12.3 Martingale inequalities ................................................................. 98

A Additional results ........................................................................... 101
   A.1 An example of the use of Radon–Nikodym derivatives ............. 101
List of Exercises

1. Exercise (Limits of monotone decreasing sequences of sets) .................................. 9
2. Exercise (Countable unions and limits of monotone increasing sequences of sets) .... 10
3. Exercise (Countable intersections and limits of monotone decreasing sequences of sets) 10
4. Exercise (Intersection of σ-algebras) ................................................................. 10
5. Exercise (Trace spaces, Çınlar Ex. I.1.15) ......................................................... 13
6. Exercise (Proof of Lemma 3.1) ................................................................. 16
7. Exercise (Proof of Proposition 3.2) ................................................................. 16
8. Exercise (σ-algebra generated by a function) .................................................. 16
9. Exercise (Measurable functions of measurable functions are measurable) ........... 17
10. Exercise (Compositions under which the class of simple functions is closed) ......... 19
11. Exercise (Plotting the dyadic functions) ........................................................... 20
12. Exercise (Restrictions and traces, Çınlar Ex. I.3.11) ....................................... 25
13. Exercise (Extensions, Çınlar Ex. I.3.12) ........................................................... 25
14. Exercise (Image measure) ................................................................. 27
15. Exercise ......................................................................................... 32
16. Exercise (Random variable from undergraduate probability) ......................... 35
17. Exercise (Random variable from your research interests) ......................... 36
18. Exercise (Insensitivity of integration) .............................................................. 42
19. Exercise (Proof of Theorem 7.9) ................................................................. 44
20. Exercise (Proof of Markov’s inequality) .......................................................... 50
21. Exercise (Chebