and [ABS24, Chapter 9][San23, Section 1.4] for a detailed account.

4.33 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 $\int_{\varphi(a)}^{\varphi(b)}$.

4.34 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

$$\int_{\varphi(a)}^{\varphi(b)} f(x)\,dx = \int_a^b f\big(\varphi(t)\big)\,\varphi'(t)\,dt. \tag{4.35}$$

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 (4.35) remains true.[7]

4.F Extension to $\mathbf{R}^n$ and general metric spaces

4.36 Rademacher's theorem. Let U be an open set in $\mathbf{R}^n$. Any function $f\colon U \to \mathbf{R}$ that is Lipschitz is differentiable a.e. in U. (The L^∞ norm[8] of ∇f is the Lipschitz constant of f.)

Global Lipschitz may be replaced by local Lipschitz, since U is σ-compact.

4.37 Aleksandrov's theorem.

[7] This is a frequent source of confusion in the undergraduate curriculum. For Riemann integrals, equation (4.35) always holds provided that f is continuous and $\varphi \in C^1$; this is just a consequence of the fundamental theorem of calculus and the chain rule. However, in many cases, we need to compute $\int_c^d f(x)\,dx$, and a choice of parametrization φ that connects between $x = c$ and d has to be made. Such a parametrization therefore has to be injective C^1, and we have

$$\int_c^d f(x)\,dx = \int_{\varphi^{-1}(a)}^{\varphi^{-1}(b)} f\big(\varphi(t)\big)\,\varphi'(t)\,dt.$$

It is noteworthy that for the general change of variables in $\mathbf{R}^n$, the injective C^1 assumption is absolutely necessary even if $f\big(\varphi(x)\big)\big|\det D\varphi(x)\big|$ is integrable. The idea is that, the geometry of $A \subseteq \mathbf{R}^n$ and an interval on the real line is just different; there is no fundamental theorem of calculus in $\mathbf{R}^n$.

[8] to be defined in the next chapter

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 $f \in \mathcal{L}^p$.

We have not yet checked that L^p is a vector space. The linearity follows from the very important inequality

$$|f + g|^p \leq (2 \max\{|f|, |g|\})^p \leq 2^p (|f|^p + |g|^p). \tag{5.1}$$

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 other function spaces.

5.2 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.)

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

5.4 Theorem. L^p is complete.

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

also holds for L^∞
the space of bounded measurable function is dense in L^p

5.6 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.

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

5.8 Theorem. Given two σ-finite measure spaces $(X, \mathcal{M}, \mu)$ and $(Y, \mathcal{N}, \nu)$, for $1 \le p < \infty$, we have

$$\overline{L^p(X) \otimes L^p(Y)} = L^p(X \times Y),$$

where $L^p(X) \otimes L^p(Y)$ is the algebraic tensor product of two L^p vector spaces,[1] and the closure is with respect to the norm of $L^p(X \times Y)$. We have made the identification $f \otimes g \leftrightarrow \big(x \mapsto f(x)g(x)\big)$.

Proof. We first show the case when μ and ν are finite. It suffices to show that $\overline{L^p(X) \otimes L^p(Y)}$ contains all the simple function $\sum_{j=1}^n c_j \mathbf{1}_{A_j}$. For $A_j \in \mathcal{M} \otimes \mathcal{N}$, there is a finite union E_j of measurable rectangles such that

$$\mu \times \nu(A_j \triangle E_j) < \epsilon$$

Therefore

$$\|\mathbf{1}_{A_j} - \mathbf{1}_{E_j}\|_p < \epsilon^{1/p},$$

and hence the simple function $\sum_{j=1}^n c_j \mathbf{1}_{A_j}$ can be approximated in L^p by $\sum_j c_j \mathbf{1}_{E_j}$. Since $\bigcup_j E_j$ can be written as a finite disjoint union of measurable rectangles, we have $\sum_j c_j \mathbf{1}_{E_j} \in L^p(X) \otimes L^p(Y)$. This proves the finite measure case.

When μ and ν are σ-finite, as usual we express X and Y into increasing unions $\bigcup_{k=1}^\infty X_k$ and $\bigcup_{k=1}^\infty Y_k$, where each X_k and Y_k has finite μ and ν measures, respectively.

Let $f \in L^p(X \times Y)$. From DCT we know we can find $k \in \mathbf{N}$ such that $\|f - f_k\|_p < \epsilon/2$, while $\|f_k - h\|_p < \epsilon/2$ for some $h \in L^p(X_k) \times L^p(Y_k)$. This completes the proof. □

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

5.9 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.

[ADM11, Remark 3.17] C_b is not dense in L^∞

5.10 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

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

[1] which means precisely that it consists of linear combinations of all functions fg, where $f \in L^p(X)$ and $g \in L^p(Y)$

This is an interpolation result for L^p spaces: if a function is both L^p and L^r, then it must be L^q for any $p < q < r$.

5.11 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 $q \geq p$, 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.

Assume $1 \leq p < q \leq \infty$.

5.12 Proposition [Fol99, Proposition 6.12]. In a finite measure space, $L^p \supseteq L^q$, with

$$\|f\|_p \leq \|f\|_q \mu(X)^{\frac{1}{p}-\frac{1}{q}}.$$

The case $\mu(X) = 1$ is nice.

It is very important to remember that the containment $L^p \supseteq L^q$ does not necessarily hold when the measure space is infinite. Consider $f = \frac{1}{x}$ on $(1, \infty)$. It is well-known that $f \notin L^1$, but $f \in L^2$.

5.13 Proposition [Fol99, Proposition 6.11]. Let A be any set, then $\ell^p(A) \subseteq \ell^q(A)$, with $\|f\|_p \geq \|f\|_q$.

Think about in $\mathbf{R}^2$, the ℓ^1-ball is contained in the ℓ^2-ball,..., and is all contained in the ℓ^∞-ball. But when thinking about abstract L^p spaces, the direction of containment is reversed. The geometry is different.

5.C Hilbert spaces and L^2

A *Hilbert space* is an inner space with a complete metric induced from the inner product. We assume the underlying field is $\mathbf{C}$ in this section for generality.

5.14 Proposition. An inner product space is a normed space with the *parallelogram law/polarization identity*:

$$\|x-y\|^2 + \|x+y\|^2 = 2\|x\|^2 + 2\|y\|^2 \quad \text{for all } x \text{ and } y.$$

5.15 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 $f(\lambda) \coloneqq \|u + \lambda v\|^2$ 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 $\frac{|\langle u,v\rangle|}{\langle u,v\rangle}$[2]. □

[2]This change-of-direction trick is a prevalent trick to extend results proved over real vector spaces to over complex vector spaces

The proof in fact shows that any real symmetric positive semidefinite bilinear form $\langle \cdot, \cdot \rangle \colon V \times V \to \mathbf{R}$ should satisfy the Cauchy–Schwarz inequality. A bilinear form is positive semidefinite means that $\langle v, v \rangle \geq 0$ with $\langle 0, 0 \rangle = 0$. (It would be positive definite if furthermore $\langle v, v \rangle = 0$ implies $v = 0$.)

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.

5.16 Projection theorem. Given a Hilbert space H and a closed convex subset Y,

(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

$$\operatorname{Re}\langle x - y, z - y \rangle \leq 0 \quad \text{for all } z \in Y. \tag{5.17}$$

(b) if Y is furthermore a closed subspace of H, then the characterization above for $\pi_Y(x)$ may be further replaced by

$$\langle x - y, z \rangle = 0 \quad \text{for all } z \in Y. \tag{5.18}$$

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 $n > m \geq 1$, 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

$$\|y_n - y_m\|^2 \leq 2\|x - y_n\|^2 + 2\|x - y_m\|^2 - 4D^2.$$

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

$$\begin{aligned} \|y - y'\|^2 &= 2\|x - y\|^2 + 2\|x - y'\|^2 - 4\left\|x - \frac{y + y'}{2}\right\|^2 \\ &\leq 2D^2 + 2D^2 - 4D^2 = 0. \end{aligned}$$

Now we want to show this y satisfies (5.17). Let $z \in Y$ be arbitrary. To get (the real part of) the inner product[3] we consider the expression

$$f(\lambda) \coloneqq \|\lambda(z - y) - (x - y)\|^2 = \|y + \lambda(z - y) - x\|^2.$$

[3] like in the proof of Cauchy–Schwarz

For all $\lambda \in [0,1]$, by convexity $y + \lambda(z-y) \in Y$, and hence $f(\lambda) \geq \|x-y\|^2$. Now expanding $f(\lambda)$ gives us

$$\lambda^2 \|z-y\|^2 - 2\lambda \operatorname{Re}\langle x-y, z-y\rangle \geq 0.$$

Hence

$$\lambda \|z-y\|^2 \geq 2 \operatorname{Re}\langle x-y, z-y\rangle \quad \text{for all } \lambda \in [0,1],$$

and take $\lambda \to 0^+$ gives us (5.17).

For the converse, now suppose (5.17) holds for some $y \in Y$, and we want to show

$$\|x-y\| \leq \|x-z\| \quad \text{for all } z \in Y.$$

We trace our steps back: first,

$$2 \operatorname{Re}\langle x-y, z-y\rangle \leq 0 \leq \|z-y\|^2.$$

It follows that

$$\|x-y\|^2 \leq \|(z-y)-(x-y)\|^2,$$

as desired.

(b) To show the second part, it suffice to prove that (5.17) and (5.18) are equivalent. Because Y is now a subspace of H, equation (5.17) is equivalent to

$$\operatorname{Re}\langle x-y, z\rangle = 0 \quad \text{for all } z \in Y.$$

Notice that

$$\operatorname{Im}\langle x-y, z\rangle = \operatorname{Re} -i\langle x-y, z\rangle = \operatorname{Re}\langle x-y, iz\rangle,$$

which completes the proof. □

5.19 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) $\operatorname{range} \pi_Y = Y$ and $\operatorname{null} \pi = Y^\perp$;

(d) $\|\pi_Y(x)\| \leq \|x\|$ for all $x \in H$.

5.20 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

$$\langle e_\alpha, e_\beta\rangle = \begin{cases} 1 & \alpha = \beta, \\ 0 & \alpha \neq \beta. \end{cases}$$

The order of α does not matter when A is countable.

Given a linearly independent countable list $\{v_j\}_{j=1}^\infty$ of vectors in H, we can always use Gram–Schmidt process to obtain an orthonormal list $\{e_j\}_{j=1}^\infty$.

5.21 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 $\pi_Y(x) = \sum_{j=1}^n \langle x, e_j\rangle e_j$.

5.22 Proposition. $\sum_{j=1}^\infty \lambda_j e_j$ converges in H if and only if $\sum_{j=1}^\infty |\lambda_j|^2 < \infty$.

Proof. It suffices to show that the partial sum $S_n = \sum_{j=1}^n \lambda_j e_j$ forms a Cauchy sequence.

$$\begin{aligned}\|S_{m+n} - S_n\|^2 &= \left\| \sum_{j=n+1}^{n+m} \lambda_j e_j \right\|^2 \\ &= \sum_{j=n+1}^{n+m} |\lambda_j|^2 \\ &= \sum_{j=1}^{n+m} |\lambda_j|^2 - \sum_{j=1}^{n} |\lambda_j|^2\end{aligned}$$

Since the partial sums $\sum_{j=1}^n |\lambda_j|^2$ form a Cauchy sequence, S_n must be Cauchy as well. □

5.23 Theorem. Let $\{e_\alpha\}_{\alpha\in A}$ be an orthonormal system, then

(a) for each $x \in H$, $\sum_{\alpha\in A} \langle x, e_\alpha\rangle^2 \le \|x\|^2$, which is known as *Bessel's inequality*.
(b) The equality above holds for each $x \in H$ if and only if the series $x = \sum_{\alpha\in A} \langle x, e_\alpha\rangle e_\alpha$ in H. This equality is known as *Parseval's identity*.

We say $\{e_\alpha\}$ is a(n) *orthonormal basis/complete orthonormal system* if and only if Parseval's identity holds for all x. This is equivalent to saying that for any $x \in H$, such that $\langle x, e_\alpha\rangle = 0$ for all α, then $x = 0$. We also have a third characterization below.

5.24 Theorem. Let $\{e_j\}_{j=1}^\infty$ be a countable orthonormal system, then it is an orthonormal basis of H if and only if $\operatorname{span}\{e_j\}$ is dense in H.

Proof. Every $x \in H$ is the limit of the finite sum $\sum_{j=1}^n \langle x, e_j\rangle e_j$, which shows that $\operatorname{span}\{e_j\}$ is dense in H. Conversely, suppose $\operatorname{span}\{e_j\}$ is dense in H. Fix x and let $\epsilon > 0$. Then for $m > n$, we have

$$\left\| x - \sum_{j=1}^m \langle x, e_j\rangle e_j \right\| \le \left\| x - \sum_{j=1}^n \langle x, e_j\rangle e_j \right\| \le \left\| x - \sum_{j=1}^n \lambda_j e_j \right\|,$$

where $\{\lambda_j\}_{j=1}^n$ are picked such that the last expression is less than ϵ. We have applied Proposition 5.21 twice. It follows that $\sum_{j=1}^\infty \langle x_j, e_j\rangle e_j = x$ for all x, as desired. □

(This suggests that in a separable Hilbert space H, it is appropriate to define an (ordinary) basis as a countable linearly independent list of vectors whose span is dense in H. However, this definition is nonstandard.)

We mention that it is not always preferred to work with normalized bases. For orthogonal bases, the above results can be appropriately carried over almost without change.

Orthonormal decomposition

5.25 Theorem. H has a countable orthonormal basis if and only if H is separable. Additionally in this case, all bases have the same cardinality.

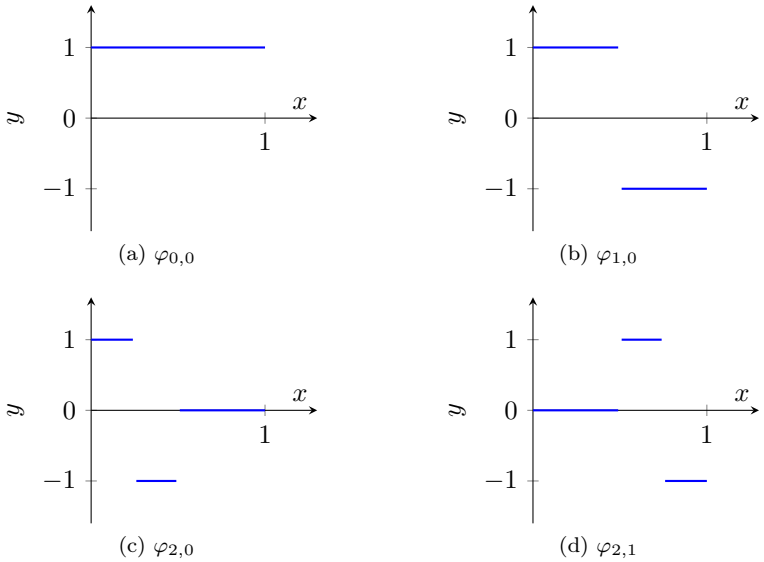


Figure 5.1: Haar basis functions, before normalization

There is a canonical choice of orthonormal basis on the Hilbert space $L^2[0,1]$. The Haar basis functions consist of $\varphi_{0,0} = 1$ and $\{\sqrt{2^{m-1}}\varphi_{m,k} : m \geq 1 \text{ and } 0 \leq k < 2^{m-1}\}$, where

$$\varphi_{m,k} = \mathbf{1}\left[\frac{2k}{2^m}, \frac{2k+1}{2^m}\right] - \mathbf{1}\left[\frac{2k+1}{2^m}, \frac{2k+2}{2^m}\right];$$

see Fig. 5.1. It is clear that any two φ's are orthogonal. Since $\|\varphi_{m,k}\|^2 = 2^{m-1}$, the Haar basis functions form an orthonormal basis on $L^2[0,1]$.

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 $L^2[0,1]$.

Given two Hilbert spaces H_1 and H_2, we have the algebraic tensor $H_1 \otimes H_2$ with the inner product

$$\langle f_1 \otimes g_1, f_2 \otimes g_2\rangle = \langle f_1, f_2\rangle_{H_1}\langle g_1, g_2\rangle_{H_2},$$

for all $f_1, f_2 \in H_1$ and $g_1, g_2 \in H_2$. It is possible to extend this inner product space to a Hilbert space by metric completion, which one can read from Appendix A. We define the *Hilbert tensor product* $H_1 \widehat{\otimes} H_2$ to be the completion of the inner product space $H_1 \otimes H_2$.[4]

5.26 Proposition. Given two separable Hilbert spaces H_1 and H_2 with orthonormal bases $\{e_j\}_{\alpha\in A}$ and $\{f_\beta\}_{\beta\in B}$ respectively. Then $\{e_\alpha \otimes f_\beta : \alpha \in A, \beta \in B\}$ is an orthonormal basis for $H_1 \widehat{\otimes} H_2$.

[4]It is noteworthy that $H_1 \widehat{\otimes} H_2$ may alternatively be seen as the space of Hilbert–Schmidt operators from H_1^* to H_2 with the Hilbert-Schmidt/Frobenius inner product. Hilbert–Schmidt operators are usually studied in compact operator theory, which is not part of the current text. One can define two natural tensor products for general Banach spaces, when we do not have the Hilbert space structure. They are the completion of the algebraic tensor with respect to two different norms, and respectively correspond to the spaces of nuclear and compact operators. For discussion, see [BS20, Section 7.10(vi)].

Proof. The fact that

$$\{e_j \otimes f_k : j, k \in \mathbf{N}\}$$

is an countable orthonormal system is clear by the definition of the inner product on $H_1 \widehat{\otimes} H_2$. In light of Theorem 5.24, it suffices to prove

$$\operatorname{span}\{e_j \otimes f_k : j, k \in \mathbf{N}\} \text{ is dense in } H_1 \otimes H_2,$$

which is dense in $H_1 \widehat{\otimes} H_2$.

Consider any $\epsilon > 0$, and any element $\sum_{i=1}^n c_i(\varphi_i \otimes \psi_i) \in H_1 \otimes H_2$. For each $\varphi_i \in H_1$, we can find some basis element $e_{i'}$ such that

$$c_i \|\varphi_i - e_{i'}\| \, \|\psi_i\| < \epsilon.$$

After $e_{i'}$ has been chosen, for $\psi_i \in H_2$, we can find some $f_{i'}$ such that

$$c_i \|e_{i'}\| \, \|\psi_i - f_{i'}\| < \epsilon.$$

Now we compute

$$\begin{aligned}
\left\| \sum_{i=1}^n c_i(\varphi_i \otimes \psi_i) - \sum_{i=1}^n c_i(e_{i'} \otimes f_{i'}) \right\| &\leq \sum_{i=1}^n c_i \|\varphi_i \otimes \psi_i - e_{i'} \otimes f_{i'}\| \\
&= \sum_{i=1}^n c_i \big(\|(\varphi_i - e_{i'}) \otimes \psi_i\| + \|e_{i'} \otimes (\psi_i - f_{i'})\| \big) \\
&\leq \sum_{i=1}^n c_i \big(\|(\varphi_i - e_{i'})\| \, \|\psi_i\| + \|e_{i'}\| \, \|(\psi_i - f_{i'})\| \big) \\
&= 2n\epsilon.
\end{aligned}$$

This shows precisely that any element in $H_1 \otimes H_2$ can be approximately by some element in the span of $\{e_j \otimes f_k\}$, and our proof is complete. □

From Theorem 5.8, we know precisely that $L^2(X) \widehat{\otimes} L^2(Y) = L^2(X \times Y)$. This now allows us to conclude that

5.27 Corollary. Let $(X, \mathcal{M}, \mu)$ and $(Y, \mathcal{N}, \nu)$ be σ-finite measure spaces with separable L^2 spaces, and let $\{e_j\}_{j=1}^\infty$ and $\{f_k\}_{k=1}^\infty$ be two orthonormal bases for $L^2(X)$ and $L^2(Y)$ respectively. Then $L^2(X \times Y, \mu \times \nu)$ admits $\big\{(x, y) \mapsto e_j(x) f_k(y) : j, k \in \mathbf{N}\big\}$ as its orthonormal basis.

5.D Duality of L^p

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

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

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 $1 < p < \infty$.

L^p spaces are reflexive for $1 < p < \infty$

$(\ell^1)^* = \ell^\infty$, then automatically $(\ell^\infty)^* = (\ell^1)^{**} \supseteq \ell^1$. The containment however is strict. Define c_0 to be the space of sequences that converges to 0. It can be proved that $c_0^* = \ell^1$ (in isomorphism).[5] (The generalization of this to general measure space will be the upcoming Riesz–Markov–Kakutani theorem (finite measures).) However, c_0 is a proper subspace of ℓ^∞, and therefore $c_0^* \subsetneq (\ell^\infty)^*$. Furthermore, by (B.13) we even have $(\ell^\infty)^* = (c_0)^\perp \oplus \ell^1$.

5.29 Exercise. Give a direct proof of $(\ell^\infty)^* = (c_0)^\perp \oplus \ell^1$.

5.E The L^0 space

We use $L^0(X, \mathcal{A}, \mu)$ to represent the space of $\mathcal{A}$-measurable functions that are identified μ-a.e. Let μ be a finite measure. On this space we have convergence in measure, and therefore it would be nice if we can metrizes this topology. This is possible, in fact, by a collection of metrics. The neighborhood base of $f \in L^0$ is given by $U_{f,\epsilon} = \{g \in L^0 : \mu\{|f(x) - g(x)| > \epsilon\} < \epsilon\}$

The topology can be defined by any metric $d(\,\cdot\,,\,\cdot\,)$ of the form

$$d(f, g) = \int_X \varphi(|f - g|)\, d\mu$$

for any bounded continuous concave increasing $\varphi\colon [0, \infty) \to \mathbf{R}$ such that $\varphi(0) = 0$ and $\varphi(t) > 0$ for $t > 0$. For example, one may take $\varphi(t) = |t| \wedge 1$ or $\varphi(t) = \frac{t}{1+t}$.

And there is also the *Ky Fan metric*,[6], which defines

$$\alpha(f, g) = \inf\{\epsilon \geq 0 : \mu\{x : |f - g| > \epsilon\} < \epsilon\}.$$

5.30 Theorem. L^0 is complete under any of the aforementioned metrics.

In some sense, lifting the separability and completeness from the image space $\mathbf{R}$ to L^0 should be expected. You have a sequence converging in $\mathbf{R}$ (which is a complete metric space), when integrated the space of functions should also be complete. Later when studying the space of probability measures on a separable metric space (with the topology of weak convergence), this lifting of (sequential) compactness, separability, and completeness will appear again.

5.F Riesz' theorems and convergence of measures

Let X be a Hausdorff space. We define $C_c(X)$ to be the space of continuous functions on X with compact support, and define $C_0(X)$ to be the space of continuous functions on X such that $\{x : |f(x)| \geq \epsilon\}$ is compact[7] for all $\epsilon > 0$.[8] Lastly, $C_b(X)$ is the space of bounded continuous functions on X. All three spaces are endowed with the uniform (or supremum) norm $\|\cdot\|_u$, and it is quite easy to verify that

[5] We may also use c_{00}, the space of sequences that are eventually zero, to replace its closure c_0.

[6] this is more well-known in probability

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

[8] Analysts sometimes asks X to be locally compact, for some technical reason from C-star algebra. The definitions still make perfect sense without such consideration.

5.31 Exercise. C_b is complete, and hence its closed subspace C_0 is complete.

5.32 Proposition. 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.

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)

5.33 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 $f\colon X \to [0,1]$ with $f = 1$ on K and $\operatorname{supp} f$ 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 $\overline{G} \subseteq \overline{B}(x;r) \subseteq U$.[9] 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. Consider the collection of all precompact open sets A with $\overline{A} \subseteq U$. By the above this collection forms an open cover of the compact set K, and therefore we have a finite subcover $\{A_1, \dots, A_m\}$ that covers K. Set $G = \bigcup_{k=1}^m A_k$, which is open. Since $\overline{G} = \bigcup_{k=1}^m \overline{A_k}$ 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 $f\colon X \to [0,1]$ 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*.

5.34 Remark. A more common (and probably natural) way to extend from the case $\{x\}$ to a general compact set K is to consider the collection $\{G_x\}_{x \in K}$ instead. However, this means that we need to choose a ball $B(x; r_x)$ around each $x \in K$, and the full axiom of choice has to be invoked. A similar situation appears in the proof of the Lebesgue's number lemma, where we are also tempted to "choose" an open neighborhood around each point to obtain an open cover; the choice can be avoided in the same way.

5.35 Proposition. Let X be any metric space. $C_c(X)$ is a dense subset of $C_0(X)$.

Proof. Let $f \in C_0(X)$. We know for any $\epsilon > 0$ that $K = \{x\colon |f(x)| \geq \epsilon\}$ and $E = \{x\colon |f(x)| \geq \epsilon/2\}$ are both compact. Meanwhile $U = \{x\colon |f(x)| > \epsilon\}$ is open, with

$$K \subseteq U \subseteq E.$$

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

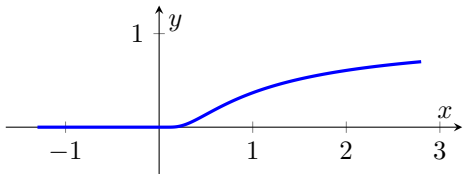


Figure 5.2: the function defined in (5.37), which is smooth at 0

It follows by the preceding lemma that there exists a continuous function $\varphi\colon E \to [0,1]$ with $f|_K = 1$ and $\operatorname{supp} f \subseteq U$. Obviously φ may be extended to the entire X by setting $\varphi|_{X-E} = 0$.

Notice that $\|f - \varphi f\|_u < \epsilon$, where $\varphi f \in C_c$. This proves the claim. □

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

5.36 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 $g_0 \coloneqq 1 - f$ 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.

Recall that the function

$$f(x) = \begin{cases} \exp(-1/x) & \text{if } x > 0, \\ 0 & \text{if } x \leq 0. \end{cases} \tag{5.37}$$

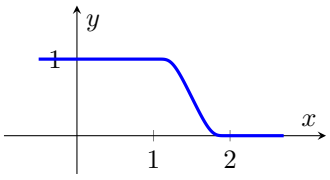


Figure 5.3: the transition function defined in (5.38), when $a = 1$ and $b = 2$

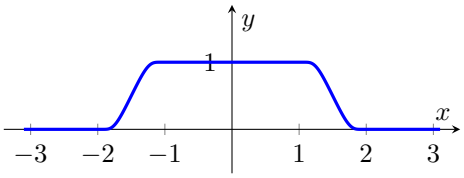


Figure 5.4: the bump function defined by h in dimension 1

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

$$g(x) = \frac{f(b-x)}{f(b-x) + f(x-a)}. \tag{5.38}$$

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 (5.37), 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.

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.F.2 Spaces of test functions

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

5.39 Definition.

(a) Let X be a metric space. A sequence $\{\mu_n\} \subseteq \mathcal{M}(X)$ is said to *converge weakly* to μ if for all $f \in C_b(X)$, we have

$$\int_X f \, d\mu_n \to \int_X f \, d\mu, \tag{5.40}$$

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

(b) If we assume furthermore that X is a locally compact and separable, then we say $\{\mu_n\}$ *converges vaguely* to μ if for all $f \in C_c(X)$, we have

$$\int_X f\,d\mu_n \to \int_X f\,d\mu. \tag{5.41}$$

It is common to see the notation μf in place of $\int_X f\,d\mu$, because we may see μ as a linear operator acting on the continuous function space, called the space of *test functions*.

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 (5.40), 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 ...

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.

5.42 Definition. Let X be a locally compact metric space. A positive *Radon measure*[10] 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.

5.43 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.

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. Hence for finite measures, Radon measures are the same as compact inner regular measures.[11] And by Theorem 1.40, if X is a σ-compact metric space, any finite Borel measure is Radon. This passes to signed and complex measures.

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

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

In notation, this topology is $\sigma\big(\mathcal{M}_{\mathrm{R}}(X), C_c(X)\big)$. Vague convergence is interesting because it is precisely this weak-star convergence (given that X is locally compact and separable.)

Some authors prefer to define vague convergence instead using C_0 test functions. There are certainly benefits of this. Recall that weak-star limit of bounded linear functionals does not have be bounded. According to Appendix C, since C_0 is Banach, any weak-star limit on C_0^* must fall in $\mathcal{M}_{\mathrm{R}} = C_0^*$. On the other hand, for $\{\mu_n\} \subseteq C_c^* = \mathcal{M}_{\mathrm{R}}$, since the space of test functions are smaller (the behavior of μ_n at infinity are no longer considered), it is possible that $\mu_n \to \mu$ in weak-star on C_c^* with $\|\mu\| = \infty$, but not weak-star on C_0^*.

The above paragraph can be really subtle and confusing, but fortunately there is no actual difference when applying either definitions. If $\sup_n \|\mu_n\| < \infty$, then it is equivalent to say $\int f\,d\mu_n \to \int f\,d\mu$ either over all $f \in C_c(X)$ or over all $f \in C_0(X)$, as a consequence of Proposition C.8. In particular there is no difference on the space of subprobability measures, which will be the only thing we are interested in.

[10] also known as the *Borel regular measure* or the *Riesz measure*

[11] This is precisely the definition given by Bogachev [Bog07; Bog18].

5.44 Riesz–Markov–Kakutani theorem (positive measures). Let X be a locally compact metric space. Given a positive linear function L on $C_c(X)$, there is a unique Radon measure μ on X such that

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

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

This μ satisfies

$$\mu(U) = \sup\{Lf : f \in C_c(X), f \prec U\} \text{ for } U \text{ open},$$

and

$$\mu(K) = \inf\{Lf : f \in C_c(X), f \geq \mathbf{1}_K\} \text{ for } K \text{ compact}.$$

It is well-known that Rudin [Rud87] proved the existence of the Lebesgue measure with this theorem. More generally, one can use this to show the existence of a Radon measure on general Riemannian manifolds.

5.45 Riesz–Markov–Kakutani theorem (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 compact inner regular Borel measure μ such that

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

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,[12] then $C_c(X) = C(X)$.

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

5.46 Corollary. $\mathcal{M}_{\mathrm{R}}(X)$ is a closed subspace of $\mathcal{M}(X)$ in the strong topology, and hence a Banach space.

Proof. In the strong topology, $\mathcal{M}_{\mathrm{R}}(X)$ is a complete subspace in $\mathcal{M}(X)$, which is complete. The vague limit of Radon measures is Radon because just because the weak-star limit on C_c must still fall in C_c. □

Notice that 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 $\mu(X) \geq 0$.

5.47 Proposition. In a locally compact metric space X with $\mu_n, \mu \in \mathcal{M}_{\mathrm{R}}^+(X)$ 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.

[12] Compact metric spaces are separable; see Proposition A.17.

5.G Convolutions and smooth approximation of functions

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.

5.48 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

5.49 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 2.45.

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 (finite measures) holds for C_c^∞ test functions when the space $X = \mathbf{R}^n$.

given $f \in C_c$, consider f^ϵ

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

fundamental theorem of Calculus of Variations

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

$$\int fg = 0$$

for all $g \in C_c^\infty(U)$, then $f = 0$ a.e. on U. In particular, $f = 0$ at all continuity points of f.[13]

5.51 Definition.

5.52 Young's inequality. For $1 \le p, q, r \le \infty$ such that $\frac{1}{p} + \frac{1}{q} = \frac{1}{r} + 1$, 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

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

The inequality is sharp only when p or q is 1. In general can improve the inequality by a better constant, see [Bar98]. Moreover, in the same article a reverse Young's inequality is also mentioned:

5.53 Theorem. For $0 < p, q, r \le 1$ such that $\frac{1}{p} + \frac{1}{q} = \frac{1}{r} + 1$, and measurable $f, g \ge 0$, we have

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

The optimal constant version is also proved in [Bar98].

Proof. We follow [Lei72], where this result was originally proved. □

Define the *convolution of two measures* μ and ν by $\mu * \nu = a_*(\mu \times \nu)$, where $a \colon \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}^n$ is the addition map. Equivalently, for all f nonnegative measurable or in L^1 that

$$\int f \, d(\mu * \nu) = \int \int f(x + y) \, d\mu(x) \, d\nu(y).$$

Because of the commutativity and associativity of products of measures, we know convolution of measures must be commutative and associative as well.

This definition is quite natural. Take $f = d\mu/dx$ and $g = d\nu/dy$, then we would like to have

$$d(\mu * \nu)(x)/dx = f * g(x) = \int_{\mathbf{R}^n} f(x - y) g(y) \, dy.$$

Therefore for all measurable A we have

$$\mu * \nu(A) = \int_A \int f(x - y) g(y) \, dy \, dx,$$

and by a change of variables we have

$$\mu * \nu(A) = \int_A \int f(x - y) \, d\nu(y) \, dx = \int_{z+y \in A} d\nu(y) \int f(z) \, dz = \mu \times \nu(z + y \in A).$$

[13] Otherwise there is an open neighborhood around some continuity point that is nonzero.

5.H Fourier transform of functions and measures

Define the *Fourier transform* of a complex integrable function f to be

$$\hat{f}(t) = \int_{\mathbf{R}^n} \exp(-it \cdot x) f(x)\, dx.$$

Define the *Fourier transform* of a complex Borel measure on $\mathbf{R}^n$ by

$$\hat{\mu}(t) = \int_{\mathbf{R}^n} \exp(-it \cdot x)\, d\mu(x)$$

Note that $|\exp(-it \cdot x)| = |\cos(-t \cdot x) + i \sin(-t \cdot x)| = 1$ for all x and t, which means the definitions above make sense.

The two are intimately related. Given any $f \in L^1(m)$, the recipe $d\nu = f\, dm$ gives us a complex Borel measures ν. In fact, Radon–Nikodym theorem tells us that all complex Borel measures absolutely continuous with respect to the Lebesgue measure can be given by such a recipe. Notice that $\|f\|_1 = \int_{\mathbf{R}^n} |f|\, dm = \|\nu\|$, so the equivalence classes $L^1(m)$ of integrable functions can be isometrically embedded into $\mathcal{M}(\mathbf{R}^n)$.

We let the reader recognize the following

$$\hat{\mu}(0) = \|\mu\|, \quad |\hat{\mu}(t)| \le 1, \quad \hat{\mu}(-t) = \overline{\hat{\mu}(t)}.$$

The Fourier transform of a measure is uniformly continuous.

Recall Corollary 2.44 to DCT discusses the continuity of the integral of parametrized functions. Here we need to further prove uniform continuity.

$$\begin{aligned} |\hat{\mu}(t+h) - \hat{\mu}(t)| &\le \left| \int_{\mathbf{R}^n} \exp(-it \cdot x)[\exp(-ih \cdot x) - 1]\, d\mu(x) \right| \\ &\le \int |\exp(-ih \cdot x) - 1|\, d\mu(x), \end{aligned} \tag{5.54}$$

where we have used $|\exp(-it \cdot x)| = 1$ to let t vanish. Now $|\exp(-ih \cdot x) - 1| \le 2$ and $\lim_{h \to 0} |\exp(-ih \cdot x) - 1| = 0$, so (5.54) converges to 0. This establishes uniform continuity.

$\|\hat{\mu}\|_u \le \|\mu\|_1$ because

$$\sup_t |\hat{f}(t)| \le \sup_t \int |\exp(-it \cdot x)|\, |f(x)|\, dx = \|f\|_1.$$

$\widehat{\mu * \nu} = \hat{\mu}\hat{\nu}$

$$\begin{aligned} \widehat{\mu * \nu}(t) &= \int_{\mathbf{R}^n} \exp\bigl(-it \cdot z\bigr)\, d(\mu * \nu)(z) \\ &= \int_{\mathbf{R}^n} \int_{\mathbf{R}^n} \exp\bigl(-it \cdot x\bigr) \exp(-it \cdot y)\, d\mu(x)\, d\nu(y). \\ &= \hat{\mu}(t)\hat{\nu}(t). \end{aligned}$$

All the properties above carry directly over to the Fourier transform of integrable functions.

5.55 Riemann–Lebesgue lemma. Fourier transform of $L^1(\mathbf{R}^n)$ falls in $C_0(\mathbf{R}^n)$. This means precisely that $\lim_{|t| \to \infty} \hat{f}(t) = 0$.

This does translate to measures: for the atomic measure δ_0, $\hat{\delta}_0(t) = 1$, independent of t.

Define the inverse Fourier transform of a function by

$$\check{f}(x) = \frac{1}{2\pi} \int_{\mathbf{R}^n} \exp(it \cdot x) f(t)\, dt,$$

and the inverse Fourier transform of a measure by

$$\check{\mu}(x) = \frac{1}{2\pi} \int_{\mathbf{R}^n} \exp(it \cdot x)\, d\mu(t).$$

The *Schwartz space* $\mathcal{S}(\mathbf{R}^n; \mathbf{C})$ plays an important role in the Fourier (and distribution) theory. It consists of all *rapidly decreasing functions* whose derivatives (including itself) are decreasing at a rate faster than any powers: we say $f \in \mathcal{S}$ if $f \in C^\infty$, and

$$p_{\beta,\alpha}(f) = \sup_x |x^\beta \partial^\alpha f(x)| < \infty \quad \text{for all multi-indices } \beta, \alpha.$$

This is equivalent to

$$p_{N,\alpha}(f) = \sup_x (1 + |x|)^N \partial^\alpha f(x) < \infty \quad \text{for all } N \in \mathbf{N}_0 \text{ and multi-indices } \alpha.$$

(Some authors may even prefer the Japanese bracket $\langle x \rangle^N = (1 + |x|^2)^{N/2}$ instead, a nowhere vanishing and smooth analogue of $|x|^N$.)

For the unfamiliar reader, on $\mathbf{R}^n$, a *multi-index* α is an n-tuple of nonnegative integers, which lets us write $x^\beta = x_1^{\beta_1} \cdots x_n^{\beta_n}$, and $\partial_x^\alpha f(x) = \partial_{x_1}^{\alpha_1} f(x) \cdots \partial_{x_n}^{\alpha_n} f(x)$. Note that $|\alpha| = \sum_{j=1}^n \alpha_j$.

The topology we endow on $\mathcal{S}$ is the seminorm topology. One can show that $p_{\beta,\alpha}$ is a seminorm for all β and α. Also $p_{N,\alpha}$ is a seminorm. In fact, one may even define $p_\alpha = \sup_{k \le N} p_{k,\alpha}$, which is again a seminorm. (In fact all these are norms: since if $p_\cdot(f) = 0$, then $f \equiv 0$. However, this is irrelevant to topology.)

We will stick to the $p_{\beta,\alpha}$ seminorms henceforth, although any choices of seminorms should lead to the exact same conclusion. Define

$$d(f, g) = \sum_{\beta,\alpha} 2^{-|\beta|-|\alpha|} \frac{p_{\beta,\alpha}(f - g)}{1 + p_{\beta,\alpha}(f - g)},$$

which metrizes the topology on $\mathcal{S}$.

