From Measure to Probability

A probabilist's survey of measure-theoretic results

Feng Cheng*

Draft as of January 15, 2025

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

Contents

Prologue

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

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

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

Shanghai, August 2024 F.C.

The prerequisite for this survey notes is a strong background in undergraduate real analysis and familiarity with elementary probability theory. Some key results about Banach spaces, Hilbert spaces, and topology will be assumed, and these can usually be found on any functional analysis texts. We have also included appendices at the end of the survey, which discuss some of these facts.

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

fecheng@math.washington.edu.

Part I

Measure theory

Chapter 1 Measure spaces

1.A Basic setup

We let X be a nonempty set in Part I.

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

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

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

1.2 Exercise. Show that

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

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

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

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

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

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

if the sequence is decreasing, then

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

Also remember that, by De Morgan's Law,

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

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 of subsets of X is an *algebra* if

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

(b) closed under complement;

(c) 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.

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

Similar definitions hold for other types of structures.

Remember the phrase “the generated is the smallest”, and this should indicate how some proofs should proceed.

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

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 “close”.

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

1.9 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.

1.10 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 , i.e., let $\{E_n\}_{n=1}^{\infty}$ be any measurable partition of $E \in \mathcal{A}$, we have

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

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

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

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

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

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

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

A σ-finite measure means it is a normal 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 facts will come up a couple of times.

1.11 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 $(X, \mathcal{A})$.

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

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

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

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

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

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

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

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

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

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

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

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

(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 $1/0$.

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 $\mu\colon \mathcal{A} \to \infty$ by

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

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

We now introduce two additional elementary results about measures, which are simple consequences from Proposition 1.13. These two results are important in probability theory, but both are indeed purely measure-theoretic.

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

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

1.14 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

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

1.15 Borel–Cantelli lemma, part 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 will be used to prove almost everywhere convergence of 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.

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

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

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

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

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

1.B Two tools from set theory

1.17 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)

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

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

1.20 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 $\mathcal{A}_0$.

We do not prove these results in this note; they are very complicated and not very interesting in the end. "The generated is the smallest" is the main idea behind these proofs though. The proof of the next result should provide the readers with a general idea what proofs of this sort look like.

This next result is 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.

1.21 Coincidence criterion [ADM11, Proposition 1.15]. Let μ_1 and μ_2 be two measures on $(X, \mathcal{A})$. Suppose we can find a π-system $\mathcal{K}$ on which the two measures agree, and $\sigma(\mathcal{K}) = \mathcal{A}$.

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 $\{X_n\} \subseteq \mathcal{K}$ such that $X_n \uparrow X$ and

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

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

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

Now consider the general assumption. We define for each n

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

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 $\sigma(\mathcal{K}_n)$.

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

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

is a σ-algebra containing $\mathcal{K}$, 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

1.22 Definition. An *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 $\mu^*(A) \le \mu^*(B)$; (monotonicity)

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

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

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 $\cup_{i\in\mathbf{N}} D_i = X$. Suppose $\ell\colon \mathcal{C} \to [0,\infty]$ with $\ell(\emptyset) = 0$. Now if we define for all $E \in \wp(X)$

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

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

Here are some forewords to what we will construct.

- Let $X = \mathbf{R}$, $\mathcal{C}$ be the collection of all left-open right-closed intervals, and $\ell\big((a,b]\big) = b-a$. This gives the Lebesgue outer measure m^* used to construct the Lebesgue measure m.
- Let $f\colon \mathbf{R} \to \mathbf{R}$ be an increasing right-continuous[9] 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.

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

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

1.24 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.D The Lebesgue measure

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

1.27 Fact. Assuming 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.

1.28 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 $\mathcal{A}_0$.

1.29 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., $\lim_{y \to x^-} F(y)$ exists at every $x \in \mathbf{R}$. More specifically,

$$\lim_{y \to x^-} F(y) = \mu(-\infty, x). \tag{1.30}$$

Regarding equation (1.30), 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 1.28, 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 $\lim_{y\to x^+} F(y)$ exists in the first place.

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

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

and use part (a).

It remains to show for every $x \in \mathbf{R}$ that (1.30) holds:

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

□

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

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

and all conclusions still hold.

In the context of part (a), the μ_F is called the *Lebesgue–Stieltjes measure* associated to F. When the function F is the identity function, μ_F is called the *Lebesgue measure* on $\mathbf{R}$, which we will denote by m in this note[10]. 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 $\mathcal{L}$.

1.31 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.

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

1.E Regularity of measures

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

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

it is *closed inner regular* if

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

it is *compact inner regular* if

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

We say a finite measure μ is *tight* if

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

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

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

1.34 Theorem [Bil99, Theorem 1.1]. For a finite measure μ on a metric space X with the Borel σ-algebra, μ is both outer regular and closed inner regular. It follows that a tight Borel measure is compact inner regular, by Proposition A.7.

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 $U_n = \{x : d(x, E) < \frac{1}{n}\}$.[11] These U_n's are open, since U_n^{c} is the continuous preimage of a closed set $[1/n, \infty)$. Also $U_n \downarrow \{x : d(x, E) = 0\}$, which is exactly E since E is closed. Therefore $\mu(U_n) \to \mu(E)$. This proves 1).

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

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

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

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

[11] See Proposition A.4 if you are not familiar with the definition of $d(\,\cdot\,, E)$.

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

1.36 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.

1.37 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.14 that K is compact. This proves the claim. □

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

- continuous functions between topological spaces are Borel measurable, and
- increasing/decreasing functions from $\mathbf{R}$ to $\mathbf{R}$ are Borel measurable.

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 $\sigma(f)$.

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.)

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

2.4 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.

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

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

Folland Ex 2.9

Baire σ-algebra

2.B Nonnegative Lebesgue integrals

Repartition function is cadlag

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

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

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

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

Fatou's lemma is usually useful when one of the two $\liminf$'s is attained.

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

2.C Signed Lebesgue integrals

2.7 Lebesgue dominated convergence theorem. Let $\{f_n\} \subseteq L^1$. If

(a) $f_n \to f$ pointwise a.e., [limit]

(b) 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

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

This next result is a special case of the above one.

We can use this to prove continuity of functions

2.8 Bounded convergence theorem. Say $\mu(X) < \infty$. Let $\{f_n\} \subseteq L^1$. If

(a) $f_n \to f$ pointwise a.e.,

(b) and there exists some $M \in \mathbf{R}^+$ such that $|f_n| \leq M$ a.e. for all n,

then $f \in L^1$ with

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

In particular, we have

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

2.9 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

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

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 $a > 0$,

$$\mu\{x : |f| \geq a\} \leq \frac{1}{a^p} \int |f|^p \, d\mu. \tag{2.10}$$

2.11 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 $\varphi\colon I \to \mathbf{R}$ such that $\varphi \circ f \in L^1$. Then

$$\varphi\Bigl(\int f \, d\mu\Bigr) \leq \int \varphi \circ f \, d\mu. \tag{2.12}$$

In particular, if φ is bounded below, then we can drop the integrability assumption on $\varphi \circ f$. If

$$\int |\varphi \circ f| = \int_{\{\varphi\circ f \geq 0\}} \varphi \circ f + \int_{\{\varphi\circ f < 0\}} -\varphi \circ f = \infty$$

and φ is bounded below, then

$$\int \varphi \circ f \, d\mu = \int_{\{\varphi\circ f \geq 0\}} \varphi \circ f - \int_{\{\varphi\circ f < 0\}} -\varphi \circ f = \infty.$$

Hence the inequality (2.12) trivially holds.

lower semicontinuous

2.D Connections to the Riemann theory

2.13 Bounded convergence theorem (Riemann integration).

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

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

2.E Modes of convergence

2.15 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 $\epsilon > 0$,

$$\lim_n \mu\{x : |f_n(x) - f(x)| > \epsilon\} = 0. \tag{2.16}$$

We say $\{f_n\}$ is

- *Cauchy/fundamental in measure* if for any $\epsilon > 0$, there exists $N \in \mathbf{N}$ such that for all $m > n \ge N$,

$$\mu\{x : |f_n(x) - f_m(x)| > \epsilon\} < \epsilon \tag{2.17}$$

Note that the "$>$" in both (2.16) and (2.17) can be replaced by "$\ge$", obviously. It suffices to use only one ϵ in (2.17) because we can always choose the smaller of two distinct ϵ's.

