Lecture notes for STAT 547C: Topics in Probability (draft)

Ben Bloem-Reddy

December 5, 2022

Contents

List of Exercises

A few words on mathematics in these notes and generally

Mathematics in these notes. I aim to be mathematically rigorous in these notes, but I will also include some (attempts at) intuitive explanations of challenging technical concepts. Appealing to intuition sometimes requires relaxing the rigor. It should be clear when I am making an intuition-based argment—if it is not, please ask.

Notation. I will mostly follow the notation of Çinlar [Çin11], who lists some frequently used notation on page ix. If I deviate from that notation, I will say so.

Lemma, Proposition, Theorem, etc. As is typically the case, only the most important results will be called theorems. Propositions are typically intermediate results that might are of interesting on their own, but whose importance—at least within this material—does not reach the lofty heights of the exalted theorems. Lemmas are intermediate or supporting results that often could be included within the body of the proof they support. Different authors have different preferences, but well-structured lemmas can enhance the conceptual clarity and readability of long or complex proofs.

Writing mathematics. Something to keep in mind as you do research is that the clarity of your writing can greatly affect how your work is received—even the very best work will be hindered by poor (or even mediocre) writing. Writing mathematics can be especially challenging. I encourage you to practice in your assignment write-ups and especially in your final project. I will post some resources that I have found helpful to my writing.

Disclaimer. The primary purpose of these notes is to stay organized during lecture. They are a **work in progress**. They **do not** replace the readings. Furthermore, they are not a model of good mathematical writing. If something is unclear, please ask. If you find a typo or an error, please let me know.

1 Sets, classes of subsets, and measurable spaces

Reading: Çinlar [Çin11], I.1.

Supplemental: Bass [Bas22], Chapters 1-2. (**Chapter 1 is strongly recommended as a brief review of analysis and set theory.**)

Learning Objectives. At the end of this section, you will be able to do the following.

- Define and use basic set properties for collections of sets.
- Define monotone sequences of sets, and show that their limits exist.
- Define σ-algebra and prove that a given collection of sets is or is not a σ-algebra.
- Define the σ-algebra generated by a collection of sets, the Borel σ-algebra of a topological space and show that the Borel σ-algebra of $\mathbb{R}$ is generated by the collection of open intervals of $\mathbb{R}$ (and other collections).
- Define measurable spaces and product spaces.
- Define p- and d-systems and monotone classes; explain the high-level structure of the proof of the monotone class theorem.

Overview. Our basic aim to start is to define a probability measure on as general a space as possible. For statistical purposes, we want the elements of the space to encode all possible outcomes of an experiment, and we want to define a function for assigning probability to those outcomes in a self-consistent way (i.e., the probability of all possible outcomes is 1, and is equal to the sum of the probabilities of each possible outcome).

We run into problems when trying to work in an uncountable space—problems that were mostly swept under the rug in undergraduate probability. In essence, there were too many points (uncountably many) that we require the probability to "measure" in a self-consistent way, purely by counting. This raises the question of what we might measure instead (and get something useful).

This section is devoted primarily to introducing and defining the relevant terms, and exploring some of their properties.

1.1 Basic set notation and operations

We'll start with some notations and definitions.

Following Çinlar's notation, let E denote a set. (For the purposes of this and the next few sections, capital letters will typically denote sets. Once we start working with random variables, capital letters will generally indicate random variables.)

Subsets. Let A be a **subset** of E, denoted $A \subset E$. $A = E$ if and only if $A \subset E$ and $A \supset E$.

The empty set is denoted by $\emptyset$.

Note that some authors use $A \subseteq E$ to denote that A is a subset of E, and $A \subset E$ to denote that it is a *proper* subset, that is, $A \neq E$. I will follow Çinlar and use $\subset$ to mean subset. Often, the distinction won't matter; if it does, then I will use $A \subsetneq E$ to denote proper subset.

Set operations. The basic set operations on subsets A, B of E are

- **union**, denoted $A \cup B$;

- **intersection**, denoted $A \cap B$;
- **complement** of B in A, denoted $A \setminus B$. E will typically be a generic "universal" or reference set, for which we use the notation B^c to denote $E \setminus B$.

Logic and set operations. Set operations encode the logical relationships `or`, `and`, and `not`, respectively. We can think of constructing a "probabilistic query"[1] as combining sets of outcomes via logical operations (e.g., "It's will rain tonight." `and` "It rained last night."). Sets and set operations are the natural mathematical objects.

Collections of (sub)sets. We will we be working extensively with collections of subsets. For an arbitrary index I, we denote a **collection** by $\mathcal{C} = \{A_i : i \in I\}$. We write

$$\bigcap_{i \in I} A_i \ , \quad \bigcup_{i \in I} A_i \tag{1.1}$$

for the union and intersection, respectively, of all sets A_i in the collection.

A collection $\mathcal{C}$ is **disjointed** if $A_i \cap A_j = \emptyset$ for all $A_i, A_j \in \mathcal{C}$, $i \neq j$.

A **partition** of a set E is a countable disjointed collection of sets whose union is E.

Closure. A collection $\mathcal{C}$ is said to be **closed** under intersections if $A \cap B \in \mathcal{C}$ whenever $A, B \in \mathcal{C}$.

$\mathcal{C}$ is **closed under countable intersections** if the intersection of every countable collection of sets in $\mathcal{C}$ is also in $\mathcal{C}$.

Closure under complements, unions, and countable unions are defined analogously.

Sequences of sets. Let $(A_n) = A_1, A_2, \ldots$ be a sequence of sets in $\mathcal{C}$. The **limit superior** and **limit inferior** are defined as

$$\limsup_n A_n = \bigcap_{n \geq 1} \bigcup_{k \geq n} A_k \tag{1.2}$$

$$\liminf_n A_n = \bigcup_{n \geq 1} \bigcap_{k \geq n} A_k \ . \tag{1.3}$$

If the two are equal, then we call it the **limit** of (A_n).

A sequence of sets is **monotone increasing** if $A_1 \subset A_2 \subset \cdots$. Likewise, a sequence of sets is **monotone decreasing** if $A_1 \supset A_2 \supset \cdots$. In both cases, the limit exists. Call it A_∞. Then we write $A_n \nearrow A_\infty$ for increasing sequences and $A_n \searrow A_\infty$ for decreasing sequences.

Exercise 1 (Limits of monotone increasing sequences of sets)**:**

Show that $A_n \nearrow \cup_m A_m$ for any monotone increasing sequence (A_n).

Solution: Evaluating (1.2)–(1.3) in both cases yields the answer. In particular, consider $\limsup_n A_n$ term-by-term:

[1] Thanks to Trevor Campbell for the analogy.

- $n = 1$: $\cup_{k \geq n} A_k = \cup_{m \geq 1} A_m$.
- $n = 2$: $\cup_{k \geq n} A_k = \cup_{m \geq 2} A_m$.
- And so on.

Now if we take the intersection of these terms, we get, for example,

$$(\cup_{m \geq 1} A_m) \cap (\cup_{m \geq 2} A_m) = \cup_{m \geq 2} A_m \ .$$

But $A_1 \subset A_2$, so $\cup_{m \geq 2} A_m = \cup_{m \geq 1} A_m$, and therefore $\limsup_n A_n = \cup_n A_n$.

Similarly, $\liminf_n A_n = \cup_n A_n$ because $\cup_{n \geq 1} \cap_{k \geq n} A_k = \cup_{n \geq 1} A_n$. (The intersection of an increasing sequence of sets is just the smallest member of the sequence.) Thus, $\limsup_n A_n = \liminf_n A_n = \lim_n A_n = \cup_n A_n$. Because (A_n) is a monotone increasing sequence, $A_n \nearrow \cup_m A_m$.

Exercise 2 (Limits of monotone decreasing sequences of sets)**:**

Show that $A_n \searrow \cap_m A_m$ for any monotone decreasing sequence.

Solution: Evaluating (1.2) in both cases yields the answer. See solution to Exercise 1.

1.2 σ-algebras

Warning! σ-algebras are notoriously non-intuitive the first (and second, and third, and ...) time you encounter them. I think it's useful to keep in mind that ultimately we want to have a properly defined notion of probability that takes as input a query (as described above) and outputs values in $[0, 1]$. We also want that probability to respect some logical relationships (e.g., exclusive outcomes cannot simultaneously occur). A σ-algebra is the mathematical formalism of what we can possibly query; a detailed description of the objects to which we will (eventually) be able to assign measure, and the algebraic rules they obey.

Most authors write "σ-algebra", but you may see "sigma-alebra". You may see both within the same text (as in Çinlar) including Greek letters in LaTeX section headings causes problems for the hyperref package.[2]

Definition. A non-empty collection $\mathcal{E}$ of subsets of E is called an **algebra** if it is closed under *finite* unions and complements. $\mathcal{E}$ is a **σ-algebra** if it is closed under *countable* unions and complements. In math:

1. $A \in \mathcal{E} \Rightarrow E \setminus A \in \mathcal{E}$
2. $A_1, A_2, \ldots \in \mathcal{E} \Rightarrow \bigcup_{n \geq 1} A_n \in \mathcal{E}$

Some authors also list closure under countable intersections. However, observe that with closure under complements and countable unions, closure under countable intersections is implied:

$$\cap_{n \geq 1} A_n = E \setminus (\cup_{n \geq 1} (E \setminus A_n)) \ . \tag{1.4}$$

[2] Aside: most of the time this can be dealt with by using the `texorpdfstring` command.

This is an element of $\mathcal{E}$ because:

1. each entry in the union is in $\mathcal{E}$ (closed under complements);
2. the union of those entries is in $\mathcal{E}$ (closed under countable unions);
3. the complement of the union is in $\mathcal{E}$ (closed under complements).

This type of recursive closure suggests an algorithm to create a σ-algebra from any collection of sets.[3]

Conversely (and somewhat less intuitively), for a given σ-algebra, we might be able to find a relatively small sub-collection that can be used to produce any of the other sets in the σ-algebra. We'll return to this idea shortly.

Why closure under complements and unions? Back to the probabilistic query analogy: if we're interested in the probability of some set A, we should also be able to query, for example, "not A" (A^c), "A or B" ($A \cup B$), "A or B and not D" ($(A \cup B) \cap D^c$). Closure under complements and unions ensures this.

Why *countable* unions? Countable unions ensure that the σ-algebra also contains the limits of all monotone sequences of its sets—a kind of completeness property that gets used extensively in proofs (implicitly, through the Monotone Class Theorem, which we will see soon).

Every monotone sequence of sets has a limit, denoted A_∞. We denote this by $A_n \nearrow A_\infty$ for increasing sequences and $A_n \searrow A_\infty$.

Lemma 1.1. *A collection of sets $\mathcal{E}$ that is closed under finite unions is also closed under countable unions if and only if it contains the limits of all monotone* increasing *sequences $A_1 \subset A_2 \subset \cdots$ of its sets.*

Exercise 3 (Countable unions and limits of monotone increasing sequences of sets)**:**

Show that a collection of sets $\mathcal{E}$ that is closed under finite unions is also closed under countable unions if and only if it contains the limits of all monotone increasing sequences $A_1 \subset A_2 \subset \cdots$ of its sets.

Solution: Assume that $\mathcal{E}$ contains the limits of all monotone increasing sequences of its sets. If $A_1, A_2, \ldots \in \mathcal{E}$ is an arbitrary sequence of sets in $\mathcal{E}$, then define $B_1 = A_1$, $B_2 = A_1 \cup A_2$, and so on: $B_n = \cup_{i \leq n} A_i$, which is in $\mathcal{E}$ because $\mathcal{E}$ is closed under finite unions. Moreover, $B_n \nearrow \cup_{m \geq 1} A_m$, which is in $\mathcal{E}$ by assumption. Hence, $\mathcal{E}$ is closed under countable unions.

Conversely, assume that $\mathcal{E}$ is closed under countable unions. Recall from Exercise 2 that $A_n \nearrow \cup_m A_m$ for any monotone increasing sequence (A_n). That is, the limit of every monotone increasing sequence is a countable union. Hence, $\mathcal{E}$ contains the limits of all monotone increasing sequences of its sets. □

Alternatively, we can work with monotone decreasing sequences.

[3] This is what Rohlin had in mind.

Exercise 4 (Countable intersections and limits of monotone decreasing sequences of sets)**:**

Show that a collection of sets $\mathcal{E}$ that is closed under finite intersections is also closed under countable intersections if and only if it contains the limits of all monotone *decreasing* sequences $A_1 \supset A_2 \supset \cdots$ of its sets.

Solution:

Assume that $\mathcal{E}$ contains the limits of all monotone decreasing sequences of its sets. If $A_1, A_2, \ldots \in \mathcal{E}$ is a countable sequence of sets in $\mathcal{E}$, then define $B_1 = A_1$, $B_2 = A_1 \cap A_2$, and so on: $B_n = \cap_{i \leq n} A_i$, which is $\mathcal{E}$ because $\mathcal{E}$ is closed under finite intersections. Moreover, $B_n \searrow \cap_{m \geq 1} A_m$, which is in $\mathcal{E}$ by assumption. Hence, $\mathcal{E}$ is closed under countable intersections.

Conversely, assume that $\mathcal{E}$ is closed under countable intersections. Recall from Exercise 2 that $A_n \searrow \cap_m A_m$ for any monotone decreasing sequence (A_n). Hence, $\mathcal{E}$ contains the limits of all monotone decreasing sequences of its sets. □

Exercise 5 (Equivalence of closure under monotone sequences in the presence of complements)**:**

Show that if $\mathcal{E}$ is closed under complements, then Exercise 3 $\iff$ Exercise 4.

We'll return to these ideas after seeing some examples.

1.3 Examples of σ-algebras

Trivial σ-algebra. Every σ-algebra on E contains at least $\emptyset$ and E. $\mathcal{E} = \{\emptyset, E\}$ is called the **trivial** σ-algebra.

Discrete σ-algebra. The largest σ-algebra is the collection of all subsets of E, denoted 2^E and called the **discrete** σ-algebra. We'll see later that if E is countable then this is the natural σ-algebra with which to define a probability space. If E is uncountable, 2^E won't work (it's too big).

Generated σ-algebra. Fix an arbitrary collection $\mathcal{C}$ of subsets of E, and consider all the σ-algebras that contain $\mathcal{C}$. At the very least, 2^E contains $\mathcal{C}$. Take the intersection of all of those σ-algebras and call the result the σ-algebra **generated** by $\mathcal{C}$, denoted $\sigma\mathcal{C}$ or $\sigma(\mathcal{C})$.

Observe (convince yourself) that $\sigma\mathcal{C}$ is the smallest σ-algebra that contains $\mathcal{C}$.

Exercise 6 (Simple discrete σ-algebra.)**:**

Let $E = \{0, 1, 2\}$. What is 2^E?

Solution: All possible subsets of E: $\{\emptyset, \{0\}, \{1\}, \{2\}, \{0, 1\}, \{0, 2\}, \{1, 2\}, \{0, 1, 2\}\}$.

Exercise 7 (Simple generated σ-algebra.)**:**

Let $E = [0, 1]$ and $\mathcal{C} = \{[0, \frac{1}{2}]\}$. What is $\sigma\mathcal{C}$?

Solution: $\sigma\mathcal{C} = \{\emptyset, [0,1], [0,\frac{1}{2}], (\frac{1}{2},1]\}$

The following proposition collects some basic relationships between collections and their generated σ-algebras.

Proposition 1.2. *Let $\mathcal{C}$ and $\mathcal{D}$ be two collections of subsets of E. Then:*

(a) If $\mathcal{C} \subset \mathcal{D}$ then $\sigma\mathcal{C} \subset \sigma\mathcal{D}$.

(b) If $\mathcal{C} \subset \sigma\mathcal{D}$ then $\sigma\mathcal{C} \subset \sigma\mathcal{D}$.

(c) If $\mathcal{C} \subset \sigma\mathcal{D}$ and $\mathcal{D} \subset \sigma\mathcal{C}$, then $\sigma\mathcal{C} = \sigma\mathcal{D}$.

(d) If $\mathcal{C} \subset \mathcal{D} \subset \sigma\mathcal{C}$, then $\sigma\mathcal{C} = \sigma\mathcal{D}$.

Proof.

(a) Since $\mathcal{C} \subset \mathcal{D}$, $\mathcal{C} \subset \sigma\mathcal{D}$. By definition, $\sigma\mathcal{C}$ is the smallest σ-algebra that contains $\mathcal{C}$, so it must be that $\sigma\mathcal{C} \subset \sigma\mathcal{D}$.

(b) The argument is the same as in part (a).

(c) Applying part (b) to $\mathcal{C} \subset \sigma\mathcal{D}$ implies that $\sigma\mathcal{C} \subset \sigma\mathcal{D}$. Applying part (b) to $\mathcal{D} \subset \sigma\mathcal{C}$ implies that $\sigma\mathcal{D} \subset \sigma\mathcal{C}$. Together these imply that $\sigma\mathcal{C} = \sigma\mathcal{D}$.

(d) Part (a) applied to $\mathcal{C} \subset \mathcal{D}$ implies that $\sigma\mathcal{C} \subset \sigma\mathcal{D}$. Part (b) applied to $\mathcal{D} \subset \sigma\mathcal{C}$ implies that $\sigma\mathcal{D} \subset \sigma\mathcal{C}$. Together these imply that $\sigma\mathcal{C} = \sigma\mathcal{D}$.

□

We will see an example of how to use these relationships in Exercise 8.

1.4 A bit of topology

To properly define the next example, we need the concept of a topological space. My reference for this section is Aliprantis and Border [AB06, Ch. 2].[4]

A **topology** τ_S on a set S is a collection of subsets of S satisfying:

1. $\emptyset, S \in \tau_S$.
2. τ_S is closed under finite intersections.
3. τ_S is closed under arbitrary (finite, countable, uncountable) unions.

A non-empty set S equipped with a topology τ_S is called a **topological space**, denoted (S, τ_S). Members of τ_S are called the **open sets** of S. (Convince yourself that your notion of an open set, likely tied to a mental model of an interval of $\mathbb{R}$, is consistent with the definition of the topology τ_S.) The complement of an open set is a **closed set**.

[4] Available as a PDF through the UBC Library at SpringerLink. (If you're not on the UBC network, you might need to go through the Library's website and sign into CWL.)

Most (if not all) spaces we encounter will be topological spaces. In fact, most (if not all) spaces we encounter will be metric spaces. A **metric** on a space S is a function $d : S \times S \to \mathbb{R}$ that is non-negative, symmetric, and satisfies:

1. $d(x, x) = 0$ for all $x \in S$.
2. $d(x, y) = 0$ implies $x = y$ for all $x, y \in S$.
3. $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in S$. (This is the triangle inequality.)

A **metric space** is a pair (S, d), where d is a metric on S.

Given a metric d, define the **open ϵ-ball** around x, $B_\epsilon(x) = \{y : d(x, y) < \epsilon\}$. A set $A \subset S$ is open in the **metric topology generated by** d if for each point $x \in A$ there is an $\epsilon > 0$ such that $B_\epsilon(x) \subset A$.

The metric $d(x, y) = |x - y|$ defines a topology on $\mathbb{R}$, and every open interval (a, b) is an open set in this topology. Every open set in $\mathbb{R}$ can be expressed as the (countable) union of disjoint open intervals [see, e.g., SS05, Thm. 1.3].

This is all the topology we need (and maybe a bit more). The main ideas are:

- A topological space is defined in terms of its open sets.
- The metric of a metric space can be used to define a topology based on open balls.
- For our purposes, the open intervals of $\mathbb{R}$ are a good mental model of a topology generated by a metric.

1.5 Borel σ-algebra

If (E, τ_E) is a topological space then the σ-algebra generated by the collection of all open subsets of E (i.e., τ_E) is called the **Borel σ-algebra**, often denoted $\mathcal{B}(E)$ or $\mathcal{B}_E$. Its elements are called the **Borel sets**.

In most applications (and everything we encounter in this class), we're working in topological spaces (and usually metric spaces). **The Borel σ-algebra is ubiquitous.**

At a high level, one of the main reasons for this ubiquity is that, in the words of Aliprantis and Border [AB06, p. 21], "topology is the abstract study of convergence and approximation". Though we won't study it in this level of detail, convergence and approximation are crucial to measure theory—and to the probability theory built on top of it. The open sets are the basic building blocks of a topological space, so it is natural to construct our system of probabilistic queries (i.e., the σ-algebra) from them.

Exercise 8 ($\mathcal{B}(\mathbb{R})$ is generated by the collection of all open intervals)**:**

Show that $\mathcal{B}(\mathbb{R})$ is generated by the collection of all open intervals of $\mathbb{R}$, $\{(a, b) \colon a, b \in \mathbb{R}\}$.

Solution: Let $\mathcal{C}_I$ denote the collection of all open intervals of $\mathbb{R}$, and $\sigma\mathcal{C}_I$ the σ-algebra generated by it. Likewise, let $\mathcal{O}$ and $\sigma\mathcal{O} = \mathcal{B}(\mathbb{R})$ be the collection of all open sets and its generated σ-algebra (i.e., the Borel σ-algebra).

Each open interval is an open set, so $\mathcal{C}_I \subset \mathcal{O}$, which by Proposition 1.2 (part (a)) implies that $\sigma\mathcal{C}_I \subset \sigma\mathcal{O}$.

Conversely, every open set is the union of at most a countable number of open intervals [see, e.g., Bas22, Prop. 1.5]. Every finite open interval is in $\sigma\mathcal{C}_I$. Now, $(a, \infty) = \lim_{n\to\infty}(a, a + n)$, so we have that $(a, \infty) \in \sigma\mathcal{C}_I$ for each $a \in \mathbb{R}$ (and likewise for intervals $(-\infty, a)$). Therefore, any open set $A \in \mathcal{O}$ is an element of $\sigma\mathcal{C}_I$, i.e., $\mathcal{O} \subset \sigma\mathcal{C}_I$. Proposition 1.2 (part (b)) implies $\sigma\mathcal{O} \subset \sigma\mathcal{C}_I$.

Hence, $\sigma\mathcal{C}_I = \sigma\mathcal{O} = \mathcal{B}(\mathbb{R})$.

This is not unique! $\mathcal{B}(\mathbb{R})$ is also generated by any of the following collections (see Çinlar, Exercise I.1.13):

- The collection of all intervals of the form $(\infty, x]$.
- The collection of all intervals of the form $(x, y]$.
- The collection of all intervals of the form $[x, y]$.
- The collection of all intervals of the form (x, ∞).

In each case (and the one above), x and y can be limited to the rational numbers $\mathbb{Q}$.

Example 1.1. Every interval of $\mathbb{R}$ is a Borel set.

Based on the previous example, the open sets generate $\mathcal{B}(\mathbb{R})$, so clearly they are Borel sets.

The semi-closed intervals are, too: $(x, y] = \cap_{n\geq 1}(x, y + 1/n)$.

The singleton sets are Borel sets. Building on the fact that the open and semi-closed intervals are Borel sets, $\{x\} = [x, y) \setminus (x, y)$. Alternatively, $\{x\} = \cap_{n\geq 1}(x - 1/n, x + 1/n)$.

The closed intervals: $[x, y] = \{x\} \cup (x, y]$.

Observe that there are many other ways we could prove that every interval of $\mathbb{R}$ is a Borel set.

1.6 Measurable spaces

A **measurable space** is a pair $(E, \mathcal{E})$, where E is a set and $\mathcal{E}$ is a σ-algebra on E. The elements of $\mathcal{E}$ are called **measurable sets** of E. When E is a topological space and $\mathcal{E} = \mathcal{B}(E)$, then the measurable sets are called the **Borel sets** of E.

1.6.1 Products of measurable spaces

Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be two measurable spaces. For $A \subset E$ and $B \subset F$, the **product**[5] of A and B is

$$A \times B = \{(x, y) : x \in A, y \in B\} . \tag{1.5}$$

If $A \in \mathcal{E}$ and $B \in \mathcal{F}$, then $A \times B$ is called a **measurable rectangle**.

[5] $A \times B$ is also called the Cartesian product of A and B.

The **product σ-algebra** on $E \times F$ is *generated by* the collection of all measurable rectangles, and is denoted $\mathcal{E} \otimes \mathcal{F}$.

The measurable space $(E \times F, \mathcal{E} \otimes \mathcal{F})$ is the **product** of $(E, \mathcal{E})$ and $(F, \mathcal{F})$, also denoted $(E, \mathcal{E}) \times (F, \mathcal{F})$.

Product spaces play a key role in probability theory for statistical models, particularly in conditioning and conditional independence.

1.7 Special systems of sets: p-systems, d-systems, and monotone classes

Systems of sets with the properties above have special names:

- A collection of subsets of E is called a **p-system** (also known as a **π-system**) if it is closed under finite intersections. ("p" for "product", an alternative to intersection.)
- A collection $\mathcal{C}$ of subsets of E is called a **d-system** (also known as a **λ-system**; "d" is for Dynkin, who introduced these systems to probability) if

  D1. $E \in \mathcal{C}$;

  D2. $A, B \in \mathcal{C}$ and $A \supset B \quad \Rightarrow \quad A \setminus B \in \mathcal{C}$; and

  D3. $(A_n) \in \mathcal{C}$ and $A_n \nearrow A_\infty \Rightarrow A_\infty \in \mathcal{C}$.

We can combine these properties with Lemma 1.1 (closure under countable unions $\iff$ closure under monotone increasing limits) to make the following observations about a collection $\mathcal{C}$ that is both a p-system and a d-system:

1. Properties D1-2 imply that $\mathcal{C}$ is closed under complements.
2. Because $\mathcal{C}$ is also closed under (finite) intersections, it is closed under (finite) unions.
3. Then by Lemma 1.1 and property D3, $\mathcal{C}$ is closed under countable unions (and therefore also under countable intersections).

We've just proven Proposition I.1.6 in Çinlar.

Proposition 1.3. *A collection of subsets of E is a σ-algebra if and only if it is both a p-system and a d-system.*

To prove the monotone class theorem, we need one more lemma, the proof of which is just checking conditions.

Lemma 1.4. *Let $\mathcal{C}$ be a d-system on E. Fix $C \in \mathcal{C}$ and let $\hat{\mathcal{C}} = \{A \in \mathcal{C} \colon A \cap C \in \mathcal{C}\}$. Then, $\hat{\mathcal{C}}$ is again a d-system.*

Exercise 9 (Trace spaces, Çinlar Ex. I.1.15)**:**

Let $(E, \mathcal{E})$ be a measurable space. Fix $D \subset E$ and let

$$\mathcal{D} = \mathcal{E} \cap D = \{A \cap D : A \in \mathcal{E}\} .$$

Show that $\mathcal{D}$ is a σ-algebra on D. It is called the **trace** of $\mathcal{E}$ on D, and $(D, \mathcal{D})$ is called the trace of $(E, \mathcal{E})$ on D. (Both are measurable spaces.)

Solution: Since we're taking intersections of sets with D, we'll always end up with subsets of D. So if $\mathcal{D}$ is a σ-algebra then it is a σ-algebra on D. We just need to check the relevant closure properties of $\mathcal{D}$. Let $B \in \mathcal{D}$. Then by definition, there is some set $A \in \mathcal{E}$ such that $B = A \cap D$. Now, $B^c = D \setminus B = D \setminus (A \cap D) = (E \setminus A) \cap D$, so $D \setminus B \in \mathcal{D}$. For countable unions: If $B_1, B_2, \cdots \in \mathcal{D}$, then $\cup_{m \geq 1} B_m = \cup_{m \geq 1}(A_m \cap D) = (\cup_{m \geq 1} A_m) \cap D$, so $\cup_{m \geq 1} B_m \in \mathcal{D}$.

The **monotone class theorem** is often useful in showing that a property of some collection $\mathcal{C}$ also holds for $\sigma\mathcal{C}$. It will be a big (useful) hammer later on in the course.

Theorem 1.5 (Monotone class theorem)**.** *If a d-system contains a p-system, then it contains the σ-algebra generated by that p-system.*

Proof. Let $\mathcal{C}$ be a p-system on E, and define $\mathcal{D}$ to be the smallest d-system on E that contains $\mathcal{C}$. That it is, $\mathcal{D}$ is the intersection of all d-systems on E that contain $\mathcal{C}$. Recall that $\sigma\mathcal{C}$ is the smallest σ-algebra containing $\mathcal{C}$, i.e., the intersection of all σ-algebras containing $\mathcal{C}$. Therefore, if we can show that $\mathcal{D}$ is a σ-algebra, then since it contains $\mathcal{C}$ it also contains $\sigma\mathcal{C}$.

$\mathcal{D}$ is defined to be a d-system, so we just need to show that it is also a p-system—Proposition 1.3 takes care of the rest. That is, we need to show that for $\mathcal{D}$ defined as above, $A \cap B \in \mathcal{D}$ whenever $A, B \in \mathcal{D}$. We will show this in three steps:

1. $A \cap B \in \mathcal{D}$ **whenever** $A, B \in \mathcal{C}$**.** Since $\mathcal{C}$ is assumed to be a p-system contained in $\mathcal{D}$, this is true.
2. $A \cap B \in \mathcal{D}$ **whenever** $A \in \mathcal{D}$ **and** $B \in \mathcal{C}$**.** Fix $B \in \mathcal{C}$ and define $\mathcal{D}_B = \{A \in \mathcal{D} \colon A \cap B \in \mathcal{D}\}$. It's straightforward to see that $\mathcal{C} \subset \mathcal{D}_B$, because in the case that $A \in \mathcal{C}$ we're just in the setting of the previous step. Moreover, by Lemma 1.4, $\mathcal{D}_B$ is a d-system. Hence, $\mathcal{D}_B$ is a d-system that contains $\mathcal{C}$, and therefore it must contain the smallest d-system containing $\mathcal{C}$: $\mathcal{D} \subset \mathcal{D}_B$. This holds for any $B \in \mathcal{C}$, and therefore $A \cap B \in \mathcal{D}$ whenever $A \in \mathcal{D}$ and $B \in \mathcal{C}$.
3. $A \cap B \in \mathcal{D}$ **whenever** $A \in \mathcal{D}$ **and** $B \in \mathcal{D}$**.** Now fix $A \in \mathcal{D}$, and define $\mathcal{D}_A = \{B \in \mathcal{D} \colon A \cap B \in \mathcal{D}\}$. By the previous part, $\mathcal{D}_A$ contains $\mathcal{C}$ because when $B \in \mathcal{C}$, we have $A \cap B \in \mathcal{D}$. By Lemma 1.4, $\mathcal{D}_A$ is a d-system. Hence, $\mathcal{D} \subset \mathcal{D}_A$. That is, $A \cap B \in \mathcal{D}$ whenever $A \in \mathcal{D}$ and $B \in \mathcal{D}$.

We've shown that $\mathcal{D}$ is a d-system and a p-system, so by Proposition 1.3, $\mathcal{D}$ is a σ-algebra. We've also shown that $\mathcal{D}$ contains $\mathcal{C}$, which means that $\sigma(\mathcal{C}) \subset \mathcal{D}$.

We've actually shown a bit more. We defined $\mathcal{D}$ to be the smallest d-system that contains $\mathcal{C}$, and since *any* σ-algebra is a d-system (Proposition 1.3), no σ-algebra that contains $\mathcal{C}$ can be smaller than $\mathcal{D}$. That is, $\mathcal{D} \subset \mathcal{E}$, for any σ-algebra $\mathcal{E}$ that contains $\mathcal{C}$. We also showed that $\mathcal{D}$ is a σ-algebra, and therefore it is the smallest σ-algebra that contains $\mathcal{C}$: $\mathcal{D} = \sigma(\mathcal{C})$.

Now, any d-system that contains $\mathcal{C}$ also contains $\mathcal{D} = \sigma(\mathcal{C})$ (by definition of $\mathcal{D}$).

□

2 Measurable functions

Reading: Çinlar, I.2.

Supplemental: Bass [Bas22], Ch. 5.1

Learning Objectives. At the end of this section, you will be able to do the following.