5.56 Proposition. The metric is complete, and hence $\mathcal{S}$ is a Fréchet space.

Proof. Suppose $d(f_m, f_n) \to 0$ as $m > n \to \infty$, then $\|f_m - f_n\|_{\beta,\alpha} \to 0$. This means that for each (β, α), the complex-valued sequence of functions $\{x^\beta \partial^\alpha f_n\}$ is uniformly Cauchy, and hence converges to some $f^{\beta,\alpha}$. By uniform convergence of derivatives, $f^{\beta,\alpha}$ has to be $x^\beta \partial^\alpha f$. Therefore $\|f_n - f\|_{\beta,\alpha} \to 0$ for all β, α, i.e., $d(f_n, f) \to 0$. □

In fact, $\mathcal{S}$ is also separable. Although there are probably other ways to approach this, we will prove it using Hermite polynomials. One should notice the close interconnections between Fourier transforms, Hermite polynomials, and Gaussian measures. (This ultimately indicates why Fourier transforms ultimately leads to the classical yet most magical proof of the central limit theorem.)

Note that $C_c^\infty \subseteq \mathcal{S}$.

For $f \in \mathcal{S}$, we have $\hat{f} \in \mathcal{S}$.

$\hat{\hat{f}}(x) = 2\pi f(-x)$

5.57 Parseval's identity. For $f, g \in \mathcal{S}(\mathbf{R}^n)$, we have

$$\int fg = 2\pi \int \hat{f}\hat{g}.$$

This implies the Plancherel's identity $\|f\|_{L^2} = \sqrt{2\pi}\|\hat{f}\|_{L^2}$.

Since $\mathcal{S}$ is dense in L^2, this allows us to extend the Fourier transform to an operator on L^2. One key question now is how does it relate to the Fourier transform on L^1. As expected, the definitions are consistent.

5.58 Theorem. For $f \in L^1 \cap L^2$, we have $\hat{f} = \mathcal{F}f$, where we used $\hat{f}$ for the Fourier transform of f in L^1 and $\mathcal{F}$ for the Fourier transform of f in L^2.

5.59 Theorem. The Fourier transform from $\mathcal{S}$ to itself is a linear homeomorphism.

We say $\varphi\colon \mathbf{R}^n \to \mathbf{C}$ is a *positive semidefinite function* if for any finite number of $t_1, \ldots, t_m \in \mathbf{R}^n$, the matrix $[\varphi(t_j - t_k)]_{1\le j,k\le m}$ is positive semidefinite. Equivalently this means that $\varphi(t) = \overline{\varphi(-t)}$, and $\sum_{j,k=1}^m \lambda_j \overline{\lambda_k}\varphi(t_j - t_k) \ge 0$ for any $(\lambda_1, \ldots, \lambda_m) \in \mathbf{C}^m$.

5.60 Proposition. $\hat{\mu}$ (and $\check{\mu}$) is positive semidefinite.

Proof. $\varphi(t) = \overline{\varphi(-t)}$ has already been proven. Now fix $t_1, \ldots, t_m \in \mathbf{R}^n$, and computation gives us

$$\begin{aligned}\sum_{j,k=1}^m \lambda_j \overline{\lambda_k}\varphi(t_j - t_k) &= \int \sum_{j,k=1}^m \lambda_j \overline{\lambda_k} \exp(it_k x)\exp(-it_j x)\,dx \\ &= \int \Big|\sum_{j=1}^m \lambda_j \exp(-it_j x)\Big|^2\,dx \ge 0\end{aligned}$$

for any λ. □

5.61 Lemma. A positive semidefinite function $\varphi\colon \mathbf{R}^n \to \mathbf{C}$ continuous at 0 is uniformly continuous everywhere.

Proof. We will show that

$$|\varphi(t+h) - \varphi(t)|^2 \le \varphi(0)[2\varphi(0) - \varphi(h) - \varphi(-h)]. \tag{5.62}$$

Consider the positive semidefinite (PSD) matrix

$$\begin{bmatrix} \varphi(0) & \varphi(t) & \varphi(t+h) \\ \varphi(-t) & \varphi(0) & \varphi(h) \\ \varphi(-t-h) & \varphi(-h) & \varphi(0) \end{bmatrix}$$

By multiplying on the left by the $P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$ and on the right by P^{-1}, we obtain its congruent matrix

$$\begin{bmatrix} \varphi(0) & \varphi(t) & \varphi(t+h) - \varphi(t) \\ \varphi(-t) & \varphi(0) & \varphi(h) - \varphi(0) \\ \varphi(-t-h) - \varphi(-t) & \varphi(-h) - \varphi(0) & 2\varphi(0) - \varphi(h) - \varphi(-h) \end{bmatrix}.$$

It is clear that the matrix remains PSD after the congruence transformation. Now the submatrix from the four corners must also be PSD, and (5.62) follows by considering the determinant. □

The following fact follows straight from the definition.

5.63 Fact. The pointwise limit of a sequence of positive semidefinite functions is positive semidefinite.

The remaining results explain our interest in the Fourier transform of measures. The proofs are purely Fourier-theoretic, but their interpretations only make sense in the context of probability. We will restate these results later using the inverse Fourier transform, which is the convention in probability theory.

5.64 Inversion formula on the real line. For $\mu \in \mathcal{M}(\mathbf{R})$, we have for any $a < b$ that

$$\frac{1}{2}\mu\{a\} + \mu(a,b) + \frac{1}{2}\mu\{b\} = \lim_{T\to\infty} \int_{-T}^{T} \frac{\exp(ibt) - \exp(iat)}{2\pi i t} \hat{\mu}(t)\, dt$$

(If a and b are not atoms of μ_F, then the expression above is equal to $\mu(a,b]$.)

In particular, this implies that on the real line, μ can be uniquely determined by $\hat{\mu}$. Take $a \to -\infty$, we have $\frac{1}{2}\{\mu(-\infty,b) + \mu(-\infty,b]\}$ on the left-hand side. Now take $b \to x^+$, we have precisely recovered the distribution function of μ.

The correspondence between μ and $\hat{\mu}$ can proved directly on $\mathbf{R}^n$ using the tools we developed. However, an explicit inversion formula is beyond our reach, since Borel measures on $\mathbf{R}^n$ are much harder to describe.

5.65 Theorem. For any $\mu \in \mathcal{M}(\mathbf{R}^n)$, there is a one-to-one correspondence between each μ and $\hat{\mu}$.

Proof. Say $\mu \neq \nu$, which implies there is some $f \in \mathcal{S}$ such that $\int f\, d\mu \neq \int f\, d\nu$. Notice the important identity

$$\begin{aligned}\int f\, d\mu &= \int\int \check{f}(t) \exp(-it\cdot x)\, dt\, d\mu(x) \\ &= \int\int \check{f}(t) \exp(-it\cdot x)\, d\mu(x)\, dt = \int \check{f}(t)\hat{\mu}(t)\, dt,\end{aligned}$$

where we have used Fubini. It follows that $\int \check{f}(t)\hat{\mu}(t)\, dt \neq \int \check{f}(t)\hat{\nu}(t)\, dt$, which implies $\hat{\mu} \neq \hat{\nu}$, thus proving the contrapositive. □

In summary, we have reduced the question of correspondence between μ and $\hat{\mu}$ to the linear homeomorphism of the Schwartz space under Fourier transform of functions. This is also the key ingredient to the next (closely connected) theorem.

5.66 Lévy's continuity theorem. Let $\mu_n \in \mathcal{M}(\mathbf{R}^n)$. Suppose $\sup_n \|\mu_n\| \leq C < \infty$. If $\hat{\mu}_n \to \varphi$ pointwise for some φ that is continuous at 0, then there exists a unique $\mu \in \mathcal{M}(\mathbf{R}^n)$ such that $\mu_n \to \mu$ vaguely, $\hat{\mu} = \varphi$, and $\|\mu\| \leq C$. (The uniqueness follows from the previous theorem.)

Proof. By the Proposition 5.60 and Lemma 5.61, we know the pointwise limit φ must be continuous everywhere.

We first find the candidate μ. In the previous proof, we showed that for all $f \in \mathcal{S}$,

$$\int f \, d\mu_n = \int \check{f}(t)\hat{\mu}_n(t) \, dt.$$

By DCT, since $\check{f} \in L^1$ and $\|\hat{\mu}_n\|_u \le C$, we have

$$\int f \, d\mu_n \to \int \check{f}(t)\varphi(t) \, dt.$$

By sequential Banach–Alaoglu theorem, there is a subsequence μ_{n_k} of the original $\{\mu_n\}$ that converges vaguely to some μ with $\|\mu\| \le C$. Expanding the definition of vague convergence, for each $f \in \mathcal{S}$, it holds that

$$\int f \, d\mu_{n_k} \to \int f \, d\mu = \int \check{f}(t)\hat{\mu}(t) \, dt.$$

Therefore

$$\int \check{f}(t)\hat{\mu}(t) \, dt = \int \check{f}(t)\varphi(t) \, dt,$$

for all $\check{f} \in \mathcal{S}$, which let us conclude that $\hat{\mu} = \varphi$, since φ is continuous. □

The above convergence in fact holds weakly, and can be established directly via tightness. This will be studied later.

Proposition C.8

[Sch17, Exercise 21.4]

5.67 Bochner's theorem. A function $\varphi\colon \mathbf{R}^n \to \mathbf{C}$ is the Fourier transform of a finite measure μ precisely when

(a) $\varphi(0) = \mu(\mathbf{R}^n)$;

(b) φ is continuous on $\mathbf{R}^n$;

(c) φ is a positive semidefinite function.

5.I Fourier series

We get the second Parseval's identity *for Fourier series*.

5.68 Parseval's identity.

5.J Stieltjes transform

5.K Laplace transform

5.L Sobolev spaces

We use U to represent an arbitrary nonempty open set in $\mathbf{R}^n$.

Hölder space
[Bog10, Chapter 2]
$C_c^\infty(\mathbf{R}^n)$ is dense in $W^{1,p}(\mathbf{R}^n)$
$W^{1,p}$ is the completion of C_c^∞ with respect to the $W^{1,p}$ norm
Define $W_0^{1,p}$ to be the closure of C_c^∞ with respect to the $W^{1,p}$ norm.

5.69 Meyers–Serrin theorem. $C^\infty \cap W^{1,p}$ is dense in $W^{1,p}(\mathbf{R}^n)$.

Weighted Sobolev spaces
For every locally finite measure on $\mathbf{R}^n$, we can set

$$\|f\| = \|f\|_{L^p(\mu)} + \sum_{j=1}^{n} \|\partial_j f\|_{L^p(\mu)} \quad \text{for all } f \in C_c^\infty,$$

and then complete C_c^∞ under this norm. This is a Banach space, but the
A well-defined weighted Sobolev norm is the one that is closable
If the density is strictly positive,
or if the density satisfies certain weak differentiability property
$f \in W^{1,1}(\mathbf{R})$ if and only if $f \in L^1(\mathbf{R})$, and has an absolutely continuous version whose derivative is integrable on $\mathbf{R}$.

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, this is an exercise using Baire category theorem.

For two separable metrizable spaces X and Y, if $f\colon X \to Y$ 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

A Hausdorff measures and dimensions

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 $\operatorname{diam} A_j$'s.

Notice that H^α_ϵ is increasing as ϵ decreases to 0. We define the *α-Hausdorff measure* of E to be

$$H^\alpha(E) = \sup_{\epsilon>0} H^\alpha_\epsilon(E) = \lim_{\epsilon\to 0^+} H^\alpha_\epsilon(E).$$

Notice that we get a Carathéodory outer measure.

We define the Hausdorff dimension of E by

$$\dim_{\mathrm{H}} E = \inf\{\alpha : H^\alpha(E) = 0\} = \inf\{\alpha : H^\alpha_\infty(E) = 0\}.$$

The ternary Cantor set on $[0, 1]$ has Hausdorff dimension $\frac{\log 2}{\log 3}$. The boundary of the Koch snowflake has Hausdorff dimension $\frac{\log 4}{\log 3}$. Sierpiński gasket

When f is Lipschitz, $\operatorname{diam}_H f(E) \leq \operatorname{diam} E$, 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

$$H^\beta\big(f(E)\big) \leq C^\beta H^{\alpha\beta}(E),$$

which gives

$$\operatorname{diam} f(E) \leq \frac{1}{\alpha}\operatorname{diam} E.$$

(This is the reason why we chose α for the exponent in the definition of Hausdorff measures.)

Given a bounded metric space X (i.e., $\operatorname{diam} X < \infty$), the lower Minkowski dimension is defined by $\liminf_{\epsilon\to 0} \frac{\log \mathcal{C}(A,\epsilon)}{\log(1/\epsilon)}$, while the upper Minkowski dimension is defined by the lim sup analog. The lower and upper Minkowski dimensions do not necessarily agree, and

Given a Borel measure μ, the *lower Minkowski content* of a Borel set A is given by

$$\mu^{\mathrm{M}+} = \liminf_{\epsilon\to 0} \frac{1}{\epsilon}\mu(A_\epsilon - A),$$

where $A_\epsilon = \{x \in X : \operatorname{dist}(x, A) < \epsilon\}$, the open ϵ-neighborhood of A. Replacing lim inf by lim sup gives us the *upper Minkowski content*. We will only use the upper Minkowski content, and usually just call it the Minkowski content.

Since $(1+\epsilon)^n\, m\big(B^n(0;1)\big) = m\big(B^n\big(0;1+\epsilon\big)\big)$, it is immediately clear that the upper and lower Minkowski content are the same $n\, m(B)$. Recall $\sigma(\partial B^n) = n\, m(B^n)$, so the usual spherical measure and the upper/lower Minkowski content are exactly the same for the Euclidean ball. We will shortly define the boundary measure for all regular-enough boundaries, and show all these definitions coincide on the ball. However the lower/upper Minkowski is defined for "any" Borel sets in a metric space.

Also recall Exercise 3.18, which now allows us to rigorously state

A.1 Euclidean isoperimetric inequality. For any bounded open set U, we have

$$m_{\mathrm{M}}(U) \geq n\, \alpha_n^{1/n} m(U)^{\frac{n-1}{n}}.$$

Since $m_{\mathrm{M}}(\partial B^n) = n\, \alpha_n$, the above inequality is sharp when U is the open ball. This means that among all such U's with the same volume, the open ball attains the minimal Minkowski lower content.

In $\mathbf{R}^n$, and consider the n-Hausdorff measure H^n. One might wonder how does it relate to the Lebesgue measure. Since H^n is translation invariant and locally finite, we have $c(n)\, H^n(A) = m(A)$ for some constant $c(n)$. It turns out that $c(n) = \alpha_n/2^n$, i.e., the volume of the ball with diameter 1. Note that in defining the Hausdorff dimensions (a geometric notion), this $c(n)$ is totally irrelevant. Some authors include the constant $c(n)$ as part of the definition of $H^n(A)$

[Tay06, Chapter 12]

We first show that $m(A) \leq \frac{\alpha_n}{2^n} H^n(A)$. This follows from

A.2 Isodiametric inequality.

$$m(A) \leq \alpha_n \left(\frac{\operatorname{diam} A}{2}\right)^n,$$

where α_n is the volume of the unit ball in $\mathbf{R}^n$.

Steiner symmetrization

Minkowski-Steiner formula

To show that $m(A) \geq \frac{\alpha_n}{2^n} H^n(A)$, all the known proofs rely on some covering lemma, used inductive way (using DC).[1]

B Topological groups and Haar measures

compact groups are unimodular

left and right Haar measures agree precisely when the group is unimodular

B.1 Theorem.

[1] This is mentioned on page [Tay06, p. 161].

C Harmonic functions

A function u defined on an open subset of $\mathbf{R}^n$ is harmonic if $\Delta u = \sum_{j=1}^n \partial_{jj} u = 0$. It is subharmonic (resp. superharmonic) if $\Delta u \geq 0$ (resp. ≤ 0). A lot of the results should be understood by their exact analogs in complex analysis. Indeed, a complex holomorphic function, when realized on $\mathbf{R}^2$, is a holomorphic function by the Cauchy–Riemann equation.

maximum principle

For $u\colon \mathbf{R}^n \to \mathbf{R}$ that is subharmonic on some open connected set U, if u attains its maximum in U, then u is constant in U; if U is bounded and u attains is maximum on $\overline{U}$, then

$$\max_{x \in \overline{U}} u(x) = \max_{x \in \partial U} u(x).$$

mean-value property

over balls

over spheres

harmonic functions are smooth, and in fact, analytic

Liouville's theorem

Harnack's inequality

Harmonic measure

Riesz–Markov–Kakutani theorem (finite measures)

We will give a geometric interpretation of Harmonic measures whening studying Brownian motions.

D Introduction to PDE

Classification of second-order partial differential equations

A general kth-order partial differential operator is given by

$$L = \sum_{|\alpha| \leq k} a_\alpha \partial^\alpha.$$

The general second-order linear PDEs in with n-variables looks like $Lu = \sum_{i,j=1}^n a_{i,j} \partial_i \partial_j u +$ first order + linear terms $= 0$. Up to a sign change (because we can always change the left-hand side by a sign), if the matrix $A = [a_{ij}]$ is positive-definite, then the equation is elliptic. If the matrix A has a zero eigenvalue, then the equation is parabolic. If A has precisely one negative eigenvalue and $n-1$ positive eigenvalues, then A is hyperbolic. These terminologies corresponds exactly to the conic sections, once we replace each ∂_i to x_i.

In addition, if A is *strongly positive definite* with parameter c, which means that $\langle Ax, x \rangle \geq c\|x\|^2$, then the differential operator L is called uniformly elliptic with parameter c.

Given the time-dependent velocity vector field $v(t,x)$, let $\rho(t,x)$ denote the density at time t. The continuity equation says

$$\partial_t \rho + \operatorname{div}_x(\rho v) = 0,$$

which precisely describes the conservation of mass.

Consider an open ball $B(x_0; r)$ around 0. The rate at which the mass escapes the ball B should equate the flux over the boundary ∂B. This translates to the equation

$$\frac{d}{dt} \int_{B(x_0;r)} \rho(t,x)\, dx = -\int_{\partial B(x_0;r)} (\rho v) \cdot n \, d\sigma \text{ at all time } t.$$

Here we have made the physical assumption that the flux density should be given by ρv. By the divergence theorem, the right-hand side becomes

$$-\int_B \operatorname{div} F \, dx.$$

Now rearranging the equation gives

$$\int_{B(x_0;r)} \partial_t \rho(t,x) + \operatorname{div}(\rho v)\, dx = 0,$$

and we obtain the continuity equation once we take $r \to 0$.

By Fourier's law of thermal conduction, the velocity field $v(t,x)$ is proportional to $-\nabla_x u(t,x)$. For our purpose let $v(t,x) = -\frac{1}{2}\nabla_x u(t,x)$. This leads to the heat equation

$$\partial_t u = \tfrac{1}{2}\Delta_x u,$$

which is a parabolic equation because Δ does not include the time variable.

The steady state of this heat equation is at $\partial_t u = 0 = \Delta_x u$, i.e., when there is no more heat diffusion, the temperature u becomes harmonic. The equation $\Delta u = 0$ is called Laplace's equation, and is the topic of first-order importance studied in PDE. Its generalization is Poisson's equation

$$\Delta u = f.$$

Both are elliptic equations, because Δ has the identity matrix as its highest order coefficient matrix.

Black–Scholes equation

$$\partial_t V + \tfrac{1}{2}\sigma^2 S^2 \partial_S^2 V = rV - rS\partial_S V.$$

perform a change of variables

backward PDE

an initial value problem for a forward PDE is equivalent to a final value problem for a backward PDE

a final value problem for a forward PDE, or equivalently, an initial value problem for a backward PDE, is ill-posed

we cannot in general find a solution $u(t,x)$ such that $u(T,x) = f(x)$ for any specified function f in general ($u(T,x)$ must follow some structure determined by the PDE)

Fundamental solution

$LF = \delta$

Green's function

symmetric Harmonic property

Poisson kernel

Dirichlet problem on the half-space

Dirichlet problem on the unit ball

Dirichlet energy

D.1 Gronwall's inequality. For $u\colon [0,T] \to [0,\infty)$ be absolutely continuous with

$$\frac{du}{dt} \le \beta u \quad \text{and} \quad u(0) = u_0.$$

for some $\beta \in L^1[0,T]$. Then

$$u(t) \le u_0 \exp\left(\int_0^t \beta \, dt\right) \quad \text{for all } 0 \le t \le T.$$

Absolute continuity is to ensure the FTC for Lebesgue integral works. Set $v(t) = \exp(-\int \beta \, dt)u(t)$, whose derivative is ≤ 0. Integrating dv/dt will then give you the inequality.

The statement remains true if we replace the "$\le$" in the condition and the conclusion both by "$\ge$".

This implies that $u(0) = 0$ implies $u \equiv 0$ on $[0,T]$.

Gradient flows

Euler–Lagrange equation

Hamilton–Jacobi equation

E Distribution theory

the space of distributions $\mathcal{D}'$, of tempered distributions $\mathcal{S}'$, and of compactly supported $\mathcal{E}'$

$\mathcal{D}'(\mathbf{R}^n)$ is defined to be the space of continuous linear functionals on $C_c^\infty(\mathbf{R}^n)$.[2] However, we have yet to specify the topology on $C_c^\infty(\mathbf{R}^n)$. seminorms on C_c^∞ induces a locally convex topology on $C_c^\infty(\mathbf{R}^n)$. We can then endow the weak-star topology on $\mathcal{D}'(\mathbf{R}^n)$.

Defining the seminorm topology on C_c^∞ is in fact a daunting task. But one has the following sequential characterization, which is ultimately what people care about. The space of distributions $\mathcal{D}'$ consists precisely of all linear functionals F on C_c^∞ such that

$$\varphi_j \to \varphi \text{ in } \mathcal{D} \text{ implies } F(\varphi_j) \to F(\varphi).$$

(This statement might remind the readers of weak-star topology. But the weak-star topology is a topology we are trying to endow on a given dual space, but here we are trying to define that dual space in the first place.)

Define $\mathcal{D}'(\mathbf{R}^n)$ to be the space of distributions, which consists of all

For any $F \in L^1_{\text{loc}}(\mathbf{R}^n)$, the linear functional

$$\varphi \mapsto \int \varphi(x) F(x) \, dx \quad \text{for all } \varphi \in C_c^\infty(\mathbf{R}^n)$$

is a distribution on $\mathbf{R}^n$. More generally, for any locally finite Borel measure μ on $\mathbf{R}^n$, the recipe

$$\varphi \mapsto \int \varphi \, d\mu \quad \text{for all } \varphi \in C_c^\infty(\mathbf{R}^n)$$

is a distribution on $\mathbf{R}^n$. In these two standard cases we just identify F and μ as distributions.

We say $F \in \mathcal{D}'(\mathbf{R}^n)$ is smooth if for every open set $U \subseteq \mathbf{R}^n$, there exists $g \in C^\infty(U)$ such that

$$F(\varphi) = \int_U \varphi(x) g(x) \, dx \quad \text{for all } \varphi \in C_c^\infty(\mathbf{R}^n) \text{ with } \operatorname{supp} \varphi \subseteq U.$$

If this g is 0 on U, then we say F is zero on U. Note that for each U the associated g, if exists, must be unique. (This is a consequence of Theorem 5.50.)

[2] The notation $\mathcal{D} = C_c^\infty$ is due to L. Schwartz.

The complement to the union of all open sets on which F is zero (resp. smooth) is called the support (resp. singular support) of the distribution F, denoted by $\operatorname{supp} F$ (resp. $\operatorname{singsupp} F$).

A distribution $F \in \mathcal{D}'(\mathbf{R}^n)$ can be extended to a distribution in $\mathcal{E}'(\mathbf{R}^n)$ if $\operatorname{supp} F$ is compact.

Cauchy principle values

distributional derivative

$$\partial^\alpha F(\varphi) = (-1)^{|\alpha|} F(\partial^\alpha \varphi)$$

Recall that for $g \in L^1_{\mathrm{loc}}$ we defined its derivative according to the integration by parts formula. The distributional derivative generalizes this idea to arbitrary distributions.

Fourier transform of distributions

F More Sobolev spaces

The differentiation operator is closed

G Functional inequalities

Poincaré inequality Let $1 \le p < \infty$ and $U \subseteq \mathbf{R}^n$ be bounded in one direction, then for every $f \in W_0^{1,p}(U)$ being zero on the boundary, we have

$$\|f\|_p \le C\|\nabla f\|_p,$$

where $C = C(n,p)$.

Let $1 \le p \le \infty$ and U be a C^1 (or Lipschitz) domain, then for all $f \in W^{1,p}(U)$, we have

$$\left\| f - \frac{1}{m(U)} \int f \right\|_p \le C\|f\|_p,$$

where $C = C(n,p)$.

Wirtinger inequality

For $f \in C_c^1[0,R]$, we have

$$\int_0^R f(x)\,dx \le \frac{R^2}{\pi^2} \int_0^R f'(x)\,dx.$$

For $f \in C^1[-R,R]$ with $f(-R) = f(R)$ and $\int f = 0$, we have

$$\int_{-R}^R f(x)\,dx \le \frac{R^2}{\pi^2} \int_{-R}^R f'(x)\,dx.$$

Rellich–Kondrachov theorem

Let U be a C^1 (or Lipschitz) domain in $\mathbf{R}^n$, and $1 \le p < n$. Define $p^* = \frac{np}{n-p} = \frac{1}{1/p - 1/n}$, then $W^{1,p}(U) \hookrightarrow L^{p^*}(U)$ is continuous and $W^{1,p}(U) \hookrightarrow L^q(U)$ is compact for every $1 \le q < p^*$.

Gagliardo–Nirenberg inequality

We say a normed space X is compactly embedded into a normed space Y if the embedding $X \hookrightarrow Y$ is a compact operator.

H Tools from vector calculus

boundary ∂U is of class C^k

H.1 Divergence theorem. For any bounded open set $U \subseteq \mathbf{R}^n$ with C^1 boundary, and a vector field $F\colon U \to TU = \mathbf{R}^n$,

$$\int_U \operatorname{div} F\, dx = \int_{\partial U} F \cdot n\, dS.$$

H.2 Green's identity. Given U with C^1 boundary, for $u, v \in C^2(\overline{U})$, we have

(a) $\int_U \Delta u\, dx = \int_{\partial U} \partial_n u\, d\sigma$
(b) $\int_U \nabla u \cdot \nabla v\, dx = -\int_U u\Delta v\, dx + \int_{\partial U} u\partial_n v\, d\sigma$
(c) $\int_U \Delta v - v\Delta u\, dx = \int_{\partial U} u \cdot \nabla v - v \cdot \nabla u\, d\sigma$.

The last one shows that Δ is a symmetric operator over the class of $L^2(m)$ functions that vanishes on the boundary.

If U is the unbounded $\mathbf{R}^n$, then we have $\langle u, \Delta v\rangle_{L^2(m)} = -\langle \nabla u, \nabla v\rangle = \langle \Delta u, v\rangle$ for all $u, v \in C_c^\infty(\mathbf{R}^n)$. This is just integration by parts. Notice that $\langle \Delta u, u\rangle = -\langle \nabla u, \nabla u\rangle \leq 0$, so some authors chose to use the negative Laplace operator, which is a positive semidefinite operator.

Euclidean isoperimetric inequality

H.3 Exercise. If the bounded open set U is further assumed to be with C^1 boundary, then

$$\sigma(\partial U) \geq n\, \alpha_n^{1/n} m(U)^{\frac{n-1}{n}}.$$

In particular, since $\sigma(\partial B^n) = n\, \alpha_n$, the above inequality is sharp when U is the open ball. This means that among all such U's with the same volume, the open ball attains the minimal surface measure.

I Differentiable manifolds and integration with differential forms

The smooth partition of unity theorem states that for a smooth manifold M and a given open cover $\{U_\alpha\}$, there is a smooth partition of unity $\{\psi_\alpha\}$ subordinate to $\{U_\alpha\}$.

We know a locally compact second countable Hausdorff space is metrizable and separable, and therefore it is paracompact by [Rud69]. We are now ready show that for a lcscH space and an open cover $\{U_\alpha\}$, there is a countable cover $\{V_j\}_{j=1}^\infty$ that is a locally finite refinement of $\{U_\alpha\}$, and a further open cover $\{W_j\}_{j=1}^\infty$ with $\overline{W_j}$ is compact and is contained in V_j.

Consider the collection of all precompact open sets A with $\overline{A} \subseteq U_\alpha$ for some α. By Urysohn's lemma for locally compact spaces, part (a), we know this collection is an open cover of M, and therefore it has a locally finite open refinement $\{V_\beta\}$ that still covers M. Repeat this argument again with $\{V_\beta\}$. and we obtain a further locally finite open refinement $\{W_j\}_{j=1}^\infty$ that covers M, where W_j is precompact and $\overline{W_j} \subseteq V_\beta$ for some β. The index set for W is made countable because M is second countable.

Now for each j, fix a choice of $\beta(j)$ such that $\overline{W_j} \subseteq V_{\beta(j)}$. Reindex $\beta(j)$ by j, and we have a countable cover $\{V_j\}_{j=1}^\infty$ (a subcover of $\{V_\beta\}$) that is a locally finite refinement of $\{U_\alpha\}$.

It remains to prove the existence of partition of unity. First, for each j, we have $g_j \in C_c^\infty(M, [0,1])$ such that $g_j = 1$ on W_j and $\operatorname{supp} g_j \subseteq V_j$. Local finiteness allows us to define $f = \sum_j g_j$ pointwise on M, and because $\{W_j\}$ is also an open cover, f is strictly positive everywhere, and hence we may define $f_j = g_j/f$, which is $C_c^\infty(M, [0,1])$ and $\sum_j f_j = 1$.

It remains to reindex. For each j, fix a choice of $\alpha(j)$ so that $\overline{V}_j \subseteq U_{\alpha(j)}$. Therefore $\operatorname{supp} f_j = \operatorname{supp} g_j \subseteq V_j \subseteq U_{\alpha(j)}$. Define

$$\psi_\alpha = \sum_{\alpha = \alpha(j)} f_j,$$

which is well-defined and smooth because $1 = \sum_{j=1}^\infty f_j$ is a finite sum at each point. Clearly $0 \le \psi_\alpha \le 1$ and $\sum_\alpha \psi_\alpha = 1$. Now

$$\operatorname{supp} \psi_\alpha = \overline{\operatorname{supp} \bigcup_{\alpha=\alpha(j)} f_j} = \bigcup_{\alpha=\alpha(j)} \overline{\operatorname{supp} f_j} = \bigcup_{\alpha=\alpha(j)} \operatorname{supp} f_j \subseteq U_\alpha,$$

where the second equality comes from the fact that V_j is locally finite. The proof is complete.

I.1 Corollary. Given a smooth manifold M, say we have a closed set A contained inside an open set U in M. Then there exists a smooth function $\varphi\colon M \to [0,1]$ such that $\varphi = 1$ on A and $\operatorname{supp} \varphi \subseteq U$.

Obviously this follows by considering the partition of unity with respect to the open cover $U, M - A$.

I.2 Smooth Urysohn's lemma. Given a closed set A and an open set U in the smooth manifold M, suppose we have a smooth function $f\colon A \to \mathbf{R}^n$, then we can extend this f to a smooth function $F\colon M \to \mathbf{R}^n$ such that $F|_A = f$ and $\operatorname{supp} F \subseteq U$.

A (C^1) Riemannian manifold is a (C^1) differentiable manifold, where at each point x we have an inner product $\langle \cdot, \cdot \rangle_x$ on the tangent space T_xM. This inner product can be expressed in local coordinates around x by

$$\langle v, w \rangle_x = v \cdot \big([g_{jk}](x) w\big),$$

where $[g_{jk}](x)$ is a positive definite matrix, and the components varies continuously with respect to x. (Here $\cdot$ is the Euclidean dot product.) This "continuous" inner product of tangent vectors at each point is called the continuous *Riemannian metric*, denoted by g. (In technical terms the Riemannian metric is a continuous symmetric covariant 2-tensor field that is positive-definite at each $x \in M$.)

Jost Section 3.3 Lee Chapter 15 Riemannian volume form justifies why $\sqrt{\det[g_{jk}]}$

what is the correct unit volume in the oriented Riemannian manifold with metric tensor g, when expressed in the local (Euclidean) coordinates $x_1, x_2, \ldots, x_n$.

For $f \in C_c(U)$, where (U, φ) is an oriented chart, we have $\int_U f \, dV_g = \int_{\varphi(U)} f \sqrt{\det[g_{jk}]}\, dx$

Note that when M is compact, we only need a finite number of charts to cover M, and by the usual partition of unity argument one can conclude that $\operatorname{Vol}(M) = \int_M 1 \, dV_g$, as a finite sum of finite integrals, must be finite. Therefore the measure μ_g induced from V_g is finite, and can be normalized to a probability measure by dividing $\operatorname{Vol}(M)$.

different choice of local coordinates would yield the same integration formula, so the integral is well-defined

For a Riemannian manifold (M, g), we can define for any $f \in C_c(M)$ a positive linear functional L by the formula $Lf = \int_M f \, dV_g$. Then by Riesz–Markov–Kakutani theorem (positive measures), we immediately obtain a Radon measure for us to perform integration with respect to arbitrary Borel measurable functions on M. (The V_g here may be taken as the Riemannian density, or the Riemannian volume form when M is oriented.)

I.3 Bochner's formula.

I.4 Lichnerowicz' formula.

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

[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.

(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 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 $F\colon \mathbf{R} \to [0,1]$ given by

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

Again we mention that the choice of "$\le$" instead of "$<$" in the definition of distribution function is a convention. In fact going back to Kolmogorov's original *Foundations of the Theory of Probability*, the distribution function is defined by $P(X < x)$.

We now slightly modify Theorem 1.34(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 1.34(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 $F\colon \mathbf{R} \to [0,1]$ 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.

[4]We have only discussed the integration of real/complex-valued functions. Some generalizations can definitely be made (to for example, Banach-valued functions/random variables), but it is beyond the scope of this survey.

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 $F\colon \mathbf{R} \to [0,1]$ 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.

7.7 Theorem. Indeed any distribution function $F\colon \mathbf{R} \to [0,1]$ 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 $\mathbf{R}$. □

Given 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 $(\Omega, \mathcal{F}, P) = \big((0,1), \mathcal{B}_{(0,1)}, m\big)$, and we define

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

It is clear to see that $X(\omega) \leq y$ if and only if $\omega \leq F(y)$. Therefore for all $y \in \mathbf{R}$ and $\omega \in (0,1)$,

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

The construction still works out perfectly if one replaces ω by $U(\omega)$, where $U \sim \text{Uniform}(0,1)$. 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 $\big((0,1), \mathcal{B}_{(0,1)}, m\big)$, via the recipe $F_\mu^{-1}(U)$.

In fact we may take U to be uniform over $[0,1]$, $(0,1]$, or $[0,1)$, whichever is the simplest for application. This is because their distributions are all the same on the real line. □

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).

7.10 Fact. We already mentioned that $F(F^{-1}(\omega)) \geq \omega$. It is also true that $F^{-1}(F(y)) \leq y$.

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 helpful in the proof of Exercise 8.25, when constructing a right-continuous candidate for a distribution function.)

We will generalize this result later. Generalization of Theorem 7.7

7.11 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 $\mu \ll \text{count}$. 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 1.14. 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

[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.

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,

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

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

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 2.46 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

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

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

[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.

Recall the layer-cake representation of integrals. It provides a nice characterization for the expectation of nonnegative random variables $g(X)$:

$$\mathrm{E}g(X) = \int_0^\infty P\big(g(X) > t\big)\,dt = \int_0^\infty P\big(g(X) \geq t\big)\,dt.$$

For general $g(X) \in L^1(P)$, we have

$$\begin{aligned}\mathrm{E}f(X) &= \int_0^\infty P\big(f(X) > t\big)\,dt - \int_{-\infty}^0 P\big(f(X) \leq t\big)\,dt \\ &= \int_0^\infty P\big(f(X) > t\big)\,dt - \int_{-\infty}^0 1 - P\big(f(X) > t\big)\,dt\end{aligned}$$

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

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

7.14 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

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

7.15 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 directly 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

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

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

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

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

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

7.17 Proposition. The following statements are equivalent.

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

(b) $A_1^{\mathrm{c}}, A_2, \dots, A_n$ 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.

7.18 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. □

7.19 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]

[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".

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

$$\mu\bigg(\bigcap_{n=1}^{\infty} B_{1/n}\bigg) = \lim_{k\to\infty} \mu\bigg(\bigcap_{n=1}^{k} B_{1/n}\bigg) = 1.$$

Since $B \coloneqq \bigcap_n B_{1/n}$ 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

7.20 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

$$\begin{aligned}\mu_{X,Y}(A_1 \times A_2) &= \int_{A_1 \times A_2} f_{X,Y}(x,y)\, dx\, dy, \\ \mu_X(A_1) \times \mu_Y(A_2) &= \int_{A_1} f_X(x)\, dx \int_{A_2} f_Y(y)\, dy \\ &= \int_{A_1} \int_{A_2} f_X(x) f_Y(y)\, dx\, dy.\end{aligned}$$

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. □

7.21 Definition. The *variance* of an L^2 random variable X is defined by

$$\begin{aligned}\operatorname{Var}(X) &= \operatorname{E}(X - \operatorname{E}X)^2 \\ &= \operatorname{E}(X^2) - 2\operatorname{E}X \cdot \operatorname{E}X + (\operatorname{E}X)^2 \\ &= \operatorname{E}(X^2) - (\operatorname{E}X)^2.\end{aligned}$$

Given two L^2 random variables X and Y, their *covariance* is defined by

$$\begin{aligned}\operatorname{Cov}(X,Y) &= \operatorname{E}\big((X - \operatorname{E}X)(Y - \operatorname{E}Y)\big) \\ &= \operatorname{E}(XY) - \operatorname{E}X \cdot \operatorname{E}Y;\end{aligned}$$

they are said to be *uncorrelated* if $\operatorname{Cov}(X,Y) = 0$, i.e.,

$$\operatorname{E}X \cdot \operatorname{E}Y = \operatorname{E}(XY);$$

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

$$\operatorname{Cov}(\mathbf{1}_A, \mathbf{1}_B) = \operatorname{E}(\mathbf{1}_{A\cap B}) - \operatorname{E}\mathbf{1}_A \operatorname{E}\mathbf{1}_B = P(A \cap B) - P(A)P(B).$$

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 $\pi\colon \mathbf{N} \to \mathbf{N}$ 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.

7.22 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.

7.23 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.

There is a succinct information-theoretic proof of Hewitt–Savage zero–one law for random variables taking values in a measurable space with a countably generated σ-field. As mentioned before, this includes any separable metric space with the Borel σ-field, which is sufficient for doing probability. See [OCo00].

exchangeable family of random variables

7.24 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.

7.25 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

$$P(X \geq a) \leq \frac{\mathrm{E}X}{a}.$$

We start with two elementary inequalities on the deviation of a random variable from its mean, assuming *only* that $\mathrm{E}X^2 < \infty$. Applying the L^2 Markov's inequality,

7.26 Chebyshev's inequality. For X with $\mathrm{E}X^2 < \infty$, we have for all $t > 0$ that

$$P(|X - \mathrm{E}X| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

And the following result gives a sharper one-side bound.

7.27 Cantelli's inequality. For X with $\mathrm{E}X^2 < \infty$, we have for all $t > 0$ that

$$P(X - \mathrm{E}X \geq t) \leq \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + t^2}.$$

Proof. It suffice to consider the case $\mathrm{E}X = 0$ since $\mathrm{Var}(X) = \mathrm{Var}(X + \mathrm{E}X)$.