2.18 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.

(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

$$|f(x) - g(x)| > \epsilon \implies |f(x) - h(x)| > \epsilon/2 \text{ or } |h(x) - g(x)| > \epsilon/2 \tag{2.19}$$

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 (2.19). 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

$$\mu\{x : |f_n(x) - f_m(x)| > 2^{-k}\} < 2^{-k} \quad \text{for all } m \geq n \geq n_k. \tag{2.20}$$

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

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

Define

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

This gives

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

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

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

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

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

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

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

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

[2] The reader might notice that we have implicitly proved and used Borel–Cantelli lemma, part 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 (2.19):

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

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

(c) Contained in the previous part.

(d) This is clearly a consequence of (2.10).

(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

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

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.1: fix $\epsilon > 0$ and consider $y_n = \mu\{x : |f_n(x) - f(x)| > \epsilon\}$. □

2.21 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.

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

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

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

$$\mu\bigl\{x : \bigl|\Psi\bigl(f_{n_k}(x), g_{n_k}(x)\bigr) - \Psi\bigl(f(x), g(x)\bigr)\bigr| > \epsilon\bigr\} \ge \epsilon. \tag{2.24}$$

Recall the construction of the subsequence in Theorem 2.18(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 (2.24). □

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.

2.25 Remark. One can prove directly that $f_n + g_n \to f + g$ 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.

2.26 Exercise. Use Theorem 2.18(c) to prove the Monotone convergence theorem and Fatou's lemma with convergence in measure.

2.F Littlewood's second and third principles

2.27 Egoroff's theorem. Say $\mu(X) < \infty$. Let f_n be a sequence of $\mathcal{A}$-measurable functions from X to $\mathbf{R}$ (or $\mathbf{C}$). 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 Bounded convergence theorem using this result.

2.28 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 $m([a,b] - F) < \epsilon$.

2.G Uniformly integrable functions

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

2.29 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.

2.30 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 2.29.

2.31 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.

2.32 Proposition.

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

2.33 Exercise. Suppose there exists some $p > 1$ such that $\sup_{n \in \mathbf{N}} \int |f_n|^p \, d\mu < \infty$, then the collection $\{f_n\}_n$ is uniformly integrable.

2.34 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.I Image measures

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

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

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

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

We state the main result below.

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

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

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

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

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

Chapter 3 Product spaces

3.A Product σ-algebras

We start with a comparison between product topologies and product σ-algebras. See [Fol99, Sections 4.1 and 4.2] for a review of bases, subbases and product topologies.

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 full-dimension 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, but in practice we mostly deal with cases where the index set I is countable. The reader should verify on their own that when 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. For any separable metric 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 $\mathcal{T}$ given by the supremum metric.

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 product of separable metric spaces, is still a separable metric space. Here we use a result from the [BS20], included as Proposition A.13 in the appendix:

> Every collection of open sets in a separable metric space contains an at most countable subcollection with the same union.

[3] This is the second edition of the book. The new proof in the third edition is very misleading, and I suspect there are many errors in the new edition.

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 $\mathcal{T} \subseteq \sigma(\mathcal{J})$. □

The direct corollary is that $\mathcal{B}(\mathbf{R}^d) = \bigotimes^d \mathcal{B}(\mathbf{R}^1)$. 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.

We reserve the discussion of two extremely important existence results about probability measures on product spaces to Section 11.A. The first of the two 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 result () says 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.B Integration on product spaces

3.7 Fubini–Tonelli theorem.

folland exercise 12

3.C Change of variables

3.D Gamma functions and polar coordinates

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

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

Since

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

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

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 σ-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.

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 we have constructed some set that attains $\inf\{\mu(A) : A \in \mathcal{A}\}$. This implies the boundedness of μ from both above and below. (Apply the argument to $-\mu$.)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Proof. □

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

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

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

4.B Radon–Nikodym theorem and Lebesgue decomposition

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

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

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

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

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

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

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.

Proof 1, using Hilbert spaces. □

Proof 2, using variational principles. □

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

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

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

(b)

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

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

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

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

Radon–Nikodym derivative with respect to counting measure

Lebesgue decomposition of a monotonic function (p344 345 Bogachev)

4.C Differentiation

4.18 Vitali covering lemma.

4.19 Definition. Hardy–Littlewood maximal function

4.20 Lebesgue differentiation theorem.

4.D Functions of bounded variations

4.E Absolutely continuous functions

4.21 Definition. Let $I \subseteq \mathbf{R}$ be an interval. A function $f\colon I \to \mathbf{R}$ is absolutely continuous if for all $\epsilon > 0$, there exists $\delta > 0$ such that

$$\sum_{i=1}^{n}(b_i - a_i) < \delta \implies \sum_{i=1}^{n}|f(b_i) - f(a_I)| < \epsilon$$

holds for any finite family of pairwise disjoint open intervals $\{(a_i, b_i)\}_{i=1}^{n}$ contained in I.

4.F Fundamental theorem of calculus

4.22 Fundamental theorem of calculus (for Lebesgue integrals). For $f\colon [a,b] \to \mathbf{R}$, the following are equivalent:

(a) f is absolutely continuous;

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

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

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

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

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

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

Bogachev 5.4.5 4.7.60

4.G Riesz' theorems, vague, and weak convergence of measures

4.23 Riesz–Markov–Kakutani theorem for compact metric spaces. Let (X,d) be a compact metric space[5], then the dual space $C(X)^*$ is isometrically isomorphic to $\mathcal{M}(X)$, i.e., for all linear functionals $L \in C(X)^*$, there is a unique $\mu \in \mathcal{M}(X)$ such that

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

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

4.24 Riesz–Markov–Kakutani theorem for l.c.(s.c.)H. spaces. Let X be an l.c.H. space, then the dual space $C_0(X)^*$ is isometrically isomorphic to $\mathcal{M}_{\mathrm{R}}(X)$, i.e., for all linear functionals $L \in C_0(X)^*$, there is a unique $\mu \in \mathcal{M}_{\mathrm{R}}(X)$ such that

$$L(f) = \int_X f\,d\mu \quad \text{for all } f \in C_0(X);$$

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

In particular, if X is further assumed to be second countable, then the space of Radon measures $\mathcal{M}_r(X)$ above can be replaced by the space of Borel measures $\mathcal{M}(X)$ instead.

finite Radon measure (Bogachev 7.1)

[5] Every compact metric space is separable.

Chapter 5 Lebesgue spaces and some Fourier theory

5.A When $1 \leq p < \infty$

5.1 Hölder's inequality.

5.2 Minkowski's inequality.

5.3 Theorem. L^p is complete.

5.4 Proposition. On a finite measure space, the equivalence class simple functions are dense in L^p hence $L^q \cap L^p$ is dense in L^p
of C_b is dense in L^p

5.B When $p = \infty$

5.5 Theorem. L^∞ is complete.

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

5.C The Hilbert space L^2

5.D Dual spaces

Chapter 6 A glimpse of Fourier analysis

6.A Fourier series

6.B Convolutions

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

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

for all x such that the integral exists.

6.1 Young's inequality.

6.C Fourier transform of functions and measures

Interlude: Between Measure and Probability

Part II

Probability

Chapter 7 Interpreting probability using measure theory

7.A Distributions

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

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

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

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

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

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

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

[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.

7.3 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* of a real-valued random variable X is defined to be a function $F\colon \mathbf{R} \to [0,1]$ given by

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

We now slightly modify Theorem 1.29(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.3), and is hence continuous a.e.

Recall Theorem 1.29(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 can state

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

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

[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.

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

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

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

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

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

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

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

There are three 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 *left-continuous inverse* 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. When F is one-to-one, then our X is just the usual inverse.

We now show $X(\omega)$ is indeed 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).

Finally, 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.3. We leave it as an exercise to reader to show that this $\overline{X}$ is right-continuous.

We will generalize this result later.

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.

[5] or *frequency function*

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 1.12. Given a random variable X, if the codomain S is a Borel subset of $\mathbf{R}$ and $\mathcal{S} = \mathcal{B}|_S$, 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 naturally be 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, as long as $\mu \ll m$.