- Define and prove the measurability of a function with respect to two σ-algebras in a number of ways (directly and by using a generating collection).
- Show that compositions of measurable functions are measurable.
- Show that continuous functions are Borel-measurable.
- Define measurable space and know some common (standard) measurable spaces.

Overview. With Section 1 in hand, most courses in probability theory move on to defining probability measures and/or random variables. We will put off probability for another two sections and remain in the realm of deterministic measure theory. When we get to probability theory, we'll see that for the most part it can be interpreted as a special case of measure theory. For example, random variables are measurable functions that take on special meaning in probability theory. However, the language of probability theory is still rooted in pre-measure-theoretic ideas, and somehow "feels" different. Çinlar [Çin11, p. 49] says,

> "... our attitude and emotional response toward one [measure theory] is entirely different from those toward the other [probability theory]. On a measure space everything is deterministic and certain, on a probability space we face randomness and uncertainty."

I tend to agree with this assessment. With this in mind, we'll continue on with measure theory—without the emotional baggage of randomness and uncertainty.

2.1 Measurability of functions

Let E and F be sets. A **function** or **mapping**[6] f from E into F is a rule that assigns an element $f(x) \in F$ to each $x \in E$. This is typically written as $f : E \to F$. For more specificity, one may also write (less commonly) $f : x \mapsto x^2 + 5$ as the function mapping $\mathbb{R}$ to $\mathbb{R}_+$. You may also see $f : E \to F, x \mapsto x^2 + 5$.

For any mapping $f : E \to F$ and subset $B \subset F$, the **inverse image** of B under f is

$$f^{-1}B = \{x \in E : f(x) \in B\} . \tag{2.1}$$

Picture on board.

As a set operation, the inverse image "plays well" with other set operations (i.e., commutes, distributes).

[6] Be aware that different sub-fields of mathematics may or may not treat these as synonymous (Çinlar does). For example, see this Math Stack Exchange thread. I will treat them as synonymous.

Lemma 2.1. *Let f be a mapping from E into F. Then,*

$$f^{-1}\emptyset = \emptyset\,, \quad f^{-1}F = E\,, \quad f^{-1}(B \setminus C) = (f^{-1}B) \setminus (f^{-1}C) \tag{2.2}$$
$$f^{-1}\bigcup_i B_i = \bigcup_i f^{-1}B_i\,, \quad f^{-1}\bigcap_i B_i = \bigcap_i f^{-1}B_i\,,$$

for all subsets B and C of F and arbitrary collections $\{B_i : i \in I\}$ of subsets of F.

Exercise 10 (Proof of Lemma 2.1)**:**

Prove Lemma 2.1.

Measurable functions. Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. A mapping $f : E \to F$ is **measurable relative to $\mathcal{E}$ and $\mathcal{F}$** if $f^{-1}B \in \mathcal{E}$ for all $B \in \mathcal{F}$. For short, we also say that f is **$\mathcal{E}/\mathcal{F}$-measurable**.

Often, $\mathcal{F}$ will be fixed or obvious (e.g., F is a topological space and $\mathcal{F} = \mathcal{B}(F)$), in which case we say that f is **measurable with respect to $\mathcal{E}$**, or **$\mathcal{E}$-measurable**.

As a further simplification, when both $\mathcal{E}$ and $\mathcal{F}$ are fixed, we may just say that f is a measurable function. (I will try to avoid this unless there is no chance for confusion, but much of the literature is written this way.)

The following proposition is very useful: if we can establish measurability with respect to a generating collection $\mathcal{F}_0$ such that $\mathcal{F} = \sigma\mathcal{F}_0$, then we have also established $\mathcal{F}$-measurability.

Proposition 2.2. *Let $\mathcal{F}_0 \subset \mathcal{F}$ such that $\sigma\mathcal{F}_0 = \mathcal{F}$. Then a mapping $f : E \to F$ is measurable relative to $\mathcal{E}$ and $\mathcal{F}$ if and only if $f^{-1}B \in \mathcal{E}$ for every $B \in \mathcal{F}_0$.*

Exercise 11 (Proof of Proposition 2.2)**:**

Prove Proposition 2.2.

Hint: Use Lemma 2.1.

2.2 Numerical functions

Some notation for variations on the real line:

- $\mathbb{R} = (-\infty, +\infty)$
- $\bar{\mathbb{R}} = [-\infty, +\infty]$
- $\mathbb{R}_+ = [0, +\infty)$
- $\bar{\mathbb{R}}_+ = [0, \infty]$

For a measurable space $(E, \mathcal{E})$, a **numerical function** on E is a mapping from E into $\bar{\mathbb{R}} := [-\infty, +\infty]$, or some subset thereof. It is **positive** if all of its values are in $\bar{\mathbb{R}}_+$. (Note that some authors use *non-negative* for this.)

A numerical function is **$\mathcal{E}$-measurable** if it is $\mathcal{E}/\mathcal{B}(\mathbb{R})$-measurable. If E is topological and $\mathcal{E} = \mathcal{B}(E)$, then $\mathcal{E}$-measurable functions are called **Borel** functions.

Positive and negative parts. For a and b in $\mathbb{R}$, we write $a \vee b$ to denote $\max\{a, b\}$ and $a \wedge b$ to denote $\min\{a, b\}$. When applied to numerical functions, the maximum is taken pointwise: $f \vee g$ is a function whose value at x is $f(x) \vee g(x)$. For a measurable space $(E, \mathcal{E})$ and a function $f : E \to \bar{\mathbb{R}}$, the **positive part** and **negative part** of f are

$$f^+ = f \vee 0\,, \quad \text{and} \quad f^- = -(f \wedge 0)\,.$$

It should be (intuitively) clear that f is $\mathcal{E}$-measurable if and only if both f^+ and f^- are. This fact is important enough that it is stated as a proposition in Çinlar [Çin11, Prop. I.2.9] because we can obtain many results for arbitrary f from the corresponding results for positive functions.

Recall from Exercise 8 that various different collections of intervals generate $\mathcal{B}(\mathbb{R})$. Recall also from Proposition 2.2 that we can establish measurability by checking measurability on a generating set. Then there is a version of the following proposition for each type of interval collection generating $\mathcal{B}(\mathbb{R})$.

Proposition 2.3. *A mapping $f : E \to \bar{\mathbb{R}}$ is $\mathcal{E}$-measurable if and only if, for every $r \in \mathbb{R}$, $f^{-1}[-\infty, r] \in \mathcal{E}$.*

σ-algebra generated by a function. Let E be a set and $(F, \mathcal{F})$ a measurable space. For $f : E \to F$, define

$$f^{-1}\mathcal{F} = \{f^{-1}B : B \in \mathcal{F}\}\,. \tag{2.3}$$

This is a σ-algebra on E, called the **σ-algebra generated by** f. $f^{-1}\mathcal{F}$ is the smallest σ-algebra on E such that f is measurable relative to it and $\mathcal{F}$.

Exercise 12 (σ-algebra generated by a function)**:**

Show that $f^{-1}\mathcal{F}$ as in Eq. (2.3) is a σ-algebra on E.

Another way to define measurability. If $(E, \mathcal{E})$ is a measurable space, then f is measurable relative to $\mathcal{E}$ and $\mathcal{F}$ if and only if $f^{-1}\mathcal{F} \subset \mathcal{E}$. This is another way of defining measurability.

Why bother? If you're confused about why we're bothering with all of this, it might be helpful to think of measurability this way: if we're given a set of values $B \subset F$ and a mapping $f : E \to F$, does $\mathcal{E}$ contain all of the possible values $x \in E$ that could have been mapped to B via f? This typically (though not always) comes down to whether $\mathcal{E}$ is large (or "fine") enough. For example, if $f : E \to \mathbb{R}$ is a non-constant mapping and $\mathcal{E}$ is the trivial σ-algebra $\{\emptyset, E\}$, then f is non-measurable (relative to $\mathcal{E}$ and $\mathcal{B}(\mathbb{R})$.) Cases that arise in practice are not typically so extreme, but it is common to ask, for two functions $f : E \to F$ and $g : E \to G$, whether g is measurable relative to $f^{-1}\mathcal{F}$ (and vice versa). In terms of probability and statistics, this is basically equivalent to whether one random variable is measurable relative to another.

Example 2.1. A **random variable** is a measurable function that we give special treatment in the context of probability and statistics. Consider the following highly idealized example of the current temperature in Vancouver. We can treat the current temperature as a random variable $X\colon \Omega \to \mathbb{R}$, where Ω is a "special" space that we can think of as encoding all states of the universe. $\omega \in \Omega$ represents a particular state of the universe, and $X(\omega)$ is the local temperature when the universe is in state ω. Measurability in this context just means that $X^{-1}B$, for $B \in \mathcal{B}(\mathbb{R})$, is in whatever σ-algebra we've defined on Ω. We will treat this in much more depth in later sections.

2.3 Compositions of functions

Let $(E, \mathcal{E})$, $(F, \mathcal{F})$, and $(G, \mathcal{G})$ be measurable spaces. Let $f : E \to F$ and $g : F \to G$. The **composition** of f and g is the mapping $g \circ f : E \to G$ defined by

$$(g \circ f)(x) = g(f(x)) \,, \quad x \in E \,. \tag{2.4}$$

In statistics, machine learning, and related fields, we often compose functions and typically do so without any second thoughts about measurability. That's okay because measurable functions of measurable functions are measurable.

Proposition 2.4. *If f is $\mathcal{E}/\mathcal{F}$-measurable and g is $\mathcal{F}/\mathcal{G}$-measurable, then $g \circ f$ is $\mathcal{E}/\mathcal{G}$-measurable.*

Exercise 13 (Measurable functions of measurable functions are measurable)**:**

Prove Proposition 2.4.

2.4 Continuous functions

Given two topological spaces (S, τ_S) and (T, τ_T), a **continuous function** is a function $f\colon S \to T$ such that $f^{-1}(A) \in \tau_S$ whenever $A \in \tau_T$. That is, the inverse image of a continuous function preserves open sets.

Theorem 2.5. *Let (S, τ_S) and (T, τ_T) be two topological spaces. with Borel σ-algebras $\mathcal{B}(S) = \sigma(\tau_S)$ and $\mathcal{B}(T) = \sigma(\tau_T)$. Then every continuous function $f\colon S \to T$ is measurable with respect to $\mathcal{B}(S)$ and $\mathcal{B}(T)$.*

Exercise 14 (Borel-measurability of continuous functions)**:**

Prove Theorem 2.5.

Hint: Recall that the Borel σ-algebra is generated by the open sets, i.e., $\mathcal{B}(S) = \sigma(\tau_S)$ and $\mathcal{B}(T) = \sigma(\tau_T)$.

Example 2.2. Mathematically, a *deep (artificial) neural network* is function, $\mathsf{NN}\colon \mathbb{R}^{d_0} \to \mathbb{R}^{d_L}$, constructed by repeated composition of functions: for $x \in \mathbb{R}^{d_0}$,

$$\mathsf{NN}(x) = f_L \circ f_{L-1} \circ \cdots \circ f_\ell \circ \cdots f_1(x) \ , \quad \text{where } f_\ell : \mathbb{R}^{d_{\ell-1}} \to \mathbb{R}^{d_\ell} \ . \tag{2.5}$$

In order to support automatic differentiation software libraries, each f_ℓ is continuous and differentiable (almost everywhere[a]), typically an affine transformation followed by a continuous pointwise nonlinearity. Continuity is sufficient for a function to be Borel-measurable, by Theorem 2.5. Observe also that composition preserves continuity, as long as the functions being composed are continuous. Combined with Proposition 2.4, this implies that NN is Borel-measurable.

[a] "Almost everywhere" means that any sets (points) of discontinuity or non-differentiability have measure zero. We will study this in detail soon.

2.5 Standard and common measurable spaces

Isomorphisms. Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. Let $f : E \to F$ be a bijection and denote by $\hat{f}$ the functional inverse: $\hat{f}(y) = x$ if and only if $f(x) = y$. If f is $\mathcal{E}/\mathcal{F}$-measurable and $\hat{f}$ is $\mathcal{F}/\mathcal{E}$-measurable, then f is an **isomorphism** of $(E, \mathcal{E})$ and $(F, \mathcal{F})$. $(E, \mathcal{E})$ and $(F, \mathcal{F})$ are said to be **isomorphic** if there exists an isomorphism between them.

Standard spaces. A measurable space $(E, \mathcal{E})$ is **standard** if it is isomorphic to $(F, \mathcal{B}(F))$ for some Borel subset F of $\mathbb{R}$. (Various other terms are used by other authors for this property. "Borel space" and "standard Borel space" are two of the most common.)

One of the biggest advantages of working with standard Borel spaces is that it's typically enough to show that some property holds on $[0, 1]$ (or $\mathbb{R}$, or any subset thereof—whatever is easiest for what you're trying to prove).

What are some standard measurable spaces? Most of the spaces we encounter in statistics and machine learning are standard Borel spaces. Some examples:

- $\mathbb{R}$, $\mathbb{R}^d$, $\mathbb{R}^\infty$ together with their Borel σ-algebras.
- If E is a **complete separable metric space** (c.s.m.s.),[7] then $(E, \mathcal{B}(E))$ is standard.
- Completeness and separability of a set E depend on the metric. If E is a topological space such that there is a metric on E which defines a topology of E and which makes E c.s.m.s., then E is a **Polish space**[8] and $(E, \mathcal{B}(E))$ is standard.
- Another for which we won't go into detail: for a separable Banach space E (of which a separable Hilbert space is a special case), $(E, \mathcal{B}(E))$ is standard.

[7] **Complete**: Every Cauchy sequence (the elements get arbitrarily close to each other) in E has a limit in E (i.e., parts of the space don't go missing). **Separable**: E has a countable dense subset (think of $\mathbb{Q}$ in $\mathbb{R}$—we can approximate anything in E arbitrarily well using only countable approximations). **Metric space**: A set with a metric on the set (i.e., a notion of distance), (E, d).

[8] From Wikipedia: "Polish spaces are so named because they were first extensively studied by Polish topologists and logicians—Sierpiński, Kuratowski, Tarski and others".

The three standard measurable spaces. A deep result from measure theory says that every standard measurable space (i.e., standard Borel space) is isomorphic to one of:[9]

- $\{1, 2, \dots, n\}$ and its discrete σ-algebra;
- $\{1, 2, \dots\}$ and its discrete σ-algebra;
- $[0, 1]$ and its Borel σ-algebra.

Notation

Çinlar [Çin11] uses $\mathcal{E}$ to denote both the σ-algebra and the class of functions that are measurable relative to it. To avoid confusion, I will use $\mathcal{E}^{\text{fn}}$ to denote the latter.

We haven't encountered them yet, but we will also need the following notation:

- $\mathcal{M}$ denotes an arbitrary collection of numerical functions;
- $\mathcal{M}_+ \subset \mathcal{M}$ denotes the positive functions in $\mathcal{M}$;
- $\mathcal{M}_b \subset \mathcal{M}$ is the collection of bounded functions in $\mathcal{M}$.

For example, $\mathcal{E}_+^{\text{fn}}$ is the class of positive $\mathcal{E}$-measurable numerical functions.

[9] This result is due to Kuratowski, *Sur une généralisation de la notion d'homéomorphie*, Fund. Math. **22** (1934), 206-220. I haven't been able to find a good textbook reference.

3 Measures, probability measures, and probability spaces

Reading: Çinlar, I.3.

Supplemental: Bass [Bas22], Ch. 3. If you're interested in how to construct measures (which we will avoid except for Lebesgue measure on $\mathbb{R}$ later on), check out Ch. 4.

Learning Objectives. At the end of this section, you will be able to do the following.

- Define a measure and a measure space.
- Prove key properties of a measure (e.g., monotonicity and sequential continuity).
- Characterize the uniqueness of measures using a generating p-system, and apply that to $\mathbb{R}$.
- Define and use the following terms: negligible/null set, and almost everywhere/surely.
- Use a measure and a measurable function to define a new measure.

Overview. Sections 1 and 2 followed Çinlar [Çin11] pretty closely. This section will continue to do so, though we will introduce a few ideas from section II.1.

3.1 Measures and measure spaces

Let $(E, \mathcal{E})$ be a measurable space. A **measure** on $(E, \mathcal{E})$ is a mapping $\mu : \mathcal{E} \to \bar{\mathbb{R}}_+$ the following properties:

a) *Zero on the empty set*: $\mu(\emptyset) = 0$.

b) *Countable additivity*: $\mu\big(\cup_n A_n\big) = \sum_n \mu(A_n)$ for every disjointed sequence (A_n) in $\mathcal{E}$.

The number $\mu(A) \in \bar{\mathbb{R}}_+$ is called the measure or **mass** of A. It is also written as μA.

A **measure space** is a triplet $(E, \mathcal{E}, \mu)$ where $(E, \mathcal{E})$ is a measurable space and μ is a measure on it.

3.1.1 Probability measures and probability spaces

A **probability measure** $\mathbb{P}$ on a measurable space $(\Omega, \mathcal{H})$ is a measure such that $\mathbb{P}(\Omega) = 1$. A **probability space** is a triple $(\Omega, \mathcal{H}, \mathbb{P})$, where $(\Omega, \mathcal{H})$ is a measurable space and $\mathbb{P}$ is a probability measure on it. Nothing more, nothing less.

The fact that the total mass of Ω is one lets us say a few more things, but mathematically there is not a major difference. However, an entirely different vocabulary—some might say an entirely different conceptual framework—is used to describe probability spaces. Recall what Çinlar says: "our attitude and emotional response toward one is entirely different from those toward the other".

The space Ω is called the **sample space** and its elements ω are called **outcomes**. The σ-algebra $\mathcal{H}$ is called the **grand history** (maybe a bit less common these days) and its elements are called **events**. Ω is often described as containing all possible states of the world/universe, each corresponding to an element ω, and the random variables we work with are functions whose value changes depending on the state of the world.

3.2 Examples

The following examples are measures that we will encounter frequently. For each of the following, let $(E, \mathcal{E})$ be a measurable space on which we will define the measure.

Example 3.1 Dirac measure. Let x be a fixed point of E. For each $A \in \mathcal{E}$, define

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} . \tag{3.1}$$

δ_x is a measure on $(E, \mathcal{E})$ called the **Dirac measure**.

Example 3.2 Counting, discrete, and empirical measures. Let D be a fixed subset of E. For each $A \in \mathcal{E}$, let $\nu_D(A)$ be the number of points in $A \cap D$ (this might be infinite). ν_D is a measure on $(E, \mathcal{E})$ called a **counting measure**. Often, D is taken to be countable, in which case

$$\nu_D(A) = \sum_{x \in D} \delta_x(A) , \quad A \in \mathcal{E} . \tag{3.2}$$

For countable D, let $m(x)$ be a positive number for each $x \in D$. Define the **discrete measure**

$$\bar{\nu}_D(A) = \sum_{x \in D} m(x) \delta_x(A) , \quad A \in \mathcal{E} . \tag{3.3}$$

It's helpful to visualize a discrete measure as a mass $m(x)$ attached to a particular point x (sometimes called an **atom** of the measure); $\bar{\nu}_D(A)$ is the sum of the mass of the atoms attached to points in A. If $(E, \mathcal{E})$ is a discrete measurable space, then every measure on it has this form.

The normalized version of the counting measure, with $m(x) = \frac{1}{|D|}$ for each $x \in D$ is the **empirical measure**

$$\hat{\nu}_D(A) = \frac{1}{|D|} \sum_{x \in D} \delta_x(A) , \quad A \in \mathcal{E} , \tag{3.4}$$

which gets its name from its use with, e.g., a sequence of observations $D = \{x_1, x_2, \ldots, x_n\}$.

Example 3.3 Lebesgue measure. A measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called the **Lebesgue measure** on $\mathbb{R}$ if $\lambda(A)$ is the length of A for every interval A. We can see that this definition makes it impossible to display $\lambda(A)$ for every Borel set A, but we can use it to integrate (which, as Çinlar says, is the main thing measures are for). For $\mathbb{R}^2$, the Lebesgue measure measures the "area"; on $\mathbb{R}^3$, the "volume". Specifically, for every half-open rectangle $[[a, b)) := [a_1, b_1) \times \cdots \times [a_n, b_n)$, the Lebesgue measure on $\mathbb{R}^n$ is the set function that assigns the value

$$\lambda^n\big([[a, b))\big) = \prod_{j=1}^{n} (b_j - a_j) \; .$$

We'll study Lebesgue measure on $\mathbb{R}^n$ in detail later in the course, in particular showing that the set function as defined on the measurable rectangles as above has a unique extension to a measure on all of $\mathcal{B}(\mathbb{R}^n)$.

The notation λ for Lebesgue measure is fairly standard, though Çinlar uses *Leb*. I will use λ.

3.3 Some properties of measures

Most of the key properties of measures and probability measures are summarized in the following proposition.

Proposition 3.1. *Measures and probability measures have the following properties:*

Property	*Measure space* $(E, \mathcal{E}, \mu)$	*Probability space* $(\Omega, \mathcal{H}, \mathbb{P})$
Norming*:*	$\mu(\emptyset) = 0$	$\mathbb{P}(\emptyset) = 0$ *and* $\mathbb{P}(\Omega) = 1$
Countable & finite additivity*:* (H_n) *disjointed* $\Rightarrow$	$\mu(\cup_n H_n) = \sum_n \mu(H_n)$	$\mathbb{P}(\cup_n H_n) = \sum_n \mathbb{P}(H_n)$
Monotonicity*:* $H \subset K \Rightarrow$	$\mu(H) \leq \mu(K)$	$\mathbb{P}(H) \leq \mathbb{P}(K)$
Sequential continuity $H_n \nearrow H \Rightarrow$ $H_n \searrow H \Rightarrow$	$\mu(H_n) \nearrow \mu(H)$	$\mathbb{P}(H_n) \nearrow \mathbb{P}(H)$ $\mathbb{P}(H_n) \searrow \mathbb{P}(H)$
Boole's inequality *(a.k.a. union bound)*	$\mu(\cup_n H_n) \leq \sum_n \mu(H_n)$	$\mathbb{P}(\cup_n H_n) \leq \sum_n \mathbb{P}(H_n)$

Proof. See Çinlar [Çin11, Prop. I.3.6 and remarks on p. 50]. □

Exercise 15 (Properties of measures)**:**

Prove that a measure μ has the monotonicity and (increasing) sequential continuity properties given above. When μ is a probability measure show that it also has the decreasing sequential continuity property.

Solution:

Exercise 16 (Restrictions and traces, Çinlar Ex. I.3.11)**:**

Let $(E, \mathcal{E})$ be a measurable space and μ a measure on it. Fix $D \in \mathcal{E}$.

a) Define $\nu(A) = \mu(A \cap D)$, $A \in \mathcal{E}$. Show that ν is a measure on $(E, \mathcal{E})$. It is called the **trace** of μ on D.

b) Let $\mathcal{D}$ be the trace of $\mathcal{E}$ on D (recall Exercise 9). Define $\nu(A) = \mu(A)$ for $A \in \mathcal{D}$. Show that ν is a measure on $(D, \mathcal{D})$. It is called the **restriction** of μ to D.

Solution:

Exercise 17 (Extensions, Çinlar Ex. I.3.12)**:**

Let $(E, \mathcal{E})$ be a measurable space, let $D \in \mathcal{E}$, and let $(D, \mathcal{D})$ be the trace of $(E, \mathcal{E})$ on D. Let μ be a measure on $(D, \mathcal{D})$ and define ν by

$$\nu(A) = \mu(A \cap D) \ , \quad A \in \mathcal{E} \ . \tag{3.5}$$

Show that ν is a measure on $(E, \mathcal{E})$. This device allows us to regard a "measure on D" as a "measure on E".

Solution:

Arithmetic of measures. The set of measures is closed under (positive) linear operations:

- If $c > 0$ and μ is a measure on $(E, \mathcal{E})$, then so is $c\mu$.
- If μ and ν are measures on $(E, \mathcal{E})$, then so is $\mu + \nu$.
- If $\mu_1, \mu_2, \ldots$ are measures on $(E, \mathcal{E})$, then so is $\sum_n \mu_n$.

You can prove these by checking the definition of a measure.

Warning! If μ is a probability measure, each of these operations will generally not result in another probability measure. But consider the following operation. For $c_1, c_2, \ldots \in \mathbb{R}_+$ such that $\sum_{n \geq 1} c_n = 1$, and $\mu_1, \mu_2, \ldots$ each probability measures, form the convex combination,

$$\mu = \sum_{n \geq 1} c_n \mu_n \ .$$

It's easy to show that μ is a probability measure. We'll see a generalization of this when we get to conditional probabilities.

Uniqueness of measures. Less straightforward is the following **uniqueness of measures** result.

Proposition 3.2. *Let $(E, \mathcal{E})$ be a measurable space. Let μ and ν be measures on it with $\mu(E) = \nu(E) < \infty$. Let $\mathcal{C}$ be a p-system that generates $\mathcal{E}$. If μ and ν agree on all elements of $\mathcal{C}$, then μ and ν are identical.*

Proof. The monotone class theorem (Theorem 1.5) demonstrates how useful it is yet again. Suppose that $\mu(A) = \nu(A)$ for every $A \in \mathcal{C}$, and $\mu(E) = \nu(E) < \infty$. We need to show that $\mu(A) = \nu(A)$ for every $A \in \mathcal{E}$.

Let $\mathcal{D} = \{A \in \mathcal{E} \colon \mu(A) = \nu(A)\}$. If we can show that $\mathcal{D}$ is a d-system, then the monotone class theorem implies that $\mathcal{E} \subset \mathcal{D}$, and the proposition is proved. □

Exercise 18 (Proof of Proposition 3.2)**:**

Show that $\mathcal{D} = \{A \in \mathcal{E} \colon \mu(A) = \nu(A)\}$ is a d-system.

Solution: See Çinlar, I.3.7.

Proposition 3.2 has the following important special case when applied to $\mathbb{R}$.

Corollary 3.3. *Let μ and ν be probability measures on $(\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$. Then $\mu = \nu$ if and only if $\mu[-\infty, r] = \nu[-\infty, r]$ for all $r \in \bar{\mathbb{R}}$.*

Atoms, purely atomic measures, diffuse measures. Suppose that the singleton sets $\{x\}$ are measurable (i.e., belong to $\mathcal{E}$) for all $x \in E$. This is true for all standard (Borel) measurable spaces. Let μ be a measure on $(E, \mathcal{E})$. A point x is called an **atom** of μ if $\mu(\{x\}) > 0$. If μ has no atoms, it is said to be **diffuse**; it is **purely atomic** if the set D of atoms is countable and $\mu(E \setminus D) = 0$. The Lebesgue measures are diffuse, and discrete measures are purely atomic.

Note that a measure may have a diffuse part and an atomic part: under quite general conditions, $\mu = \eta + \nu$, where η is a diffuse measure and ν is purely atomic [see Çin11, Prop. 3.9].

Finite and infinite measures. A measure μ on $(E, \mathcal{E})$ is said to be a **finite measure** if $\mu(E) < \infty$. Note that by monotonicity of μ, $\mu(A) < \infty$ for all $A \subset E$. Observe that if a measure is finite then it can be turned into a probability measure by normalizing by $\mu(E)$. μ is said to be **σ-finite** if there is a measurable partition (E_n) of E such that $\mu(E_n) < \infty$ for each n. An example of a σ-finite measure is Lebesgue measure on $\mathbb{R}$: we can partition $\mathbb{R}$ into the intervals $[n, n+1)$, for $n \in \mathbb{Z}$, and $\lambda([n, n+1)) = 1$ for each n. Finally, μ is called Σ-finite or s-finite if it can be written as the countable sum of finite measures: $\mu = \sum_{n \in \mathbb{N}} \mu_n$, each μ_n finite.

Observe that

$$\text{finite} \Rightarrow \sigma\text{-finite} \Rightarrow \Sigma\text{-finite},$$

but the converse implications are not necessarily true.

In this class, we will work almost exclusively with probability measures, but σ-finite measures play a big role in the background (particularly Lebesgue measure).

3.4 Negligible/null sets and completeness

Let $(E, \mathcal{E}, \mu)$ be a measure space. A measurable set B is said to be **negligible** if $\mu(B) = 0$. On a probability space $(\Omega, \mathcal{H}, \mathbb{P})$, a measurable set B with $\mathbb{P}(B) = 0$ is called a **null set**.

A measure space is said to be **complete** if every negligible set is measurable. Likewise for probability spaces and null sets. If a measure space $(E, \mathcal{E}, \mu)$ is not complete, it can be enlarged/extended to its **completion** $(E, \bar{\mathcal{E}}, \bar{\mu})$, for which $\bar{\mathcal{E}}$ contains all negligible sets and μ is extended onto $\bar{\mathcal{E}}$. (We'll take this as fact because it's really just a technical result that we won't revisit; see Çinlar [Çin11, Prop. I.3.10].)

3.5 Almost everywhere/surely

If a claim holds for all $x \in E$ except for on a negligible set, then we say that it holds for almost every x, or **almost everywhere**. If we need to indicate the measure μ that this holds with respect to, we say **μ-almost everywhere**. These are often abbreviated as a.e. and μ-a.e.

The same definitions are true for probability measures/spaces, except we say **almost surely** (or $\mathbb{P}$-almost surely or almost surely $\mathbb{P}$), abbreviated a.s. or $\mathbb{P}$-a.s.

3.6 Image measures

Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. Let ν be a measure on $(E, \mathcal{E})$ and $f : E \to F$ a $\mathcal{E}/\mathcal{F}$-measurable function. Define the mapping $\mu : \mathcal{F} \to \bar{\mathbb{R}}_+$ as

$$\mu(B) = (\nu \circ f^{-1})(B) = \nu(f^{-1}B) \,, \quad B \in \mathcal{F} \,. \tag{3.6}$$

This is well-defined because of the measurability of f, and is called the image, or **image measure**, of ν under f. Other notation includes, variously, $f_{\#}\nu$, $f_*\nu$, ν_f, and $f(\nu)$.

Another common name for the image measure is the **pushforward** measure.

Exercise 19 (Image measure)**:**

Show that $\mu = \nu \circ f^{-1}$ is a measure on $(F, \mathcal{F})$.

Solution: We just need to check that μ satisfies the definition of a measure on $(F, \mathcal{F})$:

- Positivity: For any set $B \in \mathcal{F}$, $\mu(B) = \nu(f^{-1}B) \geq 0$ because ν is a measure (and hence is positive).
- Countable additivity: For any sequence of sets $B_1, B_2, \cdots \in \mathcal{F}$,

$$\begin{aligned} \mu(\cup_n B_n) &= \nu(f^{-1} \cup_n B_n) \\ &= \nu(\cup_n f^{-1} B_n) && \text{(Lemma 2.1)} \\ &= \sum_n \nu(f^{-1} B_n) && \text{(countable additivity of } \nu\text{)} \\ &= \sum_n \mu(B_n) \,. \end{aligned}$$

Example 3.4 Optimal transport: Monge problem. Given two probability measures μ and ν on a measurable space $(E, \mathcal{E})$, the problem of **optimal transport** from μ to ν involves finding a measurable function $T\colon E \to E$ that minimizes the overall cost of transforming μ into ν. Precisely, for a cost function $c\colon E \times E \to [0, +\infty]$, the Monge problem is to solve

$$\inf_{T \text{ msbl}} \left\{ \int c(x, T(x))\mu(dx) \colon \mu \circ T^{-1} = \nu \right\} .$$

That is, we require a measurable function T such that $\mu \circ T^{-1} = \nu$ and that minimizes the total transport cost.

This is an old problem (from 1781) that resisted much progress until Kantorovich recast it into something more tractable in the 1940s. Optimal transport has seen renewed interest in mathematics, statistics, and machine learning in recent years. See, e.g., Santambrogio [San15] and Peyré and Cuturi [PC19].