By Markov's inequality,

$$P(X \geq t) = P(X + a \geq t + a) \leq \frac{\mathrm{E}(X+a)^2}{(t+a)^2} = \frac{\mathrm{Var}(X) + a^2}{(t+a)^2}$$

holds for any $a \in \mathbf{R}$. The right-hand side, as a function of a, is minimized when $a = \mathrm{Var}(X)/t$. Plugging in give us the desired bound. □

7.28 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's method*, where you set $\varphi(x) = \exp(\lambda x)$, and in the end you aim to minimize

$$\frac{1}{\exp(\lambda a)} \mathrm{E}\exp(\lambda X)$$

over all $\lambda \in \mathbf{R}^{>0}$ (so that φ is increasing).

Unfortunately $\mathrm{E}\exp(\lambda X)$ 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, $\mathrm{E}\exp(\lambda X)$, 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 $\mathrm{E}\exp(\lambda X) = \frac{\rho}{\rho - \lambda}$ for $X \sim \text{Exponential}(\rho)$ and $\lambda < \rho$; and also $\mathrm{E}\exp(\lambda Y) = \exp(\lambda^2/2)$ for $Y \sim N(0, 1)$.

The transform $\mathrm{E}\exp(\lambda X)$ 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

$$P(|X| \geq a) \leq \inf_{p \in \mathbf{N}} \frac{1}{a^p} \mathrm{E}(|X|^p),$$

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.

Markov's inequality gives an upper bound on the tail probability, with the first moment $\mathrm{E}X$. A lower bound can also be obtained, with in addition the second moment $\mathrm{E}X^2$.

7.29 Paley–Zygmund inequality. Let $X \geq 0$ with $\mathrm{E}X^2 < \infty$. For any $0 \leq \theta \leq 1$, we have

$$P(X > \theta\, \mathrm{E}X) \geq (1 - \theta)^2 \frac{(\mathrm{E}X)^2}{\mathrm{E}X^2}.$$

Proof. The case for $\theta = 1$ is trivial. We will fix $0 < \theta < 1$ first.

The key is to use Cauchy–Schwarz inequality:

$$\begin{aligned} \mathrm{E}X &= \mathrm{E}(X\mathbf{1}\{X \le \theta\,\mathrm{E}X\}) + \mathrm{E}(X\mathbf{1}\{X > \theta\,\mathrm{E}X\}) \\ &= \theta\,\mathrm{E}X + \sqrt{\mathrm{E}X^2 P(X > \theta\,\mathrm{E}X)}, \end{aligned}$$

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].

7.30 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 $t \ge 0$, we have

$$P\left(\sum_{k=1}^{n}(X_k - \mu_k) \ge t\right) \le \exp\left(-\frac{2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\right).$$

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.

7.31 Hoeffding's lemma. For a mean zero random variable Y that is a.s. bounded within $[a, b]$, we have

$$M_Y(\lambda) = \mathrm{E}\exp(\lambda Y) \le \exp\left(\frac{\lambda^2(b-a)^2}{8}\right) \tag{7.32}$$

Proof. Since $e^{\lambda x}$ is convex, we have for $x \in [a, b]$

$$e^{\lambda x} \le \frac{b-x}{b-a}e^{\lambda a} + \frac{x-a}{b-a}e^{\lambda b}.$$

Taking expectation on both sides, and we have

$$\mathrm{E}\exp(\lambda Y) \le \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} = e^{L(\lambda(b-a))}, \tag{7.33}$$

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 $L'(0) = -\gamma + \frac{\gamma e^h}{1-\gamma+\gamma e^h}\big|_{h=0} = 0$.

$$\begin{aligned} L''(h) &= \left(\frac{\gamma e^h}{1-\gamma+\gamma e^h}\right)' \\ &= \frac{\gamma e^h}{1-\gamma+\gamma e^h} - \frac{\gamma^2 e^{2h}}{(1-\gamma+\gamma e^h)^2} \\ &= t - t^2 \le 1/4 \quad \text{for any } t \in \mathbf{R}, \end{aligned}$$

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

$$L(h) = 0 + 0 \cdot h + \frac{L''(\xi)}{2} \cdot h^2 \le \frac{1}{8}h^2.$$

Now let $h = \lambda(b-a)$ and plug it back into (7.33), and we have shown (7.32). □

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]

Inequality (7.32) is tight; consider any binary random variables, with probabilities $1/2$, e.g., the Rademacher random variable.

We end this section with the *large deviation bound on the sample mean.*

7.34 Theorem. Let $X_1, \dots, X_n$ be i.i.d. and $S_n = X_1 + \cdots + X_n$. Then for $a > \mathrm{E}X_1$, we have

$$P\left(\frac{S_n}{n} \geq a\right) \leq \exp\bigl(-nI(a)\bigr),$$

where $I(a) = \sup_{t\in\mathbf{R}} at - \log M(t)$, the Legendre transform of the cumulant generating function of X. The same upper bound continues to hold when $a < \mathrm{E}X_1$:

$$P\left(\frac{S_n}{n} \leq a\right) \leq \exp\bigl(-nI(a)\bigr).$$

Proof. The case $a < \mathrm{E}X_1$ holds by replacing each X_j by $-X_j$ and a by $-a$, and observing that

$$\sup_{t\in\mathbf{R}} -at - \log \mathrm{E}\exp(-tX) = \sup_{t\in\mathbf{R}} at - \log \mathrm{E}\exp(tX).$$

□

When $M(t) < \infty$ in a neighborhood of 0, then $I(a) > 0$, so both bounds are nontrivial.

Clearly $M(t) = \infty$ for all t, then $I(a) = \infty$, and the bound becomes trivial. (More precisely a tail that decays slower than exponential)

We know from the central limit theorem that asymptotically

$$P\left(a < \frac{S_n}{\sqrt{n}} \leq b\right) \approx \Phi(b) - \Phi(a),$$

a constant, where Φ is the the standard normal distribution function. This implies that the common deviation from the sample mean is of order $\sqrt{n}$. If we consider S_n/n instead, then the deviation from $\mathrm{E}X$ becomes $\exp(-cn)$ for the constant c depending on a and $M(t)$. An exponentially decaying deviation tells us that it is very unlikely that S_n has a fluctuation of order n around $n\,\mathrm{E}X_1$, due to combined effect of the i.i.d. X_k's in the sum.

Given that the large deviation bound already contains Chernoff's method, it is quicker to directly use this bound for sum of i.i.d. random variables. For example, we may easily Hoeffding's inequality for Rademacher random variables. By Hoeffding's lemma we have

$$I(a) \geq \sup_{t\in\mathbf{R}} at - t^2/2 = t^2/2.$$

Therefore

$$P(S_n \geq na) \leq \exp\left(\frac{na^2}{2}\right).$$

[10] Hence one may extract a proof of the above inequality from there.

7.35 Exercise. Let $X_1, X_2, \dots$ be i.i.d. Poisson(μ), which has $\log M(t) \le \mu(e^t - 1)$. Verify that

$$P(S_n \ge na) \le e^{-n\mu} \left(\frac{e\mu}{a} \right)^{na}.$$

When a is much larger than μ, the probability decays at a rate approximately $a^{-a} = \exp(-a \log a)$, which is heavier than Gaussian.

7.D Miscellaneous but crucial facts and tools

7.36 Change of densities. For a random variable X and an injective C^1 map $\varphi\colon \mathbf{R}^n \to \mathbf{R}^n$, and suppose X has density with respect to the Lebesgue measure. Then

$$f_X(x) = f_{\varphi(X)}\big(\varphi(x)\big) |\det D\varphi(x)|.$$

Note

$$\int_A f_X(x)\,dx = P(X \in A) = P\big(\varphi(X) \in \varphi(A)\big) = \int_{\varphi(A)} f_{\varphi(x)}(y)\,dy.$$

By applying change of variables we immediately get the desired formula.

7.37 Definition. Fix the dimension d. The *standard Gaussian measure* on $\mathbf{R}^d$ is the measure $\gamma\colon \mathcal{B}(\mathbf{R}^d) \to [0,1]$ given by

$$\gamma(A) = \frac{1}{\left(\sqrt{2\pi}\right)^d} \int_A \exp\big(-\|x\|_2^2/2\big)\,dx.$$

The above expression without the integral is hence the *standard Gaussian density*.

For $Y \sim N(0, \Sigma)$, where Σ is positive definite, we may write $Y \overset{\mathrm{D}}{=} \Sigma^{1/2} X$, where $X \sim N(0, I_d)$. Therefore the density f_Y for Y is $f_X(\Sigma^{-1/2} y) \cdot \frac{1}{|\det(\Sigma^{1/2})|}$. Simlification gives us

$$\frac{1}{(2\pi)^{d/2} (\det \Sigma)^{1/2}} \exp\left(-\frac{y^{\mathrm{T}} \Sigma^{-1} y}{2} \right)$$

as the final density for Y.

It follows that Gaussian measures are orthogonally invariant. However, it is not translation invariant. (Lebesgue measure is the only translation invariant and locally finite Borel measure on $\mathbf{R}^d$.)

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.

7.38 Fact. For $Z \sim N(0, I_d)$, we have

$$\mathrm{E}[Zf(Z)] = \mathrm{E}[\nabla f(Z)],$$

when the expectations on the two sides are defined.

7.39 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

$$\frac{t}{t^2+1} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) \le P(Z > t) \le \frac{1}{t} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2).$$

Therefore

$$-\frac{2}{t^2}\log P(Z>t)\to 1,$$

$P(Z>t)\le \exp(-t^2/2)$ for all $t>0$, and $\le \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)$ for all $t\ge 1$.

The tail of a 1d Gaussian decays at rate $\exp(-t^2/2)$.

Let $Z\sim N(0,1)$, then for all $k\in\mathbf{N}$,

$$\mathrm{E}Z^{2k-1}=0 \quad\text{and}\quad \mathrm{E}Z^{2k}=\frac{(2k)!}{2^k k!}=(2k-1)!!.$$

Expand

$$\mathrm{E}Z^{2k-1}=\int_0^\infty P(Z>t^{-(2k-1)})\,dt-\int_{-\infty}^0 P(Z<t^{-(2k-1)})\,dt.$$

By symmetry and change of variables,

$$\int_{-\infty}^0 P(Z<t^{-(2k-1)})\,dt=\int_0^\infty P(Z>t^{-(2k-1)})\,dt,$$

and hence $\mathrm{E}Z^{2k-1}=0$.

Say X is a real *symmetric random variable*, i.e., $P(X>t)=P(X<t)$ for all $t\in\mathbf{R}$ (whether you add "$\ge$" or not does not matter). Then the same argument as above tells us that $\mathrm{E}X^{2k-1}=0$.

To get the even power, we have to directly compute.

$$\begin{aligned}\sqrt{2\pi}\,\mathrm{E}Z^n &= \int_{-\infty}^\infty x^{n-1}[x\exp(-x^2/2)]\,dx\\ &= \big[-x^{n-1}e^{-x^2/2}\big]_{-\infty}^{\infty}-\int_{-\infty}^\infty -(n-1)x^{n-2}e^{-x^2/2}\,dx\\ &= 0+(n-1)\int_{-\infty}^\infty x^{n-3}[xe^{-x^2/2}]\,dx\\ &= (n-1)!!\int_{-\infty}^\infty e^{-x^2/2}\,dx \quad\text{by repeating.}\end{aligned}$$

With $\int_{-\infty}^\infty e^{-x^2/2}\,dx=\sqrt{2\pi}$, we get our result.

For two independent $Z_1,Z_2\sim N(0,1)$, $\mathrm{E}\max\{Z_1,Z_2\}=\sqrt{1/\pi}$. This is because

$$\mathrm{E}\max\{Z_1,Z_2\}=\mathrm{E}\tfrac12\big(Z_1+Z_2+|Z_1-Z_2|\big)=\mathrm{E}\frac{1}{2}|Z_1-Z_2|,$$

and noticing that $Z_1-Z_2\sim N(0,2)$.

7.40 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. □

The coordinates of a normal random vector are independent if and only if they are uncorrelated.

Bogachev Theorem 1.4.3.

7.41 Gaussian isoperimetric inequality. For $A \subseteq \mathbf{R}^n$, then

$$\gamma(A_\epsilon) \geq$$

We now restate Borel–Cantelli lemma I.

7.42 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

$$\forall\, \epsilon > 0, P\big(|X_n(\omega) - X(\omega)| \geq \epsilon \text{ i.o.}\big) = 0.$$

(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

$$\forall\, \epsilon > 0, P\big(|X_n(\omega) - X(\omega)| < \epsilon \text{ ev.}\big) = 1.$$

7.43 Borel–Cantelli lemma II. For pairwise independent events $A_1, A_2, \ldots$, 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.

7.44 Harris' inequality. 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

$$\mathrm{E} f(X) \cdot \mathrm{E} g(X) \leq \mathrm{E}[f(X)g(X)].$$

More generally, the above inequality still holds if $X = (X_1, \ldots, X_n)$ 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.[11]

[11] also known as Fortuin–Kasteleyn–Ginibre (FKG) inequality

Proof. Let Y be an independent copy of X. Consider the expectation

$$\begin{aligned}&\mathrm{E}\big\{[f(X)-f(Y)]\cdot[g(X)-g(Y)]\big\}\\&=\mathrm{E}[f(X)g(X)]-\mathrm{E}[f(Y)g(X)]-\mathrm{E}[f(X)g(Y)]+\mathrm{E}[f(Y)g(Y)]\\&=2\,\mathrm{E}[f(X)g(X)]-2\,\mathrm{E}f(X)\cdot\mathrm{E}g(Y),\end{aligned}\tag{7.45}$$

where we have used $f(X)\perp g(Y)$ and $f(Y)\perp g(X)$.

For any outcome $\omega\in\Omega$, if $X(\omega)\geq Y(\omega)$, then by monotonicity of f and g we have

$$f(X)-f(Y)\geq 0\quad\text{and}\quad g(X)-g(Y)\geq 0,$$

which implies that

$$[f(X)-f(Y)]\cdot[g(X)-g(Y)]\geq 0.$$

The above inequality also holds when $X(\omega)<Y(\omega)$. Therefore

$$\mathrm{E}\big\{[f(X)-f(Y)]\cdot[g(X)-g(Y)]\big\}\geq 0.$$

The desired inequality then follows by using (7.45).

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

$$\mathrm{E}f(X)\cdot\mathrm{E}g(X)=\mathrm{E}f_1(X_1)\cdot\mathrm{E}g_1(X_1)$$

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. □

We say $A\in\mathcal{F}$ is an *increasing event* if $\mathbf{1}_A\colon(S,\leq)\to\{0,1\}$ is an increasing function. This means precisely that if $x_1\leq x_2$ in S, then $x_1\in A$ implies $x_2\in A$. The above theorem tells us that if A and B are two increasing events, we have

$$P(X\in A)P(X\in B)\leq P(X\in A\cap B),$$

which exactly tells us that A and B are positively correlated. Correlation structure is an important theme in statistical physics, and Harris' inequality has important consequences in domains like percolation.

We also mention that in the case $X \sim N(0, \Sigma)$, where all entries of Σ are nonnegative, then Harris' inequality remains in force despite the dependence between the coordinates. This is known as Pitt's theorem [Pit82]. We will prove this result at the very end of the text.

symmetrization technique

replace X by $X - X'$

replace X by εX

Consider the space $L^2([0,1], m)$. We know $\mathbf{R}[x]$ is dense in $C[0,1]$ with respect to the L^∞ norm, and hence dense in $C[0,1]$ with respect to the L^2 norm. Since $C[0,1]$ is dense in $L^2[0,1]$ with respect to the L^2 norm, $\mathbf{R}[x]$ must be dense in $L^2[0,1]$. We can therefore perform Gram–Schmidt process on $\{1, x, x^2, \dots\}$ to obtain an orthonormal basis for $L^2[0,1]$.

We now show this is still possible for the Gaussian measure. It turns out the orthonormal basis for $L^2(\gamma)$ that we obtain this way will be a list of polynomials. But before that, we first need to show

7.46 Theorem. $\mathbf{R}[x]$ is dense in $L^2(\mathbf{R}, \gamma)$. Equivalently, for all $k \in \mathbf{N}_0$, we have for $h \in L^2(\gamma)$,

$$\int_{\mathbf{R}} x^k h(x)\, d\gamma(x) = 0,$$

then $h = 0$ γ-a.e.

Replacing $h(x)$ by $f(x)\exp(-x^2/4)$, then $f \in L^2(m)$, and we may consider instead

$$\int_{\mathbf{R}} x^k f(x) \exp(-x^2/4)\, dx = 0,$$

and show $f = 0$ m-a.e.

Proof. To show $f = 0$ a.e., it suffices to show that the inverse Fourier transform of $f(x)\exp(-x^2/4)$

$$g(t) = \int_{\mathbf{R}} f(x) \exp(itx - x^2/4)\, dx = 0.$$

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

First note that $\check{f}(t)$ is well defined by Cauchy–Schwarz inequality and $\exp(itx - x^2/4) \in L^2(dx)$ for any $t \in \mathbf{C}$. (For any real number $R \geq |t|$, we have

$$|\exp(2itx - x^2/2)| \leq \exp(2R|x|)\exp(-x^2/2) = \exp(2R^2)\exp\left(-\frac{(|x| - 2R)^2}{2}\right),$$

which is integrable.)

Now compute

$$\frac{\partial^k}{\partial^k t} f(x)\exp(itx - x^2/4) = (ix)^k\, f(x)\exp(itx - x^2/4).$$

The derivatives are again integrable, by observing that

$$x^{2k}\exp\left(-\frac{(|x| - 2R)^2}{2}\right) \leq 2^{2k}\left[(|x| - 2R)^{2k} + (2R)^{2k}\right]\exp\left(-\frac{(|x| - 2R)^2}{2}\right),$$

which is finite (recall when we computed the even moments of Gaussians). Therefore we may differentiate under the integral sign and conclude that $g^{(k)}(0) = 0$ for all $k \geq 0$. This implies $g(t) = 0$ in a neighborhood of 0, and by the uniqueness theorem we have $g = 0$ everywhere, as desired. □

We claim that the following polynomial functions form the orthonormal basis for $L^2(\gamma)$ from applying Gram–Schmidt to $1, x, x^2, \ldots$. The *Hermite polynomial* of order $n \geq 0$ is defined by

$$\mathrm{He}_n(x) = (-1)^n \exp\left(\tfrac{x^2}{2}\right) \tfrac{d^n}{dx^n} \exp\left(-\tfrac{x^2}{2}\right) = e^{-D^2/2} x^n \tag{7.47}$$

for $D = d/dx$. By the chain rule, $\{\mathrm{He}_n\}$ is determined by the recurrence relation $\mathrm{He}_0(x) = 1$ and

$$\mathrm{He}_{n+1}(x) = x\,\mathrm{He}_n(x) - \mathrm{He}_n'(x),$$

We list the first six Hermite polynomials for reference:

$$\begin{aligned} &\mathrm{He}_0(x) = 1, && \mathrm{He}_1(x) = x, && \mathrm{He}_2(x) = x^2 - 1, \\ &\mathrm{He}_3(x) = x^3 - 3x, && \mathrm{He}_4(x) = x^4 - 6x^2 + 3, && \mathrm{He}_5(x) = x^5 - 10x^3 + 15x. \end{aligned}$$

Equivalent to (7.47), we have the generating function representation in λ

$$\exp\left(\lambda x - \tfrac{1}{2}\lambda^2\right) = \sum_{n=0}^{\infty} \frac{\mathrm{He}_n(x)}{n!} \lambda^n. \tag{7.48}$$

Completing the square on the left hand side, we have

$$\exp\left(\tfrac{1}{2}x^2 - \tfrac{1}{2}(x-\lambda)^2\right) = \exp\left(\tfrac{1}{2}x^2\right) \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \frac{d^n}{d\lambda^n} \exp\left(-\tfrac{1}{2}(x-\lambda)^2\right)\Big|_{\lambda=0}.$$

Fix x, and differentiate both sides of (7.48), we obtain the relation

$$\mathrm{He}_n'(x) = n\,\mathrm{He}_{n-1}(x).$$

This leads to a new recurrence relation: $\mathrm{He}_0(x) = 1$, $\mathrm{He}_1(x) = x$, and

$$\mathrm{He}_{n+1}(x) = x\,\mathrm{He}_n(x) - n\,\mathrm{He}_{n-1}(x). \tag{7.49}$$

that determines $\{\mathrm{He}_n\}$. Also He_n satisfies the differential equation

$$-\mathrm{He}_n''(x) + x\,\mathrm{He}_n'(x) = n\,\mathrm{He}_n(x). \tag{7.50}$$

This is a useful property that will prove to be useful near the end of the book.

This recurrence relation allow us to confirm that He_n must be a monic polynomial of degree n (if we count 1 as having degree 0), and He_{2k} should only have terms of even degrees, while He_{2k+1} should only have terms of odd degrees.

To show that $\langle \mathrm{He}_m, \mathrm{He}_n \rangle = n! \delta_{mn}$, the best idea is to use the generating function representation again. Integrating

$$\exp\left(\lambda x - \tfrac{1}{2}\lambda^2\right) \exp\left(\rho x - \tfrac{1}{2}\rho^2\right)$$

with respect to γ gives us $\exp(\lambda\rho)$. If we expand this expression using (7.48) and then integrate, we should get instead

$$\sum_{m,n} \langle \mathrm{He}_m, \mathrm{He}_n \rangle \tfrac{1}{m!n!} \lambda^m \rho^n.$$

Comparing the coefficients, we should realize that $\langle \mathrm{He}_m, \mathrm{He}_n \rangle \neq 0$ if and only if $m = n$, with $\langle \mathrm{He}_m, \mathrm{He}_n \rangle = \frac{1}{n!} \big/ \frac{1}{n!n!} = n!$.

It remains to confirm that $\{\mathrm{He}_n\}$ can indeed be obtained from $\{1, x, x^2, \dots\}$ by Gram–Schmidt (without normalization). First notice that $\{\mathrm{He}_0, \mathrm{He}_1, \dots, \mathrm{He}_k\}$ and $\{1, x, \dots, x^k\}$ are linearly independent lists with the same span for *all* $k \geq 0$. Therefore x^{k+1}, after orthogonalization, and He_{k+1} are orthogonal to the same span, so they must be the same up to scaling. Since x^{k+1}, after orthogonalization, and He_{k+1} are both monic polynomials, they must be the exactly the same.

Therefore $\{\mathrm{He}_n(x)\}$ spans the space $\mathbf{R}[x]$ over $\mathbf{R}$, and by Theorem 7.46, we conclude

7.51 Theorem. The set of vectors $\left\{\frac{\mathrm{He}_n(x)}{\sqrt{n!}}\right\}$ is an orthonormal basis of $L^2(\gamma)$.

It is easy to generalize Hermite polynomials to multiple dimensions. Define the function $\mathrm{He}_{\mathbf{k}} = \mathrm{He}_{k_1,\dots,k_n}$ on $\mathbf{R}^n$ by

$$\mathrm{He}_{k_1,\dots,k_n}(x_1, \dots, x_n) = \mathrm{He}_{k_1}(x_1) \cdots \mathrm{He}_{k_n}(x_n).$$

First of all, Corollary 5.27 immediately tells us that

$$\mathrm{He}_{k_1,\dots,k_n}, \text{ where } k_1, \dots, k_n \in \mathbf{N}_0$$

gives an orthogonal basis for $L^2(\gamma_n)$ (and can be normalized dividing $k_1! \cdots k_n!$.

We say a function $f \colon \mathbf{R}^d \to \mathbf{R}$ is *log-concave* if f can be written as the exponential of a concave function. We focus on *log-concave density functions*. Clearly the standard Gaussian measure is log-concave.

SUBTLETY WE NEED TO LET f be strictly positive on its support, or use an extended version of convex concave function

We say a probability measure μ on $(\mathbf{R}^d, \mathcal{B}^d)$ is *log-concave* if for all $A, B \in \mathcal{B}^d$ and $0 < \lambda < 1$, we have

$$\mu\big((1-\lambda)A + \lambda B\big) \leq \mu(A)^{1-\lambda} \mu(B)^{\lambda}.$$

We have already related this definition to the Brunn–Minkowski inequality, which precisely states that the above inequality is true for Lebesgue measures. Obviously for probability measures this is not always the case.

If the affine hull of $\operatorname{supp} \mu$ has dimension d, then $\frac{d\mu}{dm}$ exists and is a log-concave function. Conversely, if the affine hull of $\operatorname{supp} f$ has dimension d, then $f\,dm$ is a log-concave measure.

The converse is true by applying the Prékopa–Leindler inequality

Since we do not really care about the singular case, we will always say X follows a *log-concave distribution* if it has a log-concave density, i.e., the distribution of X is not just a log-concave measure, but also has a full-dimension support.

The Gaussian measure is log-concave because $\|x\|^2/2$ is a convex function. In fact, it is strongly log-concave because its second derivative is strictly positive. A measure that is log-concave but not strictly log-concave is the uniform measure on a convex set.

product of log-concave distributions is log-concave

the marginal distribution of a jointly log-concave pair of random variables is log-concave.

If $X \in \mathbf{R}^n$ and $Y \in \mathbf{R}^n$ are log-concave and independent, then so is $X + Y$. First, (X, Y) having density $f_X f_Y$ must be log-concave on $\mathbf{R}^{2n}$. If we can show $(X + Y, X - Y)$ is log-concave, then by marginalization we get $X + Y$ and $X - Y$ must be log-concave.

The "if" step is proved in general below. It says that under affine change of variables, log-concavity of the density function is preserved.

7.52 Proposition. Let $T\colon \mathbf{R}^d \to \mathbf{R}^d$ be an affine and invertible linear map, and let f and g be the density of X and $T(X)$ respectively. If f is log-concave, then g is also log-concave.

Proof. First

$$g(T(x)) = f(x)\frac{1}{|DT(x)|},$$

where the Jacobian $|DT(x)|$ is a positive constant independent of x, because T is linear and invertible. Now write $f(x) = e^{-\varphi(x)}$ for some convex φ and $g(y) = e^{-\psi(y)}$ for some ψ. Therefore for some real constant c,

$$e^{-\psi(T(x))} = e^{c-\varphi(x)},$$

which implies

$$\psi(T(x)) = c - \varphi(x).$$

Observe for any $0 < \lambda < 1$ and $x, y \in \mathbf{R}^d$ that

$$\begin{aligned} c - \varphi\big((1-\lambda)x + \lambda y\big) &= c - (1-\lambda)\varphi(x) - \lambda\varphi(y) \\ &= c - (1-\lambda)[c - \psi(T(x))] - \lambda[c - \psi(T(y))] \\ &= (1-\lambda)\psi(Tx) + \lambda\psi(Ty). \end{aligned}$$

Under the assumption that T is affine,

$$\begin{aligned} c - \varphi\big((1-\lambda)x + \lambda y\big) &= \psi\big(T\big((1-\lambda)x + \lambda y\big)\big) \\ &= \psi\big((1-\lambda)Tx + \lambda Ty\big). \end{aligned}$$

Combining the equations above gives the convexity of ψ, which proves that g is log-concave. □

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 Borel 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.

If $\mu - \nu \ll \rho$, then by Proposition 4.21

$$\|\mu - \nu\| = |\mu - \nu|(S) = \int_S \frac{d|\mu - \nu|}{d\rho}\, d\rho.$$

This leads to the following characterization of TV distance for discrete and continuous random variables.

Notice that two positive measures μ and ν always have at least one dominating measure $\mu + \nu$.

8.3 Fact. Say ρ is a common dominating measure μ and ν , then

$$d_{\mathrm{TV}}(\mu, \nu) = \frac{1}{2} \int_S \left| \frac{d\mu}{d\rho}(x) - \frac{d\nu}{d\rho}(x) \right| d\rho. \tag{8.4}$$

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.

Furthermore from the proof of Proposition 4.21, one can get

$$d_{\mathrm{TV}}(\mu, \nu) = \sum_{x : \mu\{x\} \geq \nu\{x\}} \mu\{x\} - \nu\{x\}$$

for discrete random variables (and similarly for continuous random variables), which can be handy at times.

The *Kullback–Leibler divergence/relative entropy* of μ with respect to ν is given by

$$D(\mu \| \nu) = \begin{cases} \int_S \log \frac{d\mu}{d\nu}\, d\mu & \text{if } \mu \ll \nu, \\ +\infty & \text{otherwise.} \end{cases}$$

First observe that $D(\mu\|\nu) \geq 0$, and $D(\mu\|\nu) = 0$ if and only if $\mu = \nu$. Clearly $D(\mu\|\nu) \neq D(\nu\|\mu)$ in general, so this is not a metric.

Assume $f = \frac{d\mu}{d\nu}$ exists. 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 still be infinite.

[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.

8.5 Fact. If $\mu \ll \nu \ll \rho$, then

$$D(\mu\|\nu) = \int_S \left(\frac{d\mu}{d\rho}\right) \log\left(\frac{d\mu/d\rho}{d\nu/d\rho}\right) d\rho.$$

Therefore if the space is discrete, then we take ρ to be the counting measure and get

$$D(\mu\|\nu) = \sum_{x\in S} \mu\{x\} \log \frac{\mu\{x\}}{\nu\{x\}}.$$

And if $(S, \mathcal{S}) = (\mathbf{R}, \mathcal{B})$, with ρ being the Lebesgue measure, then

$$D(\mu\|\nu) = \int_{\mathbf{R}} f(x) \log \frac{f(x)}{g(x)}\, dx,$$

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 $D(f\|g)$.

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.

It is easy to check that $x \mapsto x \log x$ has second derivative $1/x > 0$, and hence strictly convex. (This is an important fact in probability!) Therefore by Jensen's inequality, both the entropy and variance functionals are nonnegative. 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

$$\operatorname{Ent}_\nu \frac{d\mu}{d\nu} = D(\mu\|\nu).$$

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.

[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.

8.6 Remark. The discrete *Shannon entropy* of a p.m.f. p is given by

$$H(p) = -\sum_x p(x) \log p(x),$$

which should be seen as a measure of uncertainty of a random variable $X \sim p$. If we consider the affine transformation $\widetilde{X} = aX + b$, it is easy to see that the Shannon entropy remains unchanged.

The *differential entropy* of a density function f on $\mathbf{R}^d$ is given by

$$h(f) = -\int_{\mathbf{R}^d} f \log f \, dx,$$

which is defined to be an analog the well-known Shannon entropy. However, given $X \sim f$, it is easy to verify that

$$h(aX + b) = h(X) + \log|a|,$$

which is undesirable, and does not serve as an *intrinsic* measure of uncertainty for a continuous random variables.

However, this special integral $\int f \log f$ has appeared in our discussion of *relative* entropy. Hence despite the fundamental difference between discrete Shannon and differential entropy, it does appear in some inequalities (e.g., Blachman–Stam and Shannon–Stam inequalities) and applications (e.g., in optimal transport).

We now state a result from information theory, which provides the most important reason that differential entropy is still useful. For a mean-zero random vector X with positive-definite covariance Σ, we have $h(X) \le h\big(N(0,\Sigma)\big) = \frac{1}{2}\log\big((2\pi e)^d \det\Sigma\big)$, with equality if and only if $X \sim N(0,\Sigma)$. To prove this, expand $0 \le D(f\|\gamma_\Sigma)$, which gives

$$\begin{aligned} 0 &\le -h(f) - \int f \log \gamma_\Sigma \, dx \\ &= -h(f) - \int \gamma_\Sigma \log \gamma_\Sigma \, dx = -h(f) + h(\gamma_\Sigma) \end{aligned}$$

For $\int f \log \gamma_\Sigma = \int \gamma_\Sigma \log \gamma_\Sigma$, observe

$$\log(\gamma_\Sigma) = \log\left(\frac{1}{(2\pi)^{d/2}\det\Sigma^{1/2}}\right) - \frac{x^{\mathrm{T}}\Sigma^{-1}x}{2},$$

and also

$$\int f(x)\, x^{\mathrm{T}}\Sigma^{-1}x \, dx = \operatorname{tr}\left(\int f(x)\, xx^{\mathrm{T}} \, dx\right) \operatorname{tr}(\Sigma^{-1}) = \operatorname{tr}(\Sigma\Sigma^{-1}) = d,$$

for any f with covariance matrix Σ.

An elementary introduction to information theory can be found in [CT05]. To remain consistent with our notations from before, for a density function f, we define $\operatorname{Ent}(f) = \int f \log f \, dm = -h(f)$.

A close relative of entropy is the Fisher information, where we replace $\log f$ by $|\nabla \log f|^2$. Given $\mu, \nu \in \mathcal{P}(\mathbf{R}^d)$, the *relative Fisher information* of ν with respect to μ is given by

$$I(\nu\|\mu) = \int_S |\nabla \log f|^2 \, d\nu = -\int_S (\Delta \log f)\, d\nu = \int_S \frac{|\nabla f|^2}{f}\, d\mu = 4\int_S \left|\nabla \sqrt{f}\right|^2 d\mu$$

if $f = \frac{d\nu}{d\mu} \geq 0$ exists, and $\sqrt{f} \in H^1(\mu)$. Otherwise let $I(\nu\|\mu) = \infty$.

Given a probability density function f such that $\sqrt{f} \in H^1(dx)$, its (ordinary) *Fisher information* is given by

$$I(f) = \int_{\mathbf{R}^d} |\nabla \log f|^2 f \, dx = -\int (\Delta \log f) f \, dx = \int \frac{|\nabla f|^2}{f} \, dx = 4 \int \left|\nabla \sqrt{f}\right|^2 dx.$$

Unfortunately Fisher information will become useful only late in the text, but we state it here for convenience.

8.7 Remark. Fisher information is foremost a very important concept in parametric statistics, which we will not go into. (Check out the Cramér–Rao lower bound and the asymptotic normality of the maximum-likelihood estimator.) We do stress that the definition in statistics is quite different. The density function $f(x;\theta)$ is parametrized in terms of θ. Fix the family of density functions $x \mapsto f(x;\theta)$, the Fisher information of the real parameter θ is given by $I(\theta) = \int_{\mathbf{R}^d} |\partial_\theta \log f(x;\theta)|^2 f(x;\theta)\, dx$, which is the variance of the so-called score function $\partial_\theta f(X;\theta)$.[4] There are also some connections between Fisher information and relative entropy in the statistics, simply because the algebra in either context remains the same.

8.8 Pinsker's inequality. $d_{\mathrm{TV}}(\mu,\nu) \leq \sqrt{\frac{1}{2} D(\mu\|\nu)}$.

Proof 1. When $\mu \not\ll \nu$ the above formula is trivial, so we only consider the case $\mu \ll \nu$. An elementary way, provided in [Tsy09, Lemma 2.5(i)], is to reduce the inequality to density functions with respect to $\rho = \mu + \nu$. Set $f = \frac{d\mu}{d\rho}$ and $g = \frac{d\nu}{d\rho}$ ρ-a.e. The inequality then becomes

$$\int_S |f - g| \, d\rho \leq \sqrt{2} \int_S f \log \frac{f}{g} \, \rho. \tag{8.9}$$

It is easy to check that

$$\psi(x) = x \log x - x + 1 \quad (x \geq 0,\ 0 \log 0 = 0)$$

satisfies $\psi(0) = 0$, $\psi(1) = 0$, $\psi'(1) = 0$, $\psi''(x) = 1/x \geq 0$, and $\psi(x) \geq 0$ for all $x \geq 0$.

Now check for $x > 0$,

$$h(x) = (x-1)^2 - \left(\tfrac{4}{3} + \tfrac{2}{3}x\right)\psi(x)$$

satisfies $h(1) = 0$, $h'(1) = 0$, and $h''(x) = \frac{-4\psi(x)}{3x} \leq 0$ for $x > 0$. A Taylor series expansion would then give $h(x) \leq 0$ for all $x > 0$. When $x = 0$, $h(x) < 0$. We then conclude that

$$(x-1)^2 \leq \left(\tfrac{4}{3} + \tfrac{2}{3}x\right)\psi(x) \quad \text{for all } x \geq 0. \tag{8.10}$$

Going back to (8.9), notice that

$$\int_S |f - g| \, d\rho = \int_{\{g>0\}} \left|\frac{f}{g} - 1\right| g \, d\rho$$

[4] One can also consider parameters $\theta \in \mathbf{R}^d$, but then we need a Fisher information matrix.

since $g \geq 0$ ρ-a.e. With (8.10), the above is

$$\begin{aligned}
&\int_{\{g>0\}} g\sqrt{\Big(\frac{4}{3}+\frac{2f}{3g}\Big)\psi\Big(\frac{f}{g}\Big)}\,d\rho \\
&\leq \sqrt{\int_{\{g>0\}} \frac{4g}{3}+\frac{2f}{3}\,d\rho}\cdot\sqrt{\int g\,\psi(f/g)\,d\rho} \quad \text{by Cauchy–Schwarz} \\
&= \sqrt{\frac{4}{3}+\frac{2}{3}}\cdot\sqrt{\int_{\{fg>0\}} f\log(f/g)\,d\rho - 1 + 1} \\
&= \sqrt{2}\cdot\int_{\{fg>0\}} f\log\frac{f}{g}\,d\rho, \text{ as desired.}
\end{aligned}$$

□

The above proof might reminds people about the proof of Hoeffding's lemma. Indeed, we may use that to give a straightforward proof of Pinsker's inequality. This will be mentioned in Section 16.B.

Fix $\nu \in \mathcal{P}(S)$, and write $D(\,\cdot\,) = D(\,\cdot\,\|\nu)$. Define the function

$$\varphi(g) = \log \mathrm{E}_\nu(e^g)$$

for all bounded measurable functions $g\colon S \to \mathbf{R}$. Meanwhile define D on the entire $\mathcal{M}(S)$ by setting $D(\mu) = +\infty$ for $\mu \notin \mathcal{P}(S)$.

8.11 Donsker–Varadhan variational principle. The functions $\varphi\colon$ (bounded measurable) $\to \mathbf{R}$ and $D\colon \mathcal{M}(S) \to \mathbf{R}$ are convex conjugates of each other. This implies that $D(\,\cdot\,\|\nu)$ is convex, and

$$D(\mu\|\nu) = \sup\{\mathrm{E}_\mu g - \log \mathrm{E}_\nu(e^g) : g \text{ bounded measurable}\}.$$

If S is a metric space, then the supremum can be taken over $C_b(S)$.

weak convergence

$$D(\mu\|\nu) \leq \liminf_n D(\mu_n\|\nu).$$

8.12 Gibbs variational principle.

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

$$d_{\mathrm{H}}(\mu,\nu) = \left(\frac{1}{2}\int_S \big[\sqrt{f(x)} - \sqrt{g(x)}\big]^2\,d\rho(x)\right)^{1/2}.$$

(Do not mistaken this with the Hausdorff distance, which has the exact same notation. We will not mention Hellinger distance anywhere else in the text.) 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:

$$d_{\mathrm{H}}^2(\mu,\nu) \le d_{\mathrm{TV}}(\mu,\nu) \le \sqrt{2} d_{\mathrm{H}}(\mu,\nu).$$

This follows from a straightforward comparison with (8.4).

probability metric

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 $\int_S f\, d\mu \neq \int_S f\, d\nu$. It should be noted that $\mu - \nu(S) = 0$ is important, otherwise the supremum would easily become $+\infty$, just by taking f to be arbitrarily large positive constant.