We can define the class of distributions with densities solely in terms of their density functions. When the distribution of X is discrete, it is clear that we can specify the 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 distribution of X is continuous, then a nonnegative Borel-measurable function f satisfying

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

called a *(probability) density function* (on its own) specifies 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,

7.10 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

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

[6] to emphasize we are in the discrete setting

[7] The term “continuous” here refers to absolute continuity, and does not indicate that the density function is continuous

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 2.35 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 $\mathrm{E}X = \mathrm{E}Y$.

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

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

7.12 Jensen's inequality. Let $\mathrm{E}|X| < \infty$. Suppose I is an interval containing the range of X, and we have a convex function $\varphi\colon I \to \mathbf{R}$ such that $\varphi \circ X \in L^1$. Then

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

7.13 Lyapunov's inequality. For $p \leq q$, we have

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

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

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

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

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

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

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

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

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

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

7.15 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.

7.16 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. □

7.17 Proposition. A real-valued random variable X independent of itself must take a constant value a.s.

In special the case when $X \in L^2$ we have a simple proof: 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. The general case needs a bit more care.

Proof. For any $A \in \mathcal{B}$, we have

$$P(X \in B)P(X \in B) = P(X \in B),$$

which implies $P(X \in B) = 0$ or 1.

We now prove a more general claim that directly implies the proposition:

> a $\{0, 1\}$-valued Borel probability measure μ on a separable metric space S must be a point mass.[9]

We know every open cover has a countable subcover in S (this is Proposition A.12). Fix $\epsilon > 0$ and consider the ϵ-balls $B(x; \epsilon)$ around each $x \in S$. Now there exists a countable subcollection $\{B(x_j; \epsilon)\}_{j=1}^{m}$ that covers S, and this implies there exists one unique $j \in [m]$ 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 $B \coloneqq \cap_n B_{1/n}$ has diameter 0, B is a singleton set of measure 1. □

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

7.18 Proposition (Factorization). Let X and Y be two discrete/continuous random variables. Then X and Y are independent if and only if for all $x, y \in \mathbf{R}$,

(a) $f_{X,Y}(x, y) = f_X(x) f_Y(y)$, where the f's are density functions;

(b) $f_{X,Y}(x, y) = g(x)h(y)$ for some functions g and h.

To be precise the equalities above are up to measure zero.

Proof. We show the case when X and Y are continuous random variables on $\mathbf{R}$. For all $A_1, A_2 \in \mathcal{B}$, we have

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

[9] Hence the "real-valued random variable X" in the proposition statement may be replaced by "random variable X taking values in a separable metric space".

Part (a) now follows easily. To see the "if" direction of part (b), integrate both sides of $f_{X,Y}(x,y) = g(x)h(y)$ over $A_1 \times A_2$, we have

$$\mu_{X,Y}(A_1 \times A_2) = \int_{A_1} g(x)\,dx \int_{A_2} h(y)\,dy.$$

Consider $C = \int_{\mathbf{R}} h(y)\,dy$. We may divide h by this constant C and multiply g by this C, and assume without loss of generality that

$$\begin{aligned}
\mu_X(A_1) = \mu_{X,Y}(A_1 \times \mathbf{R}) &= \int_{A_1} g(x)\,dx, \\
\mu_Y(A_2) = \mu_{X,Y}(\mathbf{R} \times A_2) &= \int_{A_2} h(y)\,dy.
\end{aligned}$$

This completes the proof. □

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

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

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

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

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

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

and their *correlation* is defined by

$$\operatorname{Corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(y)}}.$$

As mentioned perviously, the L^2 requirement is a sufficient, but not necessary condition for covariance to always exist. This is similar to the L^1 requirement sufficient for the expectation of a random variable to always exist.

Let A and B be two events and consider two indicators $\mathbf{1}_A$ and $\mathbf{1}_B$. Notice

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

We say A and B are *positively correlated* if the covariance above is ≥ 0, i.e., $P(A \cap B) \geq P(A)P(B)$, or equivalently $P(A \mid B) \geq P(A)$. We say A and B are *negatively correlated* if the $\geq$'s are replaced by $\leq$'s. Note that the covariance and correlation are symmetric.

7.B.3 Sum of independent random variables

Fourier transform

The *tail σ-field* of a sequence of random variables $X_1, X_2, \dots$ to be

$$\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, \dots).$$

7.20 Kolmogorov's zero–one law. Let $X_1, X_2, \dots$ be a sequence of independent random variables, then any event in its tail σ-field has probability 0 or 1.

7.C Basic concentration and deviation inequalities

7.21 Generalized Markov's inequality. Let $\varphi\colon \mathbf{R} \to [0, \infty)$ be increasing. Then for any random variable X, and any $a \in \mathbf{R}$ with $\varphi(a) \neq 0$, we have

$$P(X \geq a) \leq \frac{1}{\varphi(a)} \mathrm{E}\varphi(X).$$

7.22 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

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

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

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

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 $\mathrm{E}X^2$.

7.24 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:

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

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 to respectively to the *first* and the *second moment method* in probabilistic combinatorics; see [Roc24, Chapter 2].

7.D Miscellaneous but crucial facts and tools

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

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

It is quite clear that m and γ are equivalent measures, since $\exp(\,\cdot\,)$ is nonnegative.

Bogachev Theorem 1.4.3.

The coordinates of a normal random vector are independent if and only if they are uncorrelated.

We now restate Borel–Cantelli lemma, part I.

7.26 Borel–Cantelli lemma, part 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.) Equivalently, since we are in a probability space, $X_n \to X$ a.s. is the same as saying

$$\forall\, \epsilon > 0, P\big(|X_n(\omega) - X(\omega)| < \epsilon \text{ ev.}\big) = 1.$$

7.27 Borel–Cantelli lemma, part 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

Chapter 8 Modes of convergence in probability

8.A Statistical distances

Important disclaimer. This section deals purely with comparisons of probability measures μ and ν on a given measurable space $(S, \mathcal{S})$, and has nothing to do with random variables. In practice we may want to see μ and ν indeed as probability distributions of random variables on the codomain space $(S, \mathcal{S})$. Please be very careful about this distinction.

Given two probability measure μ and ν on $(S, \mathcal{S})$, we have the signed measure $\mu - \nu \colon \mathcal{S} \to [-1, 1]$. Its total variation norm

$$\begin{aligned}\|\mu - \nu\| &= |\mu - \nu|(S) \\ &= \sup_{A \in \mathcal{S}} |(\mu - \nu)(A)| + |(\mu - \nu)(\Omega - A)| \\ &= \sup_{A \in \mathcal{S}} |\mu(A) - \nu(A)| + |1 - \mu(A) - 1 + \nu(A)| \\ &= 2 \sup_{A \in \mathcal{S}} |\mu(A) - \nu(A)|.\end{aligned}$$

The factor 2 above is usually dropped in probabilistic applications. We define the *total variation distance* between μ and ν to be

$$\|\mu - \nu\|_{\mathrm{TV}} = d_{\mathrm{TV}}(\mu, \nu) = \sup_{A \in \mathcal{S}} |\mu(A) - \nu(A)|.$$

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

$$\|\mu_n - \mu\|_{\mathrm{TV}} \to 0. \tag{8.2}$$

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

$$0 = \lim_n \|\mu_n - \mu\|_{\mathrm{TV}} = \lim_n \sup_{A \in \mathcal{S}} |\mu_n(A) - \mu(A)|,$$

which in particular implies $\mu_n(S) - \mu(S) \to 0$.

The total variation convergence given above can of course be defined for general finite/signed/complex measures, by using the total variation norm $\|\cdot\|$ in place of the half-total variation norm $\|\cdot\|_{\mathrm{TV}}$ in probability. (Of course there is no difference in definition if we change by a constant factor 2.) We do not discuss this convergence in the general setting.

The following result is a restatement of something we have proved in Section 4.A.