Example 3.5 Invariant sets, measures, etc. Let $(E, \mathcal{E})$ be a measurable space and $T\colon E \to E$ be a measurable mapping. A subset $A \subset E$ is said to be (T)-**invariant** if $T^{-1}A = A$. The collection of all invariant sets, denoted $\mathcal{I}$, is a σ-algebra (prove this!), known as the (T)-**invariant σ-algebra**. A function $f\colon E \to F$ is T-invariant if $f \circ T = f$. It is fairly straightforward to prove that a function is invariant if and only if it is measurable with respect to the invariant σ-algebra [see, e.g., Kal02, Lemma 10.3].

Let be a measure μ on $(E, \mathcal{E})$. Then T is said to be (μ-)**measure-preserving** and μ is said to be (T-)**invariant** if $\mu \circ T^{-1} = \mu$. A measure-preserving map is **ergodic** for μ if $\mu(A) \in \{0, 1\}$ for each $A \in \mathcal{I}$.

This is the launch-point for ergodic theory (essentially, studying what happens in the limit when we apply T repeatedly to a random starting point in E) and for the study of group-invariant measures (when we require μ to be invariant to a group $\mathcal{T}$ of transformations.) All sorts of cool things happen in both cases and where they overlap. For example, when μ is a $\mathcal{T}$-invariant probability measure (i.e., $\mu \circ T^{-1} = \mu$ for each $T \in \mathcal{T}$) and $\mathcal{T}$ is a group, then μ has a unique decomposition in terms of ergodic, $\mathcal{T}$-invariant measures:

$$\mu(\,\bullet\,) = \int m(\,\bullet\,)\nu(dm) .$$

This can be further related to conditioning on the invariant σ-algebra, though in this class we haven't yet built up the framework for making such claims.

Group-invariance plays a major role in my research and in many areas of statistics and ML. (Shameless self-promotion: I'm teaching a topics course (STAT 547S) in Term 2 about symmetry and invariance in statistics and ML.)

4 Probability spaces

Reading: Çinlar [Çin11], II.1, II.4.

Supplemental: Gut [Gut05], Section 2 of Chapter 2, i.e., 2.2. Skip subsection 2.2.2.

It's worth devoting a bit of time to translate from measure theory to probability theory; to make a mental model of how the abstract mathematics of the earlier sections become real when we apply probability to solve problems in statistics and other fields.

I won't spend much time discussing this in class, so please read carefully and come to class with any questions.

4.1 Probability spaces as models of reality

The inherent uncertainy in the world make probability the natural mathematical language with which to describe it. In particular, we can use probability to construct models of (parts of) reality. A traditional description would say that probability is concerned with **random experiments**: a phenomenon whose outcome is not predictable with certainty. The basic requirement for using probability to analyze uncertain phenomena is that we can describe (mathematically) the set of all possible outcomes of the phenomena, Ω.

Thus far, we've built up the mathematics to properly define a **probability model** as a probability space, $(\Omega, \mathcal{H}, \mathbb{P})$. Many courses and books on probability theory *begin* by defining a probability spaces/models axiomatically:

- Ω is a set called the **sample space**. The elements $\omega \in \Omega$ are called **outcomes**.
- A special collection of subsets, a σ-algebra, of Ω, denoted $\mathcal{H}$.[10] The sets in $\mathcal{H}$ are called **events**.[11]
- A **probability measure**, $\mathbb{P}$, satisfying:
  - $\mathbb{P}(\Omega) = 1$ (this implies that $\mathbb{P}(\emptyset) = 0$).
  - For any countable disjoint collection $(A_n) \in \mathcal{H}$, $\mathbb{P}(\cup_n A_n) = \sum_n \mathbb{P}(A_n)$ (countable additivity).

The axiomatic definition is due to Kolmogorov, who laid the foundations of measure theoretic probability in the 1930s. In particular, he showed that the measure theoretic construction based on events rather than individual outcomes solved the foundational issue of defining probability models with uncountable sample spaces. For good reason, the axioms above are often called **Kolmogorov's axioms** of probability.

A canonical example of a probability model is the following coin-tossing experiment: I flip two identical coins; what is the probability of one heads and one tails? We construct a **probability model** as follows:

- $\Omega_1 = \{(H,H), (H,T), (T,H), (T,T)\}$.

[10] Çinlar [Çin11] says that $\mathcal{H}$ is sometimes called the **grand history**; I haven't come across the term in any other texts, so I assume it's an older/outdated term.

[11] Other common notation for the σ-algebra on Ω is $\mathcal{A}$ or $\mathcal{F}$.

- $\mathcal{H}$ is all possible subsets of Ω_1, 2^{Ω_1}. For example, the event "at least one tails" is $E_{\text{tails}} = \{(H,T),(T,H),(T,T)\}$.
- $\mathbb{P}$ is up to us, a modeling choice. In this case, many would specify $\mathbb{P}$ as uniform over the elements of Ω_1.

Using this probability model, it's straightforward to calculate $\mathbb{P}(E_{\text{tails}}) = \mathbb{P}\{(H,T),(T,H),(T,T)\} = 3/4$.

Note that because Ω_1 is discrete, we don't really need any of the measure theory developed so far. Let's change that. Suppose that I had (ignoring any issues of finite precision, i.e., all numerical values are represented with infinite precision):

- a spinning wheel[12] that, upon stopping, reports the fraction $p \in (0,1)$ of 360° that it rotated (modulo 360°);
- a machine to make biased coins such that the probability of coming up heads is equal to a user-specified $p \in (0,1)$.

My experiment could take on a number of forms. Here's one: spin the wheel; input the rotation fraction p into the coin machine; flip the resulting coin two times. Now $\Omega_2 = (0,1) \times \Omega_1$ and $\mathcal{H} = \mathcal{B}((0,1)) \otimes 2^{\Omega_1}$. How might we construct a probability measure on the following $\mathcal{H}$-measurable events?

- $\{0.1\} \cap E_{\text{tails}}$
- $[0.1, 0.5] \cap E_{\text{tails}}$
- $[0.1, 0.1+\delta] \cap E_{\text{tails}}$, for any $0 < \delta < 0.9$
- E_{tails}

Because $[0,1]$ is uncountable, we need our measure theoretic machinery to define a probability measure on $\mathcal{H}$.

As a practical matter, it is cumbersome to specify a probability on every event in the product σ-algebra, $\mathcal{B}((0,1)) \otimes 2^{\Omega_1}$ (or even on generating subsets). Although you probably know how to construct a valid probability measure via conditional probability, we haven't yet developed those tools in this class. We'll return to these ideas soon.

4.2 Statistical models

Probability models are the basic building blocks of statistical models. A **statistical model** is a family of probability measure on a common measurable space $(\Omega, \mathcal{H})$,

$$\mathcal{P} = \{\mathbb{P}_\theta : \mathcal{H} \to [0,1],\ \theta \in \Theta\}\ . \tag{4.1}$$

Here, Θ is an index set for the model. If Θ is finite-dimensional (e.g., $\Theta = \mathbb{R}^d$) then $\mathcal{P}$ is called **parametric**; otherwise, it is called **non-parametric**. The term **semi-parametric** typically refers to a model whose parameter space is $\Theta_p \times \Theta_n$, where Θ_p has finite dimension and Θ_n does not.

[12] Something like this.

Example 4.1 Normal family. Let $\mathcal{N}_{(\mu,\sigma^2)}$ denote the normal distribution with mean μ and variance σ^2. We can write the parameters as $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$. A statistical model might be

$$\mathcal{P} = \{\mathcal{N}_{(\theta,1)} : \theta \in \mathbb{R}\} \quad \text{or} \quad \mathcal{P} = \{\mathcal{N}_{(\mu,\sigma^2)} : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\} .$$

Clearly, both of these models are parametric.

Example 4.2 Mixture of normals. Let $(p_k) = p_1, p_2, \ldots$ be a sequence satisfying: $p_k \geq 0$, $k \in \mathbb{N}$; $\sum_{k\geq 1} p_k = 1$. Such a sequence of length $K+1$ lies in a **K-dimensional simplex** (or K-simplex), denoted $\mathcal{S}_K$. The sequence might be infinite, in which case it lies in $\mathcal{S}_\infty$.

Define the **mixture of normals** distribution as

$$\mathcal{M}_\theta(A) = (p_k) \circ \mathcal{N}(A) = \sum_k p_k \mathcal{N}_{(\mu_k,\sigma_k^2)}(A) , \quad A \in \mathcal{B}(\mathbb{R}) .$$

The parameters of $\mathcal{M}$ are $\theta = ((p_k), (\mu_k), (\sigma_k^2))$. A statistical model using this might be

$$\mathcal{P} = \{\mathcal{M}_\theta : \theta \in \mathbb{R} \times \mathbb{R}_+ \times \mathcal{S}_K\} .$$

If K is finite, this model is parametric; if not then the model is non-parametric.

In practice, our finite data sets will use only a finite number of parameters, even in a non-parametric model. However, a larger data set can use a larger number of parameters in a non-parametric model; another way to describe non-parametric models is to say that the parameter space is **unbounded**.

Statistical estimation and inference. Frequentist estimation boils down to choosing a particular probability measure from $\mathcal{P}$ that is optimal in some sense. In the parametric setting, choosing a probability measure is equivalent to choosing a parameter. Frequentist inference is based on randomness produced by potential (but unobserved) other samples from the same process that generated the observed sample, and considering the sampling distribution of the estimated parameter (i.e., the selected probability measure).

Bayesian inference amounts to putting a **prior** probability measure on $\mathcal{P}$ and using conditional probability to determine a **posterior** probability measure on $\mathcal{P}$ after observing data. We will return to this in the context of conditioning.

5 Random variables

Reading: Çinlar [Çin11], II.1

Learning Objectives. At the end of this section, you will be able to do the following.
- Define a random variable as a "special" measurable function, and its distribution.
- Define equality almost surely, and in distribution.
- Give some examples of random variables.

Overview. Like the previous chapter, this one serves to tie the measure theory we've developed so far to the probability that we already knew. Most of it is translation, though there are a few new results in this context.

For the rest of these notes, $(\Omega, \mathcal{H}, \mathbb{P})$ will always denote the probability space with which we are working.

Let $(E, \mathcal{E})$ be a measurable space. A $\mathcal{H}/\mathcal{E}$-measurable mapping $X : \Omega \to E$ is a **random variable** taking values in $(E, \mathcal{E})$. Recall that $\mathcal{H}/\mathcal{E}$-measurability means that

$$X^{-1}A := \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{H} \quad A \in \mathcal{E} \,. \tag{5.1}$$

A more common way to denote $X^{-1}A$ is $\{X(\omega) \in A\}$, read as "the event X in A". This type of notation is called **event notation**.

It is customary to denote random variables by capital letters, and fixed values they might assume as lowercase letters, e.g., $\{X = x\}$.

When $\mathcal{E}$ is understood from context, we often say that X is E-valued.

The simplest random variables are indicator functions on sets in $\mathcal{H}$: $\mathbf{1}_H$ for $H \in \mathcal{H}$. A **simple** random variables takes only finitely many values (typically in $\mathbb{R}$); a **discrete** random variable takes on at most countably many different values.

In undergraduate probability courses, typically continuous random variables are also defined here. We will be more careful and properly define what this means soon.

Random variables, random elements. Some authors define $X : \Omega \to E$ to be a **random element** of the arbitrary $(E, \mathcal{E})$. When that measurable space is $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, the term "random variable" is applied. Çinlar [Çin11] doesn't make such a distinction, and neither will I.

5.1 Distribution of a random variable

Let X be a random variable that takes values in $(E, \mathcal{E})$. The **distribution**, or **law**, of X is the image of $\mathbb{P}$ under X,

$$\mu(A) = \mathbb{P} \circ X^{-1}(A) = \mathbb{P}(X^{-1}A) = \mathbb{P}\{X \in A\} \,, \quad A \in \mathcal{E} \,. \tag{5.2}$$

Induced probability spaces. The probability space $(\Omega, \mathcal{H}, \mathbb{P})$ is often referred to as **the background probability space** because it gets relegated to the background: the assumption of its

existence is stated at the beginning, and it is never heard from again. As such, the random variable $X(\omega)$ is written not as a function, but simply as X.

In practice, we typically work directly on the space on which X takes values. For example, we don't work with $\mathbb{P}$, defined on Ω, and then determine the image measure $\mathbb{P} \circ X^{-1}$. Rather, we define the distribution μ_X and proceed from there, forgetting $(\Omega, \mathcal{H}, \mathbb{P})$. In this case, we call $(E, \mathcal{E}, \mu_X)$ the **induced probability space**. The following fact, stated as a proposition for posterity, follows from $(E, \mathcal{E})$ being a measurable space and μ_X being a probability measure.

Proposition 5.1. *The induced space $(E, \mathcal{E}, \mu_X)$ is a probability space.*

For complicated models, we might construct hierarchies or sequences of random variables. This works under the assumption that all random quantities are measurable relative to $\mathcal{H}$ and the σ-algebra on whatever space in which they take values. A common interpretation in such models is that Ω is "the state of the world", and the random quantities included in the model are functions of different functions of that state. I like to think of the background probability space $(\Omega, \mathcal{H}, \mathbb{P})$ as a **reservoir of randomness** that we draw on to do all of our stochastic/probabilistic operations.

Stochastic equality. Let X and Y be two random variables taking values in a measurable space $(E, \mathcal{E})$. X and Y are said to be equal **almost surely** or **almost everywhere** if

$$\mathbb{P}\{\omega \in \Omega : X(\omega) = Y(\omega)\} = \mathbb{P}\{X = Y\} = 1 \, . \tag{5.3}$$

We denote this by $X \overset{\text{a.s.}}{=} Y$. Observe that this means that X and Y may differ on null sets.

X and Y are said to be **equal in distribution** if

$$\mathbb{P}\{X \in A\} = \mathbb{P}\{Y \in A\} \, , \quad A \in \mathcal{E} \, , \tag{5.4}$$

denoted $X \overset{\text{d}}{=} Y$. Note that in general, this is a much weaker form of equality: almost-sure equality implies distributional equality, but not the converse.

Distributions on $\mathbb{R}$. In light of the result Corollary 3.3 on the uniqueness of probability measures on $\bar{\mathbb{R}}$, in order to specify the distribution of a random variable X taking values in $[-\infty, \infty]$, it is enough to specify $\mu([-\infty, r])$ for all $r \in \mathbb{R}$, i.e., the **distribution function**

$$F(x) = \mu[-\infty, x] = \mathbb{P}\{X \leq x\} \, , \quad x \in \mathbb{R} \, . \tag{5.5}$$

This is also called the **cumulative distribution function**, or **cdf**. We won't dwell on them further in lecture because you've likely studied them in previous courses on statistics/probability.

Exercise 20:

Show the following: two $\mathbb{R}$-valued random variables X and Y are equal in distribution if and only if their distribution functions F_X and F_Y are equal.

Densities on $\mathbb{R}$. When the distribution function of X has the form

$$F(x) = \int_{-\infty}^{x} f(x) \lambda(dx) \, , \tag{5.6}$$

we say that X has **density function** f, typically denoted as f_X. A distribution function F that has a density is said to be **absolutely continuous** with respect to the Lebesgue measure (or simply absolutely continuous); $f = \frac{dF}{d\lambda}$ is the **Radon–Nikodym derivative** of F with respect to the Lebesgue measure. We will study these ideas in more detail in the context of integration and expectation. Note that the Lebesgue measure is so ubiquitous that $\lambda(dx)$ is typically written simply as dx.

5.2 Examples of random variables

Example 5.1 Poisson distribution. Let X be a random variable taking values in $\mathbb{N} = \{0, 1, 2, \dots\}$, with the discrete σ-algebra $2^{\mathbb{N}}$. X has **Poisson distribution** with mean $c > 0$ is

$$\mathbb{P}\{X = n\} = \frac{e^{-c}c^n}{n!} \,. \tag{5.7}$$

The corresponding distribution μ on $\mathbb{N}$ is

$$\mu(A) = \sum_{n \in A} \frac{e^{-c}c^n}{n!} \,, \quad A \subset \mathbb{N} \,. \tag{5.8}$$

Example 5.2 Gamma distribution. Let X be a random variable taking values in $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}))$. It has the **gamma distribution** with shape index a and scale parameter b if its distribution has the form

$$\mu(dx) = \lambda(dx)\frac{b^a x^{a-1} e^{-bx}}{\Gamma(a)} \,, \quad x \in \mathbb{R}_+ \,. \tag{5.9}$$

The normalizing constant $\Gamma(a)$ is the **gamma function**

$$\Gamma(a) = \int_0^\infty dx \; x^{a-1} e^{-x} \,. \tag{5.10}$$

When $a = 1$, the gamma distribution is the **exponential distribution**; when $b = 1/2$ and $a = n/2$ for some integer n, it is the χ^2**-distribution** with n degrees of freedom.

Example 5.3 Random graph. Many interesting random objects do not have a distribution function that we can write in closed form or compute easily (or at all). Rather, their distribution is induced by a constructive or **generative model**. A **graph** G is a set $\mathbf{V}$ of **vertices** and the set $\mathbf{E} \subset \mathbf{V} \times \mathbf{V}$ of **edges** between them. For example, vertices might represent users of a social network and edges represent friendship connections. A common representation of a simple graph enumerates the vertices ($\mathbf{V} = \{1, 2, \ldots, n\} := [n]$) and records a value in $\{0, 1\}$ for each possible edge in $[n] \times [b]$.[a] A random simple graph on n vertices is therefore a random variable G_n taking values in $[n] \times \{0, 1\}^{[n]\times[n]}$. This is a finite (discrete) space, so the discrete σ-algebra can be used without problems.

In all but the simplest models, specifying a probability distribution on a (possibly large) composite structure like G_n is highly non-trivial. In the case of random graphs, there are good reasons not to specify a model this way (we won't go into them). However, consider the following simple generative model:

1. Start with G_2 as two vertices connected by an edge.
2. At each $n = 3, \ldots,$ connect a new vertex via an edge to a random vertex in $\mathbf{V}(G_{n-1})$.
3. The probability that the new vertex n attaches to any existing vertex v is proportional to the number of edges that attach to v in G_{n-1}.

This is the simplest version of the widely studied **preferential attachment model**; it generates **random trees**. Its simplicity makes it amenable to probabilistic study. Observe that for any set $A \subset [n] \times \{0, 1\}^{[n]\times[n]}$, we could in theory compute $\mathbb{P}\{G_n \in A\}$. However, as n gets even moderately large, the size of $[n] \times \{0, 1\}^{[n]\times[n]}$ blows up and computing the distribution of G_n becomes unwieldy/impossible. Specifying the model as above, however, allows us to write the distribution in terms of conditional probabilities. We will revisit this example in that context.

Note that this is not (by far) the only way we might specify a distribution on random graphs. There is a vast literature on the topic, ranging from the purely probabilistic (for example, preferential attachment has been generalized in many ways for probabilistic study) to more statistics-oriented models like the so-called exponential random graph model, which specifies sufficient statistics (e.g., number of triangles, degree sequence, and so on) and uses them in an exponential family distribution. (There has been an increasing exchange of ideas from these two ends of the spectrum in recent years.)

[a] This is a **simple graph**, the simplest version of such an object; multigraphs, weighted graphs, hypergraphs, multiplex graphs, and others extend the basic idea.

Example 5.4 Function of a random variable. Let X be a random variable taking values in a measurable space $(E, \mathcal{E})$, and $(F, \mathcal{F})$ another measurable space such that the mapping $f : E \to F$ is $\mathcal{E}/\mathcal{F}$-measurable. Then the composition $Y = f \circ X$,

$$Y(\omega) = f \circ X(\omega) = f(X(\omega)) \ , \quad \omega \in \Omega \ ,$$

is a random variable taking values in $(F, \mathcal{F})$. This follows from the measurability of the composition of measurable functions (Proposition 2.4). If μ is the distribution of X, then the distribution ν of Y is the image measure

$$\nu(A) = \mathbb{P}\{Y \in A\} = \mathbb{P}\{X \in f^{-1}A\} = \mu(f^{-1}A) \ , \quad A \in \mathcal{F} \ .$$

For example, let $D_{1,n}$ be the degree of the first vertex in a preferential attachment tree with n edges, G_n. $D_{1,n}$ takes values in $\mathbb{N}$, and its distribution is much easier to analyze than the distribution of the entire random tree—much of the probabilistic analysis of preferential attachment models focuses on the properties of the entire sequence of degrees of the vertices, $(D_{1,n}, \ldots, D_{n+1,n})$.

Example 5.5 Random variable defined by a computer program. Many of the (pseudo-)random variables we know and love can be generated by transforming uniform (pseudo-)random variables. For example, let $U \sim \text{Unif}[0, 1]$ (read "U sampled from the uniform distribution on $[0, 1]$"), and let $X \sim \text{Exp}(b)$ (read "X sampled from the exponential distribution with rate parameter b"). Then we know that $X \stackrel{\text{d}}{=} -\ln(U)/b$. This is an example of a **distributional identity**.

On a computer, generating a (pseudo-)random variable with exponential distribution might call a program with the following basic structure:

```
function EXPONENTIAL(b)
    u ← −ln(rand())/b
    return u
end function
```

The function `rand()` is a sort of computerized reservoir of randomness the represents the background probability space $(\Omega, \mathcal{H}, \mathbb{P})$: anything stochastic or probabilistic that we might do on a computer relies on calls to the pseudo-random number generator. Moreover, `ln` calls the natural logarithm function, which is a built-in function in most mathematics libraries, and is evaluated via numerical methods like the Newton–Raphson algorithm.

This is a simple example of a computer program that can be viewed as a random variable. The field of **probabilistic programming** takes a broader view: *any* computer program can be viewed as a random variable whose distribution is induced by the program. Statistical inference (typically Bayesian) can be performed over the distribution of execution traces by conditioning on observed data. These programs (and research into them) is at the forefront of machine learning. Some examples are BUGS and Stan (from the statistics world), and TensorFlow Probability (formerly known as Edward), Pyro, Church, and Anglican (from the machine learning world).

Exercise 21 (Random variable from undergraduate probability)**:**

Give the full definition (i.e., set of values, σ-algebra) of a random variable you studied in under-

graduate probability. State at least one distributional identity involving that random variable.

Exercise 22 (Random variable from your research interests)**:**

Give the full definition of a random variable that is not encountered in introductory textbooks on probability, ideally one you encounter in research. If you can, state at least one distributional identity involving that random variable.

5.3 Joint distributions and independence

Let X and Y be random variables taking values in measurable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$, respectively. The pair $Z = (X, Y) : \omega \mapsto Z(\omega) = (X(\omega), Y(\omega))$ is measurable relative to $\mathcal{H}$ and the product σ-algebra $\mathcal{E} \otimes \mathcal{F}$. That is, Z is a random variable taking values in the product space $(E \times F, \mathcal{E} \otimes \mathcal{F})$. The distribution of Z is the probability measure π on the product space, called the **joint distribution** of X and Y. In order to specify π it is sufficient to specify

$$\pi(A \times B) = \mathbb{P}\{X \in A, Y \in B\} = \mathbb{P}(\{X \in A\} \cap \{Y \in B\}) \ , \quad A \in \mathcal{E}, \ B \in \mathcal{F} \ . \tag{5.11}$$

The **marginal distributions** of X and Y are

$$\mu(A) = \mathbb{P}\{X \in A\} = \pi(A \times F) \ , \ A \in \mathcal{E} \quad \text{and} \quad \nu(B) = \mathbb{P}\{Y \in B\} = \pi(E \times B) \ , \ B \in \mathcal{F} \ . \tag{5.12}$$

These terms are extended in the obvious way to arbitrary finite collections of random variables.

Independence. X and Y are said to be **independent** if their joint distribution is just the product of their marginal distributions:

$$\pi(A \times B) = \mathbb{P}\{X \in A, Y \in B\} = \mathbb{P}\{X \in A\}\mathbb{P}\{Y \in B\} = \mu(A)\nu(B) \ , \quad A \in \mathcal{E}, \ B \in \mathcal{F} \ . \tag{5.13}$$

Independence is one of the fundamental (and most useful) ideas used to construct statistical models. As a general rule, the more independence a model has, the easier it is to perform inference, computation, closed-form analysis. The trade-off for too much independence is that the model might not be able to capture properties in real data. Traditionally, good models strike a balance—but doing so is a bit of an art.

6 Approximation of measurable functions

Reading: Çinlar, I.2.

Supplemental: Bass [Bas22], Ch. 5.2; Schilling [Sch05], Ch. 8.

Learning Objectives. At the end of this section, you will be able to do the following.

- Show that for a sequence of measurable $\mathbb{R}$-valued functions $(f_n)_{n\geq 1}$, each of $\inf_n f_n$, $\sup_n f_n$, $\liminf_n f_n$, $\limsup_n f_n$, and, when it exists, $\lim_n f_n$, are measurable.
- Define simple functions and dyadic functions.
- Show that a $\mathbb{R}$-valued function is measurable if and only if it is the limit of an increasing sequence of positive simple functions.
- **Optional:** Use the monotone class theorem to prove a version that applies to so-called monotone classes of functions.

Overview. Now that we have a better idea of how the big picture fits together, we're ready to study Lebesgue integration. Just like the Riemann integral you probably studied in your calculus course, the Lebesgue integral is the limit of a sequence of better and better approximations. The first (and most important) step towards this is to build a framework for approximating measurable functions, which are the objects we will be integrating.

6.1 Measurability of limits of sequences of functions

Many of the proof techniques we will use in upcoming sections rely on analyzing sequences of numerical functions $(f_n)_{n\geq 1}$, or (f_n) for short, and their limits:

$$\inf_n f_n \ , \quad \sup_n f_n \ , \quad \liminf_n f_n \ , \quad \limsup_n f_n \ . \tag{6.1}$$

Each of these are functions defined on E *pointwise*: fix $x \in E$ and construct the sequence of numbers $(f_1(x), f_2(x), \dots)$. Sequences of real numbers and taking the inf, sup, lim inf, or lim sup should be familiar from calculus and/or analysis [see Abb15, Ch. 6 to refresh your memory]. Repeating this operation for each $x \in E$ yields the definition above. For example, $f = \inf_n f_n$ defines a function whose value at x is $f(x) = \inf_n(f_n(x))$, i.e., the infimum of $(f_1(x), f_2(x), \dots)$.

In general, $\liminf$ is dominated by $\limsup$ (i.e., $\limsup f_n(x) \geq \liminf f_n(x)$ for each $x \in E$). If they are equal, $\liminf_n f_n = \limsup_n f_n = f$, then the sequence (f_n) has the **pointwise limit** f, denoted $\lim_n f_n = f$ or, more commonly, $f_n \to f$.

Monotone sequences of functions. If (f_n) is increasing, that is, $f_1 \leq f_2 \leq \dots$ (for all $x \in E$), then $\lim_n f_n = \sup_n f_n$. The shorthand for (f_n) is increasing and has limit f is $f_n \nearrow f$. Likewise, $f_n \searrow f$ means that (f_n) is decreasing and has limit f.

Importantly, the class of measurable (numerical) functions is closed under limits.

Theorem 6.1. *Let (f_n) be a sequence of $\mathcal{E}$-measurable numerical functions. Then each of the four functions in* (6.1) *is $\mathcal{E}$-measurable. Moreover, if it exists, $\lim_n f_n$ is $\mathcal{E}$-measurable.*

Proof. We start with $f = \sup_n f_n$. For every $x \in E$ and $r \in \mathbb{R}$, observe that $f(x) \leq r$ if and only if $f_n(x) \leq r$ for all n. Thus, for each $r \in \mathbb{R}$,

$$f^{-1}[-\infty, r] = \{x : f(x) \leq r\} = \bigcap_n \{x : f_n(x) \leq r\} = \bigcap_n f_n^{-1}[-\infty, r] \,. \tag{6.2}$$

Each term in the intersection in the right-most expression is in $\mathcal{E}$ because f_n is $\mathcal{E}$-measurable. Furthermore, $\mathcal{E}$ is closed under countable intersections, so the entire right-most expression is in $\mathcal{E}$. $[-\infty, r]$ generate $\mathcal{B}(\mathbb{R})$ so by Proposition 2.3, $f = \sup_n f_n$ is $\mathcal{E}$-measurable.

Measurability of $\inf_n f_n$ follows from the identity $\inf_n f_n = -\sup_n(-f_n)$. The composition of two measurable functions is again measurable, so

$$\liminf_n f_n = \sup_m \inf_{n \geq m} f_n \,, \quad \text{and} \quad \limsup_n f_n = \inf_m \sup_{n \geq m} f_n$$

are both $\mathcal{E}$-measurable. □

As we will see in the next two sections, closure under limits gives us a powerful tool for proving certain fundamental properties of measurable functions.

6.2 Simple functions and approximating measurable functions

Let $A \subset E$. The **indicator function** is

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \,. \tag{6.3}$$

$\mathbf{1}_A$ is $\mathcal{E}$-measurable if and only if $A \in \mathcal{E}$.

Simple functions. A function f on E is said to be **simple** if it has the form

$$f = \sum_{i=1}^n a_i \mathbf{1}_{A_i} \,, \tag{6.4}$$

for some $n \in \mathbb{N}$, some real numbers $(a_i)_{i=1}^n$ and sets $(A_i)_{i=1}^n \in \mathcal{E}$. For any simple function f, there exits some $m \in \mathbb{N}$, *distinct* real numbers $(b_i)_{i=1}^m$, and a measurable partition $\{B_1, \ldots, B_m\}$ of E such that $f = \sum_{i=1}^m b_i \mathbf{1}_{B_i}$. (Convince yourself that this is true.) This is called the **canonical form** of the simple function f.

Measurability. Proposition 2.3 applied to the canonical form implies that *every simple function is $\mathcal{E}$-measurable.* Conversely, if a function f is $\mathcal{E}$-measurable and takes on finitely many values in $\mathbb{R}$, then f is simple.

Exercise 23 (Compositions under which the class of simple functions is closed)**:**

For f, g both simple functions, show that the following are also simple:

1. $f + g$

2. $f - g$
3. fg
4. f/g (with the caveat that g is nowhere equal to zero)
5. $f \vee g$
6. $f \wedge g$

Dyadic functions. Recall that **dyadic intervals** are bounded intervals of $\mathbb{R}$ whose endpoints are $\frac{j}{2^n}$ and $\frac{j+1}{2^n}$, where $j, n \in \mathbb{Z}$. Observe that any fixed finite interval (a, b) contains increasingly more dyadic intervals as n increases. In particular, $[0, n]$, $n \in \mathbb{N}$, contains $n2^n$ dyadic intervals of equal size. Define the **dyadic function** as the simple function $d_n : \mathbb{R} \to \mathbb{R}$ as

$$d_n(r) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \mathbf{1}_{[\frac{k-1}{2^n}, \frac{k}{2^n})}(r) + n\mathbf{1}_{[n,\infty]}(r) \, . \tag{6.5}$$

We can think of the functions (d_n) as a sequence of finer and finer (better and better) discrete approximations of $f(x) = x$.

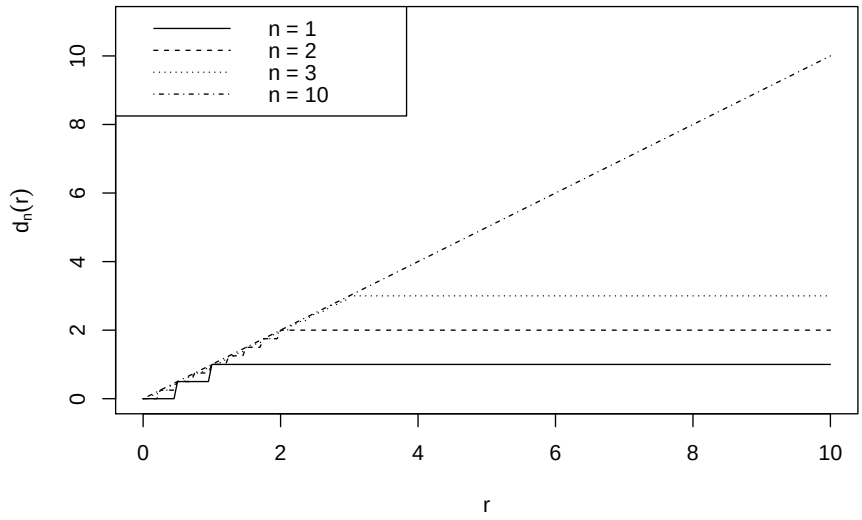

Figure 1: Dyadic functions for $n \in \{1, 2, 3, 10\}$. (Vertical jumps shown for illustration.)