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 $\mathcal{F} = \{\mathbf{1}_{(-\infty,x]} : x \in \mathbf{R}\}$.

8.B The coupling technique and Wasserstein metric

Given two probability measures μ and ν on $(\mathbf{R}, \mathcal{B})$, we say ν stochastically dominates μ, denoted by $\mu \preceq \nu$, if

$$\mu(t,\infty) \le \nu(t,\infty) \text{ for all } t \in \mathbf{R}.$$

We are interested in the case where μ and ν are realized by two real-valued random variables *defined on the same probability space* $(\Omega, \mathcal{F}, P)$, and we write $X \preceq Y$ if $\mu_X \preceq \mu_Y$.

8.13 Fact. $X \preceq Y$ is equivalent to saying for any increasing f such that $\mathrm{E}|f(X)|$ and $\mathrm{E}|f(Y)|$ are finite, we have

$$\mathrm{E}f(X) \le \mathrm{E}f(Y).$$

This is clear form layer cake representation.

8.14 Proposition. For a given joint pair (X, Y) such that $X \preceq Y$, there exists a *monotone coupling* $(\widehat{X}, \widehat{Y})$, which means that

$$\widehat{X} \le \widehat{Y} \text{ a.s., while } \mu_{\widehat{X}} = \mu_X \text{ and } \mu_{\widehat{Y}} = \mu_Y.$$

Proof. In light of Theorem 7.7, we may use the same uniform random variable to define the distribution of X and Y. The monotonicity is easy to see. □

This provides another proof of Fact 8.13.

The above result is known as *Strassen's theorem* in the general case, where X and Y take values on a finite poset with the power set σ-field, or even more generally a Polish space $(S, \mathcal{S})$ with a closed partial order. (A partial order $\preceq$ on S is *closed* if the set

$$\{(x,y) \in S \times S : x \preceq y\}$$

is closed in the product topology.) See [Roc24, Theorem 4.2.11] and [Lin99].

Given two Borel probability measures μ and ν respectively on two topological space S and T, we define the coupling space $\Pi(\mu,\nu)$ to be the space of all $\pi \in \mathcal{P}(S\times T)$ such that

$$\pi(A\times T)=\mu(A) \quad \text{and} \quad \pi(S\times B)=\nu(B),$$

i.e., the probability measures on the product space $S\times T$ whose marginals are μ and ν. This set is nonempty because of $\mu\times\nu$, and is convex: for any π_1 and π_2, $0<\lambda<1$, we have

$$(1-\lambda)\pi_1(A\times T)+\lambda\pi_2(A\times T)=\mu(A)$$

and similarly $(1-\lambda)\pi_1(S\times B)+\lambda\pi_2(S\times B)=\nu(B)$.

Let (S,ρ) be a separable metric space, and $1\le p<\infty$, the *Wasserstein distance* of order p is defined by

$$W_p(\mu,\nu)=\inf_{\pi\in\Pi(\mu,\nu)}\left[\int_{S\times S}\rho(x,y)^p\,d\pi(x,y)\right]^{1/p}. \tag{8.15}$$

Alternatively, one has the probabilistic interpretation

$$W_p(\mu,\nu)=\inf\big\{\mathrm{E}\big[\rho(X,Y)^p\big]^{1/p} : X\sim\mu, Y\sim\nu\big\}, \tag{8.16}$$

and understand it as an L^p distance between two probability measures.

To see why the two characterizations are equivalent, first of all

$$\mathrm{E}\big[\rho(X,Y)^p\big]=\int_{S\times S}\rho(x,y)^p\,d\pi(x,y).$$

for $\pi = P\circ(X,Y)^{-1}$. Secondly, $\mu = P\circ X^{-1}$ is equivalent to saying for $A\in\mathcal{S}$, we have

$$\begin{aligned}\pi(A\times S) &= P\circ(X,Y)^{-1}(A\times S)\\ &= P\circ X^{-1}(A)=\mu(A).\end{aligned}$$

(A similar result holds for ν.) This shows that $X\sim\mu$ and $Y\sim\nu$ is equivalent to $\pi\in\Pi(\mu,\nu)$, and hence the right hand sides of (8.15) and (8.16) should be the same.

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

$$\mathcal{P}_p(S)=\left\{\mu\in\mathcal{P}(S) : \int_S \rho(x_0,x)^p\,d\mu(x)<\infty \text{ for some (and hence any) } x_0\in S\right\}.$$

(Recall (5.1).) In the case where$S$ is a normed space (e.g. $\mathbf{R}^d$), we have

$$\mathcal{P}_p(\mathbf{R}^n)=\left\{\mu\in\mathcal{P}(S) : \int_S \|x\|^p\,d\mu(x)<\infty\right\},$$

the space of measures with precisely finite p-th moments.

8.17 Fact. As expected, the Wasserstein distance gives a metric on $\mathcal{P}_p(S)$.

Unfortunately the proof of this fact has to be delayed to .

For point masses δ_x and δ_y, it is easy to see

$$W_p(\delta_x, \delta_y) = \rho(x, y).$$

by the probabilistic interpretation.

$W_p\big(N(), N()\big)$

$W_1(\mu, \nu) \le W_2(\mu, \nu) \le \cdots \le W_\infty(\mu, \nu)$

$\mathcal{P}_1(S) \supseteq \mathcal{P}_2(S) \supseteq \ldots \supseteq \mathcal{P}_\infty(S)$

For a metric space with bounded metric, we have

$$W_p(\mu, \nu)^p \le \operatorname{diam}(S)^{p-1} W_1(\mu, \nu).$$

8.18 Fact. When S is Polish, the infimum in the definition of Wasserstein distance can be attained.

As in the case for L^p space, our primary focus is on $\mathcal{P}_1$ and $\mathcal{P}_2$ spaces. It is easy to check that

$$\mathcal{P}_1(S) = \left\{ \mu \in \mathcal{P}(S) : \int_S f \, d\mu < \infty \text{ for all 1-Lipschitz functions } f \right\}.$$

On the one hand, for any 1-Lipschitz function f, we have

$$\int f \, d\mu - f(x_0) = \int f(x) - f(x_0) \, d\mu(x) \le \int \rho(x_0, x) \, d\mu(x).$$

On the other hand, for each $x_0 \in S$, we may define a 1-Lipschitz function $f(x) = \rho(x_0, x)$.

8.19 Dual representation of W_1. For $\mu, \nu \in \mathcal{P}_1(S)$, we have

$$W_1(\mu, \nu) = \sup \left\{ \left| \int f \, d\mu - \int f \, d\nu \right| : f \text{ is 1-Lipschitz, } f \in L^1(\mu - \nu) \right\}.$$

Thus W_1 is an IPM.

The inequality $\ge$ should be very easy. Indeed, for any $\pi \in \Pi(\mu, \nu)$, we have for any 1-Lipschitz f that

$$\left| \int f \, d\mu - \int f \, d\nu \right| = \left| \int_{S \times S} f(x) - f(y) \, d\pi(x, y) \right| \le \int_{S \times S} \rho(x, y) \, d\pi.$$

The reverse inequality $\le$ is however quite difficult.

This may be seen as a special case of the so-called the *Kantorovich–Rubinstein duality* (KR duality henceforth).

If we consider S endowed with the discrete metric $\rho(x, y) = \mathbf{1}\{x \ne y\}$, then for $\mu, \nu \in \mathcal{P}_1(S)$, we have by definition

$$W_1(\mu, \nu) = \inf_{\substack{X \sim \mu \\ Y \sim \nu}} P(X \ne Y).$$

On the other hand notice that f is 1-Lipschitz means precisely that $\sup f - \inf f \le 1$, and hence by the KR duality, we have

$$W_1(\mu, \nu) = \sup_{0 \le f \le 1} \left| \int f \, d\mu - \int f \, d\nu \right| = d_{\mathrm{TV}}(\mu, \nu).$$

This tells us one way to bound the total variation distance between μ and ν is to find a coupling (X, Y) that marginally follows μ, ν that minimizes $P(X \ne Y)$.

8.C 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 $\mu(S) \leq 1$, and we denote the space of all Borel subprobability measures by $\mathcal{M}^{\leq 1}(S)$.

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].

8.20 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[5] $X_n \Rightarrow X$. Because of Corollary 7.6, when $S = \mathbf{R}$ we also write $F_{X_n} \Rightarrow F_X$.

Recall that vague convergence

8.21 Proposition. Weak convergence of integer-valued measures is equivalent to pointwise convergence.

8.22 Alexandroff portmanteau theorem. 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) $\limsup_n \int f \, d\mu_n \leq \int f \, d\mu$ for all USC functions bounded from above;
(d) $\liminf_n \int f \, d\mu_n \geq \int f \, d\mu$ for all LSC functions bounded from below;
(e) $\limsup_n \mu_n(F) \leq \mu(F)$ for all closed sets F;
(f) $\liminf_n \mu_n(G) \geq \mu(G)$ 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 $\mu_n, \mu \in \mathcal{M}^+(S)$.[6]

8.23 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, □

[5] sometimes even mix up and write $X_n \Rightarrow \mu$

[6] As Bogachev [Bog18] points out, "I do not know who invented such a nonsensical name for Alexandroff's theorem."

Recall that the subsequential limit F constructed above associates to a Borel measure μ_F, by Theorem 1.34(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.

8.24 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 (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. □

As an exercise, one may try to give a direct proof in the $\mathbf{R}^1$ case, by arguing directly with the distribution functions. Recall the sequential Alaoglu we used above was proved using a diagonal argument, so this will the key ingredient in this exercise as well. We knew from the classical Arzelà–Ascoli theorem that the diagonal argument is an important tool in establishing subsequential convergence.

8.25 Exercise. 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.

(We have given a hint when discussing distribution functions in Section 7.A.)

8.26 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. In particular, this includes $\mathbf{R}^d$ with the usual Euclidean norm.

The μ_F above is not in general a probability measure, for example, consider the uniform distributions over $[-n, n]$. The sequence $\{\text{Uniform}[-n, n]\}$ *itself* (and hence all of its subsequences) converges vaguely to the 0 function. To ensure that the subsequential μ_n constructed in Helly selection theorem is indeed a distribution function, we require tightness over *the entire sequence of measures* in addition. We say a family $\{\mu_\alpha\}_{\alpha\in A}$ of measures is *tight* if for each $\epsilon > 0$, there exists some compact set K_ϵ such that

$$\sup_{\alpha\in A} \mu_\alpha(S - K_\epsilon) < \epsilon.$$

Indeed this merely extends the idea of tightness of a single 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.

since we are working with subprobability measures When $S = \mathbf{R}^d$, vague convergence may be further defined by $\int f\,d\mu \to \int f\,d\mu$ for $f \in C_c^\infty(\mathbf{R}^d)$

8.27 Theorem [Sch17, Theorem 21.17]. Let S be locally compact and separable[7], 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 8.26 from compact to locally compact separable metric spaces. It also provides an accurate characterization for tightness.

8.28 Proposition. For a sequence of Borel probability measures in a locally compact separable metric space, every vague subsequential limit (which exists by Helly selection theorem) is a probability measure if and only if the whole 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 8.27. □

8.29 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 $S = \mathbf{Z}$
[Bil99, Theorem 3.4 & 3.5]
If $X_n \Rightarrow X$, then $\mathrm{E}|X| \le \liminf_n \mathrm{E}|X_n|$.
For $\{X_n\}$ uniformly integrable and $X_n \Rightarrow X$, the limit X must be integrable as well, with $\mathrm{E}X_n \to \mathrm{E}X$.

8.C.1 The topology and metric of weak convergence

When S is locally compact and separable, we know $C_c(S)$ is separable, and therefore Theorem C.9 tells us that the vague topology on $\mathcal{M}^{\le 1}(S)$ is metrizable. This induces a metric for the topology of vague convergence on $\mathcal{P}(S)$. Unfortunately, despite the fact that vague convergence and weak convergence when the limiting measure is a probability measure, their topologies on $\mathcal{P}(S)$ are still different, simply because the test functions (and hence the basic open sets that define their topologies) are different. The current subsection is subject to the study of the topology of weak convergence.

Compare with the Arzelà–Ascoli theorem for the space of continuous functions.

[7] 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.

8.30 Prohorov's theorem. Let S be a metric space (not necessarily separable). Suppose a collection of measures $\mathcal{K} \subseteq \mathcal{P}(S)$ is tight, then $\mathcal{K}$ is *sequentially precompact* in the topology of weak convergence on $\mathcal{P}(S)$, i.e., for any sequence of measures in $\mathcal{K}$ there is a subsequence that converges in $\mathcal{P}(S)$.

Proof. We follow [Kal21, Theorem 23.2]. [DaP06, Theorem 6.7] uses the diagonal argument twice □

The converse is true when S is Polish. Note this converse is just a generalization of Ulam's theorem.

8.31 Corollary. [ABS24, Corollary 2.9] Let $\mu \in \mathcal{P}(S)$ and $\nu \in \mathcal{P}(T)$ be tight (e.g., if S and T are both Polish), then the space $\Pi(\mu, \nu)$ of couplings between μ and ν is a sequentially compact subspace of $\mathcal{P}(S \times T)$.

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

$$\int_{S \times T} (\varphi, \mathrm{Id}_Y)\, d\pi = \int_S \varphi\, d\mu \quad \text{for all } \varphi \in C_b(S),$$

and similarly with respect to the marginal ν. It is then clear that the weak limit of a sequence $\{\pi_n\} \subseteq \Pi$ still falls in Π.

By Prohorov's theorem, it now suffices to show that $\Pi(\mu, \nu)$ is a tight family. 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 $\pi \in \Pi(\mu, \nu)$

$$\begin{aligned} \pi(S \times T - K_1 \times K_2) &\leq \pi\big((S - K_1) \times T\big) + \pi\big(S \times (T - K_2)\big) \\ &= \mu(S - K_1) + \nu(T - K - 2) < \epsilon, \end{aligned}$$

proving tightness. □

8.32 Theorem [Bog18, Theorem 3.1.2]. The weak topology on $\mathcal{M}^+(S)$

The weak topology can be metrized Prohorov metric on a metric space

For $\mu, \nu \in \mathcal{P}(S)$, define

$$d_{\mathrm{P}}(\mu, \nu) = \inf\{\epsilon > 0 : \nu(B) \leq \mu(B^\epsilon) + \epsilon, \mu(B) \leq \nu(B^\epsilon) + \epsilon \text{ for all } B \in \mathcal{S}\}.$$

On a separable metric space S, d_{P} metrizes the topology of weak convergence on $\mathcal{P}(S)$.

Wasserstein distance on a separable metric space

Ky Fan metric on a separable metric space metrizes convergence in probability

Convergence in probability for real random variables is the same as convergence in measure for real-valued functions. The generalization to separable metric spaces (S, ρ) is obvious. We say $X_n \to X$ in probability on S if for any $\epsilon > 0$, we have

$$P(\rho(X_n, X) > \epsilon) \to 0.$$

(Again separability of S is to ensure measurability.) Previously in Section 5.E, we have defined the Ky Fan metric $\alpha(\,\cdot\,, \cdot\,)$ on the space $L^0(\Omega, \mathcal{F}, P)$ of random variables on S by

$$d_{\mathrm{Fan}}(X, Y) = \inf\big\{\epsilon \geq 0 : P(\rho(X, Y) > \epsilon) \leq \epsilon\big\}.$$

(It is an easy exercise to show that the infimum can be attained, thanks to the inequalities ">" and $\leq$, which cannot be replaced.)

If S is Polish, then $\big(L^0(\Omega), \alpha\big)$ is Polish.

For S separable, we have

$$d_{\mathrm{P}}(\mu, \nu) = \inf_{\substack{X \sim \mu \\ Y \sim \nu}} d_{\mathrm{Fan}}(X, Y).$$

For $\mu_n \in \mathcal{P}_p(S)$ and $\mu \in \mathcal{P}(S)$, we have $\mu_n \to \mu$ in W_p if and only if $\mu_n \Rightarrow \mu$ and

$$\int \rho(x_0, x)^p \, d\mu_n(x) \to \int \rho(x_0, x)^p \, d\mu(x)$$

for some fixed $x_0 \in S$.

if and only if $\mu_n \Rightarrow \mu$ and

$$\limsup_{n \to \infty} \int_{d(x_0, x) \geq M} d(x_0, x)^p \, d\mu_n(x) \to 0 \text{ as } M \to \infty$$

For a Polish space S, we know W_p metrizes the space $\mathcal{P}_p$ with the topology of weak convergence (for any $1 \leq p < \infty$) on any subset of S with a uniform tightness condition. Notice that if the metric on S is finite, W_p metrizes $\mathcal{P}_p(S) = \mathcal{P}(S)$ with the weak topology. For a Polish space S metrized by an unbounded metric ρ, one may always replace ρ by a bounded metric $\min\{\rho, 1\}$. The topology on S is the same so weak convergence on S is still the same. Therefore W_p with respect to a bounded metric on S metrizes the entire $\mathcal{P}(S)$.

Continuing with Section 2.I and Section 4.A, we give one more useful property of the pushforward:

8.33 Proposition. For $\varphi \colon S \to S$ continuous, the pushforward $\varphi_* \colon \mathcal{P}(S) \to \mathcal{P}(S)$ is sequentially continuous when the $\mathcal{P}(S)$ is endowed with topology of weak convergence.

Proof. Say $\mu_n \Rightarrow \mu$ in $\mathcal{P}(S)$, i.e.,

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

This implies that

$$\int f \circ \varphi \, d\mu_n \to \int f \circ \varphi \, d\mu \quad \text{for all } f \in C_b(S).$$

since for continuous $\varphi \colon S \to S$, $f \circ \varphi \in C_b(S)$. It follows that

$$\int f \, d(\varphi_* \mu_n) \to \int f \, d(\varphi_* \mu),$$

i.e., $\varphi_* \mu_n \Rightarrow \varphi_* \mu$. □

8.C.2 Problem of measurability

When S is infinite, $\mathcal{B}(S)$ is not separable.

8.D Comparisons between modes of convergence

8.34 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 measures.

8.35 Theorem. When S is a separable metric space, if $X_n \to X$ a.s., then $X_n \to X$ in probability, which further implies $X_n \Rightarrow X$.

Proof. The first part was done in Theorem 2.23. □

8.36 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 $f \in C_b(S)$, $g \circ f \in C_b(S)$.

8.37 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.

8.38 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

8.39 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.E Laws of large numbers

8.40 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. □

8.41 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.

8.42 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 $\{\overline{X}_n\}$. □

If the distribution for the sequence is assumed to be i.i.d. L^4 or L^2, then much simpler proofs can be given. One should try to recover them on their own.

Let $X_1, X_2, \dots$ follow a common distribution μ on the real line, or alternatively a common distribution function F. The *empirical/sample 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 point mass of the first n sample observations. Notice that μ_n is a *random probability measure*, a random variable from Ω to $\mathcal{P}(\mathbf{R})$. This μ_n gives us the *empirical distribution function*

$$F_n(x) = \mu(-\infty, x] = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}\{X_k \le x\},$$

which is a random function defined on Ω.

8.43 Glivenko–Cantelli theorem. As $n \to \infty$,

$$\sup_x |F_n(x) - F(x)| \to 0 \quad P\text{-a.s.}$$

11.4.1 Dudley

For any Borel probability measures μ on a separable metric space S, the empirical distribution μ_n converges a.s. to μ.

Kolmogorov–Smirnov statistics and test

8.44 Dvoretzky–Kiefer–Wolfowitz–Massart inequality. For every $\epsilon > 0$,

$$P\big(\sup_x |F_n(x) - F(x)| > \epsilon\big) \le 2\exp(-2n\epsilon^2).$$

8.F 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 $M_X \colon \mathbf{R} \to \mathbf{R}$ defined by $M_X(t) = \mathrm{E}\exp(itX)$, provided that $\exp(itX)$ is integrable. Its *characteristic function* (ch.f.) is a function $\varphi_X \colon \mathbf{R} \to \mathbf{C}$ 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 $M_X(t) = \mathrm{E}\exp(t \cdot X)$ and $\varphi_X(t) = \mathrm{E}\exp(it \cdot X)$ 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

8.45 Example. For $X \sim N(0,1)$ and $t \in \mathbf{R}$, we have the $M_X(t)$ 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 $\varphi_X(t)$ 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

$$\mathrm{E}\exp(tX) = \exp(t^2/2) \tag{8.46}$$

for all $t \in \mathbf{C}$.

We wish to use the uniqueness theorem given (8.46) 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[8].

8.47 Example. The L^2 limit of a sequence of normal random variables must be normal.

8.48 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 $(-\delta, \delta)$.

8.49 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 $\varphi^{(k)}(t) = i^k \mathrm{E}X^k$.

8.50 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 $a < b$,

$$\overline{F}(b) - \overline{F}(a) = \lim_{T\to\infty} \int_{-T}^{T} \frac{\exp(-iat) - \exp(-ibt)}{2\pi it} \varphi(t)\, dt$$

8.51 Theorem (c.d.f. and ch.f. correspondence). For any real random vectors X and Y, $X =_d Y$ if and only if $\varphi_X = \varphi_Y$.

For nonnegative random variables, we may use m.g.f.

[8] Recall we defined complex exponentials as power series, and power series/polynomials are differentiable term-by-term.

8.52 Corollary. Two given random vectors X in $\mathbf{R}^m$ and Y in $\mathbf{R}^n$ are independent if and only if $\varphi_{X,Y}(s,t) = \varphi_X(s)\varphi_Y(t)$ for all $s \in \mathbf{R}^m$ and $t \in \mathbf{R}^n$.

Proof. The only if direction is easy because the characteristic function is a continuous bounded (complex-valued) function.

For the if direction, let X' and Y' be two copies of X and Y that are independent. Then we have

$$\varphi_{X,Y}(s,t) = \varphi_{X'}(s)\varphi_{Y'}(t) = \varphi_{X'}(s)\varphi_{Y'}(t) = \varphi_{X',Y'}(s,t)$$

for any $(s,t) \in \mathbf{R}^{m+n}$. This implies that $(X,Y) \overset{\mathrm{D}}{=} (X',Y')$, which means X and Y are independent as well. □

8.53 Lévy's continuity theorem. Let $\{\mu_n\} \subseteq \mathcal{P}(\mathbf{R}^d)$.

(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 converges weakly to some probability measure μ whose characteristic function is φ.

In the above theorem statement, we do not know if the pointwise limit φ is a priori the characteristic function for some μ. The "continuity at 0" assumption is precisely included to ensure that $\varphi = \check{\mu}$ for some μ. Since $\varphi(0) = \lim_n \varphi_n(0) = 1$, μ is indeed a probability measure.

8.54 Bochner's theorem. A characteristic function $\varphi\colon \mathbf{R}^d \to \mathbf{C}$ is precisely characterized by the following three properties:

(a) $\varphi(0) = 1$;

(b) φ is continuous on $\mathbf{R}^d$;

(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.

8.55 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

$$\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \Rightarrow N(0,1)$$

Proof. By scaling we may assume $\mu = 0$ and $\sigma^2 = 1$.

$$\mathrm{E}\exp(X_1 + \cdots + X_n) = \varphi_{X_1}(t)^n$$

Taylor's theorem □

8.56 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

$$\frac{X_1 + \ldots + X_n}{\sigma_n} \Rightarrow N(0,1).$$

8.57 Lyapunov condition.

moment problem

8.58 Berry–Essen bound. For $X_1, X_2, \dots$ 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

8.59 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.

8.60 Poisson limit theorem. For a sequence of $X_n \sim \mathrm{Binomial}(n, p_n)$, where $np_n \to \lambda$ for some positive constant λ, we have

$$X_n \Rightarrow \mathrm{Poisson}(\lambda).$$

Explicitly, this means given $Y \sim \mathrm{Poisson}(\lambda)$, for all $k \in \mathbf{N}_0$, we have

$$P(X_n = k) \to P(Y = k) \quad \text{as } n \to \infty.$$

8.G The moment problem

The previous section tells us the characteristic function and the moment generating function of a given distribution can capture all the moments of the distribution. With ch.f. and m.g.f., we were able to recover the original distribution, and ch.f. and m.g.f. also interact well with weak convergence of measures.

Now we ask the following question: if we know all the moments of some $\mu \in \mathcal{P}(\mathbf{R})$, can we uniquely determine μ? The answer is no in general, the most well-known example being the log-normal distribution.

Fortunately, with some additional assumptions on the moments, one can indeed uniquely determine μ. Furthermore, it is possible to get results for weak convergence of measures, when their moment sequences are converging.

Given a distribution function F, define the pth moment of F is given by $\int x^p \, dF(x)$.

8.61 Proposition. If $\int x^p \, dF_n(x)$ has a limit m_p for each p, then the sequence $\{F_n\}$ is tight.

Therefore every vague subsequential limit is weak. Every weak subsequential limit should have pth moments m_p. If there is one unique distribution with pth moments m_p, then the distribution F is uniquely determined.

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 $\Omega_1, \Omega_2, \dots$ 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.24 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

$$\mathrm{E}\left(\prod_{j=1}^{n}f(X_j)\;\middle|\;\mathcal{E}\right)=\prod_{j=1}^{n}\mathrm{E}[f(X_j)\mid\mathcal{E}].$$

This result roughly says that the conditional distribution of $X_j\mid\mathcal{E}$ becomes i.i.d. We remind that when X_j's take value in standard Borel spaces, then there is a regular conditional distribution for $X_j\mid\mathcal{E}$.

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 $P(Y\in A|\mathcal{G})$ is unique only P-a.e., hence a version of $P(Y\in A|\mathcal{G})$ 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

$$\text{the function } \omega \mapsto \nu(\omega, A) \text{ is a version of } P(Y \in A \,|\, \mathcal{G}),$$

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

$$\kappa(X(\omega), A) \text{ is a version of } P(Y \in A \,|\, X)(\omega). \tag{9.19}$$

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.

Proof. Take $S = \mathbf{R}$. Let $\nu = \delta \times \rho$ □

(a) For two discrete random variables X and Y, we want

$$\kappa(x, A) = \begin{cases} P(Y \in A \,|\, X = x) & \text{if } P(X = x) > 0, \\ \delta_{y_0}(A) & \text{if } P(X = x) = 0. \end{cases}$$

(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

$$\mathrm{E}[f(X,Y) \mid X] = \int f(X,y)\, \kappa(X, dy).$$

$$\mathrm{E}[f(X,Y) \mid \mathcal{G}] = \int f(X,y)\, \nu(dy).$$

$$\mathrm{E} f(X,Y) = \mathrm{E} \int f(X,y)\, \kappa(X, dy).$$

It is easy to verify that $\mu = \int \delta_x \, d\mu(x)$, and the point masses are the extreme points of $\mathcal{P}(S)$ in the vector space $\mathcal{M}(S)$.

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 $0 \le n \le m$, $\mathrm{E}(X_m \mid \mathcal{F}_n) = X_n$ (resp. $\le$, $\ge$).

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

$$(H \bullet X)_n = H_1(X_1 - X_0) + H_2(X_2 - X_1) + \ldots + H_n(X_n - X_{n-1}),$$

for $n \ge 1$, and $(H \bullet X)_0 = 0$.

It is useful to see that $(H \bullet -X)_n = -(H \bullet X)_n$.

9.28 Proposition.

(a) If $\{X_n\}$ is a martingale, then $\{(H \bullet X)_n\}$ is a martingale.

(b) If $\{X_n\}_n$ is a submartingale (resp. supermartingale), and $H_n \ge 0$ for all n, then $\{(H \bullet X)_n\}$ 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 $T \le M$ a.s. for some $M < \infty$ (bounded stopping time), then $\mathrm{E}X_T = \mathrm{E}X_0$ (resp. $\le \mathrm{E}X_0$).

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 $\{\mathrm{E}(X \mid \mathcal{G}) : \mathcal{G}$ is a sub-σ-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 $\mathbf{Z}^{\le 0}$-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

$$aP\big(\max_{0\le k\le n} X_k \ge a\big) \le \mathrm{E}\Big(X_n \mathbf{1}\big\{\max_{0\le k\le n} X_n \ge a\big\}\Big) \le \mathrm{E}X_n^+.$$

If $\{Y_n\}$ is a supermartingale, then for every $a > 0$, we have

$$aP\big(\max_{0\le k\le n} Y_k \ge a\big) \le \mathrm{E}Y_0 + \mathrm{E}Y_n^-.$$

Combining the two cases above, we get for a submartingale or supermartingale $\{X_k\}$, it holds that

$$aP\big(\max_{0\le k\le n} |X_k| \ge a\big) \le \mathrm{E}|X_0| + 2\,\mathrm{E}|X_n|.$$

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

$$a^p P\big(\max_{0\le k\le n} X_k \ge a\big) \le \mathrm{E}\big(\max_{0\le k\le n} X_k\big)^p \le \left(\frac{p}{p-1}\right)^p \mathrm{E}(X_n)^p.$$

Therefore if $\{Z_n\}$ is a martingale, then $\{|Z_n|\}$ is a nonnegative submartingale. Therefore we have

$$a^p P\big(\max_{0\le k\le n} |Z_k| \ge a\big) \le \mathrm{E}\big(\max_{0\le k\le n} |Z_k|\big)^p \le \left(\frac{p}{p-1}\right)^p \mathrm{E}|Z_n|^p.$$

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

$$\mathrm{E} \max_{0\le k\le n} X_k^2 \le 4\,\mathrm{E}X_n^2. \tag{9.45}$$

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

$$\begin{aligned} A_n &= \sum_{k=1}^n \mathrm{E}(X_k^2 - X_{k-1}^2 \mid \mathcal{F}_{k-1}) \\ &= \sum_{k=1}^n \mathrm{E}(X_k^2 - 2X_kX_{k-1} + X_{k-1}^2 \mid \mathcal{F}_{k-1}) \\ &= \sum_{k=1}^n \mathrm{E}[(X_k - X_{k-1})^2 \mid \mathcal{F}_{k-1}], \end{aligned}$$

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 f and F_n-measurable Y that

$$\mathrm{E}[f(X_{n+1} - X_n)Y] = 0$$

by the tower property. This of course includes the case $Y = g(X_0, X_1, \ldots, X_n)$ for some Borel measurable g.

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

$$\begin{aligned}\mathrm{E}\exp\Big(\lambda\sum_{k=1}^{n} X_k - X_{k-1}\Big) &= \mathrm{E}\Big[\exp\lambda(X_n - X_{n-1})\cdot\exp\Big(\lambda\sum_{k=1}^{n-1} X_k - X_{k-1}\Big)\Big]\\ &= \mathrm{E}\Big[\mathrm{E}\Big(\exp\lambda(X_n - X_{n-1})\cdot\exp\Big(\lambda\sum_{k=1}^{n-1} X_k - X_{k-1}\Big)\;\Big|\;\mathcal{F}_{n-1}\Big)\Big]\\ &= \mathrm{E}\Big[\exp\Big(\lambda\sum_{k=1}^{n-1} X_k - X_{k-1}\Big)\,\mathrm{E}\Big(\exp\big(\lambda(X_n - X_{n-1})\big)\Big|\,\mathcal{F}_{n-1}\Big)\Big].\end{aligned} \tag{9.54}$$

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

$$\le \mathrm{E}\Big[\exp\Big(\lambda\sum_{k=1}^{n-1} X_k - X_{k-1}\Big)\Big]\exp\Big(\frac{\lambda^2 c_n^2}{8}\Big).$$

We may repeat the procedure above by first conditioning and then applying Hoeffding's lemma, and obtain in the end that

$$\begin{aligned}\mathrm{E}\exp\Big(\lambda\sum_{k=1}^{n} X_k - X_{k-1}\Big) &\le \prod_{k=1}^{n}\exp\Big(\frac{\lambda^2 c_k^2}{8}\Big)\\ &= \exp\Big(\frac{\lambda^2\sum_{k=1}^{n} c_k^2}{8}\Big).\end{aligned}$$

Going back to (9.53), we have

$$P(X_n - X_0 \ge t) \le \exp\Big(\frac{\lambda^2\sum_{k=1}^{n} c_k^2}{8} - \lambda t\Big).$$

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

$$P(X_n - X_0 \ge t) \le \exp\Big(-\frac{2t^2}{\sum_{k=1}^{n} c_k^2}\Big),$$

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 $c_1, \dots, c_n$. This means for each $k \in [n]$, we have

$$\sup_{\substack{x_1,\dots,x_n \\ x_k' \in S_k}} |g(x_1, \dots x_n) - g(x_1, \dots x_{k-1}, x_k', x_{k+1}, \dots, x_n)| \leq c_k.$$

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. □

9.57 Efron–Stein inequality. Define $\mathrm{Var}_j(x_1, \dots, x_n) = \mathrm{Var}(x_1, \dots, X_j, \dots, x_n)$. For independent $X_1, \dots, X_n$, we have

$$\mathrm{Var}\, f(X_1, \dots, X_n) \leq \mathrm{E}\left(\sum_{j=1}^n \mathrm{Var}_j\, f(X_1, \dots, X_n)\right).$$

Let us mention one elementary application of Azuma–Hoeffding inequality. For a Erdös–Renyí graph with n vertices and Bernoulli parameter p, one can show that

$$P(|\chi - \mathrm{E}\chi| \geq t\sqrt{n}) \leq 2\exp(-t^2/2),$$

where χ is the coloring number for the graph.

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 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.

9.58 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

$$\mathbf{P}_0(T_{-1} < \infty) = \mathbf{P}_0\Big(\bigcup_{N=1}^{\infty} \{T_{-1} < T_{N-1}\}\Big) = 1.$$

However, $\mathbf{E}_0 T_{-1} = \infty$.

by monotone convergence

9.59 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 $\big\{(q/p)^{S_n}\big\}_n$ 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:

$$\mathbf{P}_0(\min_n S_n \leq -a) = \mathbf{P}_0(T_{-a} < \infty) = \left(\frac{1-p}{p}\right)^a.$$

$$\mathbf{P}_0(\max_n S_n \geq b) = \mathbf{P}_0(T_b < \infty) = 1 \quad \mathbf{E}_0 T_b = \frac{b}{2p-1}.$$

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 H.

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

$$F_{\mu_i}^{-1}(U_i)$$

[1] See also [Kal21, Theorem 4.19] for a quick introduction.

produces an independent sequence of μ_i-distributed real random variables.

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 $X \sim$ Exponential(λ) and $Y \sim$ Exponential(μ), we have

(a) $\min\{X, Y\} \sim$ 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 $N(t) \coloneqq N_1(t) + N_2(t)$ 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 $\mathbf{\Omega} = S^{\mathbf{N}}$, 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 $(\mathbf{\Omega}, \mathfrak{F}, \mathbf{P})$, 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 $(\mathbf{\Omega}, \mathfrak{F})$ 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 $X_t \colon \Omega \to \mathbf{R}$ 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

$$d(f,g) = \sup\bigl\{1 \wedge |f(t) - g(t)| : t \in [0,\infty)\bigr\}.$$

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

$$d(f,g) = \sum_{n=1}^{\infty} \frac{1}{2^n} \max\bigl\{1 \wedge |f(t) - g(t)| : t \in [0,n]\bigr\}.$$

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.)

For the functions we will consider it usually suffices to consider uniform convergence on $C[0,1]$, and by scaling we obtain uniform convergence on $C[0,N]$ for all N. However, this space is not locally compact.

Weak convergence on the metric space $C[0,1]$ cannot be determined by

10.F Other constructions of Brownian motions

modify its path so that it becomes continuous

For $0 < \alpha \leq 1$, we say $f\colon S \to \mathbf{R}$ is α-Hölder continuous if for $x, y \in S$,

$$\sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x,y)^\alpha} < \infty.$$

(If $\alpha > 1$, then the function must be bounded, and if $\alpha = 0$, then we the function is only bounded.)

It should be easy to see that

$$\text{Lipschitz} \subseteq \alpha\text{-Hölder} \subseteq \text{uniform continuity},$$

since for any $0 < \alpha \leq 1$, we get uniform continuity; and if $\alpha = 1$, then we get Lipschitz continuity. When S is a compact subset of a Euclidean space, then Lipschitz further becomes stronger than C^1, and uniform continuity is just continuity.

We can define a corresponding *local Hölder continuity*. Instead take the supremum over the entire S, we take it over every compact subset of S. Furthermore we have an even weaker notion of Hölder regularity useful in continuous processes. We say f is *α-Hölder at a point* $x_0 \in S$ if there is a ball $B(x_0;\epsilon)$ around x_0 such that

$$\sup_{\substack{x \in B(x_0;\epsilon) \\ x \neq x_0}} \frac{|f(x) - f(x_0)|}{d(x,x_0)^\alpha} < \infty.$$

Given an arbitrary function $f\colon J \to \mathbf{R}$ for an interval J, the *δ-modulus of continuity* of f is defined by

$$\sup_{0<|t-s|\le\delta} |f(t) - f(s)|.$$

10.11 Kolmogorov–Chenstov continuity lemma. Given a complete separable metric space (S, ρ), let $X\colon [0,\infty) \times \Omega \to S$ 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 2.46, 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

$$\begin{aligned} B \coloneqq \{x \in A : T^n x \notin A \text{ ev.}\} &= \bigcup_{n=1}^{\infty} \bigcap_{m \geq n} \{x \in A : T^m x \notin A\}, \\ &= \bigcap_{n=1}^{\infty} \{x \in A : T^n x \notin A\} \\ &= A \cap \left(\bigcap_{n=1}^{\infty} T^{-n}(X - A) \right). \end{aligned}$$

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) *Any i.i.d. sequence is strong mixing, and hence an ergodic sequence.* Let $X_1, X_2, \dots$ be i.i.d. $(S, \mathcal{S})$-valued. Consider the canonical space $(S^{\mathbf{N}}, \otimes^{\mathbf{N}}\mathcal{S})$ with the product measure μ. First, it is clear that $\mu(T^{-1}A) = \mu(A)$ for all cylinder sets A, and hence all measurable A. Now for any cylinder sets A and B, it is clear that for large enough n we have

$$\mu(A \cap T^{-n}B) = \mu(A)\mu(B),$$

and hence the system is ergodic. The same argument clearly applies to a two-sided i.i.d. sequence.

Alternatively, one may use Kolmogorov zero–one law. Notice that A is T-invariant means precisely that $T^{-n}A = A$ for any n. Let $X_1, X_2, \dots$ be the sequence of i.i.d. random variables, and note that $\sigma(X_1, X_2, \dots) = \otimes^{\mathbf{Z}}\mathcal{S}$. This tells us that $A \in \mathcal{T}$, and hence $\mu(A) = 0$ or 1.

The *Bernoulli shift* deals with the case where $S = \{0, 1\}$ and $\mu = \otimes \text{Bernoulli}(p)$.

(b) *Rotation on a circle.* Consider the space $([0, 1), \mathcal{B}, m)$, and define $Tx = c + x \mod 1$ for $x \in [0, 1)$.

If c is a rational number p/q (with p, q coprime), then consider the set

$$A = \left[0, \tfrac{p}{2q}\right] \cup \left[\tfrac{p}{q}, \tfrac{p}{q} + \tfrac{1}{2q}\right] \cup \cdots \cup \left[\tfrac{p(q-1)}{q}, \tfrac{p(q-1)}{q} + \tfrac{1}{2q}\right] \mod 1.$$

It is clear that this set has measure $1/2$, but is invariant under T. Hence T is not ergodic.