8.3 Fact. If μ and ν have a common dominating measure ρ, then

$$\|\mu-\nu\|_{\mathrm{TV}}=\frac{1}{2}\int_S\left|\frac{d\mu}{d\rho}(\omega)-\frac{d\nu}{d\rho}(\omega)\right|d\rho.$$

In particular, if $(S,\mathcal{S})$ is a discrete space, then

$$\|\mu-\nu\|_{\mathrm{TV}}=\frac{1}{2}\sum_{x\in S}|\mu\{x\}-\nu\{x\}|.$$

And if $(S,\mathcal{S})=(\mathbf{R},\mathcal{B})$, with ρ being the Lebesgue measure, then

$$\frac{1}{2}\int_{\mathbf{R}}|f(x)-g(x)|\,dx,$$

where $f=\frac{d\mu}{d\rho}$ and $g=\frac{d\nu}{d\rho}$ are the two probability densities[1]. In short, the total variation distance between two probability measures is half the L^1 distance between their densities.

The *Kullback–Leibler divergence/relative entropy* of μ with respect to ν is given by

$$D_{\mathrm{KL}}(\mu\|\nu)=\begin{cases}\int_S\log\frac{d\mu}{d\nu}\,d\mu & \text{if } \mu\ll\nu,\\ +\infty & \text{otherwise.}\end{cases}$$

Let $f=\frac{d\mu}{d\nu}\in L^1(\nu)$. It is very important to note that

$$\int_S\log\frac{d\mu}{d\nu}\,d\mu=\int_S f\log f\,d\nu$$

can be infinite, since $f\log f$ might not be integrable with respect to ν. Sometimes we just

8.4 Fact. If $\mu\ll\nu\ll\rho$, then

$$D_{\mathrm{KL}}(\mu\|\nu)=\int_S\left(\frac{d\mu}{d\rho}\right)\log\left(\frac{d\mu/d\rho}{d\nu/d\rho}\right)d\rho.$$

Therefore if the space is discrete, then we take ρ to be the counting measure and get

$$D_{\mathrm{KL}}(\mu\|\nu)=\sum_{x\in S}\mu\{x\}\log\frac{\mu\{x\}}{\nu\{x\}}.$$

And if $(S,\mathcal{S})=(\mathbf{R},\mathcal{B})$, with ρ being the Lebesgue measure, then

$$D_{\mathrm{KL}}(\mu\|\nu)=\int_{\mathbf{R}}f(x)\log\frac{f(x)}{g(x)}\,dx,$$

where $f=\frac{d\mu}{d\rho}$ and $g=\frac{d\nu}{d\rho}$ are the two probability densities. In this latter case we might as well write $D_{\mathrm{KL}}(f\|g)$.

[1] Of course we may consider μ and ν on some restricted subspace of $(\mathbf{R},\mathcal{B})$, but as mentioned before we drop such consideration for brevity.

Given a probability measure ν, for a nonnegative $f \in L^1(\nu)$ such that $f \log f$ is also ν-integrable, we define its *entropy functional* to be

$$\operatorname{Ent}_\nu f = \mathrm{E}_\nu(f \log f) - (\mathrm{E}_\nu f)(\log \mathrm{E}_\nu f),$$

which should be compared with the variance functional

$$\operatorname{Var}_\nu f = \mathrm{E}_\nu f^2 - (\mathrm{E}_\nu f)^2.$$

But keep in mind the entropy functional can only be applied to (ν-a.e.) nonnegative[2] functions because of the logarithm in the definition. Also note that the entropy functional is homogeneous: we have

$$\operatorname{Ent} cf = c \operatorname{Ent} f \quad \text{for } c \geq 0,$$

which is "better" than

$$\operatorname{Var} cf = c^2 \operatorname{Var} f \quad \text{for } c \in \mathbf{R}$$

in some applications.[3]

If we have another probability measure μ with $\mu \ll \nu$, then

$$\operatorname{Ent}_\nu \frac{d\mu}{d\nu} = D_{\mathrm{KL}}(\mu \| \nu).$$

If $d\mu/d\nu$ can be explicitly expressed by some function h (as discussed in Fact 8.4), then the equation above gives a simple expression for the KL divergence.

Fisher information 5.1.2 Markov diffusion operators LSI

8.5 Pinsker's inequality. $\|\mu - \nu\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mu \| \nu)}$.

Hellinger distance

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 $\int_S f \, d\mu \neq \int_S f \, d\nu$.

If we take $\mathcal{F}$ to be the collection of all indicator functions, then $d_{\mathcal{F}} = d_{\mathrm{TV}}$.

The *Kolmogorov uniform metric* is defined by

$$d_{\mathrm{K}}(\mu, \nu) = \sup_{x \in \mathbf{R}} |F_\mu(x) - F_\nu(x)| = \sup_{x \in \mathbf{R}} \big| \mu(-\infty, x] - \nu(-\infty, x] \big|,$$

which is the an IPM $d_{\mathcal{F}}$ with $\mathcal{F} = \{\mathbf{1}_{(-\infty, x]} : x \in \mathbf{R}\}$.

Let (S, d) be a Polish space, and $1 \leq p < \infty$, the *Wasserstein distance* of order p is defined by

$$W_p(\mu, \nu) = \left[\inf_{\pi \in \Pi(\mu, \nu)} \int_S d(x, y)^p \, d\pi(x, y) \right]^{1/p}.$$

[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.

Alternatively, one may write

$$W_p(\mu,\nu) = \inf\left\{\mathrm{E}\left[d(X,Y)^p\right]^{1/p} : X =_d \mu, Y =_d \nu\right\},$$

and understand it as an L^p distance between two probability measures.

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

$$\mathcal{P}_p(S) = \left\{\mu \in \mathcal{P}(S) : \int_S d(x_0,x)^p \, d\mu(x) < \infty \text{ for all } x_0 \in S\right\}.$$

8.B Weak convergence of probability measures

Let (S,d) be a metric space. We use $\mathcal{M}(S)$ for the space of finite signed/complex Borel measures on S, $\mathcal{M}_+(S)$ for the space of finite Borel measures on S, and $\mathcal{P}(S)$ for the space of Borel probability measures. A *subprobability measure* μ is a measure with $\mu(S) \leq 1$.

8.6 Definition. A sequence $\{\mu_n\} \subseteq \mathcal{M}(S)$ *converges weakly* to $\mu \in \mathcal{M}(S)$ if for all $f \in C_b(S)$, we have

$$\int_S f\, d\mu_n \to \int_S f\, d\mu, \tag{8.7}$$

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

This is the general definition of weak convergence of general Borel measures. Our attention will be restricted to the case when μ_n is a sequence of Borel probability measures.

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

In analysis one is often interested in test function classes C_0 and C_c; see Section 4.G. 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 (8.7), 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 ...

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]. 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. A brief overview of the general theory of weak convergence has been included in Appendix E, based on [Bog18].

8.8 Definition. A sequence $\{\mu_n\}$ of Borel probability measures *converges weakly* to a Borel probability measure μ if for all $f \in C_b(S)$, we have

$$\int_S f\, d\mu_n \to \int_S f\, d\mu,$$

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

If each μ_n and μ represents the distribution of some $(S, \mathcal{B}_S)$-valued random variables X_n and X, then we usually say X_n *converges to* X *in distribution*, denoted by[4] $X_n \Rightarrow X$. Because of Corollary 7.6, when $S = \mathbf{R}$ we also write $F_{X_n} \Rightarrow F_X$.

8.9 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.

Recall that vague convergence

8.10 Proposition. Weak convergence of integer-valued measures is equivalent to pointwise convergence.

8.11 Alexandroff portmanteau[5] theorem. The following statements are equivalent characterizations of the weak convergence of Borel probability measures on a metric space (S, d).

(a) $\int f\, d\mu_n \to \int f\, d\mu$ for all bounded Lipschitz functions f on S;

(b) $\int f\, d\mu_n \to \int f\, d\mu$ for all bounded uniformly continuous functions f on S;

(c) $\limsup_n \int f\, d\mu_n \le \int f\, d\mu$ for all u.s.c. functions bounded from above;

(d) $\liminf_n \int f\, d\mu_n \ge \int f\, d\mu$ for all l.s.c. functions bounded from below;

(e) $\limsup_n \mu_n(F) \le \mu(F)$ for all closed sets F;

(f) $\liminf_n \mu_n(G) \ge \mu(G)$ for all open sets G;

(g) $\lim_n \mu_n(A) = \mu(A)$ for all *continuity sets* A with respect to μ, i.e., Borel sets A with $\mu(\partial A) = 0$.