Exercise 24 (Plotting the dyadic functions)**:**

Use your scientific computing language/environment of choice to plot the dyadic functions for $n \in \{1, 2, 3, 10\}$.

Measurability. Recall from Exercise 8 that the semi-closed and closed sets are Borel sets; because d_n is a simple function, it is measurable for each n.

We can prove the following with pictures.

Lemma 6.2. *For each $n \in \mathbb{N}$, each d_n is an increasing right-continuous simple function on $\bar{\mathbb{R}}_+$, and $d_n(r)$ increases to r (that is, $d_n(r) \nearrow r$) for each $r \in \bar{\mathbb{R}}_+$ as $n \to \infty$.*

Approximating measurable functions. The dyadic function can be used to approximate any function f to arbitrary precision: define $f_n = d_n \circ f$, and observe that since $d_n(r)$ increases to r as $n \to \infty$, $d_n \circ f \nearrow f$. (Figure 6.2 illustrates this.) If f is $\mathcal{E}$-measurable, then so is $d_n \circ f$ for each $n \in \mathbb{N}$, by Proposition 2.4. This is essentially all that is needed to prove the following important result, which provides a basic tool for proving certain properties of measurable functions by reducing the problem to proving the property on simple functions and taking the limit.

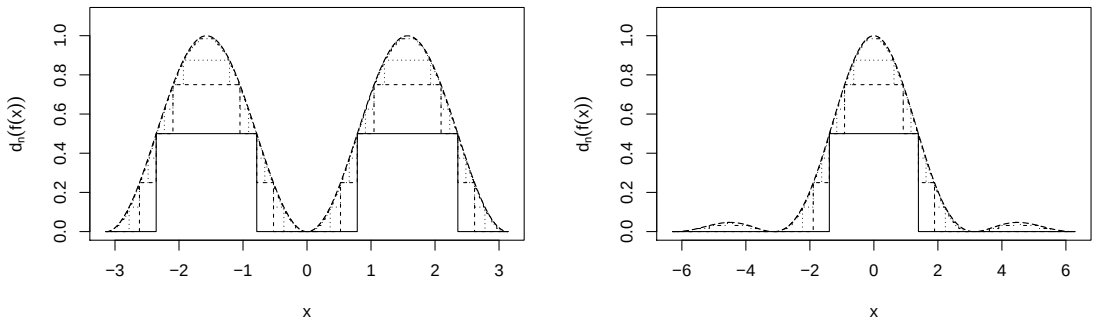

Figure 2: Approximating $f(x) = \sin^2(x)$ (left) and $f(x) = \frac{\sin^2(x)}{x^2}$ (right) for $n \in \{1, 2, 3, 6, 10\}$. (Vertical jumps shown for illustration.)

Theorem 6.3. *A positive function on E is $\mathcal{E}$-measurable if and only if it is the limit of an increasing sequence of positive simple functions.*

Proof technique. The proof of this theorem uses a technique that can be/is used to prove measurability of a positive function f (or a class of functions) in many situations:

1. Prove that all simple functions are measurable with respect to the relevant σ-algebras.
2. Approximate f by simple functions $f_n \nearrow f$ and take the limit $n \to \infty$.
3. Argue that the limit is measurable.

This technique can be extended to arbitrary functions by separating into positive and negative parts: $f = f^+ - f^-$.

Proof. If a positive function f is the limit of an increasing sequence of positive simple functions, then by Theorem 6.1, it is also measurable.

Conversely, let $f : E \to \mathbb{R}_+$ be $\mathcal{E}$-measurable. We need to show that there is an sequence (f_n) of positive simple functions increasing to f. Let d_n be as in (6.5) and put $f_n = d_n \circ f$. For each n, f_n is $\mathcal{E}$-measurable, since it is the composition of two measurable functions (Proposition 2.4). Moreover, f_n is positive and takes only finitely many values: it is positive and simple. Since $d_n(r) \nearrow r$ for each $r \in \mathbb{R}_+$, we have that $f_n(x) = d_n(f(x)) \nearrow f(x)$ for each $x \in E$. □

6.3 Monotone classes of functions*

Monotone classes of functions are used to extend the monotone class theorem (Theorem 1.5) to measurable functions, which often is more useful in practice. It's worth revisiting Section 1.7 to see how the concepts translate.

A generic collection of numerical functions on E is denoted by $\mathcal{M}$; $\mathcal{M}_+$ is the subcollection of positive functions in $\mathcal{M}$, and $\mathcal{M}_b$ is the subcollection of bounded ones. The collection $\mathcal{M}$ is a **monotone class** of functions if it satisfies:

M1. $\mathbf{1}_E \in \mathcal{M}$ (includes the constant function);

M2. $f, g \in \mathcal{M}_b$ and $a, b \in \mathbb{R} \Rightarrow af + bg \in \mathcal{M}_b$ (i.e., $\mathcal{M}_b$ is a linear space over $\mathbb{R}$); and

M3. $(f_n) \in \mathcal{M}_+$ and $f_n \nearrow f \Rightarrow f \in \mathcal{M}_+$ (closure under increasing limits).

Note how similar these properties are to the defining properties of a d-system. It's not a coincidence—monotone classes are set up to transfer the properties of d-systems to collections of numerical functions. The following result is basically a riff on the monotone class theorem, but is invoked much more frequently.

Theorem 6.4 (Monotone classes of functions)**.** *Let $\mathcal{M}$ be a monotone class of functions on E. Suppose, for some p-system $\mathcal{C}$ that generates $\mathcal{E}$, that $\mathbf{1}_A \in \mathcal{M}$ for every $A \in \mathcal{C}$. Then $\mathcal{M}$ contains all positive $\mathcal{E}$-measurable functions and all bounded $\mathcal{E}$-measurable functions.*

Proof. The overall structure is as follows:

- Show that $\mathcal{M}$ contains all simple functions. The monotone class theorem (Theorem 1.5) makes this easy.
- Apply Theorem 6.3 to any positive $\mathcal{E}$-measurable function f, which along with property M3 above, implies that $f \in \mathcal{M}_+$.
- Separate any bounded measurable function f into its positive and negative parts: $f = f^+ - f^-$. Then, apply the previous conclusion to each part separately, along with property M2 above, to conclude that $f \in \mathcal{M}_b$.

This proof pattern (simple functions, positive measurable functions, bounded measurable functions) gets used a lot. (Seriously, a lot. So much so that research articles, advanced textbooks, even Çinlar

in later chapters, often say something to the effect of, "Measurability is established by a monotone class argument.")

Simple functions. First, we'll show that $\mathbf{1}_A \in \mathcal{M}$ for all $A \in \mathcal{E}$. To that end, define $\mathcal{D} = \{A \in \mathcal{E} \colon 1_A \in \mathcal{M}\}$. This is a d-system. Let's check (the properties refer to those for d-systems in Section 1.7):

D1. $\mathbf{1}_E \in \mathcal{M}$ by the definition of monotone class, so $E \in \mathcal{D}$.

D2. For $A, B \in \mathcal{D}$, $\mathbf{1}_A - \mathbf{1}_B = \mathbf{1}_{A \setminus B} \in \mathcal{M}$ by M2, so $A \setminus B \in \mathcal{D}$.

D3. For any monotone $(A_n) \in \mathcal{D}$ such that $A_n \nearrow A$, M3 implies that $\mathbf{1}_{A_n} \nearrow \mathbf{1}_A$, so $A \in \mathcal{D}$.

Now, $\mathbf{1}_A \in \mathcal{M}$ for all $A \in \mathcal{C}$ by assumption, so $\mathcal{C} \subset \mathcal{D}$. $\mathcal{C}$ is a p-system that generates $\mathcal{E}$ by assumption and we just established that $\mathcal{D}$ is a d-system, so the monotone class theorem implies that $\mathcal{E} \subset \mathcal{D}$. Therefore, $\mathbf{1}_A \in \mathcal{M}$ for every $A \in \mathcal{E}$. Property M2 indicates that $\mathcal{M}$ therefore includes all simple functions.

Positive measurable functions. Let f be a positive $\mathcal{E}$-measurable function. By Theorem 6.3, there is a sequence of positive simple functions $f_n \nearrow f$. Since we established above that each $f_n \in \mathcal{M}_+$, property M3 implies that $f \in \mathcal{M}$ (actually, $\mathcal{M}_+$).

Bounded measurable functions. Let $f = f^+ - f^-$ be a bounded $\mathcal{E}$-measurable function. Then f^+ and f^- are each positive (bounded) $\mathcal{E}$-measurable functions and therefore each are in $\mathcal{M}$ by the previous step in the proof. By property M2, $f = f^+ - f^- \in \mathcal{M}$. □

7 Lebesgue integration

Reading: Çinlar [Çin11], I.4

Supplemental: Bass [Bas22], Ch. 6-8; Schilling [Sch05], Ch. 9.

Learning Objectives. At the end of this section, you will be able to do the following.

- Define the Lebesgue integral for simple functions, positive measurable functions, and arbitrary measurable functions.
- Integrate a function over a set.
- Prove the positivity and linearity of the Lebesgue integral for simple functions.
- State and use the monotone convergence theorem.
- Use the monotone convergence theorem to prove the linearity of the Lebesgue integral.
- Use convergence arguments in order to interchange the order of certain limiting operations.

First, a brief reminder about positive and negative parts of a function.

Positive and negative parts. For a and b in $\mathbb{R}$, we write $a \vee b$ to denote $\max\{a, b\}$ and $a \wedge b$ to denote $\min\{a, b\}$. When applied to numerical functions, the maximum is taken pointwise: $f \vee g$ is a function whose value at x is $f(x) \vee g(x)$. For a measurable space $(E, \mathcal{E})$ and a function $f : E \to \bar{\mathbb{R}}$, the **positive part** and **negative part** of f are

$$f^+ = f \vee 0 \,, \quad \text{and} \quad f^- = -(f \wedge 0) \,.$$

It should be (intuitively) clear that f is $\mathcal{E}$-measurable if and only if both f^+ and f^- are. This fact is important enough that it is stated as a proposition in Çinlar [Çin11, Prop. I.2.9] because we can obtain many results for arbitrary f from the corresponding results for positive functions.

7.1 Definition and desiderata

Let $(E, \mathcal{E}, \mu)$ be a measure space. Recall that $\mathcal{E}^{\text{fn}}$ denotes the collection of all $\mathcal{E}/\mathcal{B}(\mathbb{R})$-measurable functions, and $\mathcal{E}^{\text{fn}}_+$ is the sub-collection of positive measurable functions. Our aim in this section is to *define* the integral of a function with respect to a measure. Of course, we would like whatever we define to satisfy certain properties; the bulk of the section is devoted to proving that what we define actually has the properties we want.

To that end, denote the integral of a function f with respect to a measure μ by (these are just different ways of writing the same thing)

$$\mu f = \mu(f) = \int_E \mu(dx) f(x) = \int_E f \, d\mu \,. \tag{7.1}$$

The notation μf suggests a kind of mulitplication, and indeed integration behaves much like multiplication via the properties (which we will prove) for all $a, b \in \mathbb{R}_+$ and $f, g, f_n \in \mathcal{E}^{\text{fn}}$:

i) **Positivity**: $\mu f \geq 0$ if $f \geq 0$.

ii) **Linearity**: $\mu(af + bg) = a\mu f + b\mu g$.

iii) **Monotone convergence**: If $f_n \nearrow f$ then $\mu f_n \nearrow \mu f$.

The **integral** is defined in parts (recursively), in order to handle different types of functions:

a) **Simple and positive functions**: Let f be simple and positive, with canonical form $f = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i}$. Then the integral is defined as

$$\mu f = \sum_{i=1}^{n} a_i \mu(A_i) \ . \tag{7.2}$$

b) **Positive functions**: Let $f \in \mathcal{E}_+^{\text{fn}}$, and put $f_n = d_n \circ f$ with d_n the dyadic function defined in (6.5) (with properties given in Lemma 6.2). Then each f_n is simple and positive, and $f_n \nearrow f$. The integral μf_n is defined in part a) above; the sequence of numbers μf_n is increasing so the limit exists (we will establish this below). We define

$$\mu f = \lim_n \mu f_n \ . \tag{7.3}$$

c) **Arbitrary measurable functions**: Let $f \in \mathcal{E}^{\text{fn}}$. Then $f^+, f^- \in \mathcal{E}_+^{\text{fn}}$, and their integrals are defined by part b) above. Noting that $f = f^+ - f^-$, we define

$$\mu f = \mu f^+ - \mu f^- \ , \tag{7.4}$$

provided at least one of the terms on the RHS is finite. Otherwise, if both terms are infinite, we define μf is undefined.

We will show that this definition is the "right one", in the sense that it satisfies properties i)-iii) above, and that those properties *characterize* the integral—there is no other way to define it if we want those properties.

Exercise 25 (Some observations and immediate implications for positive simple functions)**:**

Let f, g be simple and positive. Prove that the following statements are true:

- The formula (7.2) for μf still holds even if f is not in canonical form.
- The integral is linear.
- The integral is **monotonic**: if $f \leq g$ then $\mu f \leq \mu g$.
- A sequence of increasing functions $f_1 \leq f_2 \leq \ldots$ has $\mu f_1 \leq \mu f_2 \leq \ldots$, so $\lim_n \mu f_n$ exists (it may be $+\infty$).

Solution:

What's the intuition? Recall from your calculus class that the Riemann integral of $f : \mathbb{R} \to \mathbb{R}$ (the one we know and love) is calculated by partitioning E and drawing a rectangle on each element of the partition with height equal to the value of f at the leftmost value of the element. Another set of rectangles is generated to have height equal to the value of f at the rightmost value of each element of the partition. As more and more elements are squeezed into the partition, the common limit of the two sets of rectangles is the Riemann integral. We'll see an example below where this might not work with some functions for which we'd like to be able to define an integral.

The integral defined above, generally known as the **Lebesgue integral** (not to be confused with Lebesgue measure), takes the "inverse" approach: composing d_n with f essentially partitions the "y-axis" (i.e., the range of f):

$$(d_n \circ f)(x) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \mathbf{1}_{[\frac{k-1}{2^n}, \frac{k}{2^n})}(f(x)) + n\mathbf{1}_{[n,\infty]}(f(x)) \; . \tag{7.5}$$

Applied to the sum, the Lebesgue integral then uses the definition of integral for simple functions to make a rectangle with height $2^{-n}(k-1)$ and base

$$\mu\left(f^{-1}\left[\frac{k-1}{2^n}, \frac{k}{2^n}\right)\right) = \mu\left(\{x \in E : \frac{k-1}{2^n} \leq f(x) < \frac{k}{2^n}\}\right) \; .$$

The crucial technical step is proving that the limit of this construction behaves as we would like.

7.2 Examples

Example 7.1 Integrating with discrete measures. Fix $x_0 \in E$ and consider the Dirac measure δ_{x_0}. The integral as defined above yiels $\delta_{x_0} f = f(x_0)$ for every $f \in \mathcal{E}^{\text{fn}}$. This extends to discrete measures $\mu = \sum_{x \in D} m(x)\delta_x$ for some countable set D and masses $m(x)$,

$$\mu f = \sum_{x \in D} m(x) f(x) \; ,$$

for every $f \in \mathcal{F}_+^{\text{fn}}$. Similar results hold for purely atomic measures, and for measures and functions defined on discrete spaces.

Example 7.2 Integrating with the Lebesgue measure. Suppose that E is a Borel subset of $\mathbb{R}^d$, $d \geq 1$, and that $\mathcal{E} = \mathcal{B}(E)$. Suppose that μ is the restriction of the Lebesgue measure on $\mathbb{R}^d$ to $(E, \mathcal{E})$. For $f \in \mathcal{E}^{\text{fn}}$, the **Lebesgue integral** of f on E is denoted

$$\mu f = \int_E \lambda(dx) f(x) = \int_E dx \; f(x) \; .$$

If the Riemann integral of f exists, then so does the Lebesgue integral (and they are equal). However, the Lebesgue integral exists for a larger class of functions than the Riemann integral. Consider the function on $[0, 1]$,

$$f(x) = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & x \notin \mathbb{Q} \end{cases} \; .$$

The set of discontinuities of this function is the entire interval $[0, 1]$; recall that Lebesgue's criterion for the existence of the Riemann integral for a function is that the set of discontinuities has (Lebesgue) measure zero [see Abb15, Ch. 7], and therefore the function is not Riemann-integrable. However, by our definition above, and in particular a) and the monotone convergence property (which we will prove below), the Lebesgue integral is well-defined and is equal to 0. As Bass says, "The main reason the Lebesgue integral is so much easier to work with than the Riemann integral is that it behaves nicely when taking limits" [Bas22, p. 57]. Chapter 9 of Bass [Bas22] compares the Lebesgue and Riemann integrals in detail.

7.3 Basic properties of integration

Integrability. A function $f \in \mathcal{E}^{\text{fn}}$ is said to be **integrable** if μf exists and is a real number. Thus, f is integrable if and only if $\mu f^+ < \infty$ and $\mu f^- < \infty$, or equivalently, $\mu|f| = \mu f^+ + \mu f^- < \infty$.

Integrating over a set. Let $f \in \mathcal{E}^{\text{fn}}$ and $A \in \mathcal{E}$. Then $f\mathbf{1}_A \in \mathcal{E}^{\text{fn}}$ and the **integral of f over A** is defined to be the integral of $f\mathbf{1}_A$, denoted

$$\mu(f\mathbf{1}_A) = \int_A \mu(dx) f(x) = \int_A f \, d\mu \, .$$

We would like the integrals over two disjoint sets to sum to their integral over the union of those sets; the following lemma establishes that this is the case.

Lemma 7.1. *Let $f \in \mathcal{E}_+^{\text{fn}}$. Let A and B be disjoint sets in $\mathcal{E}$ with $C = A \cup B$. Then*

$$\mu(f\mathbf{1}_A) + \mu(f\mathbf{1}_B) = \mu(f\mathbf{1}_C) \, .$$

Proof. If f is simple, the lemma follows from the linearity property established for simple functions:

$$\mu(f\mathbf{1}_A) + \mu(f\mathbf{1}_B) = \mu(f\mathbf{1}_A + f\mathbf{1}_B) = \mu(f(\mathbf{1}_A + \mathbf{1}_B)) = \mu(f\mathbf{1}_{A\cup B}) = \mu(f\mathbf{1}_C) \, .$$

For arbitrary $f \in \mathcal{E}_+^{\text{fn}}$, putting $f_n = d_n \circ f$, we get $\mu(f_n\mathbf{1}_A) + \mu(f_n\mathbf{1}_B) = \mu(f_n\mathbf{1}_C)$ for all n since the f_n are simple. Observing that $f_n\mathbf{1}_A = d_n \circ (f\mathbf{1}_A)$ (convince yourself of this), and likewise for B and C, we get

$$\mu(d_n \circ (f\mathbf{1}_A)) + \mu(d_n \circ (f\mathbf{1}_B)) = \mu(d_n \circ (f\mathbf{1}_C)) \, .$$

Taking the limit $n \to \infty$ and checking the definition (part b) yields the result. □

Exercise 26 (Dyadic function commutes with multiplication of indicator function)**:**

Show that for any $f \in \mathcal{E}_+^{\text{fn}}$ and $A \in \mathcal{E}$, $d_n \circ (f\mathbf{1}_A) = \mathbf{1}_A(d_n \circ f)$.

Solution:

Positivity and monotonicity.

Proposition 7.2. *If $f \in \mathcal{E}_+^{\text{fn}}$ then $\mu f \geq 0$. If $f, g \in \mathcal{E}_+^{\text{fn}}$ and $f \leq g$, then $\mu f \leq \mu g$.*

Exercise 27 (Proof of Proposition 7.2)**:**

Prove Proposition 7.2.

Hint: For monotonicity, first show it for positive simple functions and then for all positive measurable functions. (See the proof of Lemma 7.1 for an example of this style of proof.)

Solution: Positivity of the integral for $f \in \mathcal{E}_+^{\text{fn}}$ follows from the definition of the integral. For monotonicity, let $f_n = d_n \circ f$ and $g_n = d_n \circ g$; since d_n is an increasing function, $f \leq g$ implies

that $f_n \leq g_n$ for all n. These are both simple functions and we established monotonicity for simple functions as an immediate consequence of the definition of the integral, so $\mu f_n \leq \mu g_n$ for all n. Hence, letting $n \to \infty$ and checking the definition (part b)), we see that $\mu f \leq \mu g$.

7.4 Monotone Convergence Theorem

As Çinlar notes, this is the main theorem of integration. It says that the mapping $f \mapsto \mu f$ is continuous under increasing limits, and we may interchange integral with limit. It is useful in its own right, as it is often easier to evaluate μf_n and then take the limit, rather than the other way around.

The proof is a bit involved, but it's illuminating and worth the time.

Theorem 7.3 (Monotone Convergence (MCT))**.** *Let (f_n) be a increasing sequence of positive $\mathcal{E}/\mathcal{B}(\mathbb{R}_+)$-measurable functions. Then*

$$\mu(\lim_n f_n) = \lim_n \mu f_n \ .$$

Proof. First, let's make sure that limits on both sides of the claimed equality are well-defined. Let $f = \lim_n f_n$, which is well-defined because (f_n) is increasing. Clearly, $f \in \mathcal{E}_+^{\text{fn}}$ so μf is well-defined. Since (f_n) is increasing, Proposition 7.2 on monotonicity implies that the integrals (μf_n) form an increasing sequence of numbers. Hence, $\lim_n \mu f_n$ exists. We want to show that it is equal to μf.

We will do so by proving the following two claims:

1. $\mu f \geq \lim_n \mu f_n$.
2. $\mu f \leq \lim_n \mu f_n$.

Proof of Claim 1: Because $f \geq f_n$ for each n, monotonicity (Proposition 7.2) yields $\mu f \geq \mu f_n$ for each n. It follows that $\mu f \geq \lim_n \mu f_n$.

Proof of Claim 2: We will do this in three parts, using first an indicator function, then a simple function, then the limit of dyadic functions.

Indicator function. Fix $b \in \mathbb{R}_+$ and a set $B \in \mathcal{E}$. Suppose that $f(x) > b$ for every $x \in B$. Since the sets $\{f_n > b\} = \{x \in E : f_n(x) > b\}$ are increasing to $\{f > b\}$, the sets $B_n = B \cap \{f_n > b\}$ are increasing to B. Therefore, by the sequential continuity of μ (Proposition 3.1),

$$\lim_n \mu(B_n) = \mu(B) \ . \tag{7.6}$$

Now consider the function $f_n \mathbf{1}_B$. We have that

$$f_n \mathbf{1}_B \geq f_n \mathbf{1}_{B_n} \geq b \mathbf{1}_{B_n} \ ,$$

which (again) by monotonicity yields that

$$\mu(f_n \mathbf{1}_B) \geq \mu(b \mathbf{1}_{B_n}) = b\mu(B_n).$$

We established the limit of $\mu(B_n)$ in (7.6), so

$$\lim_n \mu(f_n \mathbf{1}_B) \geq b\mu(B) \; . \tag{7.7}$$

This is still true if $f(x) \geq b$ for all $x \in B$. To see this note that if $b = 0$ then by the positivity of the integral this is trivially true. For $b > 0$, choose a sequence of positive numbers (b_m) such that $b_m \nearrow b$. Then $f(x) > b_m$ for all $x \in B$, and (7.7) is true with b replaced by b_m. Letting $m \to \infty$, we get (7.7) again.

Simple function. Let g be a positive simple function such that $f \geq g$, with canonical representation $g = \sum_{i=1}^m b_i \mathbf{1}_{B_i}$. Therefore, $f(x) \geq b_i$ for every $x \in B_i$, and (7.7) yields

$$\lim_n \mu(f_n \mathbf{1}_{B_i}) \geq b_i \mu(B_i) \; , \quad i = 1, \ldots, m \; . \tag{7.8}$$

Furthermore, $f_n = f_n \sum_{i=1}^m \mathbf{1}_{B_i}$ (because by definition of canonical representation the sets $\{B_i\}$ form a partition of E) and therefore

$$\begin{aligned}
\lim_n \mu f_n &= \lim_n \mu(f_n \sum_{i=1}^m \mathbf{1}_{B_i}) && \text{(monotonicity)} \\
&= \lim_n \sum_{i=1}^m \mu(f_n \mathbf{1}_{B_i}) && \text{(finite additivity, Lemma 7.1)} \\
&= \sum_{i=1}^m \lim_n \mu(f_n \mathbf{1}_{B_i}) && \text{(interchange limit with finite sum)} \\
&\geq \sum_{i=1}^m b_i \mu(B_i) && \text{(by (7.8))} \\
&= \mu g && \text{(definition of integral for simple functions)} \; .
\end{aligned}$$

This holds for every positive simple function g such that $f \geq g$.

Limit of dyadic functions. Recall that by definition, $\mu f = \lim_k \mu(d_k \circ f)$. For each k, $d_k \circ f$ is a positive simple function and $f \geq d_k \circ f$. Hence, setting $g = d_k \circ f$, we have

$$\lim_n \mu f_n \geq \mu(d_k \circ f)$$

for all k. Letting $k \to \infty$, we obtain $\lim_n \mu f_n \geq \lim_k \mu(d_k \circ f) = \mu f$. □

7.5 Further properties of integration

Linearity.

Proposition 7.4 (Linearity of integration)**.** *For $f, g \in \mathcal{E}_+^{\mathrm{fn}}$ and $a, b \in \mathbb{R}_+$,*

$$\mu(af + bg) = a\mu f + b\mu g \; . \tag{7.9}$$

The same is true of integrable $f, g \in \mathcal{E}^{\mathrm{fn}}$ and arbitrary $a, b \in \mathbb{R}$.

Exercise 28 (Proof of Proposition 7.4)**:**

Prove Proposition 7.4.

Hint: Recall that we established linearity for simple f, g immediately after the definition of the integral. For general positive f, g, choose sequences of positive increasing functions $f_n \nearrow f$ and $g_n \nearrow g$, and use the Monotone Convergence Theorem.

Solution:

Proof. This proof is a nice example of how useful/powerful the MCT can be.

Suppose that $f, g \in \mathcal{E}_+^{\text{fn}}$ and $a, b \geq 0$. We established linearity for simple f, g immediately after the definition of the integral. For general positive f, g, choose sequences of positive increasing functions $f_n \nearrow f$ and $g_n \nearrow g$. Then

$$\mu(af_n + bg_n) = a\mu f_n + b\mu g_n \ .$$

Applying the Monotone Convergence Theorem to both sides,

$$\begin{aligned} \lim_n \mu(af_n + bg_n) &= \lim_n a\mu f_n + \lim_n b\mu g_n \\ \mu(a \lim_n f_n + b \lim_n g_n) &= a\mu(\lim_n f_n) + b\mu(\lim_n g_n) \\ \mu(af + bg) &= a\mu f + b\mu g \ . \end{aligned}$$

□

Insensitivity.

Proposition 7.5. *The integral has the following* ***insensitivity*** *properties:*

i) If $A \in \mathcal{E}$ is negligible (i.e., $\mu(A) = 0$), then $\mu(f\mathbf{1}_A) = 0$ for every $f \in \mathcal{E}^{\text{fn}}$.

ii) If $f, g \in \mathcal{E}_+^{\text{fn}}$ and $f = g$ almost everywhere, then $\mu f = \mu g$.

iii) If $f \in \mathcal{E}_+^{\text{fn}}$ and $\mu f = 0$, then $f = 0$ almost everywhere.

Exercise 29 (Insensitivity of integration)**:**

Prove Proposition 7.5.

Solution: See Çinlar [Çin11, Prop. I.4.13].

7.6 Characterization of the integral

We will not prove the following important result, which essentially says that the properties of positivity, linearity, and monotone convergence characterize the integral.

Theorem 7.6 (Characterization of the integral)**.** *Let $(E, \mathcal{E})$ be a measurable space. Let L be a*

mapping from $\mathcal{E}_+^{\text{fn}}$ *into* $\bar{\mathbb{R}}_+$. *Then there exists a unique measure* μ *on* $(E, \mathcal{E})$ *such that* $L(f) = \mu f$ *for every* $f \in \mathcal{E}_+^{\text{fn}}$ *if and only if:*

a) $f = 0 \Rightarrow L(f) = 0$.

b) $f, g \in \mathcal{E}_+^{\text{fn}}$ *and* $a, b \in \mathbb{R}_+ \Rightarrow L(af + bg) = aL(f) + bL(g)$.

c) $(f_n) \subset \mathcal{E}_+^{\text{fn}}$ *and* $f_n \nearrow f \Rightarrow L(f_n) \nearrow L(f)$.

7.7 More on interchanging limits and integration

Two limiting operations do not necessarily commute—they might not be interchangeable. Integration, differentiation, infinite sums are different types of limiting operations. It is common to encounter something like

$$\lim_n \int_{\mathbb{R}} f_n(x, \theta) dx \quad \text{or} \quad \frac{d}{d\theta} \int_{\mathbb{R}} f(x, \theta) dx \, .$$

Often, performing one limit operation before the other is much easier, e.g., differentiating before integrating. We will establish conditions for interchanging differentiation and integration in the context of expectations.

The MCT allows us to interchange limits with integration for increasing sequences of functions; the **Dominated Convergence Theorem** relaxes the requirement that (f_n) is increasing, at the expense of bounding the sequence with another integrable function.

Dominated functions and integration. A function f is **dominated** be a function g if $|f| \leq g$. (Note that $g \geq 0$ necessarily.) A sequence (f_n) is dominated by g if $|f_n| \leq g$ for every n.

Theorem 7.7 (Dominated Convergence Theorem)**.** *Let* (f_n) *be a sequence of* $\mathcal{E}$*-measurable functions. Suppose that* (f_n) *is dominated by some integrable function* g. *If* $\lim_n f_n$ *exists, then it is integrable and*

$$\mu(\lim_n f_n) = \lim_n \mu f_n \, .$$

See Çinlar [Çin11, pp. 25-26] for the proof, along with the following intermediate results (also useful on their own) on interchanging $\liminf/\limsup$ and integration.

Lemma 7.8 (Fatou's Lemma)**.** *Let* (f_n) *be a sequence of positive* $\mathcal{E}$*-measurable functions. Then* $\mu(\liminf_n f_n) \leq \liminf_n \mu f_n$.

Corollary 7.9. *Let* (f_n) *be a sequence of (not necessarily positive)* $\mathcal{E}$*-measurable functions. If there is an integrable function* g *such that* $f_n \geq g$ *for every* n, *then*

$$\mu(\liminf_n f_n) \leq \liminf_n \mu f_n \, .$$

If there is an integrable function g *such that* $f_n \leq g$ *for every* n, *then*

$$\mu(\limsup_n f_n) \geq \limsup_n \mu f_n \, .$$

Note that for the DCT, (f_n) does not need to be monotone; it only needs to have a limit and be dominated by an integrable function.

If (f_n) is bounded by some constant $b \in \mathbb{R}_+$ and μ is finite, then we can take $g = b$ in the DCT.

Theorem 7.10 (**Bounded Convergence Theorem**). *Let (f_n) be a sequence of $\mathcal{E}$-measurable functions, bounded by $b \in \mathbb{R}_+$. Suppose that μ is finite. If $\lim_n f_n$ exists, then it is a bounded integrable function and*

$$\mu(\lim_n f_n) = \lim_n \mu f_n \ .$$

7.8 Integration and image measures

Recall (from Section 3.6) that for a measurable function $h : F \to E$ and a measure ν on $(F, \mathcal{F})$, the image of ν under h, $\nu \circ h^{-1}$, is a measure on $(E, \mathcal{E})$. The following theorem says that we can integrate either with the image measure on $(E, \mathcal{E})$, or with ν on $(F, \mathcal{F})$.