Now we show that T is ergodic when c is an irrational number. Consider we have $A \in \mathcal{B}$ with $m(A) > 0$, and we will prove that $m(A) = 1$.

Recall Proposition 1.47. Since T is invertible, we then have

$$m(A \cap T^k I) \geq (1 - \epsilon)\, m(T^k I)$$

For any $k_1, \dots, k_n$ such that $T^{k_1}I, \dots, T^{k_n}I$ are all disjoint intervals, we have

$$m(A) \geq \sum_{j=1}^{n} m(A \cap T^{k_j} I) \geq (1 - \epsilon)\, m\left(\bigcup_{j=1}^{n} T^{k_j} I\right).$$

By the density of the left endpoints of the intervals $\{T^k I\}_{k=1}^{\infty}$, we may pick $k_1, \dots, k_n$ appropriately such that

$$m\left(\bigcup_{j=1}^{n} T^{k_j} I\right) \geq 1 - 2\epsilon,$$

where one ϵ comes from the internal gaps between $T^{k_1}I, \dots, T^{k_n}I$, while the remaining ϵ is the remaining gap that cannot be covered by translating I.

Therefore $m(A) \geq (1-\epsilon)(1-2\epsilon)$, and since ϵ is arbitrary, we conclude that $m(A) = 1$. To conclude, the shift-by-c map is ergodic on the unit interval if and only if c is rational. Equivalently, we may say that rotation-by-c map $Tx = cx$ on the unit circle is ergodic if and only if the c is a root of unity.

(c) Markov shift (delayed)

from [Bil95, Theorem 36.5]

Proof of Hewitt–Savage zero–one law. □

A stochastic process $\{X_n\}$ is *stationary* if

$$(X_1, \dots, X_n) \overset{\mathrm{D}}{=} (X_{k+1}, \dots, X_{k+n}),$$

for all n and all k. This means precisely that the distribution of the process is invariant under shift:

$$(X_1, X_2, \dots) \overset{\mathrm{D}}{=} (X_{k+1}, X_{k+2}, \dots).$$

One important fact is that studying stationary processes and studying measure-preserving systems are essentially the same. Let $(S^{\mathbf{N}}, \otimes^{\mathbf{N}}\mathcal{S}, \mu, T)$ be a measure-preserving system, where S is Polish and T is the shift operator. Let $X = (X_1, X_2, \dots)$ be a random variable on $(\Omega, \mathcal{F}, P)$ that is $(S^{\mathbf{N}}, \otimes^{\mathbf{N}}\mathcal{S})$-valued with distribution μ.[2] For any $B \in \otimes^n\mathcal{S}$, define $A = B \times S \times \cdots$, and we have

$$P\big((X_{k+1}, \dots, X_{k+n}) \in B\big) = P(\omega : T^k X \in A) = P(\omega : X \in A) = P\big((X_1, \dots, X_n) \in B\big).$$

This shows that $\{X_n\}$ is a stationary process.

Conversely, if $X = \{X_n\}$ is a stationary process on $(\Omega, \mathcal{F}, P)$, we have a probability measure $\mu = P \circ X^{-1}$ on $(S^{\mathbf{N}}, \otimes^{\mathbf{N}}\mathcal{S})$ such that

$$P\big((X_1, \dots, X_n) \in B\big) = \mu(A)$$

for any $B \in \otimes^n\mathcal{S}$ and $A = B \times S \times \cdots$. It follows that

$$\mu(A) = P\big((X_1, \dots, X_n) \in B\big) = P\big(T \circ (X_1, \dots, X_n) \in B\big) = \mu(T^{-1}A),$$

which shows that the shift T is measure-preserving on $(S^{\mathbf{N}}, \otimes^{\mathbf{N}}\mathcal{S}, \mu)$.[3]

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 $\mathrm{null}(I - U)$. 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.

[2] This is possible since S is Polish.

[3] The Polish assumption is not necessary in this direction.

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.

Set $f = X$ and $X = \mathrm{Id}$, we have

$$\frac{1}{n}\sum_{k=0}^{n-1} X(T^k \omega) \to \mathrm{E}_P(X \mid \mathcal{I}) \quad \text{a.s.}$$

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 $T\colon S \to S$ be a measurable mapping.

Often 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.

Denote the space of invariant measures by $\mathcal{P}^T(S)$. It is a convex subset of $\mathcal{P}(S)$. We know $T_*\colon \mathcal{P}(S) \to \mathcal{P}(S)$ is affine from Section 2.I. Therefore for $\mu, \nu \in \mathcal{P}^T(S)$, we have

$$T_*\big((1-\lambda)\mu + \lambda\nu\big) = (1-\lambda)T_*\mu + \lambda T_*\nu = (1-\lambda)\mu + \lambda\nu,$$

for any $0 < \lambda < 1$, which proves convexity.

If additionally T is assumed to be continuous in the topology of weak convergence, then by Proposition 8.33 we know $\mu_n \Rightarrow \mu$ implies $T_*\mu_n \Rightarrow T_*\mu$. If $\mu_n = T_*\mu_n$, we must then have $T_*\mu = \mu$. This shows that $\mathcal{P}^T(S)$ is (sequentially) closed in $\mathcal{P}(S)$.[4]

If furthermore we assume S to be compact, a nonempty set of invariant measures can be constructed.

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 8.26), $\mathcal{P}^T(S)$ 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 8.28.)

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. □

[4] We know $\mathcal{P}(S)$ can be metrized when S is separable, so no need for sequential.

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 $\mathcal{P}^T(S)$.

The following lemma precisely depicts how ergodic measures are "extreme" over the space of invariant measures. In particular, part (b) allows us to prove our theorem.

11.19 Lemma.

(a) Two distinct T-ergodic measures μ and ν must be mutually singular.

(b) If μ is ergodic and $\nu \ll \mu$, then $\mu = \nu$.

Proof. Both parts are simple applications 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 surely on the set E with $\mu(E) = 1$. It is immediate that $\nu(E) = 0 = \mu(E^{\mathrm{c}})$, which proves mutual singularity.

For the second part, rewrite (11.20) for any $A \in \mathcal{S}$,

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbf{1}_A(T^k x) \to \mu(A) \quad \mu\text{-a.s.},$$

Since $\nu \ll \mu$, we may integrate both sides with respect to $d\nu$, and obtain

$$\frac{1}{n}\sum_{k=0}^{n-1} \nu(T^{-k}A) \to \mu(A)$$

However, the left-hand side is just $\nu(A)$. Therefore $\mu = \nu$. □

The proof of part (b) is in fact easier without Birkhoff's theorem, by considering $f = \frac{d\nu}{d\mu}$, which has to be a μ-invariant function. Hence it is an a.e. constant function that integrates to 1, which implies $f = 1$ a.e.

Proof of Theorem 11.18. Suppose μ is ergodic. If μ is not an extreme point, then $\mu = (1-\lambda)\mu_1 + \lambda\mu_2$ for $0 < \lambda < 1$ and $\mu_1 \neq \mu_2$. Therefore $\mu_1 \ll \mu$ (and $\mu_2 \ll \mu$). Therefore by the above lemma, part (b), $\mu_1 = \mu$, which is a contradiction.

Suppose μ is an extreme point, but not ergodic. Then there exists a pair of two almost invariant sets $B, B^{\mathrm{c}} \in \mathcal{S}$ with $0 < \mu(B) < 1$. Therefore

$$\mu = \mu(B)\mu|_B + \mu(B^{\mathrm{c}})\mu|_{B^{\mathrm{c}}} \qquad \square$$

Suppose S is compact, then $\mathcal{P}^T(S)$ is a closed, and hence nonempty compact convex subset of $\mathcal{P}(S)$. Therefore if we are allowed the Krein–Milman theorem, then the closed convex hull of the set of T-ergodic measures are precisely $\mathcal{P}^T(S)$.

Chapter 12 Discrete-time Markov chains

12.A Markov properties

12.1 Simple Markov property. Let $G\colon \mathbf{\Omega} \to \mathbf{R}$ 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 $G_n\colon \mathbf{\Omega} \to \mathbf{R}$ 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

(Example 6.1.6 Durrett)

Let Q be recurrent, and has a stationary distribution π that is strictly positive at all states. Then the chain is irreducible if and only if it is ergodic.

Therefore if we start the irreducible positive recurrent chain from its unique stationary distribution π, then $X_0, X_1, \dots$ is a stationary sequence, and we may apply Birkhoff pointwise ergodic theorem. However, the next result shows that any arbitrary initial distributions will do.

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

Chapter 13 Continuous-time Markov chains

13.A Jump Markov chains, a primer

Let S be the countable state space of a Markov chain, with discrete transition probabilities given by the matrix M. This was the subject of our study in the last chapter, but we now wonder if we can convert the discrete-time process to a continuous-time process, while still satisfying the continuous version of Markov properties.

The answer, of course, is yes. In this section we will first define the process, and later on in the chapter we will justify that the process is indeed the *only* continuous-time Markov chain on a countable state space, and exists on a suitable probability space.

For $r(x) = 0$, $Lf(x) = 0$ For $r(x) > 0$, $L(x, y) = r(x)(M - I)(x, y)$

Let $Q_t \colon S \to S$ is now a square matrix with countable dimension

$\frac{d}{dt} Q_t = Q_t L = L Q_t$

13.1 Theorem. For each $x \in S$, the waiting time T_1 is exponentially distributed under $\mathbf{P}_x$, with some parameter $r(x) \geq 0$.

If $r(x) > 0$, then T_1 and X_{T_1} are independent. And by the case $r(x) = 0$ we mean $T_1 = \infty$ almost surely, or equivalently, the state x is an absorbing state.

In the literature, T_1 is the so-called exponential clock.

For the generator L of the semigroup Q_t,

It is important to keep in mind that for general Markov semigroups Q_t, there is no reason to believe L is a bounded operator. But the generator of a jump process is indeed bounded, and this allows us to conclude $Q_t = \exp(tL)$.

13.B The continuous-time semigroup 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.

From the transition kernel Q_t we may by abuse of notation define the Markov operator

Q_t that sends bounded measurable[1] functions to bounded measurable functions:

$$Q_t f(x) = \int_{y \in S} f(y)\, Q_t(x, dy).$$

$Q_0 = \mathrm{Id}$, , $Q_t(\mathbf{1}) = \mathbf{1}$, $f \geq 0$ implies $Q_t \geq 0$

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.

Notice that $\|Q_t f\| \leq \|f\|$ on the space of bounded Borel measurable function. To see this,

$$\begin{aligned} \|Q_t f(x)\| &= \sup_x \left| \int f(y)\, Q_t(x, dy) \right| \\ &\leq \sup_x \int |f(y)|\, Q_t(x, dy) \\ &\leq \sup_y |f(y)| = \|f\|. \end{aligned}$$

Therefore for each fixed t, Q_t is a contraction, and hence a bounded linear operator.

$Q_t f(x) = \mathbf{E}_x f(X_t)$, $\int Q_t f(x)\, d\mu(x) = \mathbf{E}_\mu f(X_t)$

positivity and mass preservation By the same proof as Jensen's inequality, we can show that given any convex function φ, for every $t \geq 0$ and every bounded measurable function f, $\varphi(Q_t f) \leq Q_t(\varphi \circ f)$

Define the dual operator Q_t^* by

$$\int Q_t f\, d\mu = \int f\, d(Q_t^* \mu)$$

for all bounded measurable f (or just simple functions, or C_b functions). We say μ is an *invariant/stationary measure* with respect to $\{Q_t\}$ if

$$\int Q_t f\, d\mu = \int f\, d\mu, \quad t \geq 0$$

for all bounded measurable f. This means precisely $Q_t^* \mu = \mu$, either viewed as measures or operators on bounded measurable functions. Also recall when the state space is discrete, $Q_t^* \mu$ were expresses as μQ_t, where μ is a row vector and Q_t is the t-th power of the transition matrix Q.

We know Q_t maps B_b into B_b. Consider Q_t as operators over richer function spaces.

A transition semigroup $\{Q_t\}$ is called a (C_0-)*Feller semigroup* if

[1] Boundedness is clear. Measurability comes from the standard simple function approximation argument.

(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 at all $x \in S$, $Q_t f(x) \to f(x)$.

It turns out that condition (b) may be replaced by the strong convergence $\|Q_t f - f\|_u \to 0$. In particular this implies that a Feller semigroup is a strongly continuous contraction semigroup on C_0, and hence we can adopt the nice results from Appendix E.

In particular, pick $f = \mathbf{1}_A$, we then have $Q_t \mathbf{1}_A(x) = Q_t(x, A) \to \mathbf{1}_A(x)$, which tells us precisely that $Q_t(x, dy) \Rightarrow \delta_x(dy)$ as $t \to 0$. This seems to be a natural and desirable condition for a Markov semigroup. Indeed almost all Markov processes we are interested in are Feller.

There are various different notions of Feller semigroups. Some authors consider the Banach space C_b instead of C_0, or even the space of uniformly continuous and bounded functions instead of C_0 (Da Prato). There are some problems with the choice C_b, and they are discussed in [BSW13, Section 1.1]

Fix x. Since Q_t is a contraction for any t, we have

$$\sup_{t \geq 0} |Q_{t+s} f(x) - Q_t f(x)| \leq |Q_s f(x) - f(x)| \to 0,$$

which shows that $t \mapsto Q_t f(x)$ is uniformly continuous over all $t \geq 0$. The same argument also shows that $t \mapsto Q_t f$ is uniformly continuous. Also since

$$|Q_{t+s} f(x) - Q_t f(y)| \leq |Q_s f(x) - f(y)| \leq |Q_s f(x) - f(x)| + |f(x) - f(y)|,$$

we have $(t, x) \mapsto Q_t f(x)$ is jointly continuous.

If $Q_t f \in D(L)$, then $\partial_t Q_t f = L(Q_t f)$.

Let $f \in D(L)$, then $Q_t f \in D(L)$, and $\partial_t Q_t f = Q_t(Lf) = L(Q_t f)$.

$Q_t f = f + \int_0^t L(Q_s f)\, ds = f + \int_0^t Q_s(Lf)\, ds$

13.2 Proposition. Suppose there exists some $x \in S$ such that $Q_t(x, \cdot) \Rightarrow \mu$, then μ is an invariant measure with respect to $\{Q_t\}$.

Proof. For any $f \in C_b$, by assumption

$$\lim_t Q_t f(x) = \lim_t \int f(y)\, Q_t(x, dy) = \int f(y)\, d\mu.$$

Fix any s, we then have

$$\int f(y)\, d\mu = \lim_{t+s} Q_{t+s} f(x) = \lim_t \int Q_s f(y)\, Q_t(x, dy) = \int Q_s f(y)\, d\mu. \qquad \square$$

Let μ be an invariant measure for $\{Q_t\}$ over bounded measurable functions. Then for all $1 \leq p < \infty$,

$$\int |Q_t f|^p\, d\mu \leq \int Q_t(|f|^p)\, d\mu = \int |f|^p\, d\mu$$

for bounded measurable f. Hence $\|Q_t f\|_p \leq \|f\|_p$ for all $1 \leq p \leq \infty$. (For a bounded measurable function f, $\sup|f|$ and $\operatorname{ess\,sup}|f|$ are the same.)

Now by density (Proposition 5.5), we can extend the contraction operator Q_t defined on bounded measurable functions to a contraction operator on the $L^p(\mu)$ spaces ($1 \leq p \leq \infty$).

13.3 Exercise. If Q_t is Feller, then Q_t defined on L^p given above is also a strongly continuous operator.

All the definitions and properties so far defined for bounded measurable functions extend naturally to L^p spaces by density. For example, if μ is an invariant measure on bounded measurable functions, then it may be taken just as an invariant measure on any L^p spaces.

We say the invariant measure μ is *reversible* with respect to Q_t if

$$\int f\, Q_t g\, d\mu = \int g\, Q_t f\, d\mu \quad \text{for all } t \geq 0 \text{ and } f, g \in L^2(\mu).$$

Equivalently,

$$\int f\, Lg\, d\mu = \int g\, Lf\, d\mu$$

$\langle Q_t f, g\rangle_\mu = \langle f, Q_t g\rangle_\mu$ and $\langle f, Lg\rangle_\mu = \langle g, Lf\rangle_\mu$

Q_t is a self-adjoint bounded operator on $L^2(\mu)$, while L is also a self-adjoint negative unbounded operator on $L^2(\mu)$ Theorem E.2

Dirichlet domain $D(\mathcal{E}) = \{f\colon \lim_{t\to 0} \int_S f(f - Q_t f)\, d\mu\}$ exists.

$D(L) \subseteq D(\mathcal{E}) \subseteq L^2(\mu)$

$\mathbf{1} \in D(L)$, with $L\mathbf{1} = 0$, and therefore $0 = \int (f)(L\mathbf{1})\, d\mu = \int Lf\, d\mu$ for any $f \in D(L)$.

Hille–Yosida theorem characterizes

13.4 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

$$\Gamma(f, g) = \tfrac{1}{2}\big[L(fg) - (Lf)g - f(Lg)\big].$$

let $L = \Delta$, then $\Gamma(f, g) = \nabla f \cdot \nabla g$

$(Q_t f)^2 \leq Q_t(f^2)$, and by taking limits, we get $2f(Lf)^2 \leq L(f^2)$.

symmetric Dirichlet form for reversible measure μ

$$\begin{aligned}
\mathcal{E}(f, g) &= \int \Gamma(f, g)\, d\mu \\
&= \frac{1}{2} \int L(fg)\, d\mu - \frac{1}{2}\left(\int (Lf)g\, d\mu + \int f(Lg)\, d\mu\right) \\
&= -\int f\, Lg\, d\mu,
\end{aligned}$$

where we used $\int f\, Lg\, d\mu = \int g\, Lf\, d\mu$ and $\int L(fg)\, d\mu = 0$. This is the integration by parts formula in the context of Markov semigroups.

In the abstract theory, one could start the formalism from the Carré du champ operators Γ, since it allows us to define a symmetric operator L. We may then extend it to be a self-adjoint operator $\widetilde{L}$ and then generate the semigroup $\exp(t\widetilde{L})$ (This is summarized in [BGL14, page 26] and discussed in detail in Chapter 3.)

It is clear that Γ is a symmetric bilinear form, and $\Gamma(f) \geq 0$. Hence $\mathcal{E}$ is bilinear and $\mathcal{E}(f) \geq 0$.

$$\frac{d}{dt}\operatorname{Var}_\mu(Q_t f) = -2\mathcal{E}(Q_t f, Q_t f)$$

Suppose μ is stationary and ergodic, then

$$\operatorname{Var}_\mu f = 2\int_0^\infty \mathcal{E}(Q_t f, Q_t f)\, dt.$$

13.5 Fact. $\operatorname{Cov}_\mu(f,g) = 2\int_0^\infty \mathcal{E}(Q_t f, Q_t g)\, dt = \int_0^\infty \mathcal{E}(f, Q_t g)\, dt$.

The first equality follows by

$$2\operatorname{Cov}_\mu(f,g) = \operatorname{Var}_\mu(f+g) - \operatorname{Var}_\mu f - \operatorname{Var}_\mu g$$

and bilinearity of $\mathcal{E}$. The second equality follows by replacing t by $t/2$, and using reversibility:

$$2\int_0^\infty \mathcal{E}(Q_{t/2} f, Q_{t/2} g)\, d(t/2) = \int_0^\infty \mathcal{E}(f, Q_t g)\, dt.$$

For a Markov process X_t with semigroup Q_t and generator L, the *Kolmogorov backward equation* is given by

$$\partial_t Q_t f = L Q_t f \quad \text{for } Q_t f \in D(L).$$

We already know that the condition $Q_t f \in D(L)$ is automatic for all $f \in D(L)$. However, we are allowed to consider functions f that are not necessarily continuous, as long as $Q_t f \in D(L)$.

If the terminal time T is fixed and given, we have

$$\partial_t Q_{T-t} f = -L Q_{T-t} f \quad \text{for } Q_{T-t} f \in D(L).$$

On the other hand, we have the *Kolmogorov forward equation*, or more commonly known as the *Fokker–Planck equation*

$$\partial_t Q_t f = Q_t L f \quad \text{for } f \in D(L).$$

Assume now the transition kernel $\{Q_t\}$ admits a density, i.e., we have

$$\frac{dQ_t(x,\cdot)}{dm} = q_t(x,\cdot)$$

for all $t > 0$ and $x \in S$, an open subset of $\mathbf{R}^n$. (Note that if $t = 0$, then the density does not exist since $\delta_x \not\ll m$.) Furthermore we assume the semigroup is Feller and $C_c^\infty \subseteq D(L)$. (We will see that this will be the case for diffusion processes.) Now the KBE becomes

$$\partial_t \int_S f(y)\, q_t(x,y)\, dy = L_x \int_S f(y)\, q_t(x,y)\, dy,$$

and the FPE becomes

$$\begin{aligned}\partial_t \int_S f(y)\, q_t(x,y)\, dy &= \int_S L_y f(y)\, q_t(x,y)\, dy \\ &= \int_S f(y)\, L_y^* q_t(x,y)\, dy\end{aligned}$$

Therefore in the weak sense, we have the following density form

$$\partial_t q_t(x,y) = L_x q_t(x,y)$$

for KBE, and

$$\partial_t q_t(x,y) = L_y^* q_t(x,y)$$

for FPE, where $t > 0$. If q_t is smooth enough, then the two equations can be understood in the classical sense.

It is important to understand these two equations in the context of differential equations. Note that the backward equation fixes the position y at some future time T, while the forward equation fixes the initial position x at time 0.

Consider a time-homogeneous Markov process with kernel $\{Q_t\}_{t\geq 0}$ and generator L. Fix a future time T, then KBE tells us that for any final position y, the density $\{q_t(x,y)\}_{0<t\leq T}$ satisfies

$$\begin{cases} \partial_t q_t(x,y) = L_x q_t(x,y), \\ q_t(x,y)\,dy \Rightarrow \delta_x(dy) \quad \text{as } t \to 0. \end{cases}$$

Alternatively, one can say that the backward density $\{\tilde{q}_t\} = \{q_{T-t}(x,y)\}_{0\leq t<T}$ satisfies

$$\begin{cases} \partial_t \tilde{q}_t(x,y) = -L_x \tilde{q}_t(x,y), \\ \tilde{q}_t(x,y)\,dy \Rightarrow \delta_x(dy) \quad \text{as } t \to T. \end{cases}$$

This form is also common in the literature. The KBE is called the "backward equation" because it fixes the terminal position y.

On the other hand, KFE/FPE tells us that for any initial position x, the forward density $\{q_t(x,\,\cdot\,)\}_{t>0}$ solves the PDE

$$\begin{cases} \partial_t u_t(y) = L^* u_t(y) \\ u_t(y)\,dy \Rightarrow \delta_x(dy) \quad \text{as } t \to 0. \end{cases}$$

But more importantly, KFE/FPE precisely characterizes the evolution of the law

$$\mu_t(x) = Q_t^* \mu(x) = q_t(x,y)\,dy\,d\mu(x)$$

of the underlying Markov process with initial distribution μ and kernel $\{Q_t\}_{t\geq 0}$, and hence the name "forward equation."

We remark that the weak convergence $q_t(x,y)\,dy \Rightarrow \delta_x(dy)$ above is just $Q_t(x,A) \to \mathbf{1}_A(x)$, which follows from the Feller property.

In the time-homogeneous case, the KBE and KFE have relatively concise statement. For non-homogeneous Markov processes, since we do no longer have Markov semigroups available, the KBE and KFE become slightly more confusing, and we give their statement below.

13.C The study of reversibility

The underlying Banach space is now chosen to be the Hilbert space $L^2(\mu)$, where μ is the invariant measure for the Markov chain.

13.6 Ergodic theorem for continuous-times jump Markov chains.

In the context of continuous reversible Markov processes, ergodicity has a second meaning and is related to the convergence of the semigroup operator. We say a Markov semigroup is ergodic if for every $f \in D(L)$,

$$Lf = 0 \implies f \text{ is a constant function.}$$

We first note that $Q_t f$ converges in L^2 to the projection of f to the subspace of all functions satisfying $Lf = 0$. This is an easy consequence of the spectral decomposition.

Therefore the definition of ergodicity tells us that $Q_t f$ converges to a constant c in L^2. Because c is the projection of f onto the kernel of L, we have $\langle f - c, 1\rangle_\mu = 0$, which implies that $c = \int f \, d\mu$.

$$Q_t f \to \int f \, d\mu \text{ in } L^2$$

Let the invariant measure $\mu = Q_t^* \mu$ have Lebesgue density w, then by the exact same calculation

$$\partial_t \int f(y)\, w(y)\, dy = \int f(y)\, L^* w(y)\, dy,$$

which implies that $0 = L^* w$. Note that L^* is the Lebesgue adjoint.

If μ is furthermore a reversible measure, with density w that is strictly positive and smooth enough, then we claim the second-order differential operator L given by

$$\begin{aligned} Lf &= \frac{1}{w} \sum_{i,j=1}^n \partial_i \big(w \, \tfrac{1}{2} a_{ij} \, \partial_j f \big) \\ &= \sum_{i,j=1}^n \tfrac{1}{2} a_{ij} \, \partial_{ij} f + \sum_{j=1}^n \frac{1}{2} \Big(\sum_{i=1}^n \partial_i \, a_{ij} + a_{ij} \, \partial_i (\log w) \Big) \partial_j f \end{aligned}$$

is symmetric for f defined on the dense domain $C_c^2 \subseteq L^2(\mu)$. This is an easy consequence of integration by parts. Conversely, since L is symmetric on $L^2(w\,dx)$, we conclude that $w\,dx$ is the invariant measure for the diffusion process

$$dX_t = \sigma(X_t)\, dB_t + b(X_t)\, dt$$

with $b_j(x) = \frac{1}{2} \big(\sum_{i=1}^n \partial_i \, a_{ij} + a_{ij} \, \partial_i (\log w) \big)$.

In the case $S = \mathbf{R}^n$ and the reversible measure μ admits a strictly positive and smooth density, the Dirichlet domain $D(\mathcal{E})$ is precisely the weighted Sobolev space $H^1(w\,dx)$. It is certainly a miracle we could obtain the same function space via two vastly different approaches.

Now take $w = e^{-V}$ and the matrix $\sigma = \sqrt{2} I_n$, then $b = -\nabla V$

13.D Spectral decomposition

[BGL14, Proposition 3.1.6]

For the positive self-adjoint operator $-L = \int_0^\infty \lambda \, d\Pi_\lambda$

$D(L)$ is the subspace of L^2 such that

$$\int (Lf)^2 \, d\mu = \int_0^\infty \lambda^2 \, d\langle \Pi_\lambda f, f \rangle < \infty,$$

while $D(\mathcal{E})$ is the subspace such that

$$\int_0^\infty \lambda \, d\langle \Pi_\lambda f, f\rangle < \infty,$$

L takes $L^2(\mu)$ to $D(L)$

When the spectrum is discrete, we have an easier characterization. Consider the simplest case $\mu = \gamma_n$. Recall the negative generator $-L = x \cdot \nabla - \Delta$ for the OU process has eigenvalues $0, 1, 2, \ldots$, and their respective eigenspaces are precisely the span of the Hermite polynomials $\mathrm{He}_0, \mathrm{He}_1, \mathrm{He}_2, \ldots$, which are orthogonal and span $L^2(\gamma)$. Write $f = \sum_{k=0}^\infty f_k$, where $f_k = \left\langle f, \frac{\mathrm{He}_k}{\sqrt{k!}} \right\rangle \frac{\mathrm{He}_k}{\sqrt{k!}}$, then

$$-Lf = \sum_{k=0}^n k f_k, \quad \mathcal{E}(f) = \sum_{k=0}^n k \|f_k\|^2, \quad Q_t f = \sum_{k=0}^\infty e^{-kt} f_k.$$

Chapter 14 Brownian motions

14.A Some sample path properties

14.1 Theorem. Almost surely, the sample paths of Brownian motions are locally α-Hölder continuous for $\alpha < 1/2$, but at no points for $\alpha > 1/2$.

14.2 Lévy's modulus of continuity. Almost surely

$$\limsup_{\delta \to 0^+} \sup_{0 \le t \le 1-\delta} \frac{|B_{t+\delta} - B_t|}{\sqrt{2\delta \log(1/\delta)}} = 1.$$

This means almost surely, $\{B_t : 0 \le t \le 1\}$ has modulus of continuity $C\sqrt{2\delta \log(1/\delta)}$ for small enough δ and some $C > 1$.

14.3 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_tl

(d) the process $B_1 - B_{1-t}$ has the same distribution as B_t on $[0, 1]$, called the *time reversal of Brownian motion*.

$\mathcal{F}_{0+} = \bigcap_{t>0} \mathcal{F}_t$
Let $\mathcal{T} = \bigcap_{s \ge 0} \sigma(B_t : t \ge s)$, the tail σ-field of the Brownian motion $\{B_t\}$.

14.4 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 $\widetilde{\mathcal{F}}_{0+} = \widetilde{\mathcal{F}}_0$.

14.5 Theorem. For any $A \in \mathcal{T}$, we have $\mathbf{P}_x(A) = 0$ or 1.

14.6 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.

14.B Markov properties

14.7 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

$$q_t(x,y) = \frac{1}{(2\pi t)^{d/2}} \exp\Big(-\frac{|y-x|^2}{2t}\Big).$$

$D(L) = \{f \in C^2(\mathbf{R}^d) : \Delta f \in C_0(\mathbf{R})\}$ and $L = \frac{1}{2}\Delta$.
For any $f \in C_b^2$, we have

$$\begin{aligned} Lf(x) &= \lim_{t\to\infty} \frac{\mathrm{E}_x f(B_t) - f(x)}{t} \\ &= \lim_{t\to\infty} \frac{\mathrm{E}_x\big[f'(x)(B_t - x) + \frac{1}{2}f''(x)(B_t-x)^2 + o\big(|B_t - x|^2\big)\big]}{t} \\ &= \lim_{t\to\infty} \frac{1}{2}f''(x) + o(t)/t = \frac{1}{2}f''(x). \end{aligned}$$

Using the multivariate Taylor's theorem, we will get one-half the Laplacian instead as the generator.

14.8 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

$$\mathbf{1}\{T<\infty\}(B_{T+t} - B_T) \quad \text{is a Brownian motion independent of } \mathcal{F}_{T+}.$$

infinitesimal peak into the future

14.9 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

14.10 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,

14.11 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

$$\mathbf{P}_0\Big(\max_{s\leq t} B_s \geq a, B_t \geq 2a - b\Big) = \mathbf{P}_0(B_t \geq 2a - b).$$

14.C A third return to random walks

$B_t^2 - t$ is a continuous martingale

$\exp(\lambda B_t - \frac{\lambda^2}{2}t)$ is a continuous martingale

14.12 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 $T_a = \inf\{t \geq 0 : B_t = a\}$

14.13 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

$$\mathbf{P}_0(B_T = -a) = \frac{b}{a+b}, \ \mathbf{P}_0(B_T = b) = \frac{a}{a+b}, \text{ and } \mathbf{E}_0 T = ab.$$

14.14 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.

14.15 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.

14.16 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.

Brownian motion is recurrent in 1D, neighborhood recurrent in 2D, and transient for dimensions ≥ 3. This matches the fact that the symmetric random walk is the degenerate Brownian motion.

14.D Introduction to Gaussian processes

We use $\|X\|_T$ to denote $\sup_{t\in T}|X_t|$.

14.17 Borell–TIS inequality. For a centered Gaussian process $\{X_t\}_{t\in T}$,

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 $\alpha < H$

14.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

$$P(T_{b_2} = t \,|\, T_{b_1} = s),$$

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 $\{B(t) - tB(1)\}_{0\leq t\leq 1}$.

14.18 Proposition. A Brownian bridge has the distribution of a Brownian motion conditioning on hitting 0 at time 1. To be precise,

$$\{B(t) - tB(1)\}_{0\leq t\leq 1} \overset{\mathrm{D}}{=} \{B(t) \,|\, B(1) = 0\}_{0\leq t\leq 1}.$$

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 $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} = 1$.

14.19 Proposition. The Brownian bridge is a continuous Gaussian process with zero mean and covariance function

$$\mathrm{E}(X_s X_t) = s(1-t) \quad \text{for all } 0 \le s \le t \le 1.$$

14.20 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

$$\sigma = \sigma_T = \sup\{s \le T : B_s = 0\},$$

and the *first passage time* after time T to be

$$\tau = \tau_T = \inf\{s \ge T : B_s = 0\}.$$

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

14.F Generalization of Brownian motions

Gaussian white noise

isonormal Gaussian process

Brownian motion B_t is an $L^2[0,\infty)$-Gaussian process, where $B_t = W(\mathbf{1}_{[0,T]})$

Wiener integral $\int_0^T f(t)\,dB_t = W(\mathbf{1}_{[0,T]} f)$, distributed according to $N\big(0, \int_0^T f(t)^2\,dt\big)$.

14.21 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 $t \ge 0$, 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 15 Stochastic calculus

15.A Continuous filtration and martingales

The *predictable/previsible σ-field* $\mathcal{P}$ is generated by left-continuous and adapted processes, and a stochastic process is said to be predictable/previsible if it is $\mathcal{P}$-measurable.

The *optional σ-field* $\mathcal{O}$ is generated by all right-continuous and adapted processes, and similarly a stochastic process is optional if it is $\mathcal{O}$-measurable.

It turns out $\mathcal{P}$ is generated by continuous and adapted processes as well. Therefore that $\mathcal{P} \subseteq \mathcal{O}$, which means every predictable process is optional.

A process is *progressively measurable* if $(s, \omega) \mapsto X_s(\omega)$ is $\mathcal{B}[0,t] \otimes \Omega$-measurable for all $t > 0$. Note that 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 $\langle M \rangle = 0$.

15.1 Proposition [LeG16, Proposition 3.4]. A left-continuous/right-continuous adapted process is progressively measurable. Therefore an optional process must be progressively measurable. (Predictable $\implies$ Optional $\implies$ Progressively Measurable)

Given a right-continuous adapted process X_t taking values in a separable metric space (S, ρ), and let U be an open set in this space. The first hitting time $\inf\{t > 0 : X_t \in U\}$ and the first entrance time $\inf\{t \geq 0 : X_t \in U\}$ are both stopping times of $\mathcal{F}_{t+}$. If the sample path of X_t is continuous, then $\inf\{t > 0 : X_t \in F\}$ and $\inf\{t \geq 0 : X_t \in F\}$ are both stopping times of $\mathcal{F}_t$.

(In fact Choquet's capacity theorem, one can show that if the filtration satisfies the usual conditions, then the entrance time to any Borel set is a stopping time.)

15.2 Doob's maximal inequality.

Let us give mention an application to reversible Markov chains, known as Rota's Lemma. For a symmetric Markov semigroup $\{Q_t\}$ with respect to μ, for any $1 < p < \infty$, there is a constant C_p such that for any measurable $f \colon S \to \mathbf{R}$, we have

$$\left\| \sup_{t \geq 0} Q_t f \right\|_p \leq C_p \|f\|_p.$$

To show this, fix $T > 0$ and look at $M_t = Q_{T-t} f(X_t) = \mathrm{E}_\mu[f(X_T) \mid X_t]$ and use Doob's maximal inequality. One can check out [BGL14, Lemma 1.6.2] for details, but it should not be hard to recover the result.

15.3 Doob's L^p inequality.

A process $\{X_t\}$ is said to be of class D if $\{X_\tau : \tau \text{ is a finite stopping time}\}$ is uniformly integrable.

15.4 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.)

15.5 Undefined Theorem Name.

A stopping time τ is predictable if there is an increasing sequence of stopping times τ_n, not equal to τ, such that $\tau_n \to \tau$. We say a stopping time σ is totally inaccessible if $P(\sigma = \tau < \infty) = 0$ for any predictable stopping time τ.

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.

Given any f bounded measurable and a Brownian motion on $[0,T]$, $Q_{T-t}f(B_t)$ is a martingale. This is an easy exercise using the Markov property and the tower property.

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

$$\sum_{j=1}^{n(p)} (M_{t_j} - M_{t_{j-1}})^2 \to \langle M \rangle_T \text{ in probability.}$$

For Brownian motion, the convergence to $\langle B \rangle_T = T$ further holds in L^2.

First we identify the expectation (which minimizes the L^2 distance):

$$\begin{aligned} \mathrm{E} \sum_{j=1}^{n(p)} (B_{t_j} - B_{t_{j-1}})^2 &= \sum_{j=1}^{n(p)} \mathrm{E}B_{t_j}^2 - \mathrm{E}B_{t_{j-1}}^2 \\ &= \sum_{j=1}^{n(p)} t_j - t_{j-1} = T. \end{aligned}$$

Now

$$\begin{aligned}\mathrm{E}\Big(\sum_{j=1}^{n(p)}(B_{t_j}-B_{t_{j-1}})^2-T\Big)^2 &= \sum_{j=1}^{n(p)}\mathrm{Var}\big((B_{t_j}-B_{t_{j-1}})^2\big)\\ &= \sum_{j=1}^{n(p)}(t_j-t_{j-1})^2\,\mathrm{E}[B_1^2-1]^2\\ &\le \|p\|\cdot T\cdot 2\to 0,\end{aligned}$$

which proves the L^2 convergence.

15.6 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 $p=\{t_0,\dots,t_{n(p)}\}$ 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 gives a Hilbert space structure on the space of continuous local martingale

For $X=M+A$ and $X'=M'+A'$, define $\langle X,X'\rangle=\langle M,M'\rangle$

A process X_t is a *continuous semimartingale* the sum of a continuous local martingale M_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.

15.7 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 $\sup_t|M_t| \le Y$ for some $Y \in L^1$ is a uniformly integrable (true) martingale.

15.8 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.

We say a process is a *Gaussian martingale* if it is both a Gaussian process and a continuous true martingale M. One can show that $\langle M\rangle_t$ is the deterministic function $\mathrm{E}(M_t^2)$. We need to verify $M_t^2 - \mathrm{E}(M_t^2)$ is a continuous martingale.

We know $(M_t - M_s)^2$ and $\mathcal{F}_s$ are uncorrelated. Since the process is Gaussian, this implies $(M_t - M_s)^2$ and $\mathcal{F}_s$ are in fact independent. Now check

$$\mathrm{E}(M_t^2 - M_s^2 \mid \mathcal{F}_s) = \mathrm{E}\big[(M_t - M_s)^2 \;\big|\; \mathcal{F}_s\big] = \mathrm{E}(M_t - M_s)^2 = \mathrm{E}(M_t^2) - \mathrm{E}(M_s^2),$$

and rearrangement proves our claim.

martingale problem

determine the generator of a Markov semigroup on C_0

15.9 Theorem. Let X_t be a Markov process adapted to $\mathcal{F}_t$, with semigroup Q_t and generator L. For $f, g \in C_0$, the following are equivalent:

(a) $f \in D(L)$ and $Lf = g$.

(b) Under $\mathbf{P}_x$, $f(X_t) - \int_0^t g(X_s)\,ds$ is an $\mathcal{F}_t$-martingale.