(sequential Banach–Alaoglu) vague convergence Prohorov's theorem

The proof of the following result resembles that of the classical Arzelà–Ascoli theorem on $\mathbf{R}^d$. The shared proof idea is to construct a desired subsequence (that converges pointwise on all rationals) by the so-called diagonal argument. The construct can be made very explicit, as we will show below.

8.12 Lemma. Let $\{F_n\}$ be a sequence of distribution functions, then there is a subsequence $\{F_{n_k}\}$ and a right-continuous increasing function F such that

$$\lim_{k\to\infty} F_{n_k}(x) = F(x)$$

for all continuity points x of F.

Proof. Let $q_1, q_2, \ldots$ be an enumeration of $\mathbf{Q}$. First $\{F_n(q_1)\}$ is a sequence in $[0, 1]$, a bounded interval, and therefore there is a subsequence that converges to $s_1 := \liminf_n F_n(q_1)$. We can construct such a subsequence by defining $F_{\nu(n)}(q_1)$ inductively for all n:

$$\nu(n) = \min\{m > \nu(n-1) : |F_m(q_1) - s_1| < 1/n\}.$$

Let $\{F_n^1\}$ be the new sequence $\{F_{\nu(n)}\}$, and in the same way one can construct its subsequence $\{F_n^2\}$ satisfying $\lim_n F_n^2(q_2) = s_2 := \liminf_n F_n^1(q_2)$. Proceeding in this fashion, we get

[4] sometimes even mix up and write $X_n \Rightarrow \mu$

[5] As Bogachev [Bog18] points out, "I do not know who invented such a nonsensical name for Alexandroff's theorem."

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 $F_1^1, F_2^2, \ldots$, which we call F_{n_k}. If we ignore the first $j-1$ terms of the diagonal sequence, this new F_{n_k} is a subsequence of $\{F_n^j\}_{n=1}^\infty$. Therefore this subsequence converges at all rational points. Let $G\colon \mathbf{Q} \to [0,1]$ be its pointwise limit:

$$G(q) = \lim_k F_{n_k}(q).$$

Take its increasing, right-continuous inverse F given by

$$F(x) = \inf\{G(q) : q \in \mathbf{Q}, q > x\}.$$

Notice for a rational q strictly between two reals x_1 and x_2, we have

$$F(x_1) \leq G(q) \leq F(x_2), \tag{8.13}$$

where the second inequality follows from the fact that G is increasing on $\mathbf{Q}$.

Now let x be a continuity point of F, then for any $\epsilon > 0$, we can find two reals $\tilde{r}$ and $\hat{r}$ with $\tilde{r} < x < \hat{r}$, such that

$$F(x) - \epsilon < F(\tilde{r}) \leq F(x) \leq F(\hat{r}) < F(x) + \epsilon.$$

Then we can find two rationals $\tilde{q}$ and $\hat{q}$ satisfying $\tilde{r} < \tilde{q} < x < \hat{q} < \hat{r}$, such that

$$F(x) - \epsilon < G(\tilde{q}) \leq F(x) \leq G(\hat{q}) < F(x) + \epsilon,$$

by (8.13). Since G is the pointwise limit of F_{n_k} on the rationals, for all k large enough we will have

$$F(x) - \epsilon < F_{n_k}(\tilde{q}) \leq F_{n_k}(x) \leq F_{n_k}(\hat{q}) < F(x) + \epsilon.$$

It follows that F_{n_k} converges to F at all continuity points x. □

Recall that the subsequential limit F constructed above associates to a Borel measure μ_F, by Theorem 1.29(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 $\mu_F(\mathbf{R}) \leq 1$.

(Our conclusion here can also be understood and generalized via very heavy machinery[6]. For every separable normed space $(S, \|\cdot\|)$, $(C_0(S), \|\cdot\|_u)$ is also a separable normed space. By the Sequential Banach–Alaoglu–Bourbaki theorem, $C_0(S)^*$ is weak-star sequentially

[6] and the stronger axiom DC; we did not use any choice axiom in our elementary proof

compact. By , the sequence $\mu_n \subseteq \mathcal{P}(S)$, which is norm bounded in $\mathcal{M}(S)$, must have a subsequential vague limit μ that is a subprobability measure.)

The μ_F above is not in general a probability measure, for example, consider the sequence of distribution functions F_n of the uniform distributions over $[-n, n]$. The sequence $\{F_n\}$ *itself* (and hence all of its subsequences) converges vaguely to the 0 function. To ensure that the subsequential F constructed in Lemma 8.12 is indeed a distribution function, we require tightness in addition.

8.14 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.

8.15 Helly selection theorem. If we assume that in Lemma 8.12 $\{F_n\}$ is a tight sequence of distribution functions, then the vague subsequential limit F constructed there is a distribution function.

generalization subprobability measure, Rd

8.16 Skorohod representation theorem (Polish space). Let (S, d) be Polish. Suppose $\mu_n \Rightarrow \mu$, then there exist X_n and X defined on a common probability space $(\Omega, \mathcal{F}, P)$, such that $X_n \sim \mu_n$, $X \sim \mu$, and $X_n \to X$ pointwise everywhere on Ω.

Redefine X_n by X outside the set of convergence Weak compactness
Prohorov metric for $S = \mathbf{Z}$

8.C Comparisons between modes of convergence

8.17 Theorem. If $\mu_n \to \mu$ in total variation, then $\mu_n \Rightarrow \mu$.

Proof. $\mu_n \to \mu$ in total variation means that for all $A \in \mathcal{S}$, $\mu_n(A) \to \mu(A)$, or equivalently,

$$\int \mathbf{1}_A \, d\mu_n \to \int \mathbf{1}_A \, d\mu.$$

Now use the standard argument to extend to all C_b functions. □

8.18 Theorem. 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 2.18. □

8.19 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 $f \in C_b(S)$, $g \circ f \in C_b(S)$.

8.20 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.

[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.

8.21 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

8.22 Slutsky's theorem. Given $X_n \Rightarrow X$ and $Y_n \Rightarrow c$, then

(a) $X_n + Y_n \Rightarrow X + c$;

(b) $X_n Y_n \Rightarrow cX$;

(c) $X_n / Y_n \Rightarrow X/c$, provided that $c \neq 0$.

8.D Laws of large numbers

8.23 L^2 weak law. Let $X_1, X_2, \dots$ 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. □

8.24 L^1 weak law. Let $X_1, X_2, \dots$ be i.i.d. and L^1 with mean μ. Then

$$\frac{X_1 + \ldots + X_n}{n} \to \mu$$

in probability.

8.25 L^1 strong law. Let $X_1, X_2, \dots$ be pairwise independent, identically distributed L^1 random variables with mean μ. We have

$$\frac{X_1 + \ldots + X_n}{n} \to \mu \quad \text{a.s.}$$

Furthermore the above convergence also holds in L^1.

Proof. The a.s. part will follow the Etemadi's classical truncation proof.