Theorem 7.11. *For every $f \in \mathcal{E}_+^{\text{fn}}$, $(\nu \circ h^{-1})f = \nu(f \circ h)$.*

Proof. Sketch of proof: Define $L : \mathcal{E}_+^{\text{fn}} \to \mathbb{R}_+$ by setting $L(f) = \nu(f \circ h)$, and check that L meets the conditions of Theorem 7.6. Thus, $L(f) = \mu f$ for some unique measure μ on $(E, \mathcal{E})$. Then, show that μ and $\nu \circ h^{-1}$ agree for all $B \in \mathcal{E}$. □

Exercise 30 (Proof of Theorem 7.11)**:**

Complete the proof of Theorem 7.11.

Solution:

Written explicitly, this theorem says that

$$\int_F \nu(dx) f(h(x)) = \int_E \mu(dy) f(y) \ . \tag{7.10}$$

This might look familiar: it is just a general version of the change-of-variables formula from calculus. When expectations and random variables are involved, it also goes by the (somewhat annoying) name "Law of the unconscious statistician". For $F = E = \mathbb{R}^d$, typically μ and ν expressed in terms of the Lebesgue measure on $\mathbb{R}^d$ and the Jacobian of the transformation h.

Schilling [Sch05, Chapters 14-15] goes into much more detail on the topic of integrating with image measures.

8 Expectation

Reading: Çinlar [Çin11], I.4, II.2

Supplemental:

Learning Objectives. At the end of this section, you will be able to do the following.

- Define expectation as a special case of integration with respect to a measure.
- Form a new measure from an indefinite integral.
- Define the Radon–Nikodym derivative, give some common examples, and use its properties in integration.
- State and use Markov's inequality and Jensen's inequality.
- Define the Laplace and Fourier transforms, and relate them to generating functions.

Random variables are measurable functions; expectations of random variables are integrals of measurable functions. Hence, we've already developed most of the necessary theory. This is really just integrating with respect to a probability measure. We'll focus on more specialized, "probabilistic" results in this section.

8.1 Integrating with respect to a probability measure

As always, we have a background probability space $(\Omega, \mathcal{H}, \mathbb{P})$. We denote the **expectation** or **expected value** of a random variable X as

$$\mathbb{E}X = \mathbb{E}[X] = \int_\Omega \mathbb{P}(d\omega) X(\omega) = \int_\Omega X d\mathbb{P} = \mathbb{P}X \ . \tag{8.1}$$

Other notation used in various fields is meant to clarify what random variable is in question, and what distribution is being used, i.e., $\mathbb{E}_X[f(X,Y)]$ or $\mathbb{E}_{X\sim P}[f(X)]$.

$\mathbb{E}$ is treated as an operator corresponding to $\mathbb{P}$.

Properties. All of the conventions and notations of integration transfer over to expectation. X is **integrable** if $\mathbb{E}X$ exists and is finite. The integral of X over the event $H \in \mathcal{H}$ is $\mathbb{E}X\mathbf{1}_H$.

Everything that we proved for integrals in Section 7 holds for expectation (e.g., positivity, monotonicity, linearity, monotone convergence, etc.).

The following result is, in Çinlar's words, the "work horse" of probabilistic computations.

Theorem 8.1. *Let X be a random variable taking values in a measurable space $(E, \mathcal{E})$. If μ is the distribution of X, then*

$$\mathbb{E}f \circ X = \mu f \tag{8.2}$$

for every $f \in \mathcal{E}_+^{\text{fn}}$. Conversely, if (8.2) holds for some measure μ and all $f \in \mathcal{E}_+^{\text{fn}}$, then μ is the distribution of X.

Proof. The first statement is true because of what we know about integration with respect to image measures (Theorem 7.11): $\mu = \mathbb{P} \circ X^{-1}$, so $\mu f = \mathbb{P}(f \circ X) = \mathbb{E}f \circ X$ for every $f \in \mathcal{E}_+^{\text{fn}}$. Conversely,

if (8.2) holds for all $f \in \mathcal{E}_+^{\text{fn}}$, taking $f = \mathbf{1}_A$ for any $A \in \mathcal{E}$, we see that

$$\mu(A) = \mu\mathbf{1}_A = \mathbb{E}\mathbf{1}_A \circ X = \mathbb{P}\{X^{-1}A\} = \mathbb{P}\{X \in A\} ,$$

which is the distribution of X. □

A common way to show that $X \stackrel{\text{d}}{=} Y$ is to show that $\mathbb{E}[f(X)] = \mathbb{E}[f(Y)]$ for every $f \in \mathcal{E}_+^{\text{fn}}$.

8.2 Changes of measure: indefinite integrals and Radon–Nikodym derivatives

Composing a measure with a function constructs a measure (the image); alternatively we might multiply a measure and a function. In particular, let $(E, \mathcal{E}, \mu)$ be a measure space and $p : E \to \mathbb{R}_+$ be a measurable function. Define

$$\nu(A) = \mu(p\mathbf{1}_A) = \int_A \mu(dx)p(x) , \quad A \in \mathcal{E} . \tag{8.3}$$

ν is a measure on $(E, \mathcal{E})$. It is called the **indefinite integral** of p with respect to μ. Like the image measure, we can choose to integrate with μ or with ν.

Proposition 8.2. *For every $f \in \mathcal{E}_+^{\text{fn}}$, $\nu f = \mu(pf)$.*

Exercise 31 (Indefinite integral yields a measure)**:**

Prove that ν as defined in (8.3) is a measure with the additional assumption that $\mu p < +\infty$.

Exercise 32 (Proof of Proposition 8.2)**:**

Prove Proposition 8.2.

In more explicit notation, (8.3) is

$$\int_E \nu(dx)f(x) = \int_E \mu(dx)p(x)f(x) , \quad f \in \mathcal{E}_+^{\text{fn}} . \tag{8.4}$$

Informally, we write

$$\nu(dx) = \mu(dx)p(x) , \quad x \in E . \tag{8.5}$$

This aligns with ideas we may have taken for granted in introductory probability. Heuristically, $\mu(dx)$ is the amount of mass μ assigns to an "infinitesimal neighborhood" dx of a point x, and similarly for $\nu(dx)$. The $p(x)$ tells us how to go between the two; it is the mass density at x of ν with respect to μ. The function p is called the **density function** of ν relative to μ, with the following notation:

$$p = \frac{d\nu}{d\mu} , \quad \text{and} \quad p(x) = \frac{d\nu}{d\mu}(x) = \frac{\nu(dx)}{\mu(dx)} . \tag{8.6}$$

When dealing with random variables X and X', we might see

$$p = \frac{\mathbb{P}(X \in dx)}{\mathbb{P}(X' \in dx)} , \quad x \in E . \tag{8.7}$$

Radon–Nikodym derivative. Above, we used some function p to construct a new measure via (8.3). What if we have two measures? Is there some p that relates the two? The **Radon–Nikodym theorem** establishes conditions for the existence (and almost sure uniqueness) of such a p. Although it may seem like something of a technical curiosity, it is extremely powerful; we will use it to build conditional probability distributions. We won't prove it in this course.

So what conditions do we need for a density function p to exist? At a minimum, μ and ν need to agree on non-null sets. In particular, a measure ν on $(E, \mathcal{E})$ is **absolutely continuous** with respect to μ (also on $(E, \mathcal{E})$) if, for every set $A \in \mathcal{E}$,

$$\mu(A) = 0 \Rightarrow \nu(A) = 0 .$$

This is denoted $\nu \ll \mu$. It is also said that μ **dominates** ν.

In the indefinite integral construction (8.3), clearly $\nu \ll \mu$. That the converse is true (for "nice" ν) is the content of the theorem.

For the purposes of stating the theorem, we need the following definition. Recall that a measure μ on $(E, \mathcal{E})$ is **σ-finite** if there exists a measurable partition (E_n) of E such that $\mu(E_n) < \infty$ for all n. Clearly, every finite measure is σ-finite.

Theorem 8.3 (Radon–Nikodym)**.** *Suppose that μ is a σ-finite measure and that $\nu \ll \mu$. Then there exists a positive $\mathcal{E}$-measurable function $p : E \to \mathbb{R}_+$ such that*

$$\int_E \nu(dx) f(x) = \int_E \mu(dx) p(x) f(x) , \quad f \in \mathcal{E}_+^{\mathrm{fn}} . \tag{8.8}$$

Moreover, p is unique up to equivalence on non-null sets: if (8.8) holds for some other function $\hat{p} \in \mathcal{E}_+^{\mathrm{fn}}$, then $\hat{p}(x) = p(x)$ for μ-almost every $x \in E$.

The proof is beyond the scope of this course. Schilling [Sch05, Ch. 19] uses martingales; the proof by Bass [Bas22, Ch. 13] is probably more immediately accessible, using a result from advanced measure theory (the Hahn Decomposition Theorem).

The Radon–Nikodym theorem is the primary technical tool for defining conditional expectation (coming soon). More immediately, we recognize its role in defining some familiar objects.

Example 8.1 Discrete c.d.f. and p.m.f. Let $-\infty < a_1 < a_2 < \ldots < \infty$ be a sequence of real numbers and let (p_n) be a sequence of positive numbers such that $\sum_n p_n = 1$. Then

$$F(x) = \begin{cases} \sum_{n=1}^{m} p_n & a_m \le x < a_{m+1} \\ 0 & -\infty < x < a_1 \end{cases} ,$$

is a cumulative distribution function consisting purely of jumps. F corresponds to some discrete probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$,

$$\mu(A) = \sum_{n: a_n \in A} p_n \ , \quad A \in \mathcal{B}(\mathbb{R}) \ .$$

Let ν be the counting measure on $2^{\mathbb{R}}$. Then $\mu \ll \nu$ and

$$\mu(A) = \int_A f(x)\nu(dx) = \sum_{a_i \in A} f(a_i) \ , \quad A \subset \mathbb{R} \ ,$$

where $f(a_i) = p_i$. f is the **probability mass function** of μ or F with respect to ν.

Example 8.2 c.d.f. and p.d.f. Let F be a distribution function (on $\mathbb{R}$), and assume that F is differentiable (in the sense from calculus). Then its derivative $f = \frac{dF}{dx}$ satisfies

$$F(x) = \int_{-\infty}^{x} f(s)\lambda(ds) \ , \quad x \in \mathbb{R} \ .$$

Let μ be the corresponding probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then

$$\mu(A) = \int_A f(s)\lambda(ds) \ , \quad A \in \mathcal{B}(\mathbb{R}) \ .$$

f is the **probability density function** of μ or F with respect to the Lebesgue measure λ.

The Radon–Nikodym derivative obeys a number of useful identities.

Proposition 8.4. *Let μ be a σ-finite measure on $(E, \mathcal{E})$. Let all other measures appearing below also be on $(E, \mathcal{E})$. The following hold*

i) If $\nu \ll \mu$ and $f \in \mathcal{E}_+^{\text{fn}}$, then

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

ii) If $\nu_1 \ll \mu$ and $\nu_2 \ll \mu$, then $\nu_1 + \nu_2 \ll \mu$ and

$$\frac{d(\nu_1 + \nu_2)}{d\mu} = \frac{d\nu_1}{d\mu} + \frac{d\nu_1}{d\mu} \ , \quad \mu\text{-a.e.}$$

iii) If η is a measure, ν is a σ-finite measure, and $\eta \ll \nu \ll \mu$, then

$$\frac{d\eta}{d\mu} = \frac{d\eta}{d\nu}\frac{d\nu}{d\mu} \ , \quad \mu\text{-a.e.}$$

*In particular, if $\nu \ll \mu$ and $\mu \ll \nu$ (in which case μ and ν are **equivalent**), then*

$$\frac{d\nu}{d\mu} = \left(\frac{d\mu}{d\nu}\right)^{-1}, \quad \mu\text{- or } \nu\text{-a.e.}$$

iv) Let μ_1, ν_1 be σ-finite measures on $(E, \mathcal{E})$ and μ_2, ν_2 σ-finite measures $(F, \mathcal{F})$, such that $\nu_1 \ll \mu_1$ and $\nu_2 \ll \mu_2$. Then $\nu_1 \times \nu_2 \ll \mu_1 \times \mu_2$ and

$$\frac{d(\nu_1 \times \nu_2)}{d(\mu_1 \times \mu_2)}(x_1, x_2) = \frac{d\nu_1}{d\mu_1}(x_1)\frac{d\nu_2}{d\mu_2}(x_2), \quad \mu_1 \times \mu_2\text{-a.e.}$$

Part i) is also known as "change of measure," and can be quite useful. An example is the following result.

Proposition 8.5. *Let X be a random variable taking values in $(E, \mathcal{E})$, and μ its distribution. If there is some $f \in \mathcal{E}_+^{\text{fn}}$ with $\mathbb{E}[f(X)] = 0$, then $f(X) = 0$ μ-a.s. Furthermore, if Y is a random variable taking values in $(E, \mathcal{E})$, with distribution $\nu \ll \mu$, then $f(Y) = 0$ ν-a.s.*

Exercise 33 (Proof of Proposition 8.5)**:**

Prove Proposition 8.5.

Solution: The first claim is a restatement of Proposition 7.5 iii). To see the second claim, note that

$$\mathbb{E}[f(Y)] = \int_E f d\nu = \int_E f \frac{d\nu}{d\mu} d\mu = \mathbb{E}[f(X)\frac{d\nu}{d\mu}(X)] = 0 ,$$

where the final equality follows from the fact that $f(X) = 0$ μ-a.s. and Proposition 7.5 i). Another application of Proposition 7.5 iii) implies that $f(Y) = 0$ ν-a.s.

8.3 Moments and inequalities

The expectation is used to define a number of special classes of functions that describe different properties of random variables.

Moments. Certain expected values have special names. Let X be a random variable taking values in $\bar{\mathbb{R}}$, with distribution μ. Then $\mathbb{E}[X^n]$ is called the n**th moment** of X. For $n = 1$, $\mathbb{E}[X]$ is the **mean**. Assuming the mean is finite, e.g., $\mathbb{E}[X] = a < \infty$, the nth moment of $X - a$ is called the n**th centered moment** of X. A familiar example is the **variance** of X, $\text{Var}[X] = \mathbb{E}[(X - a)^2]$.

Markov's inequality. One of the most useful (basic) results from probability theory is **Markov's inequality**. It is a standard tool in probability, and number of other named inequalities are derived from it. Let X be a random variable in $\mathbb{R}_+$. Then for every $c > 0$,

$$\mathbb{P}\{X > c\} \leq \frac{1}{c}\mathbb{E}[X] . \tag{8.9}$$

Exercise 34 (Proof of Markov's inequality)**:**

Prove Markov's inequality (8.9).

Hint: Use the fact that $X \geq c\mathbf{1}_{X>c}$.

Solution: Fix some $c \geq 0$. Observe that $X \geq c\mathbf{1}_{X>c}$, and therefore $\mathbb{E}[X] \geq c\mathbb{E}[\mathbf{1}_{X>c}] = c\mathbb{P}\{X > c\}$.

Exercise 35 (Chebyshev's inequality)**:**

Assume that for $X \in \mathbb{R}$, $\mathbb{E}[X]$ is finite. Apply Markov's inequality to $(X - \mathbb{E}[X])^2$ to show that

$$\mathbb{P}\{|X - \mathbb{E}[X]| > c\} \leq \frac{1}{c^2}\mathrm{Var}[X] \,. \tag{8.10}$$

Exercise 36 (Generalized Markov's inequality)**:**

Let X be a random variable in $\mathbb{R}$, and $f : \mathbb{R} \to \mathbb{R}_+$ be increasing. Show that for every $c \in \mathbb{R}$,

$$\mathbb{P}\{X > c\} \leq \frac{1}{f(c)}\mathbb{E}[f(X)] \,. \tag{8.11}$$

Jensen's inequality. Another standard tool relies on special properties of convex functions. A function $\varphi : \mathbb{R} \to \mathbb{R}$ is **convex** if $\varphi = \sup_n \varphi_n$ for some sequence of functions with the form $\varphi_n(x) = a_n + b_n x$.

Theorem 8.6 (**Jensen's inequality**). *Let X be a random variable in $\mathbb{R}$, and $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function. Then*

$$\varphi\big(\mathbb{E}[X]\big) \leq \mathbb{E}\big[\varphi(X)\big] \,. \tag{8.12}$$

The measure-theoretic version of this is (with μ the distribution of X)

$$\varphi\Big(\int_{\mathbb{R}} x\mu(dx)\Big) \leq \int_{\mathbb{R}} \varphi(x)\mu(dx) \,.$$

Exercise 37 (Proof of Jensen's inequality.)**:**

Prove Jensen's inequality (Theorem 8.6).

Example 8.3 Positivity of Kullback–Leibler divergence. Let $\nu \ll \mu$ be two probability measures on $(E, \mathcal{E})$. The **Kullback–Leibler divergence**, or KL divergence, between ν and μ is

$$\mathrm{KL}(\nu||\mu) = \int_E \ln\left(\frac{d\nu}{d\mu}\right) d\nu \; . \tag{8.13}$$

The KL divergence plays an important role in Bayesian statistics and machine learning, particularly in variational inference methods.

Assume also that $\mu \ll \nu$. Then, because $-\ln(x)$ is convex,

$$\begin{aligned} \mathrm{KL}(\nu||\mu) &= -\int_E \ln\left(\frac{d\mu}{d\nu}\right) d\nu \\ &\geq -\ln\left(\int_E \frac{d\mu}{d\nu} d\nu\right) \\ &= \ln\left(\int_E d\mu\right) = 0 \; . \end{aligned}$$

Hence, we have shown that when ν and μ are mutually absolutely continuous, $\mathrm{KL}(\nu||\mu) \geq 0$. (This is also true without the mutual absolute continuity assumption, but we can't use this technique.)

8.4 Transforms and generating functions*

Laplace and Fourier transforms. Let X be a random variable in $\mathbb{R}_+$ with distribution μ. Then for $r \in \mathbb{R}_+$, the random variable e^{-rX} takes values in $[0, 1]$, and

$$\hat{\mu}_r(X) = \mathbb{E}[e^{-rX}] = \int_{\mathbb{R}_+} e^{-rx} \mu(dx) \tag{8.14}$$

is a number in $[0, 1]$. The function $r \mapsto \hat{\mu}_r$ from $\mathbb{R}_+$ into $[0, 1]$ is called the **Laplace transform** of μ (also of X). It can be shown that the Laplace transform determines the distribution: if μ and ν are distributions on $\bar{\mathbb{R}}_+$, and $\hat{\mu}_r = \hat{\nu}_r$ for all $r \in \mathbb{R}_+$, then $\mu = \nu$.

This is also known as the **moment generating function**: If $\mathbb{E}[X] < \infty$ and $\hat{\mu}_r$ is differentiable with respect to r on $(0, +\infty)$, then (we won't prove this)

$$\lim_{r \downarrow 0} \frac{d^n}{dr^n} \hat{\mu}_r = (-1)^n \mathbb{E}[X^n] \; . \tag{8.15}$$

The **Fourier transform**, or **characteristic function**, plays a similar role for distributions on the entire real line, $\mathbb{R}$. For a random variable X taking values in $\mathbb{R}$ with distribution μ, $e^{irX} = \cos rX + i \sin rX$ is a complex-valued random variable, and the notion of expectation extends naturally:

$$\hat{\mu}^*_r = \mathbb{E}[e^{irX}] = \int_{\mathbb{R}} e^{irx} \mu(dx) = \int_{\mathbb{R}} \mu(dx) \cos rx + i \int_{\mathbb{R}} \mu(dx) \sin rx \; . \tag{8.16}$$

The Fourier transform determines the distribution.

Probability generating function. If X takes values in $\{0, 1, \dots\}$ with distribution μ, then the **probability generating function** of μ is

$$\mathbb{E}[z^X] = \sum_{n=0}^{\infty} z^n \mathbb{P}\{X = n\} \, , z \in [0, 1] \, . \tag{8.17}$$

It determines the distribution of X: in the power series expansion of the distribution, the coefficient of z^n is $\mathbb{P}\{X = n\}$ for each n.

9 Independence

Reading: Çinlar [Çin11], II.5

Supplemental:

Learning Objectives. At the end of this section, you will be able to do the following.
- Define and check independence in a number of equivalent ways.

Recall that two random variables are independent if their joint distribution factors into the product of their marginals. There is a more general notion of independence, in terms of expectations and σ-algebras.

General definition of independence. For any (countable or uncountable) set I indexing a set of σ-algebras $(\mathcal{F}_i)_{i\in I}$, denote by

$$\mathcal{F}_I = \bigvee_{i\in I} \mathcal{F}_i = \sigma\left(\bigcup_{i\in I} \mathcal{F}_i\right) \tag{9.1}$$

the σ-algebra generated by the union of the σ-algebras. (Recall that the union itself is in general not a σ-algebra.)

Let $\mathcal{F}_1, \ldots, \mathcal{F}_n$ be a sequence of sub-σ-algebras of $\mathcal{H}$. Then $\{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$ is called an **independency** if

$$\mathbb{E}[V_1 \cdots V_n] = \mathbb{E}[V_1] \ldots \mathbb{E}[V_n] \,, \tag{9.2}$$

for all positive random variables $V_1 \in \mathcal{F}_1, \ldots, V_n \in \mathcal{F}_n$.

For an arbitrary (possibly infinite) index set T, and a sub-σ-algebra $\mathcal{F}_t$ of $\mathcal{H}$ for each $t \in T$, the collection $\{\mathcal{F}_t : t \in T\}$ is an independency if every finite subset is an independency.

In general, elements—such as random variables—of an independency are said to be **independent** or **mutually independent**. This reveals what we really mean when we say the random variables X and Y are independent: σX and σY are independent.

The following test for independence of σ-algebras echoes the test for measurability of functions in Proposition 2.2: we only need to show independence on a generating subset. Recall that a **p-system** is a collection of sets that is closed under intersection.

Proposition 9.1. *Let $\mathcal{F}_1, \ldots, \mathcal{F}_n$ be sub-σ-algebras of $\mathcal{H}$, $n \geq 2$. For each $i \leq n$, let $\mathcal{C}_i$ be a p-system that generates $\mathcal{F}_i$. Then $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are independent if and only if*

$$\mathbb{P}(H_1 \cap \cdots \cap H_n) = \mathbb{P}(H_1) \cdots \mathbb{P}(H_n) \,, \tag{9.3}$$

for all H_i in $\bar{\mathcal{C}}_i = \mathcal{C}_i \cup \{\Omega\}$, $i = 1, \ldots, n$.

Exercise 38 (Partial proof of Proposition 9.1)**:**

Assume that $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are independent. Show that (9.3) holds.

See Çinlar [Çin11, Prop. II.5.2] for a proof of the converse (surprise, surprise, it uses the monotone class theorem).

9.1 Independence of random variables

For each t in some index set T, let X_t be a random variable in a measurable space $(E_t, \mathcal{E}_t)$. As defined above, the random variables X_t are **independent** if $\{\sigma X_t : t \in T\}$ is an independency. Something like a converse is true, by using the equivalence between a sub-σ-algebra and the collection of all $\mathbb{R}$-valued random variables that are measurable with respect to it. The idea at the heart of the connection is quite deep, and worth pausing to consider.

Information and determinability. A σ-algebra $\mathcal{G}$ is a **sub-σ-algebra** of $\mathcal{H}$ if $\mathcal{G} \subset \mathcal{H}$. Çinlar advocates for thinking of the sub-σ-algebra $\mathcal{G}$ both as a collection of events (i.e., measurable subsets of Ω), and as the collection of all numerical random variables that are $\mathcal{G}$-measurable. This is justified, particularly when $\mathcal{G} = \sigma X$ (recall the definition of a σ-algebra generated by a function, defined in Section 2.1, (2.3)), because of the following theorem—a version of which you proved in an assignment.

Theorem 9.2. *Let X be a random variable taking values in some measurable space $(E, \mathcal{E})$. A mapping $V : \Omega \to \bar{\mathbb{R}}$ belongs to σX (is $\sigma X/\mathcal{B}(\bar{\mathbb{R}})$-measurable) if and only if $V = f \circ X$ for some deterministic $\mathcal{E}/\mathcal{B}(\bar{\mathbb{R}})$-measurable function f.*

One interpretation is that σX is the collection of all $\mathbb{R}$-valued random variables that are measurable functions of X (and X only). In this sense, X **determines** all σX-measurable random variables. A random variable Y that is not σX-measurable is not determined by X, i.e., there is no function f such that $Y = f(X)$—there is leftover randomness. In fact, in general one can take $Y = f(X, U)$, where U is some independent "noise" variable, when Y is not determined by X.

One common analogy is that σX is a body of information, say all of the conclusions we might reach deterministically based on a measurement of the atmospheric pressure. When we study conditional expectations, we will see that conditioning on σX yields a "best estimate" of Y given the information in σX.

A word on notation. Recall that for a measurable space $(E, \mathcal{E})$, $\mathcal{E}^{\text{fn}}$ denotes the set of all numerical functions that are $\mathcal{E}/\mathcal{B}(\mathbb{R})$-measurable. When working with random variables (and σ-algebras generated by random variables), this notation can get a bit tedious. I will adopt the following short-hand: if $V : E \to \mathbb{R}$ is a random variable in $\mathbb{R}$, we say that V belongs to σX, denoted $V \in \sigma X$, if it is $\sigma X/\mathcal{B}(\mathbb{R})$-measurable.

Recall from Theorem 9.2 that a mapping $V : \Omega \to \bar{\mathbb{R}}$ belongs to σX if and only if

$$V = f \circ X \ ,$$

for some deterministic function $f \in \mathcal{E}^{\text{fn}}$.

This is used to adapt the previous result Proposition 9.1 to random variables.

Proposition 9.3. *The random variables $X_1, \ldots, X_n$ are independent if and only if*

$$\mathbb{E}[f_1(X_1) \cdots f_n(X_n)] = \mathbb{E}[f_1(X_1)] \cdots \mathbb{E}[f_n(X_n)] \ , \tag{9.4}$$

for all $f_1 \in \mathcal{E}_{1,+}^{\text{fn}}, \ldots, f_n \in \mathcal{E}_{n,+}^{\text{fn}}$. The same is true if $f_1, \ldots, f_n$ are restricted to be integrable. If each E_i is a metric space and $\mathcal{E}_i$ is its Borel σ-algebra, then the same is true if $f_1, \ldots, f_n$ are restricted to be bounded, continuous functions.

Proof. For positive measurable functions, we need to show that (9.2) holds for all positive $V_1 \in \sigma X_1, \ldots, V_n \in \sigma X_n$ if and only if (9.4) holds for all positive $f_1 \in \mathcal{E}^{\rm fn}_{1,+}, \ldots, f_n \in \mathcal{E}^{\rm fn}_{n,+}$. This is immediate from Theorem 9.2: $V_i \in \sigma X_i$ if and only if $V_i = f_i(X_i)$ for some $f_i \in \mathcal{E}^{\rm fn}_i$.

For integrable functions, write $f_i = f_i^+ - f_i^-$, with each $f_i^\pm$ positive and measurable. If $X_1, \ldots, X_n$ are independent then (9.3) extends by the linearity of the integral. Conversely, each indicator function $\mathbf{1}_{H_i}$, $H_i \in \sigma X_i$, is bounded and measurable, so if (9.3) holds for all bounded measurable function, then Proposition 9.1 is implied. Similarly for bounded, continuous functions. □

Example 9.1 Beta-gamma magic. Let X and Y be independent random gamma variables (as in Example 5.2), with parameters $(a, 1)$ and $(b, 1)$, respectively. We will show the following:

a) $Z = X + Y$ has a gamma distribution with parameters $(a + b, 1)$, denoted $\gamma_{a+b,1}$.

b) $U = X/(X + Y)$ has a beta distribution with parameters (a, b), denoted $\beta_{a,b}$. That is,

$$\mathbb{P}(U \in du) = \beta_{a,b}(du) = du \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1}(1-u)^{b-1} , \quad 0 < u < 1 . \tag{9.5}$$

c) U and Z are independent: their joint distribution $\pi(dz, du)$ is the product measure $\gamma_{a+b,1}(dz) \times \beta_{a,b}(du)$.

Let f be any positive Borel function on $\mathbb{R}_+ \times [0, 1]$ and consider

$$\begin{aligned}
\pi f &= \mathbb{E}[f(X + Y, X/(X + Y))] \\
&= \int_0^\infty dx \frac{x^{a-1}e^{-x}}{\Gamma(a)} \int_0^\infty dy \frac{y^{b-1}e^{-y}}{\Gamma(b)} f(x + y, \frac{x}{x+y}) \\
&= \int_0^\infty dz \int_0^1 \frac{z^{a+b-1}e^{-z}}{\Gamma(a)\Gamma(b)} u^{a-1}(1-u)^{b-1} f(z, u) \\
&= \int_0^\infty \gamma_{a+b,1}(dz) \int_0^1 \beta_{a,b}(du) f(z, u) ,
\end{aligned}$$

where we have made the substitution $x = zu$ and $y = (1 - u)z$; the Jacobian is equal to z.

We can translate Proposition 9.3 into a statement about the distributions of the random variables. In particular, let π be the joint distribution of $X_1, \ldots, X_n$, and $\mu_1, \ldots, \mu_n$ their marginals. Then rewriting (9.4) yields

$$\int_{E_1 \times \cdots \times E_n} \pi(dx_1, \ldots, dx_n) f_1(x_1) \cdots f_n(x_n) = \int_{E_1} \mu_1(dx_1) f_1(x_1) \int_{E_2} \cdots \int_{E_n} \mu_n(dx_n) f_n(x_n) . \tag{9.6}$$

Clearly, that this equality holds for all positive measurable $f_1, \ldots, f_n$ is equivalent to saying that $\pi = \mu_1 \times \cdots \times \mu_n$.

Proposition 9.4. *The random variables $X_1, \ldots, X_n$ are independent if and only if their joint distribution is the product of their marginal distributions.*

Finally, we establish that functions of independent random variables are also independent.

Proposition 9.5. *Measurable functions of independent random variables are independent.*

Exercise 39 (Independence of functions of independent random variables.)**:**

Prove Proposition 9.5.

9.2 Sums of independent random variables*

Sums of sequences of random variables are of constant interest in probability and statistics. We focus briefly on the distribution of the sum of two random variables. Let X and Y be $\mathbb{R}$-valued random variables with distributions μ and ν, respectively. Then, the distribution of (X, Y) is the product measure $\mu \times \nu$, and the distribution of $X + Y$, $\mu * \nu$, is given by

$$(\mu * \nu)f = \mathbb{E}[f(X + Y)] = \int_{\mathbb{R}} \mu(dx) \int_{\mathbb{R}} \nu(dy) f(x + y) \, . \tag{9.7}$$

The distribution $\mu * \nu$ is called the **convolution** of μ and ν. Of course, because $X + Y = Y + X$, $\mu * \nu = \nu * \mu$.

In many cases, there are easier ways to identify the distribution of $X + Y$ than to compute (9.7). The following example gives one.

Example 9.2 Transforms of sums of independent random variables. A common way to establish distributional identities is via Laplace and Fourier transforms. This is particularly applicable for sums of independent random variables. For example, let X and Y be independent gamma-distributed random variables with parameters (a, b) and (c, b), respectively. Then

$$\mathbb{E}[e^{-rX}] = \left(\frac{b}{b+r}\right)^a \quad \text{and} \quad \mathbb{E}[e^{-rY}] = \left(\frac{b}{b+r}\right)^c .$$

What is the distribution of $X + Y$? We have (by independence)

$$\mathbb{E}[e^{-r(X+Y)}] = \mathbb{E}[e^{-rX}]\mathbb{E}[e^{-rY}] = \left(\frac{b}{b+r}\right)^a \left(\frac{b}{b+r}\right)^c = \left(\frac{b}{b+r}\right)^{a+c} = \mathbb{E}[e^{-rZ}] \, ,$$

where Z is gamma-distributed with parameters $(a + c, b)$. Because the Laplace transform characterizes the distribution for positive random variables, we know that $Z \stackrel{\mathrm{d}}{=} X + Y$.