15.B Construction of stochastic integrals

We use $\mathbf{H}^2$ for the space of square integrable continuous local martingales starting at 0. Given $M \in \mathbf{H}^2$, we use $L^2(M) = L^2\big(\Omega \times [0,\infty), dP\,d\langle M\rangle\big)$[1] for the space of progressive process H such that

$$\int_0^t H_s^2\,d\langle M\rangle_s < \infty$$

15.B.1 The Brownian case

agrees with the Wiener integral

[1] the σ-field should be $\mathcal{B}[0,\infty) \otimes \mathcal{F}_\infty$

15.B.2 The L^2 martingale case

Itô's isometry $\|H \bullet M\|_{\mathbf{H}^2} = \|H\|_{L^2(M)}$

$$\mathrm{E}\bigg(\int H_s \, dM_s\bigg)^2 = \mathrm{E}\int H_s^2 \, d\langle M\rangle_s$$

15.10 Itô–Döblin formula. For n continuous semimartingales $X^1, \dots, X^n$ and $F \in C^2(\mathbf{R}^n)$, we have for all $t \geq 0$, it holds that

$$F(X_t^1, \dots, X_t^n) = F(X_0^1, \dots, X_0^n) + \sum_{j=1}^n \int_0^t \frac{\partial F}{\partial x^j}(X_s^1, \dots, X_s^n) \, dX_s^j + \frac{1}{2}\sum_{j,k=1}^n \int_0^t \frac{\partial^2 F}{\partial x^j \partial x^k}(X_s^1, \dots, X_s^n) \, d\langle X^j, X^k\rangle_s.$$

The first-order term is $\nabla F \cdot dX$, while the second-order term is the sum of all entries in the Hessian matrix, weighted by the covariation between the coordinates. This can be expressed as $\mathrm{tr}(F \cdot d\langle X, X\rangle)$.

When X is a continuous local martingale, if $\sum_{j,k=1}^n \frac{\partial^2 F}{\partial x^j \partial x^k}(X_s^1, \dots, X_s^n) = 0$, then $F(X_t^1, \dots, X_t^n)$ is a continuous local martingale.

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

Also check out [DaP14, Section 7.2].

If we have m coordinates $X^1, X^2, \dots, X^m$ that are finite variations processes, then it is already sufficient take F to be C^1 in those m coordinates (and C^2 in the remaining coordinates). In particular, we get the time-inhomogeneous Itô's formula: for $F \in C^{1,2}(\mathbf{R}^+, \mathbf{R}^n)$, we have

$$F(t, X_t) = F(0, X_0) + \int_0^t \partial_t F(s, X_s) \, ds + \sum_{j=1}^n \int_0^t \partial_j F(s, X_s) \cdot dX_s + \frac{1}{2}\sum_{j,k=1}^n \int_0^t \partial_j \partial_k F(s, X_s) \, d\langle X^j, X^k\rangle_s,$$

where ∂_t means the time derivative and ∂_j means the derivative with respect to the jth spatial coordinate. Check out [RY99, Section 4.3].

When $X_t = B_t$, the standard Brownian motion, if u satisfies

$$\partial_t u(t,x) + \frac{1}{2}\Delta_x u(t,x) = 0,$$

then $u(t, B_t)$ is a continuous local martingale. As we have discussed before, $u(t,x) = q_{T-t}(x, B_T)\ Q_{T-t}(x, B_T)$ satisfies the backward heat equation

This allows us to give an alternative proof that $u(t, B_t)$ is a continuous martingale on $[0, T]$. We will see a few times that for processes that naturally arises in stochastic calculus, its martingale and Markovian properties can be proved alternatively by Itô's formula.

For 1-dimensional Gaussian martingales, recall $\langle M\rangle_t = \mathrm{E}(M_t^2)$. If $d\langle M\rangle_t = h(t)\, dt$ and u satisfies $\partial_t u + \frac{1}{2}h(t)\Delta_x u = 0$, then $u(t, M_t)$ is a continuous local martingale.

15.11 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

$$d(X_t Y_t) = X_t \, dY_t + Y_t \, dX_t + d\langle X, Y \rangle_t$$

$dX_t = \sigma(t, X_t) \, dB_t + b(t, X_t) \, dt$
time-change

15.12 Dambis–Dubins–Schwarz. [LeG16, Theorem 5.13] Given a continuous local martingale M such that $\langle M \rangle_\infty = \infty$ a.s., it induces a Brownian motion $\{\beta_s\}$ adapted to a different filtration $\mathcal{F}_{\tau_s}$ such that a.s., for all $s \geq 0$, $\beta_s = M_{\tau_s}$, where

$$\tau_s = \inf\{t \geq 0 : \langle M \rangle_t \geq s\}.$$

More concisely, we have for $t \geq 0$, $\beta_{\langle M \rangle_t} = M_t$.

The τ_s is an inverse to s with respect to the quadratic variation of M. We have already seen multiple times a stopping time of this form is useful to proving results. We stress that whether one define τ_s by $\langle M \rangle_t \geq s$ or $\langle M \rangle_t > s$ does not affect the result, basically because $\langle M \rangle$ is continuous.

In particular, the above result applies directly to Gaussian martingales: any Gaussian martingales can be written as the time-changed BM $\beta_{\mathrm{E} M_t^2}$.

15.13 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

$$Y = \mathrm{E} Y + \int_0^\infty H_s \, dB_s.$$

It follows that for a (not necessarily continuous) true martingale M bounded in L^2, there exists a unique progressive process $H \in L^2(B)$ and some real constant C such that

$$M_t = C + \int_0^t H_s \, dB_s.$$

The same claim still holds if M is a (necessarily) continuous local martingale that is $L^2_{\mathrm{loc}}(B)$. The results above hold on the finite interval $[0, T]$ by the same reasoning.

In general, the exact formula for H may be expressed using Malliavin calculus. For $Y \in \mathbf{D}^{1,2}(L^2[0, T])$, a Sobolev-type regularity condition to be discussed later, we have

$$Y = \mathrm{E} Y + \int_0^T \mathrm{E}(D_t Y \mid \mathcal{F}_t) \, dB_t.$$

However, in the special case where $Y = f(B_1)$, the explicit formula for H_s is quite simple.

15.14 Proposition. For $f \in C_b^1$, we have

$$f(B_1) = \mathrm{E}_0 f(B_1) + \int_0^1 Q_{1-s} \nabla f(B_s)\, dB_s,$$

where Q is the standard Brownian semigroup.

Proof. As usual we prove the 1d case. Note $Q_{1-s} f'(B_s) = \mathrm{E}_{B_s} f'(\beta_{1-s})$, where $\beta_\cdot$ is a Brownian motion independent of B_s. Therefore we have to keep track the time and the starting position along the way.

Define $u(t,x) = \mathrm{E}_x f(B_{1-t}) = Q_{1-t} f(x)$. We have previously argued that $u(t, B_t)$ is a continuous martingale on $[0,1]$ by Itô's formula and the fact that u satisfies the backward heat equation, but did not realize $u(1, B_1) = f(B_1)$ and $u(0, B_0) = \mathrm{E}_0 f(B_1)$. Now Itô's formula lets us to write

$$u(1,B_1) = u(0,B_0) + \left(\int_0^1 \partial_t u(s,B_s) + \tfrac{1}{2} \partial_{xx} u(s,B_s)\, ds \right) + \int_0^1 \partial_x u(s,B_s)\, dB_s.$$

By construction u satisfies the backward heat equation, and therefore the above equation is simplified to

$$f(B_1) = \mathrm{E}_0 f(B_1) + \int_0^1 \frac{d}{dx} Q_{1-s} f(B_s)\, ds,$$

and we can exchange Q and $\frac{d}{dx}$ because f is nice enough. □

Given a Brownian motion B_t, its completed filtration $\mathcal{F}_t$ is automatically right-continuous. (Therefore with respect to this continuous filtration, $\mathcal{P} = \mathcal{O}$.)

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

15.15 Girsanov's theorem.

removes the drift in SDE

sufficient condition for the stochastic exponential to be a (uniformly integrable) martingale

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 $\mathrm{E}\mathcal{E}(\lambda M)_t = 1$.

$\mathcal{E}(\lambda M)_t = \exp(\lambda M_0) + \lambda \int_0^t \mathcal{E}(\lambda M_s)\, dM_s$

It is unique solution to the SDE

$$dZ_t = \lambda Z_t\, dM_t, \text{ where } M_0 = 0.$$

complex is reserved for use the Fourier inversion theorem

$$\mathcal{E}(X)_t\mathcal{E}(Y)_t = \mathcal{E}(X + Y + \langle X, Y\rangle)_t$$

For two continuous local martingales M and N, $\mathcal{E}(M)\mathcal{E}(N)$ is a continuous local martingale if $\langle M, N\rangle = 0$.

15.16 Theorem. For a continuous local martingale D that take strictly positive values, it has a *stochastic logarithm* L, in the sense that

$$D_t = \mathcal{E}(L)_t = \exp\Big(L_t - \frac{1}{2}\langle L\rangle_t\Big).$$

An explicit formula for L is given by

$$L_t = \log D_0 + \int_0^t \frac{1}{D_s}\, ds.$$

15.17 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 an m-dimensional 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

Solution to SDE is a strong Markov process that is Feller, with $C_c^2 \subseteq D(L)$

Given $\{B_t\}$ on some given probability space $(\Omega, \mathcal{F}, P)$, $\{X_t\}$ is called a *strong solution* to the SDE if it is adapted to the minimal completed filtration of B. We say the strong solutions are *pathwise unique* if any two strong solutions X and X' with $X_0 = X_0'$ a.s. must be indistinguishable.

A *weak solution* consists of three parts, the probability space $(\Omega, \mathcal{F}, P)$, a complete and right-continuous filtration $\{\mathcal{F}_t\}$, and (X_t, B_t), where $\{X_t\}$ and $\{B_t\}$ both needs to adapt to the filtration $\mathcal{F}_t$. We say the weak solutions are *weakly unique* if for two weak solutions $(\Omega, \mathcal{F}, \mathcal{F}_t, P, X_t, B_t)$ and $(\Omega', \mathcal{F}', \mathcal{F}_t', P', X_t', B_t')$ with the same initial distribution ($X_0 \overset{\mathrm{D}}{=} X_0'$), we have $\{X_t\} \overset{\mathrm{D}}{=} \{X_t'\}$.

15.18 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

$$\langle X^j, X^k\rangle_t = \delta_{jk} t \quad \text{for all components } j \text{ and } k.$$

15.19 Yamada–Watanabe.

15.C Examples of diffusion processes

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

$$\log X_t - \log X_0 = \sigma B_t + \left(\mu - \frac{\sigma^2}{2}\right)t, \tag{15.20}$$

which gives the candidate solution

$$X_t = X_0 \cdot \exp\left(\sigma B_t + \left(\mu - \frac{\sigma^2}{2}\right)t\right). \tag{15.21}$$

There is one issue. We need to show that $X_t > 0$ must holds for all $t \geq 0$, so that $\log X_t$ makes sense for all $t \geq 0$. One way is to employ the "localization trick." Define τ_0 to be the hitting time $\inf\{t \geq 0 : X_t = 0\}$. Suppose $P(\tau_0 > \infty) > 0$, then conditioning on this event, the left-hand side of (15.20) converges to $-\infty$ as $t \to \tau_0^-$, while the right-hand side converges to $\sigma B_\tau + (\mu - \frac{\sigma^2}{2})\tau$, an a.s. finite value. Therefore $P(\tau_0 = \infty) = 1$, and we can conclude that (15.21) is the one and only solution (without using Lipschitz existence and uniqueness).

Of course one may also directly appeal to Lipschitz existence and uniqueness. First use Itô–Döblin formula to verify that dX_t is indeed $\sigma X_t\, dB_t + \mu X_t\, dt$. Since we have one pathwise solution, this must be the unique solution.

One should easily check when $X_0 < 0$ or $X_0 = 0$, (15.21) remains the solution.

Ornstein–Uhlenbeck process is the solution to the classical Langevin equation

$$dX_t = \sigma dB_t - \lambda X_t\, dt \tag{15.22}$$

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.

$Q_t f(x) = \mathrm{E}_x f(X_t) = \mathrm{E} f(e^{-t}x + \sqrt{1-e^{-2t}}Z)$ for $Z \sim N(0, I_d)$

$Lf(x) = \Delta f(x) - x \cdot \nabla f(x)$

$\Gamma(f,g) = \frac{1}{2}\{\Delta(fg) - x \cdot \nabla(fg) - (\Delta f + x \cdot \nabla f)g - (\Delta g + x \cdot \nabla g)f\} = \nabla f \cdot \nabla g$

$\mathcal{E}(f,g) = \mathrm{E}_\gamma(\nabla f \cdot \nabla g)$

In $L^2(\gamma)$, $D(L) = \{f \in C^1 : \text{ the weak } Lf \in L^2(\gamma)\}$.

Now we perturb (15.22) and consider the overdamped Langevin's equation

$$dX_t = \sigma\, dB_t - [\lambda X_t + \nabla U(X_t)]dt,$$

which has an additional gradient flow drift term. Therefore compared to the classical Langevin, the overdamped process is more attracted to where $U(x)$ is small.

We focus on the only useful case $\sigma = \sqrt{2}$:

$$dX_t = \sqrt{2}\, dB_t - [\lambda X_t + \nabla U(X_t)]dt$$

The semigroup Q_t associated with this process has a unique reversible measure given by

$$d\pi(x) = Z^{-1} e^{-U(x)}\, d\gamma(x),$$

where γ is the Gaussian measure with variance $1/\lambda$, and Z is the normalizing constant for π to be a probability measure. In particular, note that when $\lambda = 1$, then γ is the standard Gaussian measure, and when $\lambda = 0$, then γ should just be the Lebesgue measure. This is not surprising, because when $U = 0$, $X_t = B_t$. The property of the Laplacian tells us that the generator $L = \frac{1}{2}\Delta$ of B_t satisfies

$$\langle Lf, g\rangle_{L^2(m)} = \langle f, Lg\rangle_{L^2(m)}$$

for $f, g \in C_c^2$.

When the potential energy U is a density function, we may sample the Gibbs measure

$$d\pi = \frac{e^{-U(x)}}{Z}\, dm$$

from the SDE

$$dX_t = \sqrt{2}\, dB_t - \nabla U(X_t)dt.$$

[DaP14, Section 12.6.5]

When $\lambda = 1$ $Lf(x) = \Delta f(x) - \big(x + \nabla U(x)\big) \cdot \nabla f(x)$

When $\lambda = 0$, $Lf = \Delta f - \nabla U \cdot \nabla f$

Linear equations

$$dX_t = CX_t\, dB_t + DX_t\, dt$$

n-dimensional squared Bessel processes

$$dX_t = 2\sqrt{X_t}\, dB_t + n\, dt$$

We now give the long-delayed proof Harris' inequality in the Gaussian case with positively correlated coordinates.

15.23 Pitt's theorem. Let $X \sim N(0, \Sigma)$ in $\mathbf{R}^d$, with all entries of Σ being nonnegative. Suppose $f, g\colon \mathbf{R}^d \to \mathbf{R}$ are increasing in each coordinate, then

$$\operatorname{Cov}\big(f(X), g(X)\big) \geq 0.$$

(Provided that $f(X), g(X)$ are nonnegative or L^2, of course.)

Proof. We follow the hints given in [Han14, Problem 2.11], and prove the special case where f, g have bounded partial derivatives (C_b^1). This allows us to interchange ∇_x and E in (15.24).

Write $X = \Sigma^{1/2} Z$, where $Z \sim N(0, I_d)$. This allows us to write

$$\operatorname{Cov}\big(f(X), g(X)\big) = \operatorname{Cov}_\gamma(f \circ \Sigma^{1/2}, g \circ \Sigma^{1/2}),$$

where γ is the standard Gaussian measure on $\mathbf{R}^d$. Using the covariance identity with respect to the OU process, we can write the above further as

$$\begin{aligned} &\int_0^\infty \mathcal{E}\Big(f \circ \Sigma^{1/2}, Q_t(g \circ \Sigma^{1/2})\Big)\, dt \\ &= \int_0^\infty \Big\langle \nabla(f \circ \Sigma^{1/2}), \nabla_x \operatorname{E}\big[g \circ \Sigma^{1/2}(e^{-t}x + \sqrt{1 - e^{-2t}}\xi)\big] \Big\rangle_\gamma \, dt, \end{aligned} \tag{15.24}$$

where $\xi \sim N(0, I_d)$.

Now $\nabla(f \circ \Sigma^{1/2}) = \nabla f(\Sigma^{1/2}) \left(\Sigma^{1/2}\right)^{\mathrm{T}}$, and

$$\nabla_x \big[g \circ \Sigma^{1/2}\big(e^{-t}x + \sqrt{1 - e^{-2t}}\xi\big)\big] = \nabla g(\star) \left(\Sigma^{1/2}\right)^{\mathrm{T}} e^{-t}$$

for

$$\star = \Sigma^{1/2}\big(e^{-t}x + \sqrt{1 - e^{-2t}}\xi\big).$$

Thus the integrand of (15.24) is equal to

$$\Big\langle \nabla f(\Sigma^{1/2}) \left(\Sigma^{1/2}\right)^{\mathrm{T}}, \left(\Sigma^{1/2}\right)^{\mathrm{T}} e^{-t} \operatorname{E}[\nabla g(\star)] \Big\rangle_\gamma = \Big\langle \Sigma\, \nabla f(\Sigma^{1/2}), e^{-t} \operatorname{E}[\nabla g(\star)] \Big\rangle_\gamma,$$

where the expectation is with respect to the ξ in "$\star$". Since $e^{-t} > 0$, Σ has nonnegative entries, and the two gradients are always nonnegative, the preceding $\langle \cdot, \cdot \rangle_\mu$ must be nonnegative, and hence (15.24) must be nonnegative, finishing the proof in the special case.

By a mollifier argument we can reduce to only assuming continuity on f and g. We can further approximate bounded increasing functions by bounded continuous increasing functions. These arguments can be found in the original [Pit82], which completes the proof. □

There is also a reverse-time SDE.

The Laplace operator is a symmetric operator with respect to the Lebesgue measure

The generator of Langevin diffusion $\Delta - \nabla \cdot \mathrm{Id}$ is symmetric with respect to the Gaussian measure

The generator of overdamped Langevin diffusion $\Delta - \nabla U \cdot \nabla$ is symmetric with respect to normalized $e^{-U}\, dm$.

15.D Applications to partial differential equations

The THSDE

$$dX_t = \sigma(X_t)\,dB_t + b(X_t)\,dt,$$

is a time-homogeneous strong Markov process

For general SDE

$$dX_t = \sigma(t, X_t)\,dB_t + b(t, X_t)\,dt$$

is a time-inhomogeneous Markov process on the original probability space, but we may consider $\{(t, X_t)\}_{t\geq 0}$ can be reduced to THSDE

$$Lf = \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}\partial_i\partial_j f + \sum_{k=1}^{n} b_k\partial_k f,$$

where a_{ij} is the (i,j)-th entry of the matrix function $a = \sigma\sigma^{\mathrm{T}}$.

Let L^* be the (Lebesgue) adjoint on the Hilbert space $L^2(m)$. This means that $\langle Lf, g\rangle = \langle f, L^*g\rangle$ for $f \in D(L)$ and $g \in D(L^*)$.

$$L^*f = \frac{1}{2}\sum_{i,j=1}^{n} \partial_i\partial_j(a_{ij}f) - \sum_{k=1}^{n} \partial_k(b_k f).$$

$\{X_t\}$ is a Feller process, and $C_c^2 \subseteq D(L)$ (Oksendal Le Gall), if we consider L to be on the Banach space C_0 or L^p $(1 \leq p \leq \infty)$

By Itô–Döblin formula, we have the following representation for $f(X_t)$.

15.25 Theorem. For $\{X_t\}$ that solves the SDE, we have for any $f \in C_c^2$ that

$$f(X_t) = f(X_0) + \int_0^t \sum_{i,j=1}^{n} \partial_i f(X_s)\sigma_{ij}(s, X_s)\,dB_s^j + \int_0^t Lf(X_s)\,ds,$$

where the second term on the right-hand side is a local martingale. It follows that

$$\mathrm{E}f(X_t) = \mathrm{E}f(X_0) + \mathrm{E}\int_0^t Lf(X_s)\,ds, \tag{15.26}$$

provided that the last term is indeed integrable.

Note that (15.26) precisely restates

$$Q_t f(x) = f(x) + \int_0^t Q_s Lf\,ds,$$

without developing the theory of continuous Markov processes and proving that Q_t is Feller. This is precisely how some textbooks chose to present the material, but we still find the Markovian interpretation of (15.26) most natural. In a moment we will see that the probabilistic solutions to various PDEs can also be developed in the Markovian way or using the martingale–Itô approach.

Hörmander's condition for the smoothness of transition density and hypoellipticity

If $a = \sigma\sigma^{\mathrm{T}}$ is uniformly elliptic

Consider the standard Langevin diffusion, which would give us $a(t,x) \equiv 2I_n$ and $b(t,x) = -x$. Therefore

$$L^* f = \Delta f + \operatorname{div}(x \cdot f),$$

and the KFE for standard Langevin diffusion becomes

$$\partial_t q_t(y) = \Delta_y q_t(y) + \operatorname{div}_y\big(y \cdot q_t(y)\big),$$

where q_t is the density of the diffusion X at time t, provided that $X_0 \sim q_0$.

Because the Gaussian measure γ and the Lebesgue measure are equivalent, the KFE $\partial_t q_t(y) = L_y^* q_t(y)$ remains true if we take q_t to be the density of the law of X_t with respect to γ. However, the adjoint L^* is now an unbounded operator on $L^2(\gamma)$.

This provides useful insight because the stationary measure of Langevin diffusion is the Gaussian measure. Consider the derivative of the relative entropy $D(X_t\|\gamma_n)$ with respect to time:

$$\begin{aligned}
\frac{d}{dt} D(X_t\|\gamma) &= \int \partial_t (q_t \log q_t)\, d\gamma \\
&= \int (\partial_t q_t) \log q_t + \partial_t q_t \, d\gamma \\
&= \langle 1 + \log q_t, L^* q_t \rangle_\gamma \\
&= -\int \nabla(1 + \log q_t) \cdot \nabla q_t \, d\gamma \\
&= -\int \frac{|\nabla q_t|^2}{q_t} \, d\gamma = -I(X_t\|\gamma).
\end{aligned}$$

exponential decay of Fisher information
de Brujin's identity For invariant measure μ
$\frac{d}{dt} \operatorname{Ent}_\mu f \le -I_\mu f$
$I(X_t\|\gamma)$
We can generalize Fisher information to general diffusion processes
[Dur96, Chapter 4] Parabolic and elliptic PDEs
and suppose for each x the SDE has a weak solution.
Let $f \in C_c^2$, the function

$$u(t,x) = Q_t f(x) = \mathrm{E}_x f(X_t).$$

solves the Cauchy problem to the parabolic PDE

$$\begin{cases} \partial_t u = Lu & \text{for } t > 0 \\ u(0,x) = f(x). \end{cases}$$

This is a direct consequence of the KBE. The result is very general because it gives a formal solution for any second-order elliptic operators, C_c^2 is clearly too restrictive as an initial value condition.

Fortunately for Brownian motion we can do much better. In that case

$$Q_t f(x) = \int \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{|y-x|^2}{2t}\right) f(y)\, dy.$$

As a convolution of a bounded measurable f with a C^∞ function, $Q_t f \in C^\infty$ as well. If we assume $f \in C_0$, then all derivatives of $Q_t f$ are also in C_0. Therefore $Q_t f \in D(L)$, and the result follows by KBE.

15.27 Fact. It turns out, by explicit computation of $Q_t f$, we can plug in and verify that for any $f \in C_b$, $u(t,x) = Q_t f$ is a solution to the PDE.

The solution is very natural from the Markovian perspective, but as previously mentioned we may also derive it from the martingale–Itô approach. The construction of the martingale is not the most intuitive.

What if the PDE is nonhomogeneous? What if the PDE has an additional potential term vu?

If the parabolic PDE is given as

$$\begin{cases} \partial_t u = Lu + g(t,x), \\ u(0,x) = f(x) \end{cases}$$

instead, then

$$u(t,x) = \mathrm{E}_x f(X_t) + \mathrm{E}_x \int_0^t g(t-s, X_s)\, ds$$

The term $\int_0^t g(t-s, X_s)\, ds$ accounts for the cumulative effect of $g(\,\cdot\,, X_s)$ from time s up to the final time t. To justify this rigorously, we have to show that

$$(\partial_t - L)\, \mathrm{E}_x \int_0^t g(t-s, X_s)\, ds = g(t,x).$$

Replacing $t-s$ by r, we have

$$\begin{aligned} (\partial_t - L) \int_0^t g(t-s, X_s)\, ds &= (\partial_t - L) \int_0^t g(r, X_{t-r})\, dr \\ &= (\partial_t - L) \int_0^t \big(Q_{t-r} g(r,\,\cdot\,)\big)(x)\, dr \\ &= \partial_t \int_0^t \big(Q_{t-r} g(r,\,\cdot\,)\big)(x)\, dr - \int_0^t \partial_t \big(Q_{t-r} g(r,\,\cdot\,)\big)(x)\, dr \\ &= \big(Q_{t-t} g(t,\,\cdot\,)\big)(x) = g(t,x), \end{aligned}$$

where we have used the Leibniz rule. For details on Leibniz rule, see [Fol23, Section 4.5]

15.28 Feynman–Kac formula. Let X_t be the solution to the SDE, and L be its generator. Suppose f is bounded measurable and $v \in C_b$, then the function

$$u(t,x) = \mathrm{E}_x \left[f(X_t) \exp\left(- \int_0^t v(t-s, X_s)\, ds \right) \right]$$

satisfies the PDE

$$\begin{cases} \partial_t u = Lu - vu \\ u(0,x) = f(x) \end{cases}$$

Combine this with previous we can obtain the nonhomogeneous Feynmann–Kac. This can be found for example on Wikipedia, but is too complicated to provide any additional insight and is hence omitted.

As we have discussed before, the initial value problem can be converted into a backward problem. Fix the terminal time T,

(a) $\tilde{u}(t,x) = u(T-t,x) = Q_{T-t}f(x) = \mathrm{E}_x f(X_{T-t})$ satisfies the backward equation

$$\begin{cases} \partial_t \tilde{u} = -L\tilde{u} & \text{for } 0 \le t < T, \\ \tilde{u}(T,x) = f(x). \end{cases}$$

(b) $\tilde{u}(t,x) = u(T-t,x) =$

$$\mathrm{E}_x f(X_{T-t}) + \mathrm{E}_x \int_0^{T-t} g(T-t-s, X_s)\, ds$$

satisfies the backward equation

$$\begin{cases} \partial_t \tilde{u} = -L\tilde{u} - g(T-t,x), \\ \tilde{u}(T,x) = f(x). \end{cases}$$

Now rewrite $\tilde{g}(t,x) = g(T-t,x)$, then the equation

$$\begin{cases} \partial_t \tilde{u} = -L\tilde{u} - \tilde{g}(t,x), \\ \tilde{u}(T,x) = f(x). \end{cases}$$

is satisfied by $\tilde{u}(t,x) = \mathrm{E}_x f(X_{T-t}) + \mathrm{E}_x \int_0^{T-t} \tilde{g}(s+t, X_s)\, ds =$

$$\mathrm{E}_x f(X_{T-t}) + \mathrm{E}_x \int_t^T \tilde{g}(s, X_{s-t})\, ds.$$

This is equivalent to writing

$$\tilde{u}(t,x) = \mathrm{E}_{X_t=x}\left[f(X_T) + \int_t^T \tilde{g}(s, X_s)\, ds \right],$$

which means that if we start the diffusion process at $X_t = x$, then $\mathrm{E} f(X_T)$ (from Lu) and $\mathrm{E}\int_t^T \tilde{g}(s, X_s)\, ds$ (from the additional constant term $\tilde{g}$) contribute to the the value of the solution $\tilde{u}$ at (t,x). This looks cleaner than the forward PDE, which contains a convolution-like term, and provides one reason why some authors prefer this backward formulation. The backward formulation is also preferred in mathematical finance.

(c) By the same derivation, the backward equation

$$\begin{cases} \partial_t \tilde{u} = -L\tilde{u} + \tilde{v}u \\ \tilde{u}(T,x) = f(x) \end{cases}$$

is satisfied by

$$\mathrm{E}_{X_t=x}\left[f(X_t) \exp\left(-\int_t^T \tilde{v}(s, X_s)\, ds \right) \right]$$

Consider the elliptic PDE

$$\begin{cases} Lu = 0 & \text{in } D \\ u(x) = g(x) & \text{on } \partial D. \end{cases}$$

We claim that it has the probabilistic solution

$$u(x) = \mathrm{E}_x g(X_{\tau_D}),$$

where $\tau_D = \inf\{t \geq 0 : X_t \notin D\}$ is the *first exit time* to the domain D.

This is the Dirichlet boundary problem for the differential operator L, which asks if there exists an function u that is L-harmonic in D and agrees with g on the boundary.

Setting $L = \frac{1}{2}\Delta$, we then obtain a formal solution to the Dirichlet problem for the Laplace equation:

$$u(x) = \mathrm{E}_x g(B_{\tau_D}).$$

Poincaré cone condition
regular boundary
a convex domain
Dirichlet problem with potential term (Bass 40.4) Schrödinger equations

Chapter 16 Special Topics

16.A Random matrices

16.A.1 Random measures

Given a probability space $(\Omega, \mathcal{F}, P)$, we say a random variable $L\colon \Omega \to \mathcal{P}(S)$ is a *random (probability) measure*. Given a sequence of random probability measures $L_n(\omega)$ and another random measure $L(\omega)$ (not necessarily with mass 1) on Ω, we say L_n converges almost surely/in probability to L if we have the almost sure convergence/convergence in probability of random variables

$$\int f\, dL_n(\omega) \to \int f\, dL(\omega) \quad \text{for all } f \in C_b(S).$$

We emphasize that $L_n(\omega)$ and $L(\omega)$ are measures defined on S, so the above indeed makes sense. Some authors use to term *weakly* almost surely/in probability to emphasize there are two layers of convergence taking place: weak convergence inside and almost sure/in-probability convergence on the outside.

almost sure convergence implies convergence in probability, but not vice versa (consider deterministic L_n and L)

16.A.2 Ensembles

A Gaussian orthogonal ensemble (GOE)

A Gaussian unitary ensemble (GUE)

16.A.3 Asymptotic laws on the spectrum of random matrices

The *empirical spectral distribution* of a self-adjoint matrix $\Sigma \in \mathbf{R}^{n\times n}$ is $\frac{1}{n}\sum_{k=1}^{n} \delta_{\lambda_k}$. We will denote it by $\mathrm{ESD}(\Sigma)$.

16.1 Semicircle law. A self-adjoint matrix with independent entries in the upper diagonal are independent is called a

We use $\mathbf{x}$ for vectors, and X for matrices. (The below needs to be corrected)

16.2 Marchenko–Pastur law. Consider the sample covariance matrix $\frac{1}{n}X_pX_p^{\mathrm{T}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_{pk}\mathbf{x}_{pk}^{\mathrm{T}}$, where $X_p \in \mathbf{R}^{p\times n}$ consists of columns $\mathbf{x}_{pk} \in \mathbf{R}^p$ ($k \in [n]$) which are i.i.d. with mean zero and finite variance. Suppose $p(n)/n \to r > 0$ as $n \to \infty$, then

$$\mathrm{ESD}\big(\tfrac{1}{n}X_pX_p^{\mathrm{T}}\big) \Rightarrow \mu_{\mathrm{MP}}$$

almost surely, where

$$d\mu_{\mathrm{MP}} = \left(1 - \tfrac{1}{r}\right)^+ \delta_0 + \frac{\sqrt{(b-x)(x-a)}}{2\pi\lambda x} \mathbf{1}_{[a,b]}(x)\, dx,$$

where $a = (1-\sqrt{r})^2$ and $b = (1+\sqrt{r})^2$.

16.3 Bai–Yin Law.

16.4 Tracy–Widom law.

16.A.4 Determinantal point processes

For $f = \exp(-\varphi)$ is a log-concave density. For all $x \in \mathbf{R}^n$, we have $f \le C\exp(-a|x|)$ for some $C > 0$ and $a > 0$. It follows that $\mathrm{E}|X|^p < \infty$ for all p.

Alternatively, we have the following result

16.5 Borell's lemma. For a log-concave random vector X, we have $P(|X| \ge t) \le C\exp(-ct)$ for some $c, C > 0$.

For a log-concave density f such that $\int xf\, dx = 0$, we have

$$f(0) \le \sup_{x\in\mathbf{R}^n} f(x) \le e^n f(0).$$

16.B Concentration of measures

For $X \sim N(0, I_n)$, the norm $\|X\|$ is concentrated around $\sqrt{n}$ [Ver18, Theorem 3.1.1]

For the uniform measure on the sphere, measure is concentrated around any equator) more generally, we have [Ver18, Lemma 5.1.6]

16.6 Gaussian Poincaré inequality. For $f \in H^1(\gamma_n)$, we have

$$\mathrm{Var}_\gamma f \le \mathrm{E}_\gamma |\nabla f|^2.$$

We know $f = \sum_{k=0}^\infty \langle f, \frac{\mathrm{He}_k}{\sqrt{k!}}\rangle \frac{\mathrm{He}_k}{\sqrt{k!}}$. Also notice that $\langle f, \mathrm{He}_0\rangle = \int f\, d\gamma$. Therefore

$$\begin{aligned}
\mathrm{Var}_\gamma f &= \|f\|^2_{L^2(\gamma)} - \langle f, \mathrm{He}_0\rangle^2 \\
&= \sum_{k=1}^\infty \tfrac{1}{k!}\langle f, \mathrm{He}_k\rangle^2 \quad \text{by Parseval's identity} \\
&\le \sum_{k=1}^\infty \tfrac{1}{(k-1)!}\langle f, \mathrm{He}_k\rangle^2 \qquad (16.7)
\end{aligned}$$

Formally differentiate $f = \sum_{k=0}^\infty \frac{1}{k!}\langle f, \mathrm{He}_k\rangle \mathrm{He}_k$ term-by-term, we obtain

$$\sum_{k=1}^\infty \frac{\langle f, \mathrm{He}_k\rangle}{\sqrt{(k-1)!}} \frac{\mathrm{He}_{k-1}}{\sqrt{(k-1)!}}. \qquad (16.8)$$

Inspired by the Sobolev space for Lebesgue measure on $\mathbf{R}^n$, we may define the *Gaussian Sobolev space*

$$H^1(\gamma) = \left\{ f \in L^2(\gamma) : \sum_{k=1}^\infty k \left\langle f, \frac{\mathrm{He}_k}{\sqrt{k!}}\right\rangle^2 < \infty \right\}.$$

For any $f \in H^1(\gamma)$, the series in (16.8) converges to ∇f (the weak derivative of f) in $L^2(\gamma)$, with

$$\|\nabla f\|^2_{L^2(\gamma)} = \sum_{k=1}^{\infty} k \left\langle f, \frac{\mathrm{He}_k}{\sqrt{k!}} \right\rangle^2,$$

which is precisely (16.7).

For details on the Gaussian Sobolev space on $\mathbf{R}^n$, we invite the readers to consult [Mal95, Chapter 5, Section 2] and [Bog98, Section 1.5].

16.9 Gaussian log–Sobolev inequality. For f that is the square of a function in $H^1(\gamma_n)$, we have

$$\mathrm{Ent}_\gamma f \le 2\,\mathrm{E}_\gamma \left|\nabla \sqrt{f}\right|^2 = \tfrac{1}{2}\,\mathrm{E}_\gamma \left(\frac{|\nabla f|^2}{f}\right) = \tfrac{1}{2}\,\mathrm{E}_\gamma \left[\nabla f \cdot \nabla(\log f)\right].$$

Proof. □

An alternative (and much more recent) proof of this using the Prékopa–Leindler inequality is given in [Gen08].

16.10 Exercise. Use the same technique to directly prove the Gaussian Poincaré inequality.

The Poincaré inequality states that

$$\mathrm{Var}_\mu f \le C\mathcal{E}(f)$$

for some constant $C > 0$ uniform probability measure over a convex body, and measures with log-concave densities [BGL14, Theorem 4.6.3]

The log-Sobolev inequality states that

$$\mathrm{Ent}_\mu f^2 \le 2C\mathcal{E}(f)$$

for some constant $C > 0$ holds for measures μ with strongly log-concave densities [Led01, Theorem 5.2]

If μ has λ-strongly log-concave density, then μ satisfies $\mathrm{LS}(1/\lambda)$.

It follows that the Poincaré constant for any strongly log-concave measure is independent of the dimension. It is still to this day open whether the Poincaré constant for any log-concave measure is independent of dimension n. The state of art, by the end of 2025, is that the Poincaré constant for log-concave measures is $O(\sqrt{\log n})$.

Log-Sobolev inequalities are stronger than Poincaré inequality. Take the function to be $1 + \epsilon f$ in the log-Sobolev inequality, and then take $\epsilon \to 0$, we should recover the Poincaré inequality

$$\mathrm{Var}_\mu f \le C\,\mathrm{E}_\mu |\nabla f|^2.$$

We know from the Gaussian case that both inequalities are tight when $C = 1$, but in general one should not expect that the Poincaré and log-Sobolev constant to differ by a multiple of 2. We will write $\mathrm{P}(C)$ and $\mathrm{LS}(C)$ to mean that a measure μ satisfy the Poincaré and log-Sobolev inequality with optimal constant C and $2C$, respectively. (Optimal constant means that infimum of all possible constants such that the inequalities hold.) Therefore, $\mathrm{LS}(C) \implies \mathrm{P}(C)$.

There is an alternative form for log-Sobolev inequalities

If we set $f = \frac{d\nu}{d\mu}$, then

$$D(\nu\|\mu) \lesssim I(\nu\|\mu),$$

where the relative Fisher information I is defined by

$$I(\nu\|\mu) = 4\int_S \Gamma(\sqrt{f})\,d\mu.$$

The constants would differ by a multiple of 4.

There is also the modified log-Sobolev inequality, which states that

$$\operatorname{Ent}_\mu f \lesssim \mathcal{E}(f, \log f)$$

for a class of functions f.

The constant C above is called the Poincaré/log-Sobolev constant. (We remark that, in the mixing time literature, the Poincaré, log-Sobolev, and modified log-Sobolev constants are defined instead by

$$\lambda = \inf\left\{\frac{\mathcal{E}(f)}{\operatorname{Var}_\mu f} : \operatorname{Var}_\mu f \neq 0\right\},$$
$$\alpha = \inf\left\{\frac{\mathcal{E}(\sqrt{f})}{\operatorname{Ent}_\mu f} : f \geq 0, \operatorname{Ent} f \neq 0\right\},$$
$$\rho = \inf\left\{\frac{\mathcal{E}(f, \log f)}{\operatorname{Ent}_\mu f} : f \geq 0, \operatorname{Ent} f \neq 0\right\},$$

which are the optimal constants in the above inequalities.)

16.11 Proposition. Given a symmetric Markov semigroup Q_t with generator L and invariant measure μ, the following are equivalent:

(a) μ satisfies $\mathrm{P}(C)$;

(b) the spectrum of $-L$ is contained in $\{0\} \cup [1/C, \infty)$.