It remains to show that $\{\overline{X}_n\}_{n \in \mathbf{N}} = \left\{\frac{X_1 + \ldots + X_n}{n}\right\}_{n \in \mathbf{N}}$ is uniformly integrable. We know each X_j, as an L^1 random variable, must be uniformly integrable. In particular, for any $\epsilon > 0$, there is some $\delta > 0$ such that for all $n \in \mathbf{N}$,

$$\begin{aligned} P(A) < \delta &\implies \mathrm{E}\big(|X_j|; A\big) < \epsilon \quad \text{for all } j \in [n] \\ &\implies \mathrm{E}\left(\left|\frac{X_1 + \ldots + X_n}{n}\right|; A\right) < \epsilon. \end{aligned}$$

Meanwhile

$$\begin{aligned}\sup_n \mathrm{E}\left|\frac{X_1+\ldots+X_n}{n}\right| &\leq \sup_n \frac{\mathrm{E}|X_1|+\ldots+\mathrm{E}|X_n|}{n}\\ &= \mathrm{E}|X_1| < \infty.\end{aligned}$$

Combining the above information gives uniformly integrability of $\{\overline{X}_n\}$. □

8.26 L^4 strong law.

8.27 L^2 strong law.

Let $X_1, X_2, \ldots$ follow a common distribution μ, or alternatively a common distribution function F. The *empirical distribution* of the first n random variables is defined to

$$\mu_n = \frac{1}{n}\sum_{k=1}^{n} \delta_{X_k},$$

which is the averaging of the first n observations. Notice that this is a random variable in terms of $X_1, \ldots, X_n$. This gives us the *empirical distribution function*

$$F_n(x) = \mu(-\infty, x] = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}\{X_k \leq x\}.$$

8.28 Glivenko–Cantelli theorem. As $n \to \infty$,

$$\sup_x |F_n(x) - F(x)| \to 0 \quad P\text{-a.s.}$$

11.4 Dudley
Kolmogorov–Smirnov statistics and test

8.29 Dvoretzky–Kiefer–Wolfowitz–Massart inequality. For every $\epsilon > 0$,

$$P\big(\sup_x |F_n(x) - F(x)| > \epsilon\big) \leq 2\exp(-2n\epsilon^2).$$

8.E Moment generating functions and characteristic functions

Integral transform converts a given problem to one which is easier to solve, and then 'inverting' to solve the original problem

For a real random variable X, its *moment generating function* (m.g.f.) is a function $M_X\colon \mathbf{R} \to \mathbf{R}$ 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.

The *cumulant generating function* is defined to be the log moment generating function.

[Bog07, Theorem 7.13.1] Bochner

8.30 Example. For $Z \sim N(0,1)$ and $t \in \mathbf{R}$, we have the $M_Z(t)$ given by

$$\begin{aligned} \mathrm{E}\exp(tX) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} e^{tx}\, dx \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\Big(-\frac{1}{2}(x-t)^2\Big)\, dx = \exp(t^2/2). \end{aligned}$$

It turns out that the $\varphi_Z(t)$ has almost the same expression (except for the sign):

$$\mathrm{E}\exp(itX) = \exp(-t^2/2)$$

for all $t \in \mathbf{R}$. It suffices to show that

$$\mathrm{E}\exp(tX) = \exp(t^2/2) \tag{8.31}$$

for all $t \in \mathbf{C}$.

We wish to use the Uniqueness theorem given (8.31) 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].

8.32 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 $(-\delta, \delta)$.

8.33 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 $\varphi^{(k)}(t) = i^k \, \mathrm{E}X^k$.

8.34 Inversion formula. 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 $a < b$,

$$\overline{F}(b) - \overline{F}(a) = \lim_{T \to \infty} \int_{-T}^{T} \frac{\exp(-iat) - \exp(-ibt)}{2\pi i t} \varphi(t)\, dt$$

[8] Recall we defined complex exponentials as power series, and power series/polynomials are differentiable term-by-term.

Note $\overline{F}(b) - \overline{F}(a) = \mu(a,b) + \frac{1}{2}(\mu\{a\} + \mu\{b\})$. In particular, if a and b are not atoms of μ_F, then the expression is equal to $\mu(a,b]$.

8.35 Theorem (c.d.f. and ch.f. correspondence). $X =_d Y$ if and only if $\varphi_X = \varphi_Y$.

Proof. One direction is obvious. Now assume $\varphi_X = \varphi_Y$, which gives

$$\overline{F}_X(b) - \overline{F}_X(a) = \overline{F}_Y(b) - \overline{F}_Y(a)$$

for all real numbers $a \leq b$. Take $a \to -\infty$ gives us $\overline{F}_X(b) = \overline{F}_Y(b)$.

We show the agreement of $\overline{F}$ implies the agreement of F. Now take F in fact to be any distribution function. For any $x \in \mathbf{R}$, consider a sequence $b_n = x + \frac{1}{n}$. Now

$$\lim_n F(b_n) = F(x)$$

and $\lim_n \mu(-\infty, b_n) = \mu(-\infty, x]$, i.e.,

$$\lim_n F(b_n -) = F(x).$$

Hence

$$\lim_n \overline{F}(b_n) = \frac{1}{2} \lim_n [F(b_n) + F(b_n -)] = F(x).$$

The conclusion now follows. □

8.36 Theorem.

(a) If $\mu_n \Rightarrow \mu$, then the corresponding ch.f.'s have $\varphi_n \to \varphi$ pointwise everywhere;

(b) if $\varphi_n \to \varphi$ pointwise, and φ is continuous at 0, then the measures μ_n associated to φ_n are tight and converges weakly to some measure μ whose characteristic function is φ.

8.37 Classical central limit theorem. Let $X_1, X_2, \ldots$ be i.i.d. L^2 random variables with variance $\sigma^2 \neq 0$, then we have

$$\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \Rightarrow N(0,1)$$

8.38 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

$$\frac{X_1 + \ldots + X_n}{\sigma_n} \Rightarrow N(0,1).$$

8.39 Lyapunov condition.

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 $\mathrm{E}(X \,|\, \mathcal{G})$.

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 $\mathrm{E}(\mathbf{1}_A \,|\, \mathcal{G})$, which we denote by $P(A \,|\, \mathcal{G})$.

Our new definitions of conditional expectation and conditional probability are very abstract, and particularly distinct from the undergraduate version, and the following example is almost included in all textbooks, which explains how our new definitions generalizes the old definitions.

9.4 Example. Let $\Omega_1, \Omega_2, \ldots$ be a countable partition of the sample space Ω, where each Ω_n has strictly positive measure. In an undergraduate class we would define

$$\mathrm{E}(X \,|\, \Omega_n) = \frac{\mathrm{E}(X; \Omega_n)}{P(\Omega_n)}$$

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 $\mathrm{E}|X| < \infty$ and $\mathrm{E}|XZ| < \infty$ (in particular, if Z is bounded $\mathcal{G}$-measurable), then we have

$$\mathrm{E}(\mathrm{E}(X \mid \mathcal{G})Z) = \mathrm{E}(XZ).$$

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):

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 L^p $(1 \leq p < \infty)$: for $X \in L^p(\Omega, \mathcal{F}, P)$,

$$\mathrm{E}\big(\big|\mathrm{E}(X \mid \mathcal{G})\big|^p\big) \leq \mathrm{E}|X|^p.$$

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 now 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.16 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. □

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 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 $X \in L^1(\mathcal{G}_2)$ or $L^+(\mathcal{G}_2)$;

(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

$$\mathrm{E}\big(f(X) \mid Y\big) = \mathrm{E}f(X)$$

for all $f \in L^1$.

9.13 Definition. Let X be nonnegative $\mathcal{F}$-measurable, then we define its *conditional expectation* given $\mathcal{G}$ to be

$$\mathrm{E}(X \mid \mathcal{G}) = \lim_{n\to\infty} \mathrm{E}(X \wedge n \mid \mathcal{G}).$$

9.B Stopping times

A *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 $X_0, X_1, \dots.$ we define its *natural filtration* by setting $\mathcal{F}_n = \sigma(X_0, X_1, \dots)$ for all $n \in \mathbf{N}_0$.

9.14 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 $S \le T$, then $\mathcal{F}_S \subseteq \mathcal{F}_T$.

(c) $\mathcal{F}_{S\wedge T} = \mathcal{F}_S \cap \mathcal{F}_T$.

9.C Discrete martingales

9.15 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)

9.16 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 $n \geq 1$, and $(H \bullet X)_0 = 0$.

9.17 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 $H_n \geq 0$ for all n, then $\{(H \bullet X)_n\}$ is a submartingale (resp. supermartingale).

9.18 Doob's decomposition.

Chapter 10 A cornucopia of ergodic theory

Given a probability space $(\Omega, \mathcal{F}, \mu)$, a *measure-preserving transformation* (MPT) T is a measurable function from $(\Omega, \mathcal{F})$ to itself such that

$$\mu(T^{-1}A) = \mu(A) \text{ for all } A \in \mathcal{F}.$$

The resulting quartet $(\Omega, \mathcal{F}, \mu, T)$ 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

$$\mu(TA) = \mu(A) \text{ for all } A \in \mathcal{F}.$$

An MPT T is said to be μ*-ergodic* (or the measure μ is said to be T*-ergodic*) if for all $A \in \mathcal{F}$, we have

$$\mu(A \triangle T^{-1}A) = 0 \implies \mu(A) = 0 \text{ or } 1.$$

A set $A \in \mathcal{F}$ 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 $A \in \mathcal{F}$, the set $B = \limsup_n T^{-n}A$ is always going to be strictly invariant.