9.3 Tail fields and Kolmogorov's 0-1 law*

We will need the following result, stating that independence survives grouping, to prove Kolmogorov's 0-1 law. In particular, let $\{\mathcal{F}_t : t \in T\}$ be an independency. For a countable partition $\{T_1, T_2, \dots\}$ of T, the subcollections $\mathcal{F}_{T_i} = \{\mathcal{F}_t : t \in T_i\}$, $i \in \mathbb{N}_+$, form a partition of the original independency.

Proposition 9.6. *Every partition of an independency is an independency.*

Tail σ-algebras. Let $(\mathcal{G}_n)$ be a sequence of sub-σ-algebras of $\mathcal{H}$. For the purposes of this section, we think of $\mathcal{G}_n$ as the information revealed by the nth trial of an experiment. Then the information about the future after n is $\mathcal{T}_n = \vee_{m>n}\mathcal{G}_m$. The σ-algebra consisting of events whose occurrences are unaffected by anything in finite time is the **tail σ-algebra**, $\mathcal{T} = \cap_n \mathcal{T}_n$.

Example 9.3 Tail σ-algebra. Let $X_1, X_2, \dots$ be random variables in $\mathbb{R}$, and $S_n = X_1 + \cdots + X_n$.

a) The event $\{\omega : \lim_n S_n(\omega) \text{ exists}\}$ belongs to $\mathcal{T}_n$ for every n, so it belongs to $\mathcal{T}$.
b) Likewise, $\{\limsup_n \frac{1}{n} S_n > b\}$ is unaffected by the first n variables, and hence it belongs to $\mathcal{T}$.
c) Conversely, $\{\limsup_n S_n > b\}$ is not in $\mathcal{T}$—we could change X_1 and change the event.
d) Let B be a Borel subset of $\mathbb{R}$. Let $\{X_n \in B \text{ i.o.}\}$, read "$X_n$ is in B **infinitely often**", be the set of ω for which $\sum_n \mathbf{1}_B \circ X_n(\omega) = +\infty$. This event belongs to $\mathcal{T}$.
e) The event $\{S_n \in B \text{ i.o.}\}$ is not in $\mathcal{T}$.

When the sequence $(\mathcal{G}_n)$ is an independency, something special happens: each of the events in $\mathcal{T}$ has probability 0 or 1.

Theorem 9.7 (Kolmogorov's 0-1 law)**.** *Let $\mathcal{G}_1, \mathcal{G}_2, \dots$ be independent. Then $\mathbb{P}(H)$ is either 0 or 1 for every event H in the tail σ-algebra $\mathcal{T}$.*

Proof. By Proposition 9.6 on partition independencies, $\{\mathcal{G}_1, \dots, \mathcal{G}_n, \mathcal{T}_n\}$ is an independency for each n, which implies that so is $\{\mathcal{G}_1, \dots, \mathcal{G}_n, \mathcal{T}\}$ for each n, since $\mathcal{T} \subset \mathcal{T}_n$. Thus, by definition $\{\mathcal{T}, \mathcal{G}_1, \mathcal{G}_2, \dots\}$ is an independency, and so is $\{\mathcal{T}, \mathcal{T}_0\}$ by Proposition 9.6 (again).

Pick any two events $H_1 \in \mathcal{T}$ and $H_2 \in \mathcal{T}_0$. Then $\mathbb{P}(H_1 \cap H_2) = \mathbb{P}(H_1)\mathbb{P}(H_2)$. Since $\mathcal{T} \subset \mathcal{T}_0$, for any $H_1 \in \mathcal{T}$, we have $H_1 \in \mathcal{T}_0$. Thus, for $H_1 \in \mathcal{T}$,

$$\mathbb{P}(H_1) = \mathbb{P}(H_1)\mathbb{P}(H_1) \,,$$

which means that $\mathbb{P}(H_1)$ equals either 0 or 1. □

As a corollary every random variable in the tail-σ-algebra (for example, $\limsup_n \frac{1}{n} S_n$) is almost surely constant.

Corollary 9.8. *Let $\mathcal{G}_1, \mathcal{G}_2, \dots$ be independent, and let V be a random variable in $\bar{\mathbb{R}}$, such that $V \in \mathcal{T}$. Then there is a constant $c \in \bar{\mathbb{R}}$ such that $V \overset{\text{a.s.}}{=} c$.*

Exercise 40:

Prove Corollary 9.8.

10 Probability distributions on real spaces

Reading: Jacod and Protter [JP04], Ch. 11-12.

Learning Objectives. At the end of this section, you will be able to do the following.

- Define the relationship between a probability measure P on $\mathbb{R}$ and its distribution function.
- Give necessary and sufficient conditions for a function $F\colon \mathbb{R} \to [0,1]$ to be the distribution of *some* probability measure on $\mathbb{R}$.

Çinlar [Çin11] leaves much of the material pertaining to probability distributions on $\mathbb{R}^n$ to exercises (or out of the book completely), but given their importance to most applications, it's important to study in more depth. We'll follow Jacod and Protter [JP04], Ch. 11-12, in a streamlined way.

The notation is slightly different here, following Jacod and Protter [JP04]. In particular, P is used to denote our "generic" probability measure.

10.1 Distribution functions and probability measures

Before proceeding, we need a couple of results that make precise the connection between distribution functions and probability measures. Recall that for a probability measure P on $\mathbb{R}$, its **distribution function** F is defined by

$$F(x) = P((-\infty, x]) \,, \quad x \in \mathbb{R} \,.$$

Clearly, a distribution function so defined will be unique to P. Conversely, we might ask when a function that "looks like" a distribution function is indeed the distribution function of a probability measure, and whether that correspondence is unique.

Theorem 10.1. *A function $F\colon \mathbb{R} \to [0,1]$ is the distribution function of a (unique) probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ if and only if it satisfies the following properties:*

(i) *F is non-decreasing: for $y > x$, $F(y) \geq F(x)$;*

(ii) *F is right continuous: if $x_n \downarrow x$ then $F(x_n) \downarrow F(x)$); and*

(iii) *$\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$.*

Proof.

Exercise 41 (Properties of a distribution function)**:**

Assume that F is a distribution function corresponding to a probability measure P. Prove that F has the properties listed above.

Solution: Assume that F is a distribution function corresponding to a probability measure P. If $y > x$ then $(-\infty, x] \subset (-\infty, y]$, so $P((-\infty, x]) \leq P((-\infty, y])$ (by monotonicity) so $F(y) \geq F(x)$, which establishes (i).

Now let (x_n) be a sequence decreasing to x. Then $\cap_{n\geq 1}(-\infty, x_n] = (-\infty, x]$, and the sequence of events $\{(-\infty, x_n] \mid n \geq 1\}$ is a decreasing sequence. By continuity under monotone limits, $P(\cap_{n\geq 1}(-\infty, x_n]) = \lim_{n\to\infty} = P((-\infty, x_n])P((-\infty, x])$. This is (ii).

Finally, $P(\mathbb{R}) = 1$ and continuity under monotone limits imply (iii).

The proof of the converse (that any function satisfying the properties is a distribution function) is conceptually simple: let $\mathcal{B}_0$ be the collection of finite disjoint unions of intervals of the form $(x, y]$, and define the set function $P\colon \mathcal{B}_0 \to [0, 1]$ by

$$P(A) = \sum_{1 \le i \le n} F(y_i) - F(x_i) \ , \quad \text{for } A = \cup_{1 \le i \le n} (x_i, y_i] \ , y_i < x_{i+1} \ .$$

If countable additivity of P so defined can be established on $\mathcal{B}_0$ (which is a p-system), then we can uniquely extend P to $\mathcal{B}(\mathbb{R})$ by Proposition 3.2. Establishing countable additivity is a somewhat lengthy exercise in real analysis; see [JP04, Thm. 7.2].

□

We can use this to express the probability of various intervals in terms of the distribution function.

Corollary 10.2. *Let F be a distribution function of the probability P on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Denote by $F(x-)$ the left limit of F at x (which exists because F is non-decreasing). For all $x < y$,*

(i) $P((x, y]) = F(y) - F(x)$;

(ii) $P([x, y]) = F(y) - F(x-)$;

(iii) $P([x, y)) = F(y-) - F(x-)$;

(iv) $P((x, y)) = F(y-) - F(x)$;

(v) $P(\{x\}) = F(x) - F(x-)$.

In particular, $P(\{x\}) = 0$ for all x if and only if F is continuous.

Try proving these yourself.

10.2 Lebesgue measure in one dimension

We just saw that a probability measure P on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is characterized by its distribution function $F(x) = P((-\infty, x])$.

For practical purposes, we need more than that, especially once we move to $n > 1$ dimensions. In particular, we need to study *Lebesgue measure* on $\mathbb{R}$ (and eventually on $\mathbb{R}^n$).

For simplicity, I will just write $\mathcal{B}$ for $\mathcal{B}(\mathbb{R})$ or for $\mathcal{B}(\mathbb{R}^n)$ when it's clear from context.

Lebesgue measure on $\mathbb{R}$ is a set function $\lambda\colon \mathcal{B} \to [0, \infty]$ that satisfies

(i) countable additivity; and

(ii) if $a, b \in \mathbb{R}$, $a < b$, then $\lambda((a, b]) = b - a$.

Theorem 10.3. *Lebesgue measure on $\mathbb{R}$ exists and is unique.*

Proof. We'll prove uniqueness first, following [JP04].

Assume that λ, Lebesgue measure on $\mathbb{R}$, exists. Fix $a < b$ and define

$$\lambda_{a,b}(A) = \frac{\lambda(A \cap (a,b])}{b-a} , \quad \text{all } A \in \mathcal{B} .$$

Then $\lambda_{a,b}$ is a probability measure on $(\mathbb{R}, \mathcal{B})$, with distribution function

$$F_{a,b}(x) = \lambda_{a,b}((-\infty, x]) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \le x < b \\ 1 & \text{if } b \le x . \end{cases} \tag{10.1}$$

Since the distribution function characterizes the probability measure, $\lambda_{a,b}$ is uniquely determined by $F_{a,b}$. We also have that

$$\lambda(A) = \sum_{n \in \mathbb{Z}} \lambda_{n,n+1}(A) , \quad \text{any } A \in \mathcal{B} , \tag{10.2}$$

and therefore λ is uniquely determined as well (since this is defined for all Borel sets).

For existence: $F_{a,b}$ in (10.1) satisfies the three criteria from Theorem 10.1 and therefore the probability measure $\lambda_{a,b}$ exists. We define λ by (10.2) and check that it satisfies the defining properties of Lebesgue measure. First, countable additivity: for $A_1, A_2, \ldots$ in $\mathcal{B}$ and pairwise disjoint,

$$\begin{aligned} \lambda(\cup_{k \ge 1} A_k) &= \sum_{n \in \mathbb{Z}} \lambda_{n,n+1}(\cup_{k \ge 1} A_k) \\ &= \sum_{n \in \mathbb{Z}} \sum_{k \ge 1} \lambda_{n,n+1}(A_k) \\ &= \sum_{k \ge 1} \sum_{n \in \mathbb{Z}} \lambda_{n,n+1}(A_k) \\ &= \sum_{k \ge 1} \lambda(A_k) . \end{aligned}$$

The interchange of infinite sums is justified by the fact that we have a double sequence that is increasing along each index (n and k). (Recall this from your homework.)

Finally, we need to show that $\lambda((a,b]) = b - a$.

Exercise 42:

Show that $\lambda((a,b]) = \sum_{n \in \mathbb{Z}} \lambda_{n,n+1}((a,b]) = b - a$.

Solution: Letting $\lfloor a \rfloor$ denote the floor of a (i.e., round down to the nearest integer), we also

have that

$$\begin{aligned}\lambda((a,b]) &= \sum_{n\in\mathbb{Z}} \lambda_{n,n+1}((a,b]) \\ &= \sum_{n\in\mathbb{Z}} \lambda_{n,n+1}((-\infty,b]) - \lambda_{n,n+1}((-\infty,a]) \\ &= (b - \lfloor b \rfloor) + (\lfloor b \rfloor - \lfloor a \rfloor) - (a - \lfloor a \rfloor) \\ &= b - a \ .\end{aligned}$$

□

All of the results for integration in Section 7 are still true when integrating with respect to λ, except that $\lambda(\mathbb{R}) = \infty$. However, (10.2) shows that Lebesgue measure is σ-finite (there is a disjoint partition $\cup_{n\geq 1} A_n = \mathbb{R}$ such that $\lambda(A_n) < \infty$ for each n). Most of the results from Section 7 have a σ-finite version (Çinlar [Çin11] has them).

10.3 Lebesgue measure in n dimensions

Now consider

$$\mathbb{R}^n = \underbrace{\mathbb{R} \times \mathbb{R} \cdots \times \mathbb{R}}_{n \text{ of these products}} \ .$$

This is just an n-dimensional product space, and we can generate its Borel σ-algebra, $\mathcal{B}^n$ using n-dimensional measurable rectangles or, in this case, measurable "quadrants" or rays,

$$\prod_{i=1}^{n} (-\infty, a_i] \ , \quad a_i \in \mathbb{Q} \ .$$

We end up with $\mathcal{B}^n = \sigma(\bigtimes_{i=1}^n \mathcal{B})$, the smallest σ-algebra generated by the n-fold Cartesian product of the Borel sets on $\mathbb{R}$. Our measurable space is then $(\mathbb{R}^n, \mathcal{B}^n)$.

The **n-dimensional distribution function** of a probability measure on $(\mathbb{R}^n, \mathcal{B}^n)$ is defined as

$$F(x_1, \ldots, x_n) = P\left(\prod_{i=1}^{n} (-\infty, x_i] \right) \ . \tag{10.3}$$

Characterizing P via its distribution function on $n \geq 2$ dimensions is more subtle than on 1 dimension and n-dimensional distribution functions aren't very common. We'll use densities instead, but first we need to extend Lebesgue measure to $\mathbb{R}^n$.

The **Lebesgue measure** on $(\mathbb{R}^n, \mathcal{B}^n)$ is defined on Cartesian product sets $A_1 \times A_2 \times \cdots \times A_n$ by

$$\lambda_n \left(\prod_{i=1}^{n} A_i \right) = \prod_{i=1}^{n} \lambda(A_i) \ , \tag{10.4}$$

where λ is the one-dimensional Lebesgue measure defined in the beginning of the section.

Later on in Theorem 12.5 (Tonelli–Fubini), we will see that we can extend λ_n from product sets uniquely to all of $(\mathbb{R}^n, \mathcal{B}^n)$, preserving countable additivity. (Intuitively, since the measurable rectangles generate the product σ-algebra, we could use the monotone class theorem to extend λ uniquely to all of $(\mathbb{R}^n, \mathcal{B}^n)$.) The resulting Lebesgue measure is also characterized by

$$\lambda_n\left(\prod_{i=1}^n (a_i, b_i]\right) = \prod_{i=1}^n (b_i - a_i)\ , \quad \text{all} -\infty < a_i < b_i < \infty\ . \tag{10.5}$$

10.4 Densities

If f is a Borel measurable function on $\mathbb{R}$ that is Lebesgue integrable, we write its integral as $\int f(x)dx$. Recall that f is integrable if and only if both its positive and negative parts have finite integral.

For $n \geq 2$, we write

$$\int f(x)dx = \int f(x_1, \ldots, x_n)dx_1 dx_2 \cdots dx_n\ .$$

A probability measure P on $(\mathbb{R}^n, \mathcal{B}^n)$ has a **density** or Radon–Nikodym derivative (with respect to Lebesgue measure) if f is a nonnegative Borel measurable function on $\mathbb{R}^n$ satisfying

$$P(A) = \int_A f(x)dx = \int f(x)\mathbf{1}_A(x)dx \tag{10.6}$$

$$= \int f(x_1, \ldots, x_n)\mathbf{1}_A(x_1, \ldots, x_n)dx_1 \cdots dx_n\ , \tag{10.7}$$

for all $A \in \mathcal{B}^n$. If $n = 1$, we can use sets of the form $(-\infty, x]$,

$$P((-\infty, x]) = \int_{-\infty}^x f(y)dy = \int_{-\infty}^{+\infty} f(y)\mathbf{1}_{(-\infty,x]}(y)dy\ . \tag{10.8}$$

If P is the distribution of a random variable X, then we say that f is the density of X.

Not all probability distributions have densities with respect to Lebesgue measure! For $n = 1$, Eq. (10.8) implies that the distribution function F is continuous, but there are some continuous F that do not have densities (and non-continuous F certainly do not).

Densities, when they exist, characterize the probability distribution.

Theorem 10.4 (Combines Jacod and Protter [JP04], Thm. 11.3 and Thm. 12.1)**.** *A nonnegative Borel measurable function $f\colon \mathbb{R}^n \to \mathbb{R}_+$ is the density of a probability measure on $(\mathbb{R}^n, \mathcal{B}^n)$ if and only if it satisfies $\int f(x)dx = 1$. In this case it entirely determines the probability measure, and any other Borel measurable function f' such that $\lambda_n(\{f \neq f'\}) = 0$ is also a density for the same probability measure.*

Conversely, a probability measure on $(\mathbb{R}^n, \mathcal{B}^n)$ determines its density (when it exists) up to a set of Lebesgue measure zero.

Following [JP04], we'll prove this for $n = 1$ only. The proof for $n \geq 2$ is analogous but more involved.

Proof. Let f be the density of a probability measure P. We have

$$P((-\infty, x]) = \int_{-\infty}^{x} f(y)dy \; .$$

Letting $x \nearrow +\infty$, we have

$$\int_{-\infty}^{+\infty} f(y)dy = \lim_{x\to\infty} \int_{-\infty}^{x} f(y)dy = \lim_{x\to\infty} P((-\infty, x]) = 1 \; .$$

For sufficiency (f determines P), we could just argue via the distribution function and Theorem 10.1. Alternatively, the following longer proof generalizes to $n \geq 2$.

Let f be a positive Borel function on $\mathbb{R}$ with $\int f(x)dx = 1$. For every Borel set A we put

$$P(A) = \int f(x)\mathbf{1}_A(x)dx \; . \tag{10.9}$$

This defines a function $P\colon \mathcal{B} \to \mathbb{R}_+$ with $P(\mathbb{R}) = 1$. It is also countably additive: If $A_1, A_2, \ldots$ are pairwise disjoint, then

$$\begin{aligned} P(\cup_{i\geq 1} A_i) &= \int f(x)\mathbf{1}_{\cup_{i\geq 1} A_i}(x)dx = \int \left(\sum_{i\geq 1} f(x)\mathbf{1}_{A_i}(x) \right) dx \\ &= \sum_{i\geq 1} \int f(x)\mathbf{1}_{A_i}(x)dx = \sum_{i\geq 1} P(A_i) \; . \end{aligned}$$

We used the disjointedness of the sets for the second equality, and monotone convergence for the third.

If we take $A = (-\infty, x]$ then $P(A) = \int_{-\infty}^{x} f(y)dy$, so P admits f as a density.

Exercise 43 (Equivalence of densities)**:**

Suppose that f' is another function such that $\lambda_n(\{f \neq f'\}) = 0$ (i.e., the two functions are equal almost everywhere). Show that f' is also a density of P.

Solution: Using the insensitivity of the integral, we can break up any integral as

$$\begin{aligned} P(A) = \int f(x)\mathbf{1}_A(x)dx &= \int f(x)\mathbf{1}_A(x)\mathbf{1}_{\{f=f'\}}(x)dx + \int f(x)\mathbf{1}_A(x)\mathbf{1}_{\{f\neq f'\}}(x)dx \\ &= \int f(x)\mathbf{1}_A(x)\mathbf{1}_{\{f=f'\}}(x)dx + 0 \\ &= \int f'(x)\mathbf{1}_A(x)\mathbf{1}_{\{f=f'\}}(x)dx + \int f'(x)\mathbf{1}_A(x)\mathbf{1}_{\{f\neq f'\}}(x)dx \\ &= \int f'(x)\mathbf{1}_A(x)dx \; . \end{aligned}$$

This shows that f' is also a density of P.

Now we need to show that P determines f up to a set of Lebesgue measure zero.

Suppose that f and f' are both densities for P. Then they both induce the same distribution function and therefore they both satisfy (10.9). So if we choose some $\epsilon > 0$ and set $A = \{x\colon f(x) + \epsilon \le f'(x)\}$, then if $\lambda(A) > 0$ we have

$$P(A) + \epsilon\lambda(A) = \int (f(x) + \epsilon \mathbf{1}_A(x))dx \le \int f'(x)dx = P(A) \, .$$

Therefore, it cannot be that $\lambda(A) > 0$, and hence $\lambda(\{f + \epsilon \le f'\}) = 0$. Since $\{f + \epsilon \le f'\}$ increases to $\{f < f'\}$ as $\epsilon \searrow 0$, we arrive at $\lambda(\{f < f'\}) = 0$. Likewise, by the same argument with f, f' reversed, $\lambda(\{f > f'\}) = 0$. Therefore, $f = f'$ almost everywhere.

□

Differentiability. Since f and F satisfy $F(x) = \int_{-\infty}^{x} f(y)dy$, it might look like F is differentiable with derivative f. This is true at all points x at which f is continuous. It turns out that F is differentiable (Lebesgue) almost everywhere—in general, not at all points $x \in \mathbb{R}$. In most actual examples (i.e., things we care about), if F has a density then it is piecewise differentiable and we can take $f = dF/dx$ wherever it exists, and $f = 0$ elsewhere.

Expectation rule. If the random variable X has density f then we know that $\mathbb{E}[g(X)] = \int g(x)f(x)dx$. You essentially already proved this on a homework. (See Corollary 11.1 in [JP04].)

10.5 Marginal, joint, "conditional" densities

We'll now consider how a density on $\mathbb{R}^n$ might be turned into a density on $\mathbb{R}^{n-1}$, $\mathbb{R}^{n-2}$, and so on. For simplicity we'll work explicitly only with $n = 2$, but the arguments generalize to $n > 2$ quite easily.

Let X be an $\mathbb{R}^2$-valued random variable with components Y and Z, i.e., $X = (Y, Z)$.

Theorem 10.5. *Assume that $X = (Y, Z)$ has a **joint density** f on $(\mathbb{R}^2, \mathcal{B}^2)$, $(y, z) \mapsto f(y, z)$. Then:*

*(a) Each of Y and Z have a **marginal density** on $(\mathbb{R}, \mathcal{B})$ given by*

$$f_Y(y) = \int_{-\infty}^{\infty} f(y, z)dz \, , \quad \text{and} \quad f_Z(z) = \int_{-\infty}^{\infty} f(y, z)dy \, . \tag{10.10}$$

(b) Y and Z are independent if and only if

$$f(y, z) = f_Y(y)f_Z(z) \, , \quad \lambda_2\text{-a.e.}$$

(c) The following formula defines another ("conditional") density on $\mathbb{R}$ at every point $y \in \mathbb{R}$ such that $f_Y(y) \neq 0$:

$$f_{Z|Y=y}(z) = \frac{f(y, z)}{f_Y(y)} \, . \tag{10.11}$$

In part (c), "conditional" is in quotes because strictly speaking, we haven't formulated how to handle conditional probabilities, especially $P(A \mid Y = y)$ when $P(Y = y) = 0$. But operationally, it behaves as expected and we can integrate with the "conditional" density and recover probabilities like $P(Z \in A, Y \in B)$. We'll revisit this after developing the framework for conditional probability.

Proof. (a) For each Borel set $A \in \mathcal{B}$,

$$\begin{aligned} P(Y \in A) = P(Y \in A, X \in \mathbb{R}) &= \int \int_{A \times \mathbb{R}} f(y,z) dy \; dz \\ &= \int_A dy \int_{\mathbb{R}} f(y,z) dz = \int_A dy \; f_Y(y) \; . \end{aligned}$$

Since this holds for each $A \in \mathcal{B}$ and densities on $\mathbb{R}$ are characterized by (10.9), $f_Y(y)$ as defined in (10.10) is a density of Y. The same proof holds for f_Z.

(b) First, suppose that $f(y,z) = f_Y(y) f_Z(z)$. Then it is easy to show the independence of Y and Z:

$$\begin{aligned} P(Y \in A, Z \in B) &= \int \int \mathbf{1}_{A \times B} f(y,z) dy \; dz \\ &= \int \mathbf{1}_A(y) f_Y(y) dy \int \mathbf{1}_B(z) f_Z(z) dz \\ &= P(Y \in A) P(Z \in B) \; . \end{aligned}$$

This holds for all Borel sets A, B so Y and Z are independent.

Now suppose that Y and Z are independent. Define the collection of sets

$$\mathcal{C} = \left\{ C \in \mathcal{B}^2 \colon \int \int_C f(y,z) dy \; dz = \int \int_C f_Y(y) f_Z(z) dy \; dz \right\} \; .$$

Since Y and Z are independent, if $C = A \times B$ for $A \in \mathcal{B}, B \in \mathcal{B}$, then

$$\begin{aligned} P((Y,Z) \in C) &= \int \int_C f(y,z) dy \; dz \\ &= P(Y \in A, Z \in B) = P(Y \in A) P(Z \in B) = \mathbb{E}[\mathbf{1}_A(Y)] \mathbb{E}[\mathbf{1}_B(Z)] \\ &= \int_A f_Y(y) dy \int_B f_Z(z) dz \; , \end{aligned}$$

by Proposition 9.4. Therefore, our collection $\mathcal{C}$ contains all measurable rectangles $C = A \times B$, $A \in \mathcal{B}, B \in \mathcal{B}$. This is a p-system that generates $\mathcal{B}^2$. Moreover, one can show that $\mathcal{C}$ is a d-system (show this!). Therefore, the monotone class theorem (Theorem 1.5) tells us that $\mathcal{C} = \mathcal{B}^2$. So, for any $C \in \mathcal{B}^2$,

$$P(X \in C) = \int_C f(y,z) dy \; dz = \int_C f_Y(y) f_Z(z) dy \; dz \; .$$

By the uniqueness of the density (Theorem 10.4), we have that $f(y,z) = f_Y(y) f_Z(z)$ λ_2-a.e.

(c) Since $f_Y(y) > 0$, we have

$$\begin{aligned}\int f_{Z|Y=y}(z)dz &= \int_{\mathbb{R}} \frac{f(y,z)}{f_Y(y)}dz \\ &= \frac{1}{f_Y(y)}\int_{\mathbb{R}} f(y,z)dz = \frac{1}{f_Y(y)} f_Y(y) = 1\ .\end{aligned}$$

Since $f_{Z|Y=y}(z)$ is nonnegative, Borel measurable, and integrates to 1, it is a density.

□

Exercise 44:

Confirm that the class $\mathcal{C}$ in the proof of Theorem 10.5(b) is a d-system.

Exercise 45 ("Conditional" density is Borel measurable)**:**

Show that $f_{Z|Y=y}(z)$ is Borel measurable.

10.6 Densities of transformed random variables

Suppose X has density f. Let $Y = g(X)$. Does Y have a density? If so, what is it? Jacod and Protter [JP04] go over the one-dimensional case in detail on pages 80–83; we'll skip ahead to the multidimensional case but there will be some one-dimensional examples here and in the homework.

In order to get a handle on the density of Y, we need a result from calculus. Recall that if g is a differentiable function from an open set $G \subset \mathbb{R}^n$ into $\mathbb{R}^n$, then its **Jacobian matrix** $J_g(x)$ at $x \in G$ is $J_g(x) = \frac{\partial g}{\partial x}(x)$, or

$$J_g(x)_{ij} = \frac{\partial g_i}{\partial x_j}(x)\ . \tag{10.12}$$

The **Jacobian** at x is the determinant of $J_g(x)$. If it is not zero then g is invertible on a neighborhood of x, and the Jacobian inverse g^{-1} at $y = g(x)$ is the inverse of the Jacobian of g at x: $J_{g^{-1}}(g(x)) = J_g^{-1}(x)$.

Theorem 10.6 (Jacobi's Transformation Formula)**.** *Let G be an open set in $\mathbb{R}^n$ and let $g\colon G \to \mathbb{R}^n$ be continuously differentiable (i.e., it has a continuous derivative). Suppose g is injective on G and its Jacobian never vanishes. Then for f measurable and such that the product $f\mathbf{1}_g(G)$ is positive or integrable with respect to Lebesgue measure,*

$$\int_{g(G)} f(y)dy = \int_G f(g(x))|\det(J_g(x))|dx\ , \tag{10.13}$$

where $g(G) = \{y \in \mathbb{R}^n\colon$ there exists $x \in G$ with $g(x) = y\}$.

The proof of this theorem is beyond the scope of the course; see Schilling [Sch05, Ch. 15] for details. In essence, Jacobi's formula comes from finding the Radon–Nikodym derivative of the image measure $\lambda \circ g^{-1}$ with respect to λ.

We can specialize the theorem to random variables.

Theorem 10.7. *Let $X = (X_1, \dots, X_n)$ have joint density f_X. Let $g\colon \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable and injective, with non-vanishing Jacobian. Then $Y = g(X)$ has density*

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y))|\det(J_{g^{-1}}(y))| & \text{if } y \text{ is in the range of } g \\ 0 & \text{otherwise.} \end{cases} \tag{10.14}$$

This can be extended to certain situations with non-injective but smooth g; see Corollary 12.1 in [JP04].

Examples. There quite a few examples in [JP04]: Chapter 11 covers $n = 1$, and Chapter 12 covers $n > 2$. Working with these things often requires some tricks, so going through the examples in the book is worthwhile. Time permitting, we'll go through some in class.

11 Conditional expectation

Reading: Çinlar [Çin11], Chapter IV.1

Supplemental:

Learning Objectives. At the end of this section, you will be able to do the following.

- Define conditional expectation given σ-algebras, and in particular given σ-algebras generated by random variables.
- Understand conditional expectation as a kind of projection.
- Define and use the conditional determinism and repeated conditioning properties of conditional expectation.

Conditioning is one of the most important concepts in modern probability, especially for applications in statistics and machine learning. Martingales and Markov processes are very difficult (or impossible) to even define without conditioning, and the starting point of Bayesian inference is conditioning on data. Hierarchical models, latent variable models, stochastic processes—none of these work without conditioning. Besides model construction, conditioning is the best tool we have for making inference, model simplification, and generally making our lives as statisticians / machine learners / data analysts easier.

11.1 Conditional expectations

For two $\mathbb{R}$-valued random variables X and Y, a heuristic notion of the conditional expectation $\mathbb{E}[X \mid Y = y]$ is an estimate of X given the information contained in the event $\{Y = y\}$, for some fixed value y. This is a pretty good heuristic, and one to keep in mind as we develop a more technical view of conditional expectation.

Some intuition. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space, $\mathcal{F}$ a sub-σ-algebra of $\mathcal{H}$, and X a $\mathbb{R}_+$-valued random variable. Çinlar reminds us to regard $\mathcal{F}$ both as a collection of events and as the collection of all $\mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable random variables. Let H be an event in $\mathcal{H}$ and assume that $\mathbb{P}(H) > 0$. Fix some $\omega \in \Omega$, and suppose that all we know is that $\omega \in H$. Given this information, our best estimate of $X(\omega)$ should be the "average" over H,

$$\mathbb{E}[X \mid H] = \frac{1}{\mathbb{P}(H)} \int_H \mathbb{P}(d\omega) X(\omega) = \frac{1}{\mathbb{P}(H)} \mathbb{E}[X \mathbf{1}_H] \, .$$

The quantity $\mathbb{E}[X \mid H]$ is a *number* called the **conditional expectation** of X given the event H. Note that we're conditioning on an *event*, not a random variable. In general, we will condition on events or collections of events (i.e., σ-algebras). Conditioning on a random variable Y is really short-hand for conditioning on σY.