This is precisely the reason why the Poincaré inequality is sometimes called the spectral gap inequality. Notice that when the spectrum is discrete, this says precisely that the Poincaré constant is precisely the inverse of the smallest strictly positive eigenvalue of $-L$.

We note that since $-L\mathbf{1} = 0$, 0 is always an eigenvalue. (The spectral gap is sometimes defined as the gap between the largest and second largest eigenvalue of L.) Now let $\lambda > 0$ be an eigenvalue for $-L$, with an eigenfunction $f \in D(L)$. Then

$$\mathcal{E}(f, f) = \langle -Lf, f\rangle_\mu \geq \lambda \int f^2\,d\mu = \lambda \operatorname{Var}_\mu f.$$

Here we used $\int f\,d\mu = 0$, since f (associated to the eigenvalue λ) is orthogonal to the eigenfunction $\mathbf{1}$ associated to the eigenvalue 0. Therefore the Poincaré constant must be $\geq 1/\lambda$ for any nonzero eigenvalue λ.

Conversely, suppose all positive eigenvalues are $\geq 1/C$. Then

$$\mathcal{E}(f, f) = \int_{1/C}^\infty \lambda\,d\langle \Pi_\lambda f, f\rangle \geq \frac{1}{C}\int_{1/C}^\infty d\langle \Pi_\lambda f, f\rangle = \frac{1}{C}\int f^2\,d\mu,$$

proving that the Poincaré constant must be $\leq C$.

The (largest possible) domain of functions for P/LS to hold should be some weighted Sobolev space with respect to the measure, if the space is well-defined. Therefore it usually

suffices to take the class of functions to be C_c^∞, whose completion is the Sobolev space. Also C_b^1 is dense in the Sobolev space.

Dirichlet domain

hypercontractivity

[Han14, Lemma 3.13]

16.12 Herbst's lemma. Given $\mu \in \mathcal{P}(\mathbf{R}^n)$, if

$$\operatorname{Ent}_\mu(e^{\lambda F}) \le \frac{\lambda^2\sigma^2}{2} \operatorname{E}_\mu \exp(\lambda F)$$

for all $\lambda \in \mathbf{R}$, then

$$\operatorname{E}_\mu \exp(\lambda F) \le \frac{\lambda^2\sigma^2}{2}$$

for all $\lambda \in \mathbf{R}$. If we replace F by $X \sim \mu$, then this is saying precisely that X is subgaussian with proxy variance σ^2.

It suffices to prove the above for $\lambda > 0$ and then replace F by $-F$ to get the case when $\lambda < 0$.

Suppose $F \in C_b^1$ and L-Lipschitz, then $e^{\lambda F} \in C_b^1$. Therefore if μ is $\operatorname{LS}(C)$, then

$$\operatorname{Ent}_\mu(e^{\lambda F}) \le 2C\lambda^2 L^2 \operatorname{E}_\mu \exp(\lambda F).$$

where L comes from $\|\nabla F\|_u$, which is precisely the Lipschitz constant of F (this is Rademacher's theorem). It follows that

$$\operatorname{E}_\mu \exp(\lambda F) \le 2C\lambda^2 L^2.$$

This yields the concentration of Lipschitz function, after appropriately extending F to all Lipschitz functions by a mollifier argument (see Chafaï & Joseph Lehec Theorem 3.1).

16.13 Lipschitz concentration. For $\mu \in \mathcal{P}(\mathbf{R}^n)$ that is $\operatorname{LS}(C)$ for some smooth enough functions (C_c^∞ should be good on $\mathbf{R}^n$), then any L-Lipschitz function $F\colon \mathbf{R}^n \to \mathbf{R}$ is integrable, and for every $t > 0$, we have

$$\mu(F \ge \operatorname{E}_\mu F + t) \le \exp\left(-\frac{t^2}{2CL^2}\right).$$

If μ satisfies Poincaré's inequality, then we instead have the weaker subexponential concentration for any Lipschitz function F:

$$\mu(F \ge \operatorname{E}_\mu F + t) \lesssim \exp\left(-\frac{t}{\sqrt{C}L}\right).$$

Both the Poincaré and log-Sobolev inequality gives us a dimension-free bound for the concentration of Lipschitz function; however, log-Sobolev gives us subgaussian concentration of Lipschitz functions, while Poincaré only gives us subexponential concentration. This is expected since log-Sobolev implies Poincaré.

The most classical result, of course, is the

16.14 Gaussian Lipschitz concentration. Consider $Z \sim \gamma_n$, we then have

$$P\big(F(Z) \ge EF(Z) + t\big) \le \exp\left(-\frac{t^2}{2L^2}\right),$$

since the log-Sobolev constant for γ_n is $1/2$.

We say a real random variable X is mean-zero subgaussian with proxy variance σ^2 if

$$\mathrm{E}\exp(\lambda X) \le \exp\left(\frac{\sigma^2\lambda^2}{2}\right).$$

Recall when $X \sim N(0,\sigma^2)$, the above inequality is precisely an equality. The random variables such that Hoeffding's lemma (and hence Hoeffding's inequality) holds are precisely $\frac{b-a}{2}$-subguassian. On the one hand,

16.15 Theorem. Let S be a separable metric space, for $X \sim \mu \in \mathcal{P}_1(S)$, the following are equivalent

(a) $F(X)$ is subgaussian with proxy variance σ^2 for all 1-Lipschitz F;
(b) $W_1(\mu,\nu) \le \sqrt{2\sigma^2 D(\nu\|\mu)}$ for all $\nu \in \mathcal{P}_1(S)$.

Pinsker's inequality can be recovered. We know that when S is endowed with discrete metric $\rho(x,y) = \mathbf{1}\{x \ne y\}$, $W_1(\mu,\nu) = d_{\mathrm{TV}}(\mu,\nu)$. On the other hand, by Hoeffding's lemma, we know $F(X)$ is subgaussian with proxy variance $\frac{1}{4}(\sup F - \inf F) \le 1/4$. Hence

$$d_{\mathrm{TV}}(\mu,\nu) \le \sqrt{\frac{1}{2}D(\nu\|\mu)}.$$

16.16 Theorem [Led01, Theorem 6.2].

16.17 Theorem [Led01, Corollary 6.4].

tensorization
If μ satisfies $\mathrm{P}(C)$, then the product measure $\mu \times \cdots \times \mu$ satisfies $\mathrm{P}(C)$.

16.18 Proposition [Led01, Proposition 6.3].

Talagrand's inequality

16.19 Theorem [Led01, Theorem 6.5]. Let $\mu,\nu \in \mathcal{P}_2(\mathbf{R}^n)$. If μ is a strongly log-concave measure with parameter c, we have

$$W_2(\mu,\nu) \le \sqrt{\frac{2}{c}D(\nu\|\mu)}.$$

This is in fact a consequence of log-Sobolev inequality, and the Herbst argument
Note that the standard Gaussian measure γ_n satisfies $c = 1$, and therefore

$$W_2(\mu,\gamma_n) \le \sqrt{2D(\gamma_n\|\mu)}.$$

Ledoux defined W_2 using the quadratic cost $\frac{\|x-y\|^2}{2}$ instead

16.20 HWI inequality. For $\mu,\nu \in \mathcal{P}_2(\mathbf{R}^n)$, we have

$$D(\nu\|\mu) \le W_2(\mu,\nu)\sqrt{I(\nu\,\|\,\mu)} - \frac{c}{2}W_2^2(\mu,\nu).$$

16.21 Definition. A random variable X is *subgaussian* if there is a constant $K_{\_} > 0$ such that

(a) $P(|X| > t) \le 2\exp\left(-t^2/K_1^2\right)$ for all $t \ge 0$; [tail condition]

(b) $\big(\mathrm{E}|X|^p\big)^{1/p} \le K_2\sqrt{p}$ for all $p \ge 1$; [moment condition]

(c) $\mathrm{E}\exp(\lambda^2 X^2) \le \exp(K_3^2\lambda^2)$ for all $|\lambda| \le \frac{1}{K_3}$; [$X^2$ MGF condition]

(d) $\mathrm{E}\exp(X^2/K_4^2) \le 2$; [boundedness condition]

and additionally, if X is mean-zero, then

(e) $\mathrm{E}\exp(\lambda X) \le \exp(K_5^2\lambda^2)$ for all $\lambda \in \mathbf{R}$. [mean zero X MGF condition]

If X is subgaussian, then the constants $K\_$'s above differ by absolute constants. This is Vershynin's definition.

In a metric space (S, ρ), we define the concentration function of

16.B.1 Talagrand's generic chaining argument

16.C Functional inequalities of Markov processes

Poincaré inequality and exponential convergence to ergodicity

16.D Stochastic localization

16.22 Localization lemma.

16.E 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

$$t_{\mathrm{mix}}(\epsilon) = \inf\{t \ge 0 : d(t) \le \epsilon\}$$

relaxation time
$d_{\mathrm{TV}}(\mu,\nu) \le d_{\mathrm{TV}}(\mu,\rho) + d_{\mathrm{TV}}(\nu,\rho)$
submultiplicativity of coupling distance

16.23 Theorem. $d(s+t) \le 2d(s)d(t)$

16.24 Fekete's lemma. For a subadditive function $f\colon [0,\infty) \to \mathbf{R}$, 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]. Taking $-f$ in place of f gives us a result for superadditive function.

$$t_{\text{mix}}(\epsilon) \geq t_{\text{rel}} \log\left(\frac{1}{2\epsilon}\right)$$

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 $n \log n$.

16.F Models from statistical mechanics

Gibbs random field.

ferromagnetic Ising model on a finite graph $G = (V, E)$ with parameter $\beta \geq 0$

Let $S = \{-1, 1\}^V$ be the space of configurations, which basically means that each vertex of the graph is assigned a spin -1 or 1. We want to define a probability distribution P on the space of random configurations S, which takes the interaction of spins between adjacent vertices into account.

Define the probability measure π on the finite set S by

$$\pi\{\sigma\} = \frac{1}{Z_\beta} \exp\big(-\beta H(\sigma)\big),$$

where

$$H(\sigma) = -\sum_{v \sim w} \sigma(v)\sigma(w)$$

is known as the *potential energy/Hamiltonian*, and Z_β is the normalizing constant

$$Z_\beta = \sum_{\sigma \in S} \exp\big(-\beta H(\sigma)\big),$$

also known as the *partition function*. The measure π is called the *Gibbs measure*. In general we can generalize

The parameter β is the reciprocal of the temperature in physics. For β close to 0, π is closed be being uniform, and for β large, we should expect larger $\pi(\sigma)$ for distributions with lower energy.

The partition function is impossible to compute when S is large, which means that it is impossible to find the exact π. However in computer science we are interested in developing fast sampling algorithms with these distributions. A sampling algorithms from a given distribution using a Markov chain converging to this distribution is called a *Markov chain Monte Carlo* method.

Metropolis chain

Glauber Dynamics

Curie–Weiss model on n spins

$$H(\sigma) = -\frac{J}{2n} \sum_{1 \leq i,j \leq n} \sigma_i \sigma_j - h \sum_{k=1}^{n} \sigma_j,$$

where $J > 0$ is a constant and $h \in \mathbf{R}$ represent the external magnetic field. The difference between the CW model and the Ising model is that in the latter we are considering interactions between neighbors on a graph, but in the former we are considering interactions between all the spins, and the underlying graph is irrelevant.

[1] If discrete-time, we can let the chain to be 1/2-lazy

16.F.1 Bernoulli bond percolation

We use $\omega \in \{0,1\}^{E^d}$ for each configuration on the E^d grid, which is in fact a random variable that takes value in $\{0,1\}^{E^d}$. Using ω for a random outcome can indeed be confusing initially, but at the end of the day what we really care about is how the configurations are distributed. The probability space we will take is $(\Omega, \mathcal{F}, P_p)$, where $\Omega = \{0,1\}^{E^d}$, $\mathcal{F}$ is the product σ-field, and $P_p = \otimes^{E^d}$ Bernoulli(p), the distribution of the configurations when the edges are open with i.i.d. Bernoulli(p).

Notice that each vertex of the grid must be contained in exactly one connected component with open edges, which is often called an *open cluster*. Let this open cluster be be denoted

Define

$$\theta(p) = P_p(\text{the origin is contained in an infinite open cluster}),$$

and

$$p_c(d) = \sup\{p : \theta(p) = 0\}.$$

By ergodicity or Kolmogorov zero–one law, one can conclude that $\theta(p) = 0$ or 1.

$\theta(p)$ is an increasing function in p. This is a straightforward consequence from the coupling method. (We can alter the underlying probability space without changing P_p.) Let $p_1 \geq p_2$ be the edge probabilities of two configurations. We use the same uniform random variable to determine the two Bernoulli random variables Bernoulli(p_1) and Bernoulli(p_2) attached to each side. Under this coupling, then an edge in configuration 1 is open if an edge in configuration 2 is open. This shows precisely that $\theta(p_1) \geq \theta(p_2)$.

$p_c(d+1) \leq p_c(d)$, and in fact the strict inequality holds for $d \geq 1$

$0 < p_c(d) < 1$ for $d \geq 2$

increasing event A and B,

$$P_p(A)P_p(B) \leq P_p(A \cap B)$$

For any increasing $L^2(P_p)$ function, we have

$$\mathrm{E}_p f \, \mathrm{E}_p g \leq \mathrm{E}_p fg$$

Let N be the number of infinite open clusters. P_p is ergodic, $P_p(N = k) = 0$ or 1 for any $k \in \mathbf{N}_0 \cup \{\infty\}$. This implies the number of infinite clusters must be constant a.s. In fact, N can only be $0, 1$.

First one can show that $P_p(2 \leq N < \infty) = 0$. For the sake of contradiction suppose $P_p(N = k) = 1$ for some $k \geq 2$. There must exist some square box B_n centered at the origin such that

$$P_p(\text{all } k \text{ infinite cluster intersects } B_n) > 0.$$

Since the probability that all edges in B_n are open is positive, and

$$P_p(\text{all } k \text{ infinite cluster intersects } B_n, \text{ all edges in } B_n \text{ open}) = P(N = 1) > 0,$$

which contradicts our assumption.

By the Burton–Keane trifurcation argument, we cannot have infinite open cluster.

percolation at critical value

$\theta(p_c) = 0$ for $d = 2$ and $d \geq 19$

16.F.2 First passage percolation

16.G Large deviation theory

Let $I\colon S \to [0,\infty]$ be a LSC function and $\{r_n\}$ be an increasing sequence of positive real numbers that goes to $+\infty$. We say the sequence of Borel probability measures μ_n satisfies the large deviation principle (LDP) with rate function I and normalization r_n if for any closed sets F and open sets G,

$$\limsup_n \frac{1}{r_n} \log \mu_n(F) \le -\inf_{x\in F} I(x), \tag{16.25}$$

$$\liminf_n \frac{1}{r_n} \log \mu_n(G) \ge -\inf_{x\in G} I(x). \tag{16.26}$$

We write $\mathrm{LDP}(\mu_n, r_n, I)$.

If (16.25) is replaced by

$$\limsup_n \frac{1}{r_n} \log \mu_n(K) \le -\inf_{x\in K} I(x), \tag{16.27}$$

for any compact set K, then we say the weak large deviation principle holds, and write $\mathrm{wLDP}(\mu_n, r_n, I)$ instead.

This is equivalent to saying that for any Borel set A,

$$\begin{aligned} -\inf_{x\in \operatorname{Int} A} I(x) &\le \liminf_n \frac{1}{r_n} \log \mu_n(A) \\ &\le \limsup_n \frac{1}{r_n} \log \mu_n(A) \le -\inf_{x\in \overline{A}} I(x). \end{aligned}$$

We say A is an *I-continuity set* if $-\inf_{x\in \operatorname{Int} G} I(x) = -\inf_{x\in \overline{G}} I(x)$. When A is such a set,

$$\lim_n \frac{1}{n} \log \mu_n(A) \to -\inf_{x\in A} I(x)$$

If S is regular (e.g. a metric space), then there is at most one rate function I that satisfies $\mathrm{LDP}(\mu_n, r_n, I)$.

We say a rate function I is *tight* if $\{x : I(x) \le c\}$ is a compact subset of S for any $c \in \mathbf{R}$.

We say the sequence $\{\mu_n\} \subseteq \mathcal{P}(S)$ is *exponentially tight with normalization* $\{r_n\}$ if for each $0 < b < \infty$, there exists a compact set b such that

$$\limsup_n \frac{1}{r_n} \log \mu_n(K_b^c) \le -b.$$

exponential tightness and (16.27) together implies (16.25), and to show LDP, it suffices to show wLDP and then establish exponential tightness. In addition, the rate function I is tight.

16.28 Cramér's theorem. Let $X_1, X_2, \dots$ be be a sequence of i.i.d. $\mathbf{R}^d$-valued random variables with distribution π. Let μ_n be the distribution of the sample mean $\frac{X_1+\cdots+X_n}{n}$. Then $\mathrm{wLDP}(\mu_n, n, I)$ holds for $I(a) = \sup_{t\in\mathbf{R}^d}\{\langle a,t\rangle - \log M_{X_1}(t)\}$, the Legendre dual of the cumulant generating function of X.

If in addition there exists a neighborhood around 0 such that $M_{X_1}(t) < \infty$, then LDP holds and I is a tight rate function. (In the special $d = 1$ then LDP also holds, but I may not be tight.)

Any interval (unbounded, open/closed, half-open) are all I-continuity sets because of convexity. In the case $a > \mathrm{E}X_1$, we have $\inf_{x>a} I(x) = \inf_{x\ge a} I(x) = I(a)$. In the $\mathbf{R}^d$ case, unbounded intervals may be replaced half-spaces, while bounded intervals may be replaced by convex bodies.

Convexity and LSC of $I(x)$ tell us that there exists a sequence $\{x_n\} \subseteq \operatorname{Int} A$ converging to $x \in \overline{A}$ such that $I(x_n) \to I(x)$.

16.29 Sanov's theorem. Given a Polish space S, let $X_1, X_2, \dots$ be a sequence of i.i.d. S-valued random variables with distribution $\pi \in \mathcal{P}(S)$. Define the empirical sample distribution

$$L_n = \frac{1}{n}\sum_{k=1}^{n} \delta_{X_k},$$

which is a $\mathcal{P}(S)$-valued random variable. Let ν_n be the distribution of L_n (i.e., $P \circ L_n^{-1}$), which is a measure defined on $\mathcal{P}(S)$. Then

$$\mathrm{LDP}\big(\nu_n, n, D(\,\cdot\,\|\pi)\big).$$

The interpretation of this theorem is that under the true distribution that π, the probability that the sample distribution is some other ν_n decays at the rate $\exp\big(-D(\nu_n\|\pi)n\big)$. This provides a quantitative comparison between the sampling distribution and the true distribution, a very natural question in statistics. The probability that the sampling distribution lies in some set in $\mathcal{P}(S)$ corresponds to the distribution of L_n. Therefore the distribution of sample distribution indeed makes sense.

16.H Optimal transport

Let $c\colon S \times T \to [0,\infty]$ be the cost function, usually assumed to be at least LSC. Given $\mu \in \mathcal{P}(S)$ and $\nu \in \mathcal{P}(T)$, the classical *Monge's formulation* of the optimal transport problem considers

$$\inf_{L_*\mu=\nu} \int_S c\big(x, L(x)\big)\, d\mu(x)$$

The measurable maps from $(S,\mathcal{S})$ to $(T,\mathcal{T})$ are called *transport maps*, and any transport map that achieves the infimum is called a *Monge map*.

It turned out Monge's formulation can be extremely hard to study directly, and it became necessary to study another formulation. *Kantorovich's formulation* of the optimal transport problem wants us to find

$$\inf_{\pi\in\Pi(\mu,\nu)} \int_{S\times T} c(x,y)\, d\pi(x,y),$$

where $\Pi(\mu,\nu)$ is the space of all couplings between μ and ν, also called the space of *transport plans*. We have mentioned in the main text that this set is nonempty and convex. Any transport map that achieves the infimum is called an optimal coupling. We will refer to the two formulations as MOT and KOT respectively. KOT is a more general formulation because if a Monge map L exists, then an optimal coupling must π exists of the form

$$\frac{d\pi(x,y)}{d\mu(x)} = \mathbf{1}_{y=L(x)}(x),$$

or equivalently, $\pi = (\mathrm{Id} \times L)_* \mu$. To put into words, KOT allows transportation beyond a point-to-point scheme.

Under the assumption that $(S, \mathcal{S})$ and $(T, \mathcal{T})$ are Polish, and the cost function is LSC, then the optimal coupling in KOT is attained. This boils down to the so-called *direct method in the calculus of variations*. This method is based on the following result:

16.30 Proposition [San15, Box 1.1]. Let S be a compact metric space, and $f\colon S \to (-\infty, \infty]$ be LSC, then $\min_x f(x)$ can be obtained.

Define $\mathcal{C}\colon \Pi(\mu, \nu) \to [0, \infty]$ by $\mathcal{C}\pi = \int c(x, y)\, d\pi$, the evaluation of the cost of a transport plan. This is a positive linear functional.

16.31 Theorem. If $c\colon S \times T \to [0, \infty]$ is LSC, then the function(al) $\mathcal{C}\colon \Pi(\mu, \nu) \to [0, \infty]$ is also LSC, when $\Pi(\mu, \nu) \subseteq \mathcal{P}(S \times T)$ endowed with the topology of weak convergence is sequentially compact.

If S and T are compact and c is continuous (even be negative-valued), the above result follows straight from Riesz–Markov–Kakutani theorem (finite measures) and sequential Banach–Alaoglu theorem, a technique we should already be familiar by now; see [San15, Theorem 1.4].

Combine this with Corollary 8.31, we can immediately conclude that the infimum in KOT can be achieved.

16.32 Gluing lemma for transport plans. For transport plans $\pi_1 \in \Pi(S, T)$ and $\pi_2 \in \Pi(T, R)$, there is a product measure π on $S \times T \times R$ such that its marginals on $S \times T$ and on $T \times R$ are π_1 and π_2 respectively.

Proof that W_p is a metric. □

Proof of the dual representation of W_1. □

16.33 Duality of the Kantorovich problem. Let S and T be Polish spaces, and consider Kantorovich's formulation. We have

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{S \times T} c(x, y)\, d\pi = \sup \left\{ \int f\, d\mu + \int g\, d\nu \right\},$$

where the supremum is over all $f \in C_b(S)$ and $g \in C_b(T)$ such that $f(x) + g(y) \le c(x, y)$.

It is possible to also take $f \in L^1(\mu)$ and $g \in L^1(\nu)$, or $f \in \mathrm{Lip}_b(S)$ and $g \in \mathrm{Lip}_b(T)$. (By this notation we mean bounded Lipschitz functions.)

Indeed, this may remind the readers of linear programming. Consider the primal problem

$$\text{minimize } c \cdot x \quad \text{subject to } Ax = b, x \ge 0$$

and the dual problem

$$\text{maximize } b \cdot y \quad \text{subject to } A^{\mathrm{T}} y \le c.$$

Let μ and ν to be discrete measures with finite support in the context of the previous result. They naturally corresponds to a matrix A of size 2-by-n, and we take $b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Also let c be the cost function. It should then be clear that duality of the Kantorovich problem in finite dimension is precisely the aforementioned linear programming problem.

For x and y that satisfies the constraints in the primal and the dual problem, we have the result that x and y solve the primal and dual problem if and only if the equilibrium equation

$$x_j > 0 \implies \sum_{i=1}^{m} y_i a_{ij} = c_j$$

for all $j \in [n]$. This is called the complementary slackness condition because the slackness of x_j from $x_j = 0$ implies the tightness of $\sum_{i=1}^{m} y_i a_{ij}$, and the slackness of the latter that implies the tightness of the former.

c-concave functions

16.34 Brenier's theorem. Consider KOT with the quadratic cost $c(x, y) = |x - y|^2$ on $\mathbf{R}^n$, and let μ and ν be absolutely continuous.[2] The optimal coupling $\pi \in \Pi(\mu, \nu)$ must be of the form $\pi = (\mathrm{Id} \times L)_* \mu$, where L is unique, and must be the a.e. gradient of a convex function φ. Meanwhile,

$$\operatorname{supp} \nu = \overline{\nabla \varphi(\operatorname{supp} \mu)}.$$

This means precisely that Monge's problem can be obtained by the transport map $\nabla \varphi$. Also, it should not be hard to show that $\nabla \varphi^*$, being the inverse of $\nabla \varphi$, is the optimal transport map from ν to μ. This shows that the Monge map L is in fact almost bijective.

Brenier's theorem naturally leads to the solution of a fully nonlinear elliptic PDE, known as the *Monge–Ampère equation*. Let $\frac{d\mu}{dm} = f$ and $\frac{d\nu}{dm} = g$. Given $L_* \mu = \nu$, by change of densities, we get

$$f(x) = g\big(\nabla \varphi(x)\big) \big| \det \nabla^2 \varphi(x) \big|$$

since by convexity $\nabla^2 \varphi \succeq 0$, rearranging gives us

$$\det \nabla^2 \varphi(x) = \frac{f(x)}{g\big(\nabla \varphi(x)\big)}.$$

Note that everything above holds only for a.e. x.

McCann interpolation $\lambda_t = \big((1-t)\mathrm{Id} + tL\big)_* \mu$ for $0 \le t \le 1$. This is also called the constant speed geodesic, because

$$W_2(\lambda_s, \lambda_t) = |t - s| W_2(\mu, \nu)$$

for any $0 \le s, t \le 1$. For comparison, one can check that for δ_x and δ_y,

$$W_2(\delta_x, \delta_y) = \sqrt{t - s}\, W_2(\sigma_s, \sigma_t),$$

if we set $\sigma_t = (1-t)\sigma_x + t\sigma_y$, the *straight-line* interpolation between point x and point y.

monotone transport for two

infimum convolution

displacement convexity for three types of energies

internal energy

potential energy

interaction energy

16.35 Benamou–Brenier.

[2] The cost $|x - y|^2/2$ is chosen in the proof for computation. Also absolute continuity with respect to the Lebesgue measure can be replaced by absolute continuity with respect to the H^{d-1} measure.

16.H.1 Otto's calculus

16.H.2 Entropy-regularized optimal transport

Define

$$\mathrm{EOT}_\epsilon(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \frac{1}{2}\int \|y-x\|^2\, d\pi + \epsilon\,\mathrm{Ent}(\pi).$$

This is the optimization problem that aims to minimize a strictly convex function over the convex set $\Pi(\mu,\nu)$. The minimum can thus be uniquely attained, which is known as the *Schrödinger bridge*. Notice that $\mathrm{EOT}_0(\mu,\nu) = \frac{1}{2}W_2^2(\mu,\nu)$, and one should expect the Schrödinger bridge to converge to the Brenier solution as $\epsilon \to 0$.

McKean–Vlasov process

16.H.3 Martingale optimal transport

16.I Mathematical finance

arbitrage free
- fundamental theorem of asset pricing
- complete
- discount factor
- risk-neutral measure
- no free lunch with vanishing risk (NFLVR)
- self-financing

16.J Local times

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)[3] 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 $f\colon \mathbf{R} \to \mathbf{R}$, the set of discontinuities is countable.

A.6 Proposition. Given a set A in a metric space (X, d), the function $\operatorname{dist}(\,\cdot\,, A)\colon X \to [0, \infty)$ given by

$$\operatorname{dist}(x, A) = \inf\{d(x, y) : y \in A\}$$

is a continuous function. Also $\operatorname{dist}(x, A) = 0$ 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.

We discuss the *completion of a metric space* X. Consider the set of all Cauchy sequences in X. We identify two Cauchy sequences $\{x_n\}$ and $\{y_n\}$ if $d(x_n, y_n) \to 0$, which gives an equivalence relation $\sim$. The completion $\widehat{X}$ of X is defined to be the set of all Cauchy sequences in X quotient this equivalence relation. It is clear to see that $\widehat{X}$ is *the* complete metric space that contains X as a dense subset.

[3] to ensure that the sequential limit must be unique; actually not necessary for this proposition

If X is a normed vector space, then we can define a norm on $\widehat{X}$ by the recipe $\|[x_n]\|_{\widehat{X}} = \lim_n \|x_n\|_X$. If X is an inner product space, then we can define an inner product on $\widehat{X}$ by the recipe $\langle [x_n], [y_n] \rangle_{\widehat{X}} = \lim_n \langle x_n, y_n \rangle_X$. It should be an easy exercise to check that $\|\cdot\|_{\widehat{X}}$ and $\langle \cdot, \cdot \rangle_{\widehat{X}}$ are indeed a norm and an inner product. Hence we may complete an normed space to be a Banach space, and a inner product space to be a Hilbert space.

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.

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 $f\colon X \to [0,1]$ 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

$$f(x) = \frac{\operatorname{dist}(x, A)}{\operatorname{dist}(x, A) + \operatorname{dist}(x, B)}.$$

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 $f\colon X \to [0,1]$ 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.

Therefore compact metric spaces are separable because they are Lindelöf. A topological space is *σ-compact* if it can be written as a countable union of compact sets. It follows that a σ-compact metric space must be separable. Conversely, a *locally compact* separable metric space must be σ-compact.

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.

In the Euclidean space we know compactness is equivalent to closed and bounded. In a metric space we need something stronger than boundedness for the equivalence to hold. We say a subset of a metric space is *totally bounded* if it can be covered by a finite ϵ-net for any $\epsilon > 0$. This is stronger than being bounded in general.

A.20 Theorem (Characterization of compactness in metric spaces). A subset A of a metric space X is totally bounded if and only if every sequence has a Cauchy subsequence.

Therefore A is compact if and only if it is sequentially compact if and only if it is totally bounded and complete.[4]

Sequentially compact implies completeness, because a given Cauchy sequence has a convergent subsequence, and we can then use triangular inequality. (This is exactly how we prove real Cauchy sequence to be convergent.)

There is a close connection between precompactness (relative compactness) and total boundedness. Say A is precompact in X, then every sequence in A has a subsequence that converges in $\overline{A}$, and this subsequence is Cauchy. Hence A is totally bounded.

It should be an easy exercise to check that if A is totally bounded in X, then so is $\overline{A}$. To get a converse we need to we assume in addition that X is complete. then $\overline{A}$ must be complete as well. Therefore $\overline{A}$ is compact the theorem above, showing that A is precompact. In summary,

[4] In the usual proof of both directions of this result, DC is used. The alternative proof that only requires CC is given in [Her06, Proposition 3.26].

A.21 Proposition. Precompactness implies total boundedness, and is equivalent to the latter in a complete metric space.

This proves the first part of

A.22 Generalized Arzelà–Ascoli theorem. Let X be a compact Hausdorff[5] space, then the collection $\mathcal{F}$ is precompact in the metric space $C(X)$ if and only if totally bounded if and only if it is equicontinuous and pointwise bounded.

$\sup_{f\in\mathcal{F}}|f(x)|<\infty$ for each $x\in X$
pointwise bounded and equicontinuous implies uniformly bounded.

Every equicontinuous and pointwise bounded sequence of continuous functions on X has a uniformly convergent subsequence. The classical Arzelà–Ascoli theorem states the special case that a sequence of equicontinuous and uniformly bounded functions contains a uniformly convergent subsequence. One should know/recognize that this can be directly proved using the diagonal argument, a technique used throughout the whole text.

A.23 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.24 Theorem. Let X and Y be metric spaces, with Y being complete. Let D be a dense subspace of X, and $f\colon D\to Y$ 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.23.

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.

[5] Hausdorff is not necessary, but usually stated.

A.25 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.26 Mean value inequality for $\mathbf{R}^d$-valued functions [Rud76, Theorem 5.19]. Let $f\colon [a,b] \to \mathbf{R}^d$ be continuous, and f be differentiable in (a,b), then there exists $x \in (a,b)$ such that

$$|f(b) - f(a)| \leq (b-a) \sup_{a<x<b} |f'(x)|.$$

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.27 Mean value inequality for $\mathbf{C}$-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.28 Uniform convergence of derivatives. 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 $f' = g$.[6]

The key part of the proof is the use of the mean value theorem on $f_n' - f_m'$.

A.29 Fact.

A.30 Tychonoff's theorem. Arbitrary product of compact topological spaces is compact.

A.31 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.32 Exercise. Give a direct proof of Tychonoff's theorem for the countable product of compact metric spaces, using a metric on this product.

A.33 Theorem. The countable product of sequentially compact spaces is sequentially compact.

[6] [Rud76, Theorem 7.17]; also see Theorem 8.15 and Remark 8.16 in [Kra22].

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. Given a real vector space $(V, \leq)$, where "$\leq$" is a partial order that obeys vector addition and scalar multiplication:

$$x \leq y \implies \begin{cases} x + z \leq y + z & \text{for } z \in V, \\ \lambda x \leq \lambda y & \text{for } \lambda \in \mathbf{R}^{\geq 0}. \end{cases}$$

We say $f \in V^*$ is a *positive linear functional* if $x \geq 0$ implies $f(x) \geq 0$.

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 in a normed space if and only if the normed space is finite-dimensional.

Therefore a normed space is locally compact if and only if it is finite-dimensional. Hence an infinite-dimensional separable Banach space is a Polish space that is not locally compact.

The "if" direction follows by the fact that the finite dimensional normed space is homeomorphic to $\mathbf{R}^n$.[7] For the contrapositive of the only if direction, we use Riesz' lemma to inductively choose a sequence in the infinite-dimensional space that has distance $\geq 1/2$ from the existing finite-dimensional subspace.

B.7 Fact. Let E be a dense subspace of a normed space X, then E^* and X^* can be isometrically identified in a natural way since a continuous function[8] is uniquely determined by its value on a dense subset.

B.8 Proposition. If Y is complete, then $\mathcal{L}(X, Y)$ is complete. In particular the dual space of any normed space is complete.

B.9 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[9] 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 $x \in X$.

Note we can always define the seminorm p by $p(x) = \|x\| \, \|f\|_{E^*}$, and it becomes immediately clear that an extension F of f always exists and can be made so that the norm is preserved: $\|F\|_{X^*} = \|f\|_{E^*}$.

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 C, 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 if p_0 is a linear functional, then $p = |p_0|$ is a seminorm, and since p_0 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 are the existence of nontrivial linear functionals that satisfy certain properties.

B.10 Corollary. Let X be a normed space.

(a) Let E be a closed proper subspace of X. Take any $x \in X - E$, then there exists $f \in X^*$ such that $f(x) = \inf_{v \in E} \|x - v\| = \delta \neq 0$, $f|_E \equiv 0$, and $\|f\| = 1$.

(b) For $x \neq 0_X$, there exists $f \in X^*$ such that and $f(x) = \|x\|$ and $\|f\| = 1$.

[7] This in fact holds for more general spaces as well, see Proposition C.2.

[8] consider $f \in X^*$ and $\frac{|f(x)|}{\|x\|}$ in our context

[9] Note that seminorms are a subclass of sublinear functionals that are always nonnegative. The absolute value signs that pop up later are expected.

(c) For any $f \in X^*$, there exists $x, y \in X$ such that $f(x) \neq f(y)$.

Proof. Part (b) and (c) follow quickly from (a). For part (a), apply Hahn–Banach theorem to the subspace $E + \mathbf{F}x$ with the linear functional $f\colon y + \lambda x \mapsto \lambda\delta$ and the dominating seminorm $p = \|\cdot\|$. □

We discuss some elementary use of this corollary. What happens if in Fact B.7, E is not dense in X? This is a conceptually important question.

B.11 Fact. For any closed proper subspace E of X, then we know every linear functional on E can be extended to a linear functional on X. Furthermore, by Corollary B.10(a), there exists a nontrivial linear functional on X that vanishes on E. Therefore E^* is also properly contained in X^*. In fact, it is true that the first isomorphism holds

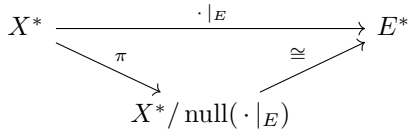
The E above can be replaced by any subspace of X that is not dense, since we can always take its closure.

Actually this is not the best we can get. As you might have already felt, the first isomorphism should hold; see the proof of the next proposition.

It follows by Corollary B.10(b)(c) that the hat map[10] $\hat{\cdot}\colon X \to X^{**}$ 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. The primary example of a reflexive space is the Hilbert space, the L^p space, and the Sobolev spaces $W^{k,p}$ and $W_0^{1,p}$ when $1 < p < \infty$.

B.12 Proposition. The closed subspace of a reflexive space is reflexive.

Proof. Part (a) of Corollary B.10 tells us that given a closed subspace E of a normed space X, there exists some linear functional on X that vanishes on E. We can characterize all linear functionals on E by these linear functionals on X that vanishes on E: according to the first isomorphism theorem



we have

$$E^* \cong X^*/E^\perp, \tag{B.13}$$

where $E^\perp = \{f \in X^* : f(E) = 0\}$ is called the *annihilator* of E in X, a closed subspace of X.[11] Since E is closed in X, E^* is closed in X^* as well, and hence

$$\begin{aligned} E^{**} &\cong (X^*/E^\perp)^* \cong (E^\perp)^\perp && \text{(B.14)}\\ &= \{\hat{x} \in X^{**} : \hat{x}(f) = 0 \text{ for all } f \in E^\perp\} \\ &\cong \{x \in X : f(x) = 0 \text{ for all } f \in E^\perp\} \\ &= \overline{E} = E. && \text{(B.15)} \end{aligned}$$

This shows that the isometric embedding is surjective. □

[10] denoted by the letter J as well

[11] The notation suggests that it bears connection with the *orthogonal complement* for Hilbert spaces.

Equation (B.15) is [BS18, Corollary 2.3.24], while the two isomorphisms (B.13) and (B.14) are rigorously proved in [BS18, Corollary 2.3.26]. In fact both are isometric isomorphisms. (The finite-dimensional case is stated in [Axl24, Section 3F, Exercises 31 & 33].)

B.16 Theorem. Let X be a normed space and E be a subspace. Then

(a) $\{x \in X : f(x) = 0 \text{ for all } f \in E^\perp\} = \overline{E}$.
Hence E is dense in X if and only if $E^\perp = \{0\}$.

(b) the linear map
$$X^*/E^\perp \to E^* : [f] \to f|_E$$
is an isometric isomorphism.

(c) If furthermore E is a closed subspace, then
$$(X/E)^* \to E^\perp : \Lambda \mapsto \Lambda \circ \pi$$
is an isometric isomorphism. Here $\pi : X \to X/E$ is the quotient map.

Bühler and Salamon [BS18, Theorem 2.4.4] also provides a proof of Proposition B.12 that works directly with the elements in the spaces. The same theorem also proves that

B.17 Proposition. A Banach space X is reflexive if and only if X^* is reflexive.

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.18 Uniform boundedness principle. 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 $\sup_\alpha \|T_\alpha\| < \infty$.[12]

B.19 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.20 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.

B.21 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.22 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 (which includes any locally compact second countable Hausdorff 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.

[12] also known as the *Banach–Steinhaus theorem*

B.23 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 Weak topologies and topological vector spaces

Some motivation is needed before we start the main material of this section.

$f\colon X \to Y$ 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 $\mathcal{F} = \{f_\beta\colon X \to Y_\beta\}_{\beta\in B}$ if and only if $f(x_\alpha) \to f(x)$ for all $f \in \mathcal{F}$. This is true for both nets and sequences.

In particular, this applies 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.