10.1 Definition. An MPDS $(\Omega, \mathcal{F}, \mu, T)$ is said to be *strong mixing* if for all $A, B \in \mathcal{F}$,

$$\lim_n \mu(A \cap T^{-n}B) = \mu(A)\mu(B); \tag{10.2}$$

it is said to be *weak mixing* if for all $A, B \in \mathcal{F}$,

$$\lim_n \frac{1}{n} \sum_{k=0}^{n-1} \bigl|\mu(A \cap T^{-k}B) - \mu(A)\mu(B)\bigr| = 0, \tag{10.3}$$

i.e., $\bigl|\mu(A \cap T^{-k}B) - \mu(A)\mu(B)\bigr|$ converges to 0 in the Cesàro sense.

Hence strong mixing implies weak mixing. In fact weak mixing further implies the system is ergodic. Let $A = B \in \mathcal{F}$ be strictly invariant, then we may replace $T^{-k}B$ by B in (10.3) and get $\mu(B) = \mu(B)^2$.

Notice that the above argument remains true if we remove the $|\cdot|$ in the definition (10.3) of weak mixing. It turns out that

$$\lim_n \frac{1}{n} \sum_{k=0}^{n-1} \mu(A \cap T^{-k}B) = \mu(A)\mu(B) \quad \text{for all } A, B \in \mathcal{F}$$

is in fact equivalent to the saying that the system is ergodic. But the converse requires the Birkhoff 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.

Dyadic transformation

strongly ergodic completely positive entropy isomorphic to Bernoulli shift

occurrence time recurrence time sojourn time

10.4 Poincaré recurrence theorem. $\mu(\{x \in E : T^n x \in E \text{ i.o.}\}) = 1.$

10.5 Von Neumann mean ergodic theorem.

10.6 Birkhoff ergodic theorem.

induced transformation

Chapter 11 Setting up random processes

11.A Product probability measures

It is noteworthy that both results use the axiom of dependent choice in the proof.

11.1 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 $\bigotimes_{n=1}^{\infty} \mathcal{F}_n$.

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. □

11.2 Daniell–Kolmogorov extension/existence theorem.

11.3 Nelson extension theorem [Fol99, Theorem 10.18].

Ionesco–Tulcea extension theorem
Sometimes we want to work with explicit constructions

11.B Poisson processes

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)

11.4 Poisson limit theorem.

11.C Random walks

There are two different perspectives on may look at random walks.

Chapter 12 Markov processes

Chapter 13 Brownian motions

Epilogue

Appendices

A Helpful results from analysis and topology

A.1 Proposition. In a given 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.2 Proposition (Sequential criterion for limits and continuity).

A.3 Proposition. For an increasing function $f\colon \mathbf{R} \to \mathbf{R}$, the set of discontinuities is countable.

A.4 Proposition. Given a set A in a metric space (X, d), the function $d(\,\cdot\,, A)\colon X \to [0, \infty)$ given by

$$d(x, A) = \inf\{d(x, y) : y \in A\}$$

is a continuous function. Also $d(x, A) = 0$ if and only if $x \in \overline{A}$.

A.5 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.6 Proposition. Infinite subset of a compact set has a limit point.

A.7 Proposition. Intersection of a closed set and a compact set is compact.

A.8 Proposition. Compact subsets of a Hausdorff space are closed.

A.9 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 $f(A) = \{0\}$ and $f(B) = \{1\}$.

If X is a metric space (which is necessarily normal), then this is easy. We may just take

$$f(x) = \frac{d(x, A)}{d(x, A) + d(x, B)}.$$

Here is a sketch of the standard proof of this important result in topology. Based on normality, we may inductively dyadically choose (i.e., using DC) an increasing sequence of sets $U_{j/2^n}$ that "lie between" A and B:

$$A \subseteq U_{1/2^n}, \quad \ldots \quad, \overline{U}_{(j-1)/2^n} \subseteq U_{j/2^n}, \quad \ldots \quad, \overline{U}_{(2^n-1)/2^n} \cap B = \emptyset.$$

One can show that the function $f\colon X \to [0,1]$ 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 proof of the following proposition.

A.10 Proposition. Every second countable regular space is normal.

A.11 Urysohn metrization theorem. Every second countable regular space is metrizable.

In particular, every lcscH space is metrizable.

A.12 Proposition.

(a) A second countable space is separable; the converse is also true when we are in a metric space.

(b) A second countable space is Lindelöf, the converse is also true when we are in a metric space.

A.13 Proposition [BS20, Theorem 1.2.13]. Every collection of open sets in a separable metric space contains an at most countable subcollection with the same union.

A.14 Theorem (Characterization of compactness in metric spaces). A subset of a metric space is compact if and only if it is sequentially compact if and only if it is totally bounded and complete.

A.15 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.16 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.15.

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. □

We emphasize X and D here have the same metric structure.

A.17 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.18 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

$$|f(b) - f(a)| \leq (b-a) \sup_{a<x<b} |b-a|.$$

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.19 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.20 Uniform convergence of derivatives [Rud76, Theorem 7.17][1]. Let $f_n\colon (a,b) \to \mathbf{R}$ be a sequence of differentiable functions that converges pointwise to f. If f_n' converges uniformly to some function g, then $f_n \to f$ uniformly and also $f' = g$.

The key part of the proof is the use of the mean value theorem on $f_n' - f_m'$.

[1] Also see Theorem 8.15 and Remark 8.16 in [Kra22].

B Banach spaces

Let X and Y be two normed spaces in this section.

B.1 Example. $(C_b(X), \|\cdot\|_u)$ $(C_0(X), \|\cdot\|_u)$

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.

B.2 Proposition. For $T \in \mathcal{L}(X,Y)$, then T is bounded if and only if it is continuous if and only if it is continuous at 0_X.

B.3 Fact. A bounded linear operator is Lipschitz continuous.

B.4 Proposition. If Y is complete, then $\mathcal{L}(X,Y)$ is complete. In particular the dual space of any normed space is complete.

B.5 Uniform boundedness principle.

B.6 Open mapping theorem.

B.7 Closed graph theorem.

B.8 Baire category theorem.

C Hilbert spaces

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

C.1 Proposition. An inner product space (resp. Hilbert space) is a normed space (resp. Banach space) with the *parallelogram law*:

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

C.2 Cauchy–Schwarz inequality. On an inner product space V, we have

$$|\langle u, v\rangle| \le \|u\|\|v\|,$$

with equality if and only if one is a scalar multiple of the other.

Proof. Expand the nonnegative expression $f(\lambda) := \|u+\lambda v\|^2$ for all $\lambda \in \mathbf{R}$, which contains the desired (real part of the) inner product and has discriminant ≤ 0. □

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 through [Axl24, Section 6], and understand why these results should be true.

C.3 Projection theorem. Given a Hilbert space H and a closed convex subset Y,

(a) for each $x \in H$ there exists a unique

$$y = \operatorname*{arg\,min}_{z \in Y} \|x - z\|,$$

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

$$\operatorname{Re}\langle x - y, z - y\rangle \leq 0 \quad \text{for all } z \in Y. \tag{C.4}$$

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

$$\langle x - y, z\rangle = 0 \quad \text{for all } z \in Y. \tag{C.5}$$

Proof.

(a) Let $D = \inf_{z \in Y} \|x - z\|$, and since Y is close, we may choose a sequence $\{y_n\}$ such that $\|x - y_n\| \to D$ from above. Our goal is to show that it is a Cauchy sequence, and hence converges.

For $n > m \geq 1$, 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 (C.4). Let $z \in Y$ be arbitrary. To get (the real part of) the inner product[2] we consider the expression

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

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

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

Hence

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

[2] like in the proof of Cauchy–Schwarz

and take $\lambda \to 0^+$ gives us (C.4).