Let's continue with building our intution. Suppose that the sub-σ-algebra $\mathcal{F}$ is generated by a measurable partition (H_n) of Ω. We can think of the parts of the partition as the highest resolution picture we have of Ω. For fixed ω, consider our estimate of $X(\omega)$ given the information $\mathcal{F}$: we can tell which of the events $H_1, H_2, \ldots$ includes ω, and therefore our estimate

$$\bar{X}(\omega) = \mathbb{E}[X \mid \mathcal{F}] = \sum_n \mathbb{E}[X \mid H_n] \mathbf{1}_{H_n}(\omega) \, . \tag{11.1}$$

If we do this for each $\omega \in \Omega$, we have defined a *random variable* $\bar{X} : \Omega \to \mathbb{R}_+$. This is the conditional expectation of X given $\mathcal{F}$.

Observe that because $\bar{X} = \mathbb{E}[X \mid \mathcal{F}]$ is a random variable, already important technical considerations are indicated. For example: Is it unique? Under what conditions is it integrable?

In order to generalize to arbitrary $\mathcal{F}$, two properties from the example above are key. Firstly, $\bar{X}$ is $\mathcal{F}$-measurable; *it is determined by the information in* $\mathcal{F}$. Second, $\mathbb{E}[VX] = \mathbb{E}[V\bar{X}]$ for every V that belongs to $\mathcal{F}_+$. The second property is known in its general form as the *projection property*. In our intuition-building setting, we can think of the set of indicator functions $(\mathbf{1}_{H_n})_{n\geq 1}$ as a set of basis functions for all positive $\mathcal{F}$-measurable functions. $\bar{X}$ as defined by (11.1) is then the projection of X onto that basis. Using Hilbert spaces of measurable functions, it is possible to extend these ideas to the general case. Although it is more intuitive, it requires covering background material on Hilbert spaces, so we'll take a shorter approach below.

Before continuing, let's check that these two properties hold for our intution-building scenario.

Exercise 46:

Show that $\bar{X}$ in Eq. (11.1) belongs to $\mathcal{F}$.

Solution: $\bar{X}$ is $\mathbb{R}_+$-valued, we we just need to show measurability with respect to a p-system of sets that generates $\mathcal{B}(\mathbb{R})$. To that end, take $0 < a < b < +\infty$, and consider the open interval (a, b). Assume that Then $\bar{X}^{-1}(a, b)$ is the union of all sets $H_n \in \mathcal{F}$ such that $\mathbb{E}[X\mathbf{1}_{H_n}]/\mathbb{P}(H_n) \in (a, b)$. Note that there will be at most a countable number of such sets. Because $\mathcal{F}$ is generated by the partition, that union is also in $\mathcal{F}$, and therefore $\bar{X}$ is $\mathcal{F}$-measurable (i.e., it belongs to $\mathcal{F}$).

Exercise 47:

Show that for $\bar{X}$ in Eq. (11.1), $\mathbb{E}[VX] = \mathbb{E}[V\bar{X}]$ for every V that belongs to $\mathcal{F}_+$.

Hint: First, fix n and let $V = \mathbf{1}_{H_n}$ and show the desired equality. Then, extend that to arbitrary $V \in \mathcal{F}_+$ using the MCT, observing that for this particular $\mathcal{F}$, all such $V = \sum_n a_n \mathbf{1}_{H_n}$.

Solution: Following the hint, fix n and let $V = \mathbf{1}_{H_n}$. Then

$$\mathbb{E}[V\bar{X}] = \mathbb{E}[\mathbf{1}_{H_n} \sum_k \mathbb{E}[X\mathbf{1}_{H_k}]\mathbf{1}_{H_k}/\mathbb{P}(H_k)] = \mathbb{E}[\mathbf{1}_{H_n} X]\mathbb{E}[\mathbf{1}_{H_n}]/P(H_n) = \mathbb{E}[\mathbf{1}_{H_n} X] = \mathbb{E}[VX] \, .$$

For simple functions V, the same is easy to check. For $\mathcal{F}$ generated by a (countable) partition, every $V \in \mathcal{F}_+$ must be of the form $V = \sum_n a_n \mathbf{1}_{H_n}$ (if it takes on an uncountable number of different values, $\mathcal{F}$ is not fine enough to resolve them). So, using the MCT twice,

$$\begin{aligned} \mathbb{E}[V\bar{X}] &= \mathbb{E}\left[\bar{X} \lim_n \sum_{i=1}^n a_i \mathbf{1}_{H_i}\right] = \lim_n \mathbb{E}\left[\bar{X} \sum_{i=1}^n a_i \mathbf{1}_{H_i}\right] \\ &= \lim_n \mathbb{E}\left[X \sum_{i=1}^n a_i \mathbf{1}_{H_i}\right] = \mathbb{E}\left[X \lim_n \sum_{i=1}^n a_i \mathbf{1}_{H_i}\right] = \mathbb{E}[VX] \, . \end{aligned}$$

These two properties will be used to define conditional expectation.

Definition. Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$. The **conditional expectation of X given $\mathcal{F}$**, denoted $\mathbb{E}[X \mid \mathcal{F}]$, is defined as follows:

i) For X in $\mathcal{H}_+$, $\mathbb{E}[X \mid \mathcal{F}]$ is any random variable $\bar{X}$ that satisfies

   a) *measurability*: $\bar{X}$ belongs to $\mathcal{F}_+$;

   b) *projection*: $\mathbb{E}[VX] = \mathbb{E}[V\bar{X}]$ for every V that belongs to $\mathcal{F}_+$.[13]

   Then we write $\mathbb{E}[X \mid \mathcal{F}] = \bar{X}$ and call $\bar{X}$ a **version** of $\mathbb{E}[X \mid \mathcal{F}]$.

ii) For arbitrary $X \in \mathcal{H}$, if $\mathbb{E}[X]$ exists, then we define

$$\mathbb{E}[X \mid \mathcal{F}] = \mathbb{E}[X^+ \mid \mathcal{F}] - \mathbb{E}[X^- \mid \mathcal{F}] \, .$$

   Otherwise, if $\mathbb{E}[X^+] = \mathbb{E}[X^-] = +\infty$, then $\mathbb{E}[X \mid \mathcal{F}]$ is left undefined.

Remark. Observe that for $X \in \mathcal{H}_+$, the projection property is equivalent to the condition that

$$\mathbb{E}[\mathbf{1}_H X] = \mathbb{E}[\mathbf{1}_H \bar{X}] \, , \quad H \in \mathcal{F} \, .$$

To see this, we can use the MCT on both sides of the equality to extend from indicators to simple functions, and then to arbitrary positive $V \in \mathcal{F}$.

Uniqueness and versions. If Y and Z are random variables belonging to $\mathcal{F}_+$, and if $\mathbb{E}[\mathbf{1}_H Y] = \mathbb{E}[\mathbf{1}_H Z]$ for every $H \in \mathcal{F}$, then $Y \overset{\text{a.s.}}{=} Z$. That is, Y and Z are *equivalent up to null sets*. To see this, let $H_{q,r}$ be of the form $\{Y < q < r < Z\}$, for some rational numbers $q < r$. Then

$$q\mathbb{P}(H_{q,r}) = \mathbb{E}[q\mathbf{1}_{H_{q,r}}] > \mathbb{E}[Y\mathbf{1}_{H_{q,r}}] = \mathbb{E}[Z\mathbf{1}_{H_{q,r}}] > \mathbb{E}[r\mathbf{1}_{H_{q,r}}] = r\mathbb{P}(H_{q,r}) \, ,$$

but this can only be true if $\mathbb{P}(H_{q,r}) = 0$ for all $q < r \in \mathbb{Q}_+$, and likewise for the reverse ordering of $Z < q < r < Y$. Therefore, $Y \overset{\text{a.s.}}{=} Z$. We use this fact to establish uniqueness of conditional expectation.

Proposition 11.1. *Let X be a random variable belonging to $\mathcal{H}_+$. Then the conditional expectation defined above is unique up to almost-sure equivalence.*

Proof. Consider two versions of $\mathbb{E}[X \mid \mathcal{F}]$ for $X \geq 0$, $\bar{X}$ and $\bar{X}'$. Then both versions belong to $\mathcal{F}_+$ and $\mathbb{E}[VX] = \mathbb{E}[V\bar{X}] = \mathbb{E}[V\bar{X}']$ for all $V \in \mathcal{F}_+$. Therefore, $\bar{X} \overset{\text{a.s.}}{=} \bar{X}'$.

Conversely, if $\mathbb{E}[X \mid \mathcal{F}] = \bar{X}$ and if $\bar{X}' \in \mathcal{F}_+$ and $\bar{X}' \overset{\text{a.s.}}{=} \bar{X}$, then $\bar{X}'$ satisfies the measurability and projection properties in the definition and hence is a version of $\mathbb{E}[X \mid \mathcal{F}]$. □

The uniqueness extends to arbitrary X for which $\mathbb{E}[X]$ exists via an integrability argument (see Çinlar, IV.1.6f).

[13] "Projection" is intentional: an alternative definition of conditional expectation is via projection in Hilbert spaces. See Çinlar [Çin11].

We see that "the conditional expectation" actually refers to an equivalence class of random variables (versions), but for probabilistic purposes, their almost-sure equivalence is strong enough to justify the definite article. In light of this, some authors may say $\mathbb{E}[X \mid \mathcal{F}] = \bar{X}$ almost surely.

Integrability. With $V = 1$, if $X \in \mathcal{H}_+$, then $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{F}]]$. Therefore, if X is integrable then so is $\mathbb{E}[X \mid \mathcal{F}]$.

For general X belonging to $\mathcal{H}$, the previous argument applies separately to X^+ and X^-. Hence, if X is integrable then so is $\mathbb{E}[X \mid \mathcal{F}] = \bar{X}$, and the projection property can be expressed as

$$\mathbb{E}[V(X - \bar{X})] = 0 \quad \text{for every bounded } V \in \mathcal{F}.$$

Intuitively, the random variable $(X - \bar{X})$ is "orthogonal" to the subspace of $\mathcal{F}$-measurable functions, so that the inner product with $V \in \mathcal{F}$ is zero: $\mathbb{E}[V(X - \bar{X})] = 0$.

Existence. The following proof uses the Radon–Nikodym derivative to show the existence of conditional expectations. A different proof relying on projections in Hilbert spaces can be found in Çinlar (and elsewhere).

Theorem 11.2. *Let $X \in \mathcal{H}_+$. Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$. Then $\mathbb{E}[X \mid \mathcal{F}]$ exists and is unique up to almost sure equivalence.*

Proof. For each event H in $\mathcal{F}$, define

$$P(H) = \mathbb{P}(H) \quad \text{and} \quad Q(H) = \int_H \mathbb{P}(d\omega) X(\omega) \,. \tag{11.2}$$

That is, P is the restriction of $\mathbb{P}$ to $\mathcal{F}$, and Q is obtained by averaging X over the sets in $\mathcal{F}$. On the measurable space $(\Omega, \mathcal{F})$, P is a probability measure and Q is a measure that is absolutely continuous with respect to P. Hence, by Theorem 8.3 (Radon–Nikodym), there exists $\bar{X}$ belonging to $\mathcal{F}_+$ such that for every $H \in \mathcal{F}$,

$$\begin{aligned}
\mathbb{E}[X\mathbf{1}_H] &= \int_\Omega \mathbb{P}(d\omega) X(\omega) \mathbf{1}_H(\omega) && \text{(definition)} \\
&= \int_\Omega Q(d\omega) \mathbf{1}_H(\omega) && \text{(Equation (11.2))} \\
&= \int_\Omega P(d\omega) \bar{X}(\omega) \mathbf{1}_H(\omega) && \text{(Radon–Nikodym theorem)} \\
&= \int_\Omega \mathbb{P}(d\omega) \bar{X}(\omega) \mathbf{1}_H(\omega) = \mathbb{E}[\bar{X}\mathbf{1}_H] && \text{(Equation (11.2)) .}
\end{aligned}$$

Recall (from the remark above) that this is equivalent to $\mathbb{E}[XV] = \mathbb{E}[\bar{X}V]$ for all V belonging to $\mathcal{F}_+$. This shows that $\bar{X}$ is a version of $\mathbb{E}[X \mid \mathcal{F}]$. Uniqueness was established in Proposition 11.1 above. □

Properties similar to expectation. For the most part, conditional expectations behave much like expectations, with the caveat that one should remember that $\mathbb{E}[X \mid \mathcal{F}]$ is a random variable, so statements must be modified with probabilistic qualifiers like "almost surely". See Çinlar [Çin11, Remark IV.1.9] for a detailed explanation.

Proposition 11.3. *Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$, $X, Y, (X_n)_{n\geq 1}, (Y_n)_{n\geq 1}$ be $\bar{\mathbb{R}}$-valued random variables, and a, b, c be $\mathbb{R}$-valued constants. Furthermore, assume that all of the conditional expectations in the following claims exist. Then the following properties hold almost surely:*

- ***Monotonicity**: $X \leq Y \Rightarrow \mathbb{E}[X \mid \mathcal{F}] \leq \mathbb{E}[Y \mid \mathcal{F}]$.*
- ***Linearity**: $\mathbb{E}[aX + bY + c \mid \mathcal{F}] = a\mathbb{E}[X \mid \mathcal{F}] + b\mathbb{E}[Y \mid \mathcal{F}] + c$.*
- ***Monotone Convergence Theorem**: $X_n \geq 0, X_n \nearrow X \Rightarrow \mathbb{E}[X_n \mid \mathcal{F}] \nearrow \mathbb{E}[X \mid \mathcal{F}]$.*
- ***Dominated Convergence Theorem**: $X_n \to X$, $|X_n| \leq Y$ with Y integrable $\Rightarrow \mathbb{E}[X_n \mid \mathcal{F}] \to \mathbb{E}[X \mid \mathcal{F}]$.*
- ***Jensen's inequality**: f convex $\Rightarrow \mathbb{E}[f(X) \mid \mathcal{F}] \geq f(\mathbb{E}[X \mid \mathcal{F}])$.*

The proof is left as an exercise.

Special properties. In addition to its expectation-like properties, the conditional expectation has two special properties that capture how it behaves with more or less information.

The first special property is **conditional determinism**: if a random variable W is determined by $\mathcal{F}$, i.e., it is $\mathcal{F}$-measurable, then it should be treated as a deterministic number in the presence of the information contained in $\mathcal{F}$.

The second special property is **repeated conditioning**. Let $\mathcal{F} \subset \mathcal{G}$ be two sub-σ-algebras of $\mathcal{H}$. At a high level, $\mathcal{F}$ contains less information than $\mathcal{G}$. Çinlar says to "think of $\mathcal{F}$ as the information a fool has, and $\mathcal{G}$ as that a genius has: the genius cannot improve on the fool's estimate, but the fool has no difficulty worsening the genius's. In repeated conditioning, the fool wins all the time."

Theorem 11.4. *Let $\mathcal{F}$ and $\mathcal{G}$ be sub-σ-algebras of $\mathcal{H}$. Let W and X be random variables such that $\mathbb{E}[X]$ and $\mathbb{E}[WX]$ exist. Then the following hold:*

*a) **Conditional determinism**: $W \in \mathcal{F} \Rightarrow \mathbb{E}[WX \mid \mathcal{F}] = W\mathbb{E}[X \mid \mathcal{F}]$.*

*b) **Repeated conditioning**: $\mathcal{F} \subset \mathcal{G} \Rightarrow \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{F}] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{F}] \mid \mathcal{G}] = \mathbb{E}[X \mid \mathcal{F}]$.*

We give the proofs for positive W and X; the general case follows from the usual arguments.

Exercise 48 (Proof of conditional determinism)**:**

Prove Theorem 11.4 part a) for positive W and X.

Solution:

Proof. a) Suppose $X \in \mathcal{H}_+$ and $W \in \mathcal{F}_+$. Then $\bar{X} = \mathbb{E}[X \mid \mathcal{F}] \in \mathcal{F}_+$ and for every $V \in \mathcal{F}_+$,

$$\mathbb{E}[V \cdot (WX)] = \mathbb{E}[(VW) \cdot X] = \mathbb{E}[(VW) \cdot \bar{X}] = \mathbb{E}[V \cdot (W\bar{X})] \,,$$

by the projection property and the fact that $VW \in \mathcal{F}_+$. Hence, $W\bar{X} = W\mathbb{E}[X \mid \mathcal{F}]$ is a version of $\mathbb{E}[WX \mid \mathcal{F}]$. □

Proof. b) Since the random variable $\mathbb{E}[X \mid \mathcal{F}]$ belongs to $\mathcal{F}_+$ by definition, and $\mathcal{F} \subset \mathcal{G}$, we have that $\mathbb{E}[X \mid \mathcal{F}]$ belongs to $\mathcal{G}_+$. Therefore, by conditional determinism, $\mathbb{E}[\mathbb{E}[X \mid \mathcal{F}] \mid \mathcal{G}] = \mathbb{E}[X \mid \mathcal{F}]$, which is the second equality above.

For the first equality: Observe that $\bar{X} = \mathbb{E}[X \mid \mathcal{F}]$ belongs to $\mathcal{F}_+$. Now let $V \in \mathcal{F}_+$. By definition, $\mathbb{E}[VX] = \mathbb{E}[V\bar{X}]$. Because $V \in \mathcal{F}_+$, it also belongs to $\mathcal{G}_+$. If we define $Y = \mathbb{E}[X \mid \mathcal{G}]$, then $\mathbb{E}[VY] = \mathbb{E}[VX]$. But we also had $\mathbb{E}[VX] = \mathbb{E}[V\bar{X}]$, which shows that $\mathbb{E}[VY] = \mathbb{E}[V\bar{X}]$, and thus $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{F}] = \mathbb{E}[X \mid \mathcal{F}]$. To recap: for $V \in \mathcal{F}_+$, $\bar{X} = \mathbb{E}[X \mid \mathcal{F}]$, and $Y = \mathbb{E}[X \mid \mathcal{G}]$,

$$\begin{aligned} \mathbb{E}[VX] = \mathbb{E}[V\bar{X}] = \mathbb{E}[V\mathbb{E}[X \mid \mathcal{F}]] \qquad & \text{(projection property for } X \text{ onto } \mathcal{F}) \\ = \mathbb{E}[VY] = \mathbb{E}[V\mathbb{E}[X \mid \mathcal{G}]] \qquad & \text{(projection property for } X \text{ onto } \mathcal{G})\ . \end{aligned}$$

By the uniqueness of conditional expectation, this implies that $\mathbb{E}[X \mid \mathcal{F}]$ is a version of $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{F}]$. □

Conditional expectations given random variables. Let Y be a random variable taking values in some measurable space $(E, \mathcal{E})$. Recall from Theorem 9.2 that the σ-algebra generated by Y, σY, contains all $\bar{\mathbb{R}}$-valued random variables of the form $f \circ Y$, for some measurable f. For X belonging to $\mathcal{H}$, the conditional expectation given Y is defined to be $\mathbb{E}[X \mid \sigma Y]$. Similarly, for a collection of random variables $\{Y_t : t \in T\}$, the conditional expectation of X given the collection is $\mathbb{E}[X \mid \{Y_t : t \in T\}]$. Observe that these are really the same definition because we can always define $Y = (Y_t)_{t \in T}$.

The following result is an immediate consequence of these definitions and Theorem 9.2. It says that $\mathbb{E}[X \mid \sigma Y] = f(Y)$ for some measurable f.

Theorem 11.5. *Let X belong to $\mathcal{H}_+$, and let Y be a random variable taking values in some measurable space $(E, \mathcal{E})$. Then, every version of $\mathbb{E}[X \mid \sigma Y]$ has the form $f(Y)$ for some $\mathcal{E}$-measurable function $f : E \to \bar{\mathbb{R}}_+$. Conversely, $f(Y)$ is a version of $\mathbb{E}[X \mid \sigma Y]$ if and only if*

$$\mathbb{E}[f(Y)\, h(Y)] = \mathbb{E}[X \cdot h(Y)]\ , \quad \textit{for every } h \textit{ belonging to } \mathcal{E}_+. \tag{11.3}$$

Independence. If X and Y are independent, then conditioning on σY tells us nothing about X.

Proposition 11.6. *Suppose that X and Y are independent random variables taking values in $(D, \mathcal{D})$ and $(E, \mathcal{E})$, respectively. If f belongs to $\mathcal{D}_+$ and g to $\mathcal{E}_+$, then*

$$\mathbb{E}[f(X)\, g(Y) \mid \sigma Y] = g(Y)\mathbb{E}[f(X)]\ .$$

Exercise 49 (Independence in conditional expectation)**:**

Prove Proposition 11.6.

Hint: Check the defining properties.

Solution: First, note that $\mathbb{E}[f(X)]$ is a (constant) real number, say a, so $\mathbb{E}[f(X)\, g(Y) \mid \sigma Y] = a \cdot g(Y)$, which is clearly σY-measurable.

For projection, let V be a positive σY-measurable random variable, so that by Theorem 9.2, $V = h(Y)$ for some $h \in \mathcal{D}_+$. By the definition of independence,

$$\mathbb{E}[h(Y)\, g(Y)\, f(X)] = \mathbb{E}[h(Y)\, g(Y)]\mathbb{E}[f(X)] = \mathbb{E}[h(Y)\, (g(Y)\mathbb{E}[f(X)])] \,,$$

which satisfies the projection property.

11.2 Conditional probabilities and distributions

Conditional expectations are (almost) all we need to define conditional probabilities. For a sub-σ-algebra $\mathcal{F}$ of $\mathcal{H}$, for each event $H \in \mathcal{H}$,

$$\mathbb{P}[H \mid \mathcal{F}] = \mathbb{E}[\mathbf{1}_H \mid \mathcal{F}] \,, \tag{11.4}$$

is called the **conditional probability** of H given $\mathcal{F}$. In more elementary settings (i.e., undergraduate probability), conditional probability is defined in terms of events. For events G and H, the conditional probability of H given G is defined to be any number $\mathbb{P}(H \mid G) \in [0, 1]$ satisfying

$$\mathbb{P}(G \cap H) = \mathbb{P}(G)\mathbb{P}(H \mid G) \,. \tag{11.5}$$

When $\mathbb{P}(G) > 0$, this is unique.

However, a more general version will define conditional probability in terms of σ-algebras, allowing us to handle events with zero measure, such as conditioning on the event that a $\mathbb{R}$-valued random variable takes a particular value. For this, we need some (interesting!) technical tools.

12 Kernels and product spaces

Reading: Çinlar [Çin11], Chapter I.6

Supplemental:

We need a bit of mathematical background to study conditional probability. In particular, we need some tools for moving between measurable spaces.

Transition kernels. Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be two measurable spaces, and let K be a mapping from $E \times \mathcal{F}$ into $\bar{\mathbb{R}}_+$. Then K is called a **transition kernel** from $(E, \mathcal{E})$ into $(F, \mathcal{F})$ if

a) the mapping $x \mapsto K(x, B)$ is $\mathcal{E}$-measurable for every set $B \in \mathcal{F}$; and

b) the mapping $B \mapsto K(x, B)$ is a measure on $(F, \mathcal{F})$ for every $x \in E$.

If the mapping $B \mapsto K(x, B)$ is a *probability measure* on $(F, \mathcal{F})$ for every $x \in E$, then K is a **probability transition kernel**.

If you've ever studied Markov chains on discrete spaces, you've seen a particular (common) version of the latter. When $E = F = \{1, \dots, m\}$ is equipped with its discrete σ-algebras, the probability transition kernel is specified by the $K(x, \{y\})$. This is often denoted by the m-by-m matrix of positive numbers P, with

$$K(x, B) = \sum_{y \in B} K(x, \{y\}) = \sum_{y \in B} P_{x,y} \ , \quad B \subset \{1, \dots, m\} \ .$$

This special case (of discrete E, F) informs the choice of notation Kf and μK below.

For general E and F, it can be awkward/difficult to define K with the second argument taking sets $B \in \mathcal{F}$. A common way to overcome this difficulty is to obtain a transition kernel by integrating a function $k : E \times F \to \mathbb{R}_+$ against a finite measure ν on $(F, \mathcal{F})$:

$$K(x, B) = \int_B \nu(dy) k(x, y) \ , \quad x \in E, \ B \in \mathcal{F} \ . \tag{12.1}$$

Obtaining measures and functions from transition kernels. One of the most useful properties of transition kernels is that they can be used to obtain functions and measures from other functions and measures.

Theorem 12.1. *Let K be a transition kernel from $(E, \mathcal{E})$ into $(F, \mathcal{F})$. Then:*

i)

$$Kf(x) = \int_F K(x, dy) f(y) \ , \quad x \in E \ , \tag{12.2}$$

defines a function Kf that belongs to $\mathcal{E}_+$ for every function f belonging to $\mathcal{F}_+$.

ii)

$$\mu K(B) = \int_E \mu(dx) K(x, B) \ , \quad B \in \mathcal{F} \ , \tag{12.3}$$

defines a measure on $(F, \mathcal{F})$ for each measure μ on $(E, \mathcal{E})$.

iii)

$$(\mu K)f = \mu(Kf) = \int_E \mu(dx) \int_F K(x, dy) f(y) \tag{12.4}$$

for every measure μ on $(E, \mathcal{E})$ and function f belonging to $\mathcal{F}_+$.

Exercise 50 (Function)**:**

Let $f \in \mathcal{F}_+$. Since, for fixed x, $B \mapsto K(x, B)$ is a measure, $Kf(x)$ is the integral of f with respect to the measure $B \mapsto K(x, B)$. By varying x, this defines a function Kf.

Prove that Kf is in $\mathcal{E}_+$.

Proof sketch of ii) and iii). ii) and iii) are proven at the same time, by using the characterization theorem for the integral, Theorem 7.6. Define $L : \mathcal{F}_+ \to \bar{\mathbb{R}}_+$ by setting $L(f) = \mu(Kf)$. Clearly, if $f = 0$ then $L(f) = 0$. Linearity follows from the linearity of integration with respect to the measure $B \mapsto K(x, B)$ for each x and by the linearity of integration with respect to μ. Similarly, the MCT applied to the measures $B \mapsto K(x, B)$ and to μ characterize $L(f)$ as the integral νf, for some unique ν. Taking $f = \mathbf{1}_B$ shows that $\mu K = \nu$.

See Çinlar [Çin11, Theorem I.6.3] for the full proof. □

Products of kernels. We can construct a transition kernel from the product of two kernels. Specifically, let K be a transition kernel from $(E, \mathcal{E})$ into $(F, \mathcal{F})$, and L from $(F, \mathcal{F})$ into $(G, \mathcal{G})$. Then their **product transition kernel** KL from $(E, \mathcal{E})$ into $(G, \mathcal{G})$ is defined by

$$KL(x, B) = \int_F K(x, dy) L(y, B) \ , \quad x \in E, B \in \mathcal{G} \ .$$

An equivalent way of defining this (or any transition kernel) is by defining $(KL)f$ for every $f \in \mathcal{E}_+$. (We won't show this; see Çinlar [Çin11, Remark I.6.4].)

Markov kernels. A transition kernel from $(E, \mathcal{E})$ into $(E, \mathcal{E})$ is a **transition kernel** on $(E, \mathcal{E})$. Such a kernel K is a **Markov kernel** if $K(x, E) = 1$ for every x, and a **sub-Markov kernel** if $K(x, E) \leq 1$ for every x.

Transition kernels can be recursively multiplied to obtain a sequence of kernels,

$$K^0 = I \ , \quad K^1 = K \ , \quad K^2 = KK \ , \quad K^3 = KK^2 \ , \dots \ ,$$

where I is the identity kernel on $(E, \mathcal{E})$ defined by $I(x, A) = \mathbf{1}_A(x)$, $x \in E, A \in \mathcal{E}$.

If K is Markov, so is K^n for every integer $n \geq 0$.

Kernels finite and bounded. As with measures, there are various notions of "well-behavedness" relating to boundedness. For example, when viewed as a measure, a kernel K from $(E, \mathcal{E})$ into $(F, \mathcal{F})$ is said to be **finite** if $K(x, F) < \infty$ for each x, and **σ-finite** if the measure $B \mapsto K(x, B)$ is σ-finite for each x. When viewed as a function, K is said to be bounded if $x \mapsto K(x, F)$ is bounded. There are further definitions, but for the purposes of these lecture notes, we will assume (mostly

for convenience) that K is finite as a measure and bounded as a function. See Çinlar [Çin11, p. 40] for the details.

Functions on product spaces. The two most important results are Theorems 12.4 and 12.5 below. Their importance will become clear when we apply them in the context of conditioning. In order to prove them, we need two intermediate results, the first of which has to do with the measurability of **sections**. Specifically, let f be function defined on a product space, $f : E \times F \to G$. Then for a fixed $x \in E$, the mapping $f_x : F \to G,\ y \mapsto f(x, y)$ is the **section** of f at x; and likewise for $f_y : E \to G$.

As usual, we are particularly interested in $G \subset \mathbb{R}$ and $\mathcal{G} = \mathcal{B}(G)$, in which case we use our usual notation $f \in \mathcal{E} \otimes \mathcal{F}$ to say that f is $\mathcal{E} \otimes \mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable.

Lemma 12.2 (Measurable sections)**.** *Let $f \in \mathcal{E} \otimes \mathcal{F}$. Then the sections are measurable. That is, $x \mapsto f(x, y)$ belongs to $\mathcal{E}$ for each $y \in F$, and $y \mapsto f(x, y)$ belongs to $\mathcal{F}$ for each $x \in E$.*

Exercise 51 (Measurable sections)**:**

Prove Lemma 12.2. (See Çinlar [Çin11, Exercise I.2.2] for a hint.)

Note that the converse is not necessarily true: it is possible for each of the sections to be measurable, but f is not $\mathcal{E} \otimes \mathcal{F}$-measurable. Stronger conditions on the sections are needed; for example, left- or right-continuity of $x \mapsto f(x, y)$ for each $y \in F$. [See Çin11, Exercise I.6.28].

For functions defined on a product space, the following result generalizes the operation $K : \mathcal{F}^{\text{fn}} \to \mathcal{E}^{\text{fn}},\ f \mapsto Kf$ of Theorem 12.1.

Proposition 12.3. *Let K be a finite kernel from $(E, \mathcal{E})$ into $(F, \mathcal{F})$. Then, for every positive function f that belongs to $\mathcal{E} \otimes \mathcal{F}$,*

$$Tf(x) = \int_F K(x, dy) f(x, y)\ , \quad x \in E\ ,$$

defines a function Tf that belongs to $\mathcal{E}_+$. Moreover, the transformation $T : (\mathcal{E} \otimes \mathcal{F})_+^{\text{fn}} \to \mathcal{E}_+^{\text{fn}}$ is linear and continuous under increasing limits. That is,

a) $T(af + bg) = aTf + bTg$ for positive f and g in $(\mathcal{E} \otimes \mathcal{F})_+^{\text{fn}}$ and $a, b \in \mathbb{R}_+$; and

b) $Tf_n \nearrow Tf$ for every positive sequence $(f_n) \subset (\mathcal{E} \otimes \mathcal{F})_+^{\text{fn}}$ with $f_n \nearrow f$.

Some remarks. This theorem holds even if we relax the finiteness condition on K, but the proof becomes more involved. See Çinlar [Çin11, Proposition 6.9].

For the purposes of stating the proposition, I reverted back to the $(\mathcal{E} \otimes \mathcal{F})_+^{\text{fn}}$ notation, to avoid confusion about what T operates on. I'll switch back to $(\mathcal{E} \otimes \mathcal{F})_+$ now.

The proof highlights the properties of K as a measure for each $x \in E$.

Proof. Let f be a positive function belonging to $(\mathcal{E} \otimes \mathcal{F})_+$. By Lemma 12.2, for each $x \in E$ the section $f_x : y \mapsto f(x, y)$ belongs to $\mathcal{F}_+$. Therefore, $Tf(x)$ is the integral of f_x with respect to the measure $K_x : B \mapsto K(x, B)$. Thus, $Tf(x)$ is a well-defined number for each $x \in E$, and the linearity

property a) follows from the linearity of integration with respect to K_x for each x; the continuity under increasing limits property b) follows from the MCT for the measures K_x.