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 $\langle\cdot,\cdot\rangle\colon X \times Y \to \mathbf{F}$. 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 two new types of convergence on vector spaces and their dual spaces. We will start by discussing them given a normed space X.

For $\{x_n\} \subseteq X$, we say $x_n \to x$ weakly (i.e., converges in the weak topology) if for all $f \in X^*$, $f(x_n) \to f(x)$.

For $\{f_n\} \subseteq X^*$, we say $f_n \to f$ in weak-star (i.e., converges in the weak-star topology) if for all $x \in X$, $\hat{x}(f_n) = f_n(x) \to \hat{x}(f) = f(x)$.

Both limits are unique, but for very different reasons. Say $x_n \to x$ and y weakly, then $x \neq y$ if and only if there exists f that $f(x) \neq f(y)$. This is a clear consequence of the Hahn–Banach theorem.

On the other hand, say $f_n \to f$ and $f_n \to g$ in weak-star, then $f(x) = g(x)$ for all $x \in X$, and hence $f = g$. Be very aware that this unique limit f does not have to be a *bounded* linear functional, i.e., an element in X^*. Of course, if $\sup_n \|f_n\| < \infty$, the weak-star limit f has to be bounded. (This is true automatically when X is Banach, by the uniform boundedness principle: for each x, since $\{f_n(x)\}$ is a convergent sequence, it is bounded, and hence $\sup_n \|f_n\| < \infty$.)

One can also impose the weak topology on X^*. When X is a reflexive Banach space, $\sigma(X^*, X)$ and $\sigma(X^*, X^{**})$ coincides, but this is not true in general.

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 at 0 is made of locally convex sets

$$\big\{x \in X : p_j(x) < \epsilon \text{ for all } j \in [n]\big\}$$

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].[13] The weak topology is just a canonical example of the general seminorm topologies on locally convex 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

$$d(x, y) \coloneqq \sum_{j=1}^{\infty} 2^{-j} \frac{p_j(x - y)}{1 + p_j(x - y)}.$$

Of course, $d(x, y) \to 0$ if and only if $p_j(x, y) \to 0$ for all j. 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 $\mathcal{S}(\mathbf{R}^n)$ from Fourier analysis is the primary example.

[13] 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.

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

$$T \mapsto f(Tx) \text{ over } x \in X, f \in Y^*.$$

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.

C.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.

C.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. □

.

C.3 Proposition [Rud91, Theorem 1.22]. A topological vector space is locally compact if and only if it is finite-dimensional.

C.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.

We need to show that in an infinite-dimensional normed space X, 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 $F\colon X \to \mathbf{F}^n$ given by

$$F(x) = \big(f_1(x), \dots, f_n(x)\big).$$

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.

If X is a finite-dimensional normed space, we claim that the metric ball $\{x\colon \|x\| < \epsilon\}$ around 0_X is weakly open. Take X to be $\mathbf{R}^n$ with supremum norm, then

$$\{x\colon \|x\| < \epsilon\} = \bigcap_{j=1}^{n} \{x\colon |x_j| < \epsilon\} = \bigcap_{j=1}^{n} \{x\colon e_j^*(x) < \epsilon\},$$

where $\{e_j^*\}$ is the dual basis with respect to the standard basis $\{e_j\}$. Since the dual basis elements are continuous linear functionals, the proof is complete. □

C.5 Proposition. Suppose X and Y are topological spaces defined by seminorms $\{p_\alpha\}_{\alpha\in A}$ and $\{q_\beta\}_{\beta\in B}$ respectively. Say T is a linear map, then T is continuous if and only if for each $\beta \in B$, there exists $\alpha_1, \dots, \alpha_k$ such that

$$q_\beta(Tx) \le C[p_{\alpha_1}(x) + \cdots + p_{\alpha_k} x].$$

C.6 Proposition. Suppose $x_n \to x$ weakly, then $\sup_n \|x_n\| < \infty$, and $\|x\| \le \liminf_n \|x_n\|$.

C.7 Proposition.

C.8 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.

Separability of the space and the weak and weak-star topology on the closed unit ball plays an important role in some results.

C.9 Theorem [Bre11, Theorem 3.28]. Let X be a normed[14] space, then the closed unit ball

$$\{f \in X^* : \|f\| \le 1\}$$

in X^* is metrizable in the weak-star topology if and only if X is separable.

[14] Brezis [Bre11] assumes that X to be Banach space, but this assumption is used nowhere in the proof. Also dense normed subspace of a Banach space has the same dual, and dense subset of a separable space is always separable.

For the more important "if" direction, one can define a norm $[f] = \sum_{n=1}^{\infty} 2^{-n}|f(x_n)|$, where $\{x_n\}$ is a fixed countable dense subset in X. The topology induced from this norm is the same as the weak-star topology restricted to the closed ball.

One can find an alternative proof of the "if" direction from [BS20, Theorem 6.10.23]. It uses the following result, which is invoked throughout our text.

C.10 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 $\sup_n |f_n(x_1)| \leq C\|x_1\|$, $\{f_n(x_1)\}$ lives in a bounded interval, there is a subsequence $\{f_{\nu(n)}(x_1)\}$ that converges.[15] 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.24, 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

$$f_n^n(x_j) - C\epsilon \leq f_n^n(x) \leq f_n^n(x_j) + C\epsilon$$

Let f satisfy $f(x_j) = \lim_n f_n^n(x_j)$ for all j, then taking limits we have

$$f(x_j) - C\epsilon \leq \liminf_n f_n^n(x) \leq \limsup_n f_n^n(x) \leq f(x_j) + C\epsilon.$$

It follows that

$$\limsup_n f_n^n(x) - \liminf_n f_n^n(x) \leq 2C\epsilon,$$

[15] This can be done explicitly by letting

$$\nu(n) = \min\{m > \nu(n-1) : |f_m(x_1) - s_1| < 1/n\},$$

where $s_1 \coloneqq \liminf_n f_n(x_1)$.

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

$$|f(x)| = \lim_n |f_n^n(x)| \le C.$$

for $x \in X$ with unit norm, which shows that $f \in X^*$ with $\|f\| \le C$, as desired. □

One important consequence of sequential Banach–Alaoglu theorem is a sequential characterization of reflexive Banach spaces. This essentially explains why (separable) reflexive Banach spaces are desirable in some applications.

C.11 Eberlein–Shmulian theorem. If a Banach space X is reflexive, then it is weakly sequentially compact in the weak topology $\sigma(X, X^*)$.

Proof. First assume X is in addition separable, then X^{**} is separable and hence X^* is separable. Also X^* is reflexive. By the sequential Banach–Alaoglu theorem, we know the weak-star topology $\sigma(X^{**}, X^*)$ on X^{**} is weak-star sequentially compact, and hence X is weakly sequentially compact.

Now we drop the assumption that X is separable. Given a norm-bounded sequence $\{x_n\}$, define $E = \overline{\operatorname{span}\{x_n\}_{n=1}^\infty}$, which as the closure of a separable space is separable. By Proposition B.12 we know E is also reflexive, and hence the preceding paragraph tells us that $\{x_n\}$ has a weakly convergence subsequence with the same norm bound. □

The converse turns out to be true as well, but we do not discuss here.

Going back to sequential Banach–Alaoglu theorem, from Theorem C.9, we know when the normed space is separable, the weak-star sequential compactness and weak-star compactness coincide. In fact, we have

C.12 Banach–Alaoglu theorem. For a (not necessarily separable) normed vector space X, every closed and bounded-in-norm subset of X^* is weak-star compact.

Be aware that weak-star compactness and weak-star sequentially compactness are not the same when X is nonseparable. In fact, neither of the two implies the other.[16]

The proof of the *topological* Banach–Alaoglu theorem is nonconstructive (it requires Tychonoff's theorem for arbitrary product of the unit interval). The result is also less interesting because we are generally more interested in convergence of sequences. However, the statement do provide some additional insights to weak topology and reflexive spaces. The sequential version may also be proved from the topological version below by metrizing the closed unit ball in the dual space (again by Theorem C.9), but we do not recommend this approach.

C.13 Exercise. The converse of Banach–Alaoglu theorem is also correct when X is a Banach space. (Hint: use the uniform boundedness principle)

Here is the topological characterization of reflexive Banach spaces, which is quite difficult to prove.

C.14 Kakutani's theorem. A Banach space is reflexive if and only if the closed unit ball is compact in the weak topology.

We ask the reader to consult [Bre11, Theorem 3.17, 3.18, & 3.19] or [BS18, Theorem 3.4.1] for dedicated discussion of .

A set S in a vector space is called *balanced* if $\lambda S \subseteq S$ for all $|\lambda| \le 1$.

[16] There are explicit counterexamples.

D 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.

D.1 Theorem. For $T \in \mathcal{L}(H)$, there is a unique $T^* \in \mathcal{L}(H)$ such that $\langle T^*x, y\rangle = \langle x, Ty\rangle$ for all $x, y \in H$. This T^* is known as the *adjoint* of T, which has the following properties:

(a) $\|T^*\| = \|T\|$, $\|T^*T\| = \|T\|^2$, $T^{**} = T$,

$$(\operatorname{range} T)^\perp = \operatorname{null} T^* \quad \text{and} \quad (\operatorname{null} T)^\perp = \overline{\operatorname{range} T^*}.$$

This T^* is called the Hilbert adjoint

Keep in mind that there is a canonical isomorphism between a Hilbert space and its dual, but no such canonical isomorphism between a Banach space and its dual. This is the key reason why we have to develop the notion of adjoint differently for Hilbert spaces and general normed spaces.[17]

Let X and Y be two normed spaces. For $T \in \mathcal{L}(X, Y)$, define its norm adjoint $T^*\colon Y^* \to X^*$ by

$$T^*f(x) = f(Tx) \quad \text{for all } f \in Y^* \text{ and } x \in X.$$

This T^* turns out to be bounded and also satisfies $\|T^*\| = \|T\|$. (One direction uses Hahn–Banach.)

In the case Y and X are Hilbert spaces, by Riesz representation theorem (Hilbert space), we can make the identification between $f \in Y^*$ and some $y \in Y$ via $f(x) = \langle x, y\rangle$. Under the definition of the Banach adjoint, we have

$$\langle Tx, y\rangle = f(Tx) = T^*f(x) = \langle x, T^*y\rangle,$$

which is fortunately consistent with the Hilbert space adjoint notation. However, in the case of a complex Hilbert space we have

$$(\lambda T)^* = \bar{\lambda} T^*,$$

compared to

$$(\lambda T)^* = \lambda T^*$$

for a Banach space. Because we have a conjugate linear isomorphism in Riesz' theorem, such subtlety appears.

For X that is a finite-dimensional inner product space, with respect an orthonormal basis T^* is just the conjugate transpose of T: let * be the notation for the conjugate transpose, then

$$\langle Tx, y\rangle = y^*Tx = y^*(T^*)^*x = \langle x, T^*y\rangle.$$

unitary operators

annihilators

Closed range theorem for Banach spaces

An unbounded operator from Banach space X to Y is a linear map $T\colon D(T) \to Y$ only defined on a normed subspace of X.

[17] because they belong to different categories

For unbounded operators on Hilbert spaces, self-adjoint operators are stronger than symmetric operators. An densely defined operator T is symmetric if

$$\langle Tx, y\rangle = \langle x, Ty\rangle \quad \text{for all } x, y \in D(T),$$

and T is self-adjoint if furthermore $D(T) = D(T^*)$, i.e., $T = T^*$ as operators.

The adjoint T^* is defined as follows. We declare $y \in D(T^*)$ if there exists $z \in H$ such that $\langle Tx, y\rangle = \langle x, z\rangle$ for all $x \in D(T)$, and we say $z = T^*y$.

Given a linear map $T\colon D(T) \subseteq X \to X$, its resolvent set is defined as

$$\rho(T) = \{\lambda \in \mathbf{C} : (\lambda I - T) \text{ is invertible}\},$$

which consists of the regular points of T. (By invertible we mean $\lambda I - T$ maps $D(T)$ one-to-one onto X.) When λ is a regular point, $(\lambda I - T)^{-1}$ is called the resolvent of the operator T, which we denote by $R_\lambda(T)$. By Corollary B.20 we know when T is bounded the resolvent must be bounded as well.

We stress that many people define R_λ to be $(T - \lambda I)^{-1}$ instead, which can lead to slightly different conclusions. Our choice is consistent with the semigroup literature.

The resolvent satisfies an important identity (attributed to Hilbert). For $\lambda, \mu \in \rho(T)$,

$$R_\lambda(T) - R_\mu(T) = (\mu - \lambda)R_\mu(T)R_\lambda(T) = (\mu - \lambda)R_\lambda(T)R_\mu(T).$$

This can be easily verified as follows:

$$(\mu I - T)[R_\lambda(T) - R_\mu(T)](\lambda I - T) = (\mu - \lambda)I = (\lambda I - T)[R_\lambda(T) - R_\mu(T)](\mu I - T),$$

and multiply R_μ and R_λ on the two sides of the expression to get cancellation.

[BS20] 7.8.6 For a self-adjoint bounded operator T 7.10.8 .9 10.3.3

projection-valued measure and resolutions

D.2 Spectral decomposition of self-adjoint operators.

E Semigroups

On the Banach space X, the family of bounded operators $\{T_t\}_{t\geq 0}$ is a strongly continuous one-parameter semigroup[18] if it satisfies the semigroup properties

(a) $T_0 = I$;

(b) $T_{t+s} = T_t \circ T_s$;

and the continuity of $t \mapsto T_t$ in the strong operator topology:

(c) $\|T_t x - x\| \to 0$ as $t \to 0$.

We are often interested in a strongly continuous semigroup that satisfies in addition

(d) $\sup_t \|T_t\| \leq 1$,

[18] It is also called a C_0 semigroup, where C_0 stands for strong continuity and can be confusing. We will now use this name.

which is called a strongly continuous contraction semigroup. In this section all semigroups will be assumed to be strongly continuous by default, but not necessarily contraction. However, all the results should be understood in the context of contraction semigroups, the only case we are interested in. One can make the above definitions for the one-parameter groups $\{T_t\}_{t\in\mathbf{R}}$. Note that for semigroups any limit at zero is one-sided, but for groups the limit should be two-sided.

As a consequence of the uniform boundedness principle, since $\sup_{t\in[0,1]}\|T_t x\| \le \infty$, we can define $M = \sup_{t\in[0,1]}\|T_t\| < \infty$. It follows that

$$\|T_t\| \le M\|T_1\|^{\lfloor t\rfloor} \le M\max\{M,1\}^t = Me^{\beta t}$$

for some $\beta \ge 0$. This allows us to define a new semigroup $S_t = e^{-\beta t}T_t$ that is uniformly bounded by M.

These computations are trivially true for contraction semigroups (where $M = 1$ and $\beta = 0$), but it still inspires us to look at the λ-resolvent

$$R_\lambda x = \int_0^\infty e^{-\lambda x}T_t x\,dt$$

defined for any $\lambda \in \mathbf{C}$ such that $\operatorname{Re}\lambda > \beta$. (When $\lambda > 0$ this is just the Laplace transform.) The improper integral here is in the Riemann sense, and the limit is in norm. It evaluates how much $T_t x$ grows over its trajectory.

The infinitesimal generator of a strongly continuous semigroup $\{T_t\}_{t\ge0}$ is defined by

$$Lx = \lim_{t\to0}\frac{T_t x - x}{t}$$

in norm, whenver the limit exists at x. In general, the operator L is an unbounded operator, and we would like to characterize $D(L)$.

[BS20, Sections 10.5 & 10.6]

semigroups induced from bounded operators

For a *bounded* operator L, its exponential $T_t = \exp(tL) = \sum_{k=0}^\infty \frac{t^k L^k}{k!}$ is a semigroup whose generator is L

A is a self-adjoint operator on H with $A \ge 0$, $\{\exp(-tA)\}_{t\ge0}$ is a strongly continuous semigroup with generator $-A$.

For self-adjoint operator on H, $\{\exp(itA)\}_{t\in\mathbf{R}}$ is a strongly continuous group of unitary operators with generator iA.

One have to mention the famous Schrödinger equation from quantum physics here: on a complex separable Hilbert space, we have

$$\widehat{H}\psi = i\hbar\,\partial_t\psi,$$

which describes the evolution of the quantum state $\psi(t)$ over time. Here $\widehat{H}$ is an unbounded self-adjoint operator (usually semibounded from below), called the Hamiltonian observable of the system.

We know that $U_t = \exp\big(-it\frac{\widehat{H}}{\hbar}\big)$ defines a one-parameter unitary group over $t \in \mathbf{R}$. We now verify $\psi(t) = U_t\psi(0)$, which we hinted above.

First, U_t has generator $-i\frac{\widehat{H}}{\hbar}$. Now $\partial_t U_t\psi(0) = -i\frac{\widehat{H}}{\hbar}U_t\psi(0)$, which implies the Schrödinger equation by replacing $\psi(t) = U_t\psi(0)$.

We remark that that the Schrödinger equation is not just physically sensible and mathematically consistent, but also mathematically somewhat necessary. If we mandate that the quantum state must evolve according to a unitary operator U_t, then the operator must be of the form $\exp(it\frac{-\widehat{H}}{\hbar})$, and the Schrödinger equation follows from there.

L is defined precisely on the image of the resolvent, i.e., $R_\lambda(X) = D(L)$. $(\lambda I - L)$ and R_λ are inverse operators to each other, and hence the name resolvent and its notation.

$\lim_{\substack{\lambda>0\\ \lambda\to\infty}} \lambda R_\lambda x = x$

The next proposition is very essential.

The generator L is closed and densely defined

If $T_t x \in D(L)$, then $\partial_t T_t x = L(T_t x)$.

If furthermore $x \in D(L)$, then $T_t x \in D(L)$, and hence $\partial_t T_t x = T_t(Lx) = L(T_t x)$.

It follows that $T_t x = x + \int_0^t L(T_s x)\,ds = x + \int_0^t T_s(Lx)\,ds$

$$\begin{aligned} LT_t x &= \lim_{h\to 0} \frac{T_{t+h}x - T_t x}{h} = \partial_t T_t x \\ &= \lim_{h\to 0} \frac{T_t(T_h x - x)}{h} \\ &= T_t\Big(\lim_{h\to 0} \frac{T_h x - x}{h}\Big) = T_t L x, \end{aligned}$$

because $T_t(\,\cdot\,)$ is continuous.

E.1 Proposition. Strongly continuous semigroups are uniquely determined by their generators. For two strongly continuous operator semigroups $\{S_t\}_{t\geq 0}$ and $\{T_t\}_{t\geq 0}$ on the Banach space X with the same generator, we have $S_t = T_t$ for all $t \geq 0$. (And hence the name generator.)

E.2 Theorem. For a contraction strongly continuous semigroup on a Hilbert space, L is self-adjoint and $L \leq 0$.

Conversely, if L is self-adjoint and $L \leq 0$, then $\{\exp(tL)\}_{t\geq 0}$ is a strongly continuous contraction semigroup of self-adjoint operators with generator L.

E.3 Stone's theorem. Given a strongly continuous group of unitary operators $\{U_t\}$ on a Hilbert space, then its generator L must be iA for some self-adjoint operator A. This means that $U_t = \exp(itA)$.

This result is related to Schrödinger's equation.

An operator L generates a strongly continuous contraction semigroup implies that every $\lambda > 0$ belongs to the resolvent set of L, and

$$\|R_\lambda\| = \|(\lambda I - L)^{-1}\| \leq 1/\lambda.$$

In addition, recall that L is closed and densely defined. We now state the converse of the result.

E.4 Hille-Yoshida theorem. Suppose L is densely defined and for every $\lambda > 0$, the operator $\lambda I - L$ has a bounded inverse $R_\lambda \colon X \to D(L)$ that satisfies $\|R_\lambda\| \leq 1/\lambda$, then L can generate a strongly continuous contraction semigroup.

It is usually easier to show that the range of $\lambda I - L$ is only dense in X, and in that case one may use the closure of L to generate the semigroup.

An unbound operator on a Banach space X is densely defined if for any $x \in D(L)$ and $\lambda > 0$ we have

$$\|(\lambda I - L)x\| \geq \lambda \|x\|.$$

This is clearly true for contraction semigroups. Just like for Hille–Yoshida, we have the corresponding converse.

E.5 Lumer–Phillips theorem. Suppose L is densely defined and dissipative, and the operator $\lambda I - L$ is surjective for some (and hence all) $\lambda > 0$, then L can generate a strongly continuous contraction semigroup.

It is usually easier to show that the range to $\lambda I - L$ is only dense in X, and in that one may use the closure of L to generate the semigroup.

F Convex geometry, optimization, and analysis

Let X be a nonempty vector or topological space, and let $f\colon X \to \overline{\mathbf{R}}$ throughout this section.

Beyond the vanilla convex functions, there are two stronger notions of convexity that are often useful. Let X be a vector space, and consider $\varphi\colon X \to (-\infty, \infty]$. The function is *convex* if

$$\varphi\big((1-t)x + ty\big) \leq (1-t)\varphi(x) + t\varphi(y) \quad \text{for all } 0 < t < 1 \text{ and } x, y \in X,$$

and is *strictly convex* if the inequality is strict. The domain of a convex function φ is $\{x \in X : \varphi(x) < +\infty\}$. If X is furthermore a normed space, then the function φ is strongly convex with parameter $\lambda > 0$ if

$$\varphi\big((1-t)x + ty\big) + \frac{\lambda}{2}(1-t)t\|y-x\|^2 \leq (1-t)\varphi(x) + t\varphi(y)$$

for all $0 < t < 1$ and $x, y \in X$. This means precisely that there is a second order gap between the linear interpolation and the graph itself.

It should be easy to verify that φ is λ-strongly convex if and only if $x \mapsto \varphi(x) - \frac{\lambda}{2}\|x\|^2$ is convex. We also remark that if λ is set to 0, then this is just the vanilla convexity.

If we assume $\varphi \in C^1(\mathbf{R}^n)$, then strong convexity is equivalent to

$$\varphi(y) \geq \varphi(x) + \langle \nabla\varphi(x), y - x\rangle + \frac{\lambda}{2}\|y-x\|^2. \tag{F.1}$$

In fact, it is also equivalent to a strong monotonicity of the gradient

$$\langle \nabla\varphi(y) - \nabla\varphi(x), y - x\rangle \geq \lambda\|y-x\|^2. \tag{F.2}$$

To see this equivalence, switching the x and y in (F.1) gives

$$\varphi(x) \geq \varphi(y) + \langle \nabla\varphi(y), x - y\rangle + \frac{\lambda}{2}\|y-x\|^2,$$

Once we add it back to (F.1), inequality (F.2) follows. For the other direction,

$$\begin{aligned}
\varphi(y) - \varphi(x) &= \int_0^1 \frac{1}{t}\big\langle \nabla\varphi\big(x + t(y-x)\big), t(y-x)\big\rangle\, dt \\
&\geq \int_0^1 \frac{1}{t}\Big(t\big\langle \nabla\varphi(x), y-x\big\rangle + \lambda t^2\|y-x\|^2\Big)\, dt \\
&= \langle \nabla\varphi(x), y-x\rangle + \frac{\lambda}{2}\|y-x\|^2.
\end{aligned}$$

The fundamental theorem of calculus is necessary here, because we need to take the growing gradient condition over the entire straight line between x and y into account.

If we assume further that $\varphi \in C^2(\mathbf{R}^n)$, then strong convexity precisely means the $D^2\varphi(x) \succeq \lambda I_n$ uniform over all x. This is a simple consequence of the second-order Taylor's theorem:

$$\varphi(y) = \varphi(x) + \langle \nabla\varphi(x), y - x\rangle + \frac{1}{2}\big\langle \nabla^2 f(\xi)(y-x), y-x\big\rangle,$$

where ξ is an intermediate point on the straight line between x and y.

It is easy to show that strict convexity of $\varphi\colon \mathbf{R} \to \mathbf{R}$ is equivalent to saying that φ' is strictly increasing. However, from strict convexity we cannot conclude $\varphi''(x) > 0$ for all x, e.g., consider $\varphi(x) = x^4$ at the origin. The converse remains true, i.e., $\varphi'' > 0$ implies strict convexity.

The *epigraph* of f, denoted by $\operatorname{epi} f$, 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.

F.3 Fact. Let X be convex. The function f is convex if and only if its epigraph is convex.

A function $f\colon X \to (-\infty, +\infty]$ 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 $f\colon X \to [-\infty, +\infty)$ 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) $\operatorname{epi} f$ is a closed in $X \times \mathbf{R}$.

geometric consequence of the Hahn–Banach 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

F.4 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

$$f(x) \leq t - \epsilon \text{ for all } x \in A \quad \text{and} \quad f(x) \geq t + \epsilon \text{ for all } x \in B.$$

F.5 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.

F.6 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.

Given a normed vector space X, and a function $\varphi\colon X \to (-\infty, \infty]$ we define the *Legendre convex dual* of φ to be the function $\varphi^*\colon X^* \to (-\infty, \infty]$ given by

$$\varphi^*(f) = \sup_{x \in X} f(x) - \varphi(x).$$

If we have $\psi\colon X^* \to (-\infty, \infty]$, then we define

$$\psi^*(x) = \sup_{f \in X^*} f(x) - \psi(f).$$

(Technically we can define the ψ^* on the larger X^{**} to make the definition consistent over any normed vector space, but we chose not to.)

By definition, we have the following

F.7 Fenchel–Young inequality. For every $x \in X$ and $f \in X^*$, we have

$$f(x) \le \varphi(x) + \varphi^*(f).$$

To motivate the notion of convex conjugate, we may look at the following toy example. Consider the space $X = \mathbf{R}$ and $f(x) = px$ for any fixed slope $p \in \mathbf{R}$. We consider a function $\varphi \in C^2(\mathbf{R})$ such that $\varphi'' > 0$. This implies that $\varphi'\colon \mathbf{R} \to \mathbf{R}$ is strictly increasing, and φ must be strictly convex.

By definition,

$$\varphi^*(p) = \sup_x px - \varphi(x).$$

The supremum is attained at the x where $p = \varphi'(x)$, by taking derivative. Therefore

$$\varphi^*\big(\varphi'(x)\big) = \varphi'(x)x - \varphi(x).$$

Now taking derivatives on both sides, we have

$$(\varphi^*)'\big(\varphi'(x)\big) \cdot \varphi''(x) = \varphi'(x) + x\varphi''(x) - \varphi'(x), \tag{F.8}$$

which implies that

$$(\varphi^*)' \circ \varphi' = \mathrm{Id},$$

or equivalently $(\varphi^*)'(p) = (\varphi')^{-1}(p)$. If we show that $\varphi^{**} = \varphi$, then we may conclude that the derivatives of φ^* and φ are indeed inverse to each other.

Replacing $p = \varphi'(x)$ in (F.8), we now get

$$\varphi^*(p) = p\,(\varphi')^{-1}(p) - \varphi\big((\varphi')^{-1}(p)\big) = p\,(\varphi^*)'(p) - \varphi\big((\varphi^*)'(p)\big).$$

Using this, we can simplify

$$\varphi^{**}\big((\varphi^*)'(p)\big) = (\varphi^*)'(p)\,p - \varphi^*(p) = \varphi\big((\varphi^*)'(p)\big).$$

Since $(\varphi^*)' = (\varphi')^{-1}$ is bijective, we conclude that $\varphi^{**} = \varphi$, which proves also that φ^* and φ are genuinely inverses to each other, as desired.

The key takeaway from this computation is that the convex conjugates φ^* are designed so that the "derivatives" of φ and φ^* becomes inverses to each other. For the general case of φ defined on normed vector spaces, we have the notion of *subdifferential* $\partial\varphi(x)$ for any $x \in \operatorname{Dom} f$, which is the set

$$\{f \in X^* : \varphi(y) \geq \varphi(x) + f(y - x) \text{ for all } y \in X\}.$$

Notice that

$$f \in \partial\varphi(x) \iff f(x) = \varphi(x) + \varphi^*(f) \iff x \in \partial\varphi^*(f).$$

To see this, we now have the reverse of Fenchel–Young inequality:

$$\varphi(x) + \varphi^*(f) = \sup_y \varphi(x) + f(y - x) - \varphi(y - x) \leq \sup_y \varphi(y) - \varphi(y - x) \leq f(x).$$

F.9 Exercise. For $f_1 \in \partial\varphi(x_1)$ and $f_2 \in \partial\varphi(x_2)$, we have the monotonicity

$$(f_2 - f_1)(x_2 - x_1) \geq 0.$$

If φ and φ^* are differentiable (Lipschitz), then

$$\nabla u^* \circ \nabla u(x) = x \quad \text{and} \quad \nabla u \circ \nabla u^*(y) = y$$

everywhere (almost everywhere).

F.10 Fact. $\partial f(x) = \{\nabla f(x)\}$ at all points where $\nabla f(x)$ exists, which is understood to be

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

for all $y \in X$. (This is known as Gateaux derivative, which is just the normal gradient when $X = \mathbf{R}^n$. It makes sense over any locally convex TVS.)

F.11 Fenchel–Moreau theorem. Let $\varphi\colon X \to (-\infty, \infty]$ be a convex LSC function such that $\varphi \not\equiv +\infty$. (Such a function is also called to be *proper*.) We have $\varphi^{**} = \varphi$.

Expanding the definition, this is precisely

$$\varphi(x) = \sup_{f \in X^*} f(x) - \varphi^*(f)$$

This tells us that the set of

$$\text{all convex LSC function not identically } +\infty$$

is exactly the set of

$$\text{all supremums of affine functions that are not identically } +\infty.$$

(Since $f \in X^*$, it is an affine function.)

F.12 Fenchel–Rockafellar theorem. Let $\varphi, \psi\colon X \to (-\infty, \infty]$ be convex. Suppose there is $x_0 \in X$ such that $\varphi(x_0) < +\infty$, $\psi(x_0) < +\infty$, while φ and ψ are continuous at x_0. Then

$$\inf_{x\in X} \varphi(x) + \psi(x) = \max_{f\in X^*} -\varphi^*(-f) - \psi^*(f).$$

For extended convex functions, we do not necessarily have continuity.

If $C \subseteq X$ has nonempty interior, then $\overline{\operatorname{Int} C} = \overline{C}$.

We already know convex sets. A subset A of X is *affine* if for all $\lambda \in \mathbf{R}$ and $x, y \in A$,

$$(1-\lambda)x + \lambda y \in A.$$

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

$$p \neq (1-\lambda)x + \lambda y \quad \text{for any } 0 < \lambda < 1.$$

Let A be a subset of a vector space X, and Z be another vector space. A map $f\colon A \to f(A) \subseteq Z$ is *affine* if for any $x, y \in A$ and $\lambda \in \mathbf{R}$ such that $(1-\lambda)x + \lambda y \in A$,

$$f\big((1-\lambda)x + \lambda y\big) = (1-\lambda)f(x) + \lambda f(y).$$

In particular, affine maps take convex sets to convex images (convexity-preserving).

Given a set of points S in a vector space X, the *convex hull* $\operatorname{conv} S$ 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 $x_j \in S$, $0 \le \lambda_j \le 1$, 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 $\operatorname{conv} S$.

We can define *affine hull* similarly, without restricting λ_j to be nonnegative.

F.13 Krein–Milman theorem. A nonempty compact convex subset of a locally convex topological vector space is equal to the closed convex hall of its extreme points (which always exist).

A set $C \subseteq X$ is a cone if $x \in C$ implies $\lambda x \in C$ for all $\lambda > 0$.

[Bre11, Corollary 3.22 & 3.23]

F.14 Theorem.

direct method of calculus of variations

F.15 Radon's theorem. Any set of $d+2$ points in $\mathbf{R}^d$ can be partitioned into two subsets whose convex hulls intersect.

F.16 Carathéodory's theorem. Given some set $S \subseteq \mathbf{R}^d$, for any point in $\operatorname{conv} S$, it is the convex combination of at most $d+1$ points of S.

F.17 Helly's theorem. Let $A_1, A_2, \ldots, A_n$ be convex subsets of $\mathbf{R}^d$, where $n \ge d+1$. 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 $A_\gamma \cap A'$.)

F.18 Lemma. For $F\colon X \times Y \to [-\infty, \infty]$, we have

$$\sup_{x\in X} \inf_{y\in Y} F(x,y) \le \inf_{x\in X} \sup_{y\in Y} F(x,y).$$

G Proof of the two extension theorems

G.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

$$\mathcal{K}_E = \{F \subseteq X : E \cap F \in \Gamma\} \tag{G.2}$$

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 $A_1, A_2, \dots$ 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 (G.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.

G.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 $B_n \coloneqq \bigcup_{j=1}^{n} A_j \in \mathcal{M}$. 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

$$\mathcal{K} \coloneqq \{A \subseteq X : A^{\mathrm{c}} \in \mathcal{M}\}$$

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 $\mathcal{K}$,

$$\left(\bigcup_{j=1}^{\infty} A_j\right)^{\mathrm{c}} = \bigcap_{j=1}^{\infty} A_j^{\mathrm{c}} \in \mathcal{M}.$$

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 $B \supseteq A$.

H Existence theorems for probability measures on product spaces

It is noteworthy that all results here use the axiom of dependent choice in the proof.

H.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]

H.2 Ionesco-Tulcea existence theorem. For any sequence of measurable spaces $\{(S_n, \mathcal{S}_n)\}$ and kernels $\mu_n \colon S_1 \times \cdots \times S_{n-1} \to S_n$ 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

$$(X_1, \ldots, X_n) \sim \mu_1 \times \cdots \times \mu_n.$$

H.3 Nelson extension theorem [Fol99, Theorem 10.18].

I 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

$$\log(n+1) \leq \sum_{x=1}^{n} \frac{1}{x} \leq \log(n) + 1$$

I.1 Coupon collector's problem.

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.

[Axl24] Sheldon Axler. *Linear Algebra Done Right*. 4th ed. Springer, 2024.

[Bar98] F. Barthe. "Optimal young's inequality and its converse: a simple proof". *Geometric & Functional Analysis GAFA* 2 (Apr. 1998), pp. 234–242.

[Bas11] Richard F. Bass. *Stochastic Processes*. Cambridge University Press, Cambridge, 2011, pp. xvi+390.

[BGL14] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer International Publishing, 2014.

[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.

[Bog10] Vladimir I. Bogachev. *Differentiable Measures and the Malliavin Calculus*. American Mathematical Society, Providence, RI, 2010, pp. xvi+488.

[Bog18] Vladimir I. Bogachev. *Weak Convergence of Measures*. American Mathematical Society, 2018.

[Bog98] Vladimir I. Bogachev. *Gaussian Measures*. American Mathematical Society, Providence, RI, 1998, pp. xii+433.

[Bre11] Haïm Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer New York, 2011, pp. xiv+599.

[BS18] Theo Bühler and Dietmar A. Salamon. *Functional Analysis*. American Mathematical Society, Providence, RI, 2018, pp. xiv+466.

[BS20] Vladimir I. Bogachev and Oleg G. Smolyanov. *Real and Functional Analysis*. Springer International Publishing, 2020.

[BSW13] Björn Böttcher, René Schilling, and Jian Wang. *Lévy matters. III*. Lévy-type processes: construction, approximation and sample path properties, With a short biography of Paul Lévy by Jean Jacod, Lévy Matters. Springer, Cham, 2013, pp. xviii+199.

[Coh13] Donald L. Cohn. *Measure Theory*. 2nd ed. Birkhäuser/Springer, New York, 2013.

[CT05] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, John Wiley & Sons, 2005.

[DaP06] Giuseppe Da Prato. *An Introduction to Infinite-Dimensional Analysis*. Revised and extended from the 2001 original by Da Prato. Springer-Verlag, Berlin, 2006, pp. x+209.

[DaP14] Giuseppe Da Prato. *Introduction to Stochastic Analysis and Malliavin Calculus*. 3rd ed. Edizioni della Normale, Pisa, 2014, pp. xviii+279.

[Dud02] R. M. Dudley. *Real Analysis and Probability*. Revised reprint of the 1989 original. Cambridge University Press, Cambridge, 2002, pp. x+555.

[Dur19] Rick Durrett. *Probability: Theory and Examples*. 5th ed. Cambridge University Press, 2019.

[Dur96] Richard Durrett. *Stochastic calculus*. A practical introduction. CRC Press, Boca Raton, FL, 1996, pp. x+341.

[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.

[Fol23] Gerald B. Folland. *Advanced Calculus*. 2nd ed. self-published, 2023, pp. xi+465.

[Fol99] Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. 2nd ed. John Wiley & Sons, 1999.

[Gen08] Ivan Gentil. "From the Prékopa-Leindler inequality to modified logarithmic Sobolev inequality". en. *Annales de la Faculté des sciences de Toulouse : Mathématiques* 2 (2008), pp. 291–308.

[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.

[KS91] Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. 2nd Ed. Springer-Verlag New York, 1991, pp. xxiv+470.

[Lan86] Robert Lang. "A note on the measurability of convex sets". *Arch. Math. (Basel)* 1 (1986), pp. 90–92.

[Led01] Michel Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, Providence, RI, 2001, pp. x+181.

[LeG16] Jean-François Le Gall. *Brownian Motion, Martingales, and Stochastic Calculus*. Springer, 2016, pp. xiii+273.

[LeG22] Jean-François Le Gall. *Measure Theory, Probability, and Stochastic Processes*. Springer International Publishing, 2022.

[Lei72] L. Leindler. "On a Certain Converse of Hölder's Inequality". *Linear Operators and Approximation / Lineare Operatoren und Approximation*. Birkhäuser Basel, 1972, pp. 182–184.

[Lew86] Jonathan W. Lewin. "A Truly Elementary Approach to the Bounded Convergence Theorem". eng. *The American mathematical monthly* 5 (1986), pp. 395–397.

[Lin99] Torgny Lindvall. "On Strassen's Theorem on Stochastic Domination". *Electron. Comm. Probab.* (1999), pp. 51–59.

[Mal95] Paul Malliavin. *Integration and Probability*. With the collaboration of Hélène Airault, Leslie Kay and Gérard Letac, Edited and translated from the French by Kay, With a foreword by Mark Pinsky. Springer-Verlag, New York, 1995, pp. xxii+322.

[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.

[OCo00] N. O'Connell. *Information-theoretic proof of the Hewitt-Savage zero-one law*. Tech. rep. Technical Report. Bristol, U.K.: Hewlett–Packard Laboratories, June 2000.

[Par67] K. R. Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press, 1967.

[Pit82] Loren D. Pitt. "Positively correlated normal variables are associated". *Ann. Probab.* 2 (1982), pp. 496–499.

[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.

[Roy88] H. L. Royden. *Real analysis*. Third. Macmillan Publishing Company, New York, 1988, pp. xx+444.

[Rud69] Mary Ellen Rudin. "A new proof that metric spaces are paracompact". *Proc. Amer. Math. Soc.* (1969), p. 603.

[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.

[RY99] Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion*. Third. Springer-Verlag, Berlin, 1999, pp. xiv+602.

[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.

[San23] Filippo Santambrogio. *A Course in the Calculus of Variations—Optimization, Regularity, and Modeling*. Springer, Cham, 2023, pp. xxi+338.

[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.

[Tsy09] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. Springer, New York, 2009, pp. xii+214.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, 2018, pp. xiv+284.

[Vil15] Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, Providence, RI, 2015, pp. xxii+378.

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