For the converse, now suppose (C.4) holds for some $y \in Y$, and we want to show

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

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 (C.4) and (C.5) are equivalent. Because Y is now a subspace of H, equation (C.4) 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. □

C.6 Proposition. For H and its closed subspace Y, π_Y has the following properties:

(a) $\pi_Y \in \mathcal{L}(H)$;

(b) $\pi_Y^2 = \pi_Y$;

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

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

C.7 Riesz representation theorem. 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 $\|f\| = \|v\|$.

An *orthonormal system* $\{e_\alpha\}_{\alpha\in I}$ 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 I is countable.

C.8 Proposition. Suppose we have a finite orthonormal system $\{e_j\}_{j=1}^n$ that spans Y. If $Y \subseteq H$. Then the projection of any $x \in H$ is explicitly $\pi_Y(x) = \sum_{j=1}^n \langle x, e_j\rangle e_j$.

C.9 Proposition. $\sum_{\alpha\in I}\langle x, e_\alpha\rangle e_\alpha$

C.10 Theorem. Let $\{e_\alpha\}_{\alpha\in I}$ be an orthonormal system, then

(a) $\sum_{\alpha\in I}\langle x, e_\alpha\rangle^2 \leq \|x\|^2$, which is known as *Bessel's inequality*;

(b) the equality above holds if and only if the series $x = \sum_{\alpha\in I}\langle x, e_\alpha\rangle e_\alpha$ in H.

Orthonormal decomposition
Parseval's identity
Gram–Schmidt process

C.11 Theorem (complete orthonormal system). $\{e_\alpha\}_{\alpha\in I}$ is an orthonormal basis of H if and only if $\operatorname{span}\{e_\alpha\}$ is dense in H.

C.12 Theorem. H has a countable orthonormal basis if and only if H is separable. Additionally in this case, all bases have the same cardinality.

D Weak and weak-star topologies on normed spaces

initial topology and net convergence $f\colon X\to Y$ 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 $\mathcal{F}=\{f_\beta : X\to Y_\beta\}_{\beta\in B}$ if and only if $f(x_\alpha)\to f(x)$ for all $f\in\mathcal{F}$. This is true for both nets and sequences.

convergence in product spaces

Bogachev 1.6.5 6 8

$x_n\to x$ weakly (i.e., converges in the weak topology) if and only if for all $f\in X^*$, $f(x_n)\to f(x)$

$f_n\to f$ weakly (i.e., converges in the weak-star topology) if and only if for all $x\in X$, $\hat{x}(f_n)=f_n(x)\to\hat{x}(f)=f(x)$

The *strong operator topology* on $\mathcal{L}(X,Y)$ is generated by the evaluation maps $\{T\mapsto Tx : x\in X\}$, where Y is endowed with the norm topology. $T_n\to T$ in the strong operator topology if and only if $T_nx\to Tx$.

The *weak operator topology* on $\mathcal{L}(X,Y)$ is generated by the maps $\{T\mapsto f(Tx) : x\in X, f\in Y^*\}$ to the dual space of Y. $T_n\to T$ in the weak operator topology if and only if for all $x\in X$ and $f\in Y^*$, $f(T_nx)\to f(Tx)$, which is equivalent to saying that $T_nx\to T$ weakly.

D.1 Sequential Banach–Alaoglu–Bourbaki theorem. For a separable normed vector space X, every bounded sequence in X^* has a weak-star convergent subsequence (i.e., X^* is weak-star sequentially compact).

or the generalized Helly selection theorem, for the reasons we discussed after Lemma 8.12

D.2 Banach–Alaoglu–Bourbaki theorem. For a normed vector space X, every closed and bounded subset of X^* is weak-star compact.

metrizability

D.3 Tychonoff's theorem. Arbitrary product of compact topological spaces is compact.

D.4 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].
If the product is finite, then no choice is needed.

D.5 Exercise. Give a direct proof of Tychonoff's theorem for the countable product of compact metric spaces, using metrization.

D.6 Theorem. The countable product of sequentially compact spaces is sequentially compact.

Every weakly convergent sequence is norm bounded

E Weak convergence of general measures

F Topological groups and Haar measures

G A gentle introduction to optimal transport

H 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

$$\log(n+1) \leq \sum_{x=1}^{n} \frac{1}{x} \leq \log(n)+1$$

H.1 Proposition. For two independent random variables $X \sim \text{Exponential}(\lambda)$ and $Y \sim \text{Exponential}(\mu)$, these hold.

(a) $\min\{X,Y\} \sim \text{Exponential}(\lambda+\mu)$;

(b) $P(X \leq Y) = \frac{\lambda}{\lambda+\mu}$;

(c) $\min\{X,Y\}$ and $\{X \leq Y\}$ are independent.

Proof. $P(X>t, Y>t) = P(X>t)P(Y>t) = e^{-(\lambda+\mu)t}$

$$\begin{aligned} P(X-Y \leq t) &= \int_{-\infty}^{\infty} f_X(t+y) f_{-Y}(-y)\,dy \\ &= \lambda\mu e^{-\lambda t} \int_{-\infty}^{\infty} e^{(-\lambda-\mu)y} \mathbf{1}\{y \geq 0, t+y \geq 0\}\,dy \\ &= \frac{\lambda\mu}{\lambda+\mu} e^{-\lambda t} e^{(-\lambda-\mu)y}\Big]_{y=\min\{0,-t\}}^{\infty} \end{aligned}$$

Hence

$$P(X-Y \leq t) = \begin{cases} \frac{\lambda\mu}{\lambda+\mu} e^{-\lambda t} & \text{if } t \geq 0, \\ \frac{\lambda\mu}{\lambda+\mu} e^{\mu t} & \text{if } t < 0. \end{cases}$$

$$P(X>t, Y>t, X \leq Y) = P(X>t, Y>t)P(X \leq Y)$$

LHS $= \int_t^\infty \mu e^{-\mu y} \int_t^y \lambda e^{-\lambda x}\,dx\,dy$

□

The converse is also true Poisson thinning

H.2 Birthday problem.

H.3 Pólya's urn.

H.4 Gambler's ruin.

H.5 Coupon collector's problem.

H.6 Simple random walks on $\mathbf{Z}^d$.

H.7 Bernoulli bond percolation.

Bibliography

[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.

[Bil99] Patrick Billingsley. *Convergence of Probability Measures.* 2nd ed. John Wiley & Sons, 1999.

[Bog07] Vladimir I. Bogachev. *Measure Theory.* Springer Berlin Heidelberg, 2007.

[Bog18] Vladimir I. Bogachev. *Weak Convergence of Measures.* American Mathematical Society, 2018.

[BS20] Vladimir I. Bogachev and Oleg G. Smolyanov. *Real and Functional Analysis.* Springer International Publishing, 2020.

[Fal19] Neil Falkner. "Hahn's Proof of the Hahn Decomposition Theorem, and Related Matters". *The American Mathematical Monthly* 3 (Mar. 2019), pp. 264–268.

[Fol99] Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications.* 2nd ed. John Wiley & Sons, 1999.

[Her06] Horst Herrlich. *Axiom of Choice.* Springer Berlin Heidelberg, 2006.

[Kal02] Olav Kallenberg. *Foundations of Modern Probability.* 2nd ed. Springer New York, 2002.

[Kra22] Steven G. Krantz. *Real Analysis and Foundations.* 5th Ed. CRC Press, Boca Raton, FL, 2022.

[Par67] K. R. Parthasarathy. *Probability Measures on Metric Spaces.* Academic Press, 1967.

[RF23] Halsey Royden and Patrick M. Fitzpatrick. *Real Analysis.* 5th ed. Pearson, 2023.

[Roc24] Sébastien Roch. *Modern Discrete Probability: An Essential Toolkit.* Cambridge University Press, 2024.

[Rud76] Walter Rudin. *Principles of Mathematical Analysis.* 3rd ed. McGraw-Hill, 1976.

[Rud87] Walter Rudin. *Real and Complex Analysis.* 3rd ed. McGraw-Hill, 1987.

[Sae96] Sadahiro Saeki. "A Proof of the Existence of Infinite Product Probability Measures". *The American Mathematical Monthly* 8 (Oct. 1996), pp. 682–683.

[Sch17] René L. Schilling. *Measures, Integrals and Martingales.* 2nd ed. Cambridge University Press, 2017.

List of Definitions