We assumed that K is finite, i.e., $K(x, F) < \infty$ for each $x \in E$, so clearly $x \mapsto K(x, F)$ is bounded. Boundedness of K implies that Tf is well-defined and bounded for each bounded $f \in \mathcal{E} \otimes \mathcal{F}$. $\mathcal{E}$-measurability follows from the usual indicator→simple function→positive function→measurable function chain of arguments. We will just show that functions of the form $T\mathbf{1}_{A\times B}$ for $A \in \mathcal{E}, B \in \mathcal{F}$ are $\mathcal{E}$-measurable. Specifically,

$$T\mathbf{1}_{A\times B}(x) = \int_F K(x, dy)\mathbf{1}_{A\times B}(x, y) = \mathbf{1}_A(x)K(x, B) \ ,$$

is a product of two $\mathcal{E}$-measurable functions ($A \in \mathcal{E}$ and $x \mapsto K(x, B)$ is $\mathcal{E}$-measurable by definition), so $T\mathbf{1}_{A\times B}(x)$ is $\mathcal{E}$-measurable.

Now, because of properties a) and b), similar identities can be shown for simple functions and increasing limits of simple functions (i.e., positive measurable functions). □

Measures on product spaces. For our purposes, this is the key construction, and something we've been working towards in the last few subsections. Specifically, we know from experience that specifying joint probability distributions with any non-trivial structure over more than a few random variables is very hard. In practice, probability models are often constructed one random quantity at a time, through a chain or hierarchy of conditional distributions. This mode of model-specification is ubiquitous in probabilistic modeling. The following result shows that this very general method is valid, in the sense that the resulting joint probability distribution is unique.

Theorem 12.4 (Marginal-conditional-joint construction)**.** *Let μ be a finite measure on $(E, \mathcal{E})$, and K a finite transition kernel from $(E, \mathcal{E})$ into $(F, \mathcal{F})$. Then π is the unique (finite) measure on the product space $(E \times F, \mathcal{E} \otimes \mathcal{F})$ satisfying,*

$$\pi(A \times B) = \int_A \mu(dx)K(x, B) \ , \quad A \in \mathcal{E}, B \in \mathcal{F} \ . \tag{12.5}$$

Moreover,

$$\pi f = \int_E \mu(dx) \int_F K(x, dy)f(x, y) \ , \quad f \in (\mathcal{E} \otimes \mathcal{F})_+ \ . \tag{12.6}$$

Proof. Clearly, π defined by (12.5) is a measure on the product space that inherits the finiteness of μ and K. To show that it is unique, assume that $\hat{\pi}$ is another measure satisfying (12.5). Then

$$\pi(A \times B) = \hat{\pi}(A \times B) \ , \quad A \in \mathcal{E}, B \in \mathcal{F} \ .$$

Thus, π and $\hat{\pi}$ agree on the set of measurable rectangles $A \times B$, which is closed under unions (hence, a p-system) and which generates $\mathcal{E} \otimes \mathcal{F}$. Thus, by Proposition 3.2, $\pi = \hat{\pi}$.

To prove (12.6), note that by Proposition 12.3 it can be written as

$$\pi f = \int_F \mu(dx)Tf(x) = \mu(Tf) \ ,$$

with Tf belonging to $\mathcal{E}_+$. Now Theorem 7.6 can be used, with $L(f) = \mu(Tf)$, to show that the equality holds. □

Note that this theorem holds under more general conditions on μ and K, but in statistics we are typically dealing with probability measures and probability transition kernels, so this is sufficient.

Product measures. In the special case that K has the special form $K(x, B) = \nu(B)$ for all $x \in E$, for some (finite) measure ν on $(F, \mathcal{F})$, then π is the **product measure** $\mu \times \nu$. In this case, the following result (Fubini's theorem) says that we can integrate in whichever order is convenient.

Theorem 12.5 (Fubini)**.** *Let μ and ν be finite measures on $(E, \mathcal{E})$ and $(F, \mathcal{F})$, respectively. There exists a unique finite measure π on $(E \times F, \mathcal{E} \otimes \mathcal{F})$ such that, for each $f \in (\mathcal{E} \otimes \mathcal{F})_+$,*

$$\pi f = \int_E \mu(dx) \int_F \nu(dy) f(x, y) = \int_F \nu(dy) \int_E \nu(dx) f(x, y) \ .$$

Proof. Let $h : E \times F \to F \times E$ be the transposition mapping $(x, y) \mapsto (y, x)$. This is clearly $\mathcal{E} \otimes \mathcal{F}/\mathcal{F} \otimes \mathcal{E}$-measurable. Now, for sets $A \in \mathcal{E}, B \in \mathcal{F}$,

$$\pi \circ h^{-1}(B \times A) = \pi(A \times B) = \mu(A)\nu(B) = \hat{\pi}(B \times A) \ ,$$

which implies that $\hat{\pi} = \pi \circ h^{-1}$ via Proposition 3.2, and that π is unique. Let $\hat{f}(y, x) = f(x, y)$. Then

$$\hat{\pi}\hat{f} = (\pi \circ h^{-1})\hat{f} = \pi(\hat{f} \circ h) = \pi f$$

since $\hat{f} \circ h(x, y) = \hat{f}(y, x) = f(x, y)$. □

13 Conditional probabilities and distributions

Reading: Çinlar [Çin11], Chapter IV.2-3

Supplemental:

Recall that before our detour into kernels and product spaces, we defined the conditional probability as

$$\mathbb{P}[H \mid \mathcal{F}](\omega) = \mathbb{E}[\mathbf{1}_H \mid \mathcal{F}](\omega) \;, \quad H \in \mathcal{H},\ \omega \in \Omega \;, \tag{13.1}$$

where $\mathcal{F}$ is a sub-σ-algebra of $\mathcal{H}$, and I am making explicit the dependence on ω. Now let $Q(H)$ be a version of $\mathbb{P}[H \mid \mathcal{F}]$ for each $H \in \mathcal{H}$, assuming that $Q(\emptyset) = 0$ and $Q(\Omega) = 1$. So-defined, $Q(H)$ is a random variable belonging to $\mathcal{F}$. We denote its value at $\omega \in \Omega$ by $Q_\omega(H)$.

The mapping $Q : (\omega, H) \mapsto Q_\omega(H)$ looks like a transition probability kernel from $(\Omega, \mathcal{F})$ into $(\Omega, \mathcal{H})$: the mapping $\omega \mapsto Q_\omega(H)$ is $\mathcal{F}$-measurable for each $H \in \mathcal{H}$ (by definition of conditional expectation); and by the monotone convergence property of Proposition 11.3,

$$Q_\omega(\cup_n H_n) = \sum_n Q_\omega(H_n) \;, \quad (H_n) \text{ disjointed in } \mathcal{H} \;, \tag{13.2}$$

for *almost* every ω. The almost is a limitation: we would need to pin down the null sets (and therefore the almost sure event Ω_h) for which (13.2) holds, but this generally depends on the sequence $h = (H_n)$. We might run into some serious technical difficulties in specifying the set of ω in Ω for which $H \mapsto Q_\omega(H)$ is a probability measure. This set is $\Omega_0 = \cap_h \Omega_h$, where the intersection is taken over all disjointed sequences $h \subset \mathcal{H}$. As Çinlar writes, "Ω_0 is generally a miserable object."

Despite these challenges, it is often possible to pick versions of $Q(H)$ such that $H \mapsto Q_\omega(H)$ is a probability measure for all $\omega \in \Omega$ (i.e., $\Omega_0 = \Omega$).

Regular versions. Let $Q(H)$ be a version of $\mathbb{P}[H \mid \mathcal{F}]$ for every $H \in \mathcal{H}$. Then $Q : (\omega, H) \mapsto Q_\omega(H)$ is said to be a **regular version** of the conditional probability $\mathbb{P}[\,\bullet \mid \mathcal{F}]$ if Q is a transition probability kernel from $(\Omega, \mathcal{F})$ into $(\Omega, \mathcal{H})$.

We also call Q a **regular conditional probability**. In the literature, these versions are studied and used almost exclusively. The main reason for their ubiquity is the following.

Proposition 13.1. *Suppose that $\mathbb{P}[\,\bullet \mid \mathcal{F}]$ has a regular version Q. Then*

$$QX : \omega \mapsto Q_\omega X = \int_\Omega Q_\omega(d\omega') X(\omega') \tag{13.3}$$

is a version of $\mathbb{E}[X \mid \mathcal{F}]$ for every random variable X whose expectation exists.

Proof. It is sufficient to prove this for positive X belonging to $\mathcal{H}$. By Theorem 12.1 applied to the transition kernel Q and the function X, we have that QX belongs to $\mathcal{F}_+$. We just need to check the projection property. That is, we need to show that for V belonging to $\mathcal{F}_+$,

$$\mathbb{E}[VX] = \mathbb{E}[V\ QX] \;.$$

Fix V. For $X = \mathbf{1}_H$, this follows from the definition of $Q(H)$ as a version of $\mathbb{E}[H \mid \mathcal{F}]$. This extends to simple and then arbitrary positive random variables by the linearity and monotone convergence properties of the operators $X \mapsto QX$ (Proposition 11.3) and $Z \mapsto \mathbb{E}[Z]$ (Section 7). □

Conditional distributions. Let Y be a random variable taking values in a measurable space $(E, \mathcal{E})$, and let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$. Then the **conditional distribution** of Y given $\mathcal{F}$ is *any* transition probability kernel $L : (\omega, B) \mapsto L_\omega(B)$ from $(\Omega, \mathcal{F})$ into $(E, \mathcal{E})$ such that

$$\mathbb{P}[Y \in B \mid \mathcal{F}](\omega) = L_\omega(B) \;, \quad \omega \in \Omega,\ B \in \mathcal{E} \;. \tag{13.4}$$

If $\mathbb{P}[\,\bullet \mid \mathcal{F}]$ has a regular version Q, then

$$L_\omega(B) = Q_\omega\{Y \in B\} \;, \quad \omega \in \Omega,\ B \in \mathcal{E} \;, \tag{13.5}$$

defines a version L of the conditional distribution of Y given $\mathcal{F}$. In general the problem is to find a regular version of $\mathbb{P}[\,\bullet \mid \mathcal{F}]$ restricted to σY. The following standard existence result does so under fairly general conditions, but requires some regularity conditions on $(E, \mathcal{E})$.

Theorem 13.2. *Let Y be a random variable taking values in $(E, \mathcal{E})$. If $(E, \mathcal{E})$ is a standard measurable space, then there exists a version of the conditional distribution of Y given $\mathcal{F}$ (13.4). Moreover, $\mathbb{P}[\,\bullet \mid \mathcal{F}]$ has a regular version.*

Some remarks. The standard measurable space is key: we need to be able to map to $\mathbb{R}$ and take advantage of the properties of distribution functions. Specifically, the proof is constructive and looks like a much more involved version of that of **??**. There are a number of technical details related to measurability that require careful attention, but the basic idea is very similar: construct the distribution function by mapping $(E, \mathcal{E})$ into $([0,1]), \mathcal{B}([0,1]))$, and then construct the kernel from the distribution function. See Çinlar [Çin11, Theorem IV.2.10].

The second claim, that $\mathbb{P}[\,\bullet \mid \mathcal{F}]$ has a regular version, follows from the first: define $Y(\omega) = \omega$ for all $\omega \in \Omega$. Then $(E, \mathcal{E}) = (\Omega, \mathcal{H})$ and the conditional distribution of Y given $\mathcal{F}$ is the regular version of $\mathbb{P}[\,\bullet \mid \mathcal{F}]$ given by (13.4).

Conditioning on random variables. In most applications, we are interested in conditioning on a random variable. As with conditional expectations, this means conditioning on the σ-algebra generated by the random variable.

The following result shows how to do so via a transition probability kernel K, giving proper meaning to the conditional distribution of Y given $X = x$ as,

$$\mathbb{P}[\,\bullet \mid X = x] = K(x, \bullet) \;. \tag{13.6}$$

Theorem 13.3. *Suppose X is a random variable taking values in a measurable space $(D, \mathcal{D})$ and Y is a random variable taking values in a standard measurable space $(E, \mathcal{E})$. Then there is a transition probability kernel K from $(D, \mathcal{D})$ into $(E, \mathcal{E})$ such that*

$$L_\omega(B) = K(X(\omega), B) \;, \quad B \in \mathcal{E} \;,$$

is a version of the conditional distribution of Y given $\mathcal{F} = \sigma X$.

Proof. Applying Theorem 13.2 with $\mathcal{F} = \sigma X$ implies the existence of a regular version

$$L_\omega(B) = Q_\omega\{Y \in B\} = \mathbb{P}[Y \in B \mid \mathcal{F}](\omega) = \mathbb{E}\big[\mathbf{1}_{\{Y\in B\}} \;\big|\; \mathcal{F}\big](\omega) \;, \quad \omega \in \Omega,\ B \in \mathcal{H} \;,$$

where the second equality is just the definition re-stated as a reminder. Furthermore, it makes clear that $L_\omega(B)$ belongs to $\mathcal{F}_+$ (by definition of conditional expectation). On the other hand, we know from Theorem 9.2 that a mapping $V : \Omega \to \bar{\mathbb{R}}$ belongs to σX if and only if $V = f \circ X$ for some $\mathcal{D}$-measurable $f : D \to \bar{\mathbb{R}}$. Therefore, we can define $K(X(\omega), B) = L_\omega(B)$ for each $B \in \mathcal{E}$ and check that K so defined is a transition probability kernel. By regularity of Q_ω, $B \mapsto K(X(\omega), B)$ is almost surely a probability measure on $(E, \mathcal{E})$, and by the previous arguments $\omega \mapsto K(X(\omega), B)$ belongs to σX. □

Disintegration. In the previous section, we constructed a joint measure π from a measure μ and a transition kernel K, with (12.6)

$$\pi f = \int_D \mu(dx) \int_E K(x, dy) f(x, y) \;, \quad f \in (\mathcal{D} \otimes \mathcal{E})_+ \;, \tag{13.7}$$

and π satisfying

$$\pi(A \times B) = \int_A \mu(dx) K(x, B) \;, A \in \mathcal{D},\ B \in \mathcal{E} \;. \tag{13.8}$$

A natural question is whether a measure π on $(D \times E, \mathcal{D} \otimes \mathcal{E})$ has a **disintegration** into components, in which case we would write (informally),

$$\pi(dx, dy) = \mu(dx) K(x, dy) \;, \quad x \in D,\ y \in E \;. \tag{13.9}$$

The probabilistic interpretation is as follows: if π is the joint distribution of X and Y, then $\mu(dx)$ is "the probability that X is in the small set dx centered at x" and $K(x, dy)$ is "the conditional probability that Y is in the small set dy, given that X is equal to x".

The disintegration of π is the converse of the construction of π in Theorem 12.4. The following result is the exact converse of that theorem, except that we also require here that $(E, \mathcal{E})$ be standard.

Theorem 13.4 (Disintegration)**.** *Let π be a probability measure on the product space $(D \times E, \mathcal{D} \otimes \mathcal{E})$, and suppose that $(E, \mathcal{E})$ is standard. Then there exist a probability measure μ on $(D, \mathcal{D})$ and a transition probability kernel K from $(D, \mathcal{D})$ into $(E, \mathcal{E})$ such that* (13.8) *holds.*

The theorem has some implications. One is that for every f belonging to $(\mathcal{D} \otimes \mathcal{E})_+$,

$$\mathbb{E}[f(X, Y) \mid \sigma X] = \int_E K(X, dy) f(X, y) \;, \tag{13.10}$$

which further implies that

$$\mathbb{E}[f(X, Y)] = \int_D \mu(dx) \int_E K(x, dy) f(x, y) \;. \tag{13.11}$$

Proof of Theorem 13.4. This can be cast as a special case of Theorem 13.2: Let $W = D \times E$, $\mathcal{W} = \mathcal{D} \otimes \mathcal{E}$, $\mathbb{P} = \pi$. On the probability space $(W, \mathcal{W}, \pi)$, define the random variable $X(w) = x$ and $Y(w) = y$, for $w = (x, y) \in W$. Let μ be the distribution of X: $\mu(A) = \pi(A \times E)$, $A \in \mathcal{D}$. Since Y takes values in a standard measurable space $(E, \mathcal{E})$, there is a regular version L of the conditional distribution of Y given $\mathcal{F} = \sigma X$.

Observe that $\mathcal{F}$ consists of measurable rectangles of the form $A \times E$, $A \in \mathcal{D}$, and therefore a random variable V belongs to $\mathcal{F}_+$ if and only if $V(w) = V(x, y) = v(x)$, for some mapping $v : D \to \bar{\mathbb{R}}_+$ belonging to $\mathcal{D}_+$.

Now, by the same argument we used in the proof of Theorem 13.4, this implies that $L_w(B) = K(X(w), B)$, where K is a transition probability kernel from $(D, \mathcal{D})$ into $(E, \mathcal{E})$. Therefore, using the projection property of $L_w(B)$, and writing $\mathbb{E}_\pi$ for expectation with respect to π,

$$\pi(A \times B) = \mathbb{E}_\pi[\mathbf{1}_A \circ X \ \mathbf{1}_B \circ Y] = \mathbb{E}_\pi[\mathbf{1}_A \circ X \ K(X, B)] = \int_D \mu(dx) \mathbf{1}_A(x) K(x, B) \ .$$

This establishes (13.8), and also (13.7) for $f = \mathbf{1}_{A \times B}$; it is extended to measurable f by the usual arguments. □

13.1 Conditional independence

This is an important generalization of independence (Section 9), which is recovered in the special case of conditioning on a trivial σ-algebra.

Let $\mathcal{F}, \mathcal{F}_1, \ldots, \mathcal{F}_n$ be sub-σ-algebras of $\mathcal{H}$. Then $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are said to be **conditionally independent** given $\mathcal{F}$ if, for all positive random variables $V_1, \ldots, V_n$ belonging to $\mathcal{F}_1, \ldots, \mathcal{F}_n$, respectively,

$$\mathbb{E}[V_1 \cdots V_n \mid \mathcal{F}] = \mathbb{E}[V_1 \mid \mathcal{F}] \cdots \mathbb{E}[V_n \mid \mathcal{F}] \ . \tag{13.12}$$

Recall that both sides of this equality are random variables, so there is an implicit "$\mathbb{P}$-almost surely" here.

Comparing the definition (13.12) to that of independence (9.2), it is apparent that the only difference is the substitution of $\mathbb{E}[\bullet \mid \mathcal{F}]$ for $\mathbb{E}[\bullet]$. All of the results about independence (i.e., Propositions 9.1, 9.3, 9.4 and 9.6) have conditionally independent counterparts. If $\mathcal{F}$ is trivial—that is, $\mathcal{F} = \{\emptyset, \Omega\}$ and we condition on what amounts to nothing—then conditional independence given $\mathcal{F}$ is the same as independence.

Our heuristic in Section 9 was that $\mathcal{F}_1$ is independent from $\mathcal{F}_2$ if information from $\mathcal{F}_1$ is useless for estimating random variables belonging to $\mathcal{F}_2$. An analogous heuristic can be used for conditional independence: given the information in $\mathcal{F}$, the further information in $\mathcal{F}_1$ is useless for estimating quantities belonging to $\mathcal{F}_2$.

For convenience, let $\mathcal{F}_1 \perp\!\!\!\perp_{\mathcal{F}} \mathcal{F}_2$ denote that $\mathcal{F}_1$ and $\mathcal{F}_2$ are conditionally independent given $\mathcal{F}$. Furthermore, conditioning on multiple σ-algebras means that we condition on the σ-algebra generated by the union, e.g.,

$$\mathbb{E}[X \mid \mathcal{F}, \mathcal{G}] = \mathbb{E}[X \mid \sigma(\mathcal{F} \cup \mathcal{G})] = \mathbb{E}[X \mid \mathcal{F} \vee \mathcal{G}] \ .$$

I will use the expression on the left-hand side for convenience.

Proposition 13.5. *The following are equivalent:*

a) $\mathcal{F}_1 \perp\!\!\!\perp_{\mathcal{F}} \mathcal{F}_2$.

b) $\mathbb{E}[V_2 \mid \mathcal{F}, \mathcal{F}_1] = \mathbb{E}[V_2 \mid \mathcal{F}]$ *for every positive* $V_2 \in \mathcal{F}_2$.

c) $\mathbb{E}[V_2 \mid \mathcal{F}, \mathcal{F}_1] \in \mathcal{F}$ *for every positive* $V_2 \in \mathcal{F}_2$.

Proof. Assume V, V_2, V_2 are positive and belong to $\mathcal{F}, \mathcal{F}_1, \mathcal{F}_2$, respectively. First, consider a), which is equivalent to (by conditional independence and then the conditional determinism property),

$$\mathbb{E}[V_1 V_2 \mid \mathcal{F}] = (\mathbb{E}[V_1 \mid \mathcal{F}])(\mathbb{E}[V_2 \mid \mathcal{F}]) = \mathbb{E}[(V_1 \mathbb{E}[V_2 \mid \mathcal{F}]) \mid \mathcal{F}] \; .$$

Therefore,

$$\mathbb{E}[V V_1 V_2] = \mathbb{E}[\mathbb{E}[V V_1 V_2 \mid \mathcal{F}]] = \mathbb{E}[V \mathbb{E}[V_1 V_2 \mid \mathcal{F}]] = \mathbb{E}[V \mathbb{E}[(V_1 \mathbb{E}[V_2 \mid \mathcal{F}]) \mid \mathcal{F}]] = \mathbb{E}[V V_1 \mathbb{E}[V_2 \mid \mathcal{F}]] \; .$$

On the other hand,

$$\mathbb{E}[V V_1 V_2] = \mathbb{E}[\mathbb{E}[V V_1 V_2 \mid \mathcal{F}, \mathcal{F}_1]] = \mathbb{E}[V V_1 \mathbb{E}[V_2 \mid \mathcal{F}, \mathcal{F}_1]] \; .$$

Therefore,

$$\mathbb{E}[V V_1 \mathbb{E}[V_2 \mid \mathcal{F}, \mathcal{F}_1]] = \mathbb{E}[V V_1 \mathbb{E}[V_2 \mid \mathcal{F}]] \; .$$

Now, random variables of the form $V V_1$ generate $\mathcal{F}, \mathcal{F}_1$ (i.e., all random variables in $\mathcal{F}, \mathcal{F}_1$ have that form for some $V \in \mathcal{F}$ and $V_1 \in \mathcal{F}_1$), so this shows a) $\iff$ b).

By the measurability property of conditional expectations, b) $\Rightarrow$ c). Conversely, if c) holds, then (by the conditional determinism implied by c) and then by repeated conditioning with $\mathcal{F} \subset (\mathcal{F} \vee \mathcal{F}_1)$),

$$\mathbb{E}[V_2 \mid \mathcal{F}, \mathcal{F}_1] = \mathbb{E}[\mathbb{E}[V_2 \mid \mathcal{F}, \mathcal{F}_1] \mid \mathcal{F}] = \mathbb{E}[V_2 \mid \mathcal{F}] \; ,$$

so c) $\Rightarrow$ b). □

I have found the following results useful in my research. The first follows from the previous proposition, via the correspondence between simple functions and measurable functions.

Proposition 13.6. *For any* σ*-algebras* $\mathcal{F}, \mathcal{F}_1, \mathcal{F}_2$*, we have* $\mathcal{F}_1 \perp\!\!\!\perp_{\mathcal{F}} \mathcal{F}_2$ *if and only if,*

$$\mathbb{P}[H \mid \mathcal{F}, \mathcal{F}_1] = \mathbb{P}[H \mid \mathcal{F}] \quad \textit{for each event } H \in \mathcal{F}_2 \; .$$

Corollary 13.7. *For any* σ*-algebras* $\mathcal{F}, \mathcal{F}_1, \mathcal{F}_2$*, we have* $\mathcal{F}_1 \perp\!\!\!\perp_{\mathcal{F}} \mathcal{F}_2$ *if and only if* $(\mathcal{F}, \mathcal{F}_1) \perp\!\!\!\perp_{\mathcal{F}} \mathcal{F}_2$.

Exercise 52 (Equivalence of conditional independence relationships)**:**

Prove Corollary 13.7.

Proposition 13.8 (Chain rule)**.**

$$\mathcal{F} \perp\!\!\!\perp_{\mathcal{G}} (\mathcal{F}_1, \mathcal{F}_2, \dots) \quad \textit{if and only if} \quad \mathcal{F} \perp\!\!\!\perp_{(\mathcal{G}, \mathcal{F}_1, \dots, \mathcal{F}_n)} \mathcal{F}_{n+1} \textit{ for } n \geq 0.$$

In particular,

$$\mathcal{F} \perp\!\!\!\perp_{\mathcal{G}} (\mathcal{F}_1, \mathcal{F}_2) \quad \textit{if and only if} \quad \mathcal{F} \perp\!\!\!\perp_{\mathcal{G}} \mathcal{F}_1 \textit{ and } \mathcal{F} \perp\!\!\!\perp_{(\mathcal{G}, \mathcal{F}_1)} \mathcal{F}_2 \; . \tag{13.13}$$

Exercise 53 (Chain rule for conditional independence)**:**

Prove (13.13).

The previous results specified how to check for conditional independence via σ-algebras. An alternative is to check via an independent randomization. The proof relies on an application of **??**.

Proposition 13.9. *Let X, Y, Z be random variables taking values in standard measurable spaces $(E, \mathcal{E}), (F, \mathcal{F}), (G, \mathcal{G})$, respectively. Then $X \perp\!\!\!\perp_Y Z$ if and only if $X \overset{\text{a.s.}}{=} f(Y, U)$ for some $(\mathcal{F} \otimes \mathcal{B}([0,1]))/\mathcal{E}$-measurable function $f : F \times [0,1] \to E$ and a uniform random variable U that is independent of Y and Z.*

13.2 Statistical sufficiency

[14]

Conditioning and conditional independence have many important applications in statistics (see the classic papers by Dawid [Daw79; Daw80]). Sufficiency is one of the most important. It continues to play an important rule in modern statistics and machine learning.

Recall that a statistical model for a random variable X taking values in a sample space $(E, \mathcal{E})$ is a family $\mathcal{P}$ of probability measures on $(E, \mathcal{E})$. A statistic **statistic** is a measurable function $S : E \to F$ into some measurable space $(F, \mathcal{F})$. (In practice, S is typically Polish and therefore standard.) A statistic S is called **sufficient** for a model $\mathcal{P}$ if all probability measures $P \in \mathcal{P}$ have the same conditional distribution of X given S. For standard $(E, \mathcal{E})$, we showed above that this means there is some transition probability kernel K from $(F, \mathcal{F})$ into $(E, \mathcal{E})$ such that

$$P[X \in \bullet \mid S = s] = K(s, \bullet) \,, \quad \text{for each } P \in \mathcal{P} \,. \tag{13.14}$$

Example 13.1 Sufficiency in coin-flipping. Suppose $X = (X_1, \ldots, X_n)$ is a sequence of n i.i.d. flips of a biased coin with probability p of heads, encoded as 1 for heads, 0 for tails. As you may know from an introductory statistics class, $S(X) = \sum_{i=1}^n X_i$ is a sufficient statistic. In this case, $K(s, \bullet)$ is the uniform distribution supported on all binary sequences of length n that have s 1's.

As we know, conditioning on S means conditioning on σS, and therefore any measurable function f of S is also a sufficient statistic because f belongs to σS. Hence, a sufficient statistic typically is not unique; let $\mathcal{S}_{\mathcal{P}}$ be the set of sufficient statistics for $\mathcal{P}$. There is a large literature from classical statistics on finding a **minimal sufficient statistic**, which is a statistic T such that for every sufficient statistic S there is some measurable function f such that $T = f \circ S$. Modulo some technical caveats, we see (via Theorem 9.2) that a minimal sufficient σ-algebra can be thought of as $\sigma T = \bigcap_{S \in \mathcal{S}_{\mathcal{P}}} \sigma S$.

Sufficient statistic-kernel pairs. It is helpful to think of sufficiency in terms of the pair (S, K). A statistic S may be sufficient for two different models $\mathcal{P}$ and $\mathcal{P}'$, but may fail to be sufficient for $\mathcal{P} \cup \mathcal{P}'$. If the kernel K is the same in both cases, then (S, K) is sufficient for the union. More

[14] This section borrows heavily from the lecture notes of Peter Orbanz, http://www.gatsby.ucl.ac.uk/~porbanz/teaching/G6106S16/NotesG6106S16_9May16.pdf.

importantly, if both S and K are specified, there is a uniquely defined set for which (S, K) is sufficient,

$$\mathcal{P}(S, K) := \{P : P[\bullet \mid S] = K(S, \bullet)\} .$$

The work of Lauritzen [Lau74], Diaconis and Freedman [DF84], and Diaconis [Dia88] explored the relationship between symmetry and sufficiency, and showed that $\mathcal{P}(S, K)$ is a convex set. Under suitable conditions, the extreme points of $\mathcal{P}(S, K)$ are those measures which are of the form $K(s, \bullet)$ for some $s \in F$. Therefore, every $P \in \mathcal{P}(S, K)$ has the representation

$$P(\bullet) = \int_F K(s, \bullet)\nu_P(ds) ,$$

for some probability measure ν_P on $(F, \mathcal{F})$. Representations of this type provide the foundations for much of Bayesian statistics. For example, the famous theorem of de Finetti is obtained as a special case.

13.3 Construction of probability spaces

Transition kernels can be used to prove the existence of basically all probability spaces that we encounter. The main results to that end are Ionescu-Tulcea's Theorem and Kolmogorov's Extention Theorem. The latter is commonly invoked to show the existence of stochastic processes. We won't study these results here, but Çinlar [Çin11, Section IV.4] covers them in detail. You're now equipped to understand the proofs and you should see them once in your life, so I encourage you to read that section on your own.

References

[Abb15] S. Abbott. *Understanding Analysis*. 2nd. Springer New York, 2015.

[AB06] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer-Verlag, 2006.

[Bas22] R. F. Bass. *Real analysis for graduate students*. Version 4.3. 2022. URL: https://bass.math.uconn.edu/real.html.

[Çin11] E. Çinlar. *Probability and Stochastics*. Springer New York, 2011.

[Daw79] A. P. Dawid. "Conditional Independence in Statistical Theory". In: *J. Royal Stat. Soc. Ser. B* 41.1 (1979), pp. 1–31.

[Daw80] A. P. Dawid. "Conditional Independence for Statistical Operations". In: *The Annals of Statistics* 8.3 (1980), pp. 598–617.

[Dia88] P. Diaconis. "Sufficiency as statistical symmetry". In: *Proceedings of the AMS Centennial Symposium*. Ed. by F. Browder. American Mathematical Society, 1988, pp. 15–26.

[DF84] P. Diaconis and D. Freedman. "Partial Exchangeability and Sufficiency". In: *Proc. Indian Stat. Inst. Golden Jubilee Int'l Conf. Stat.: Applications and New Directions*. Ed. by J. K. Ghosh and J. Roy. Indian Statistical Institute, 1984, pp. 205–236.

[Gut05] A. Gut. *Probability: A Graduate Course*. Springer New York, 2005.

[JP04] J. Jacod and P. Protter. *Probability Essentials*. 2nd. Springer-Verlag Berlin Heidelberg, 2004.

[Kal02] O. Kallenberg. *Foundations of Modern Probability*. 2nd. Springer-Verlag New York, 2002.

[Lau74] S. L. Lauritzen. "Sufficiency, Prediction and Extreme Models". In: *Scandinavian Journal of Statistics* 1.3 (1974), pp. 128–134.

[PC19] G. Peyré and M. Cuturi. "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* 11.5-6 (2019), pp. 355–607. eprint: arXIv1803.00567. URL: https://arxiv.org/pdf/1803.00567.pdf.

[San15] F. Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhäuser Cham, 2015.

[Sch05] R. L. Schilling. *Measures, Integrals and Martingales*. Cambridge University Press, 2005.

[SS05] E. M. Stein and R. Shakarchi. *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*. Princeton University Press, 2005.

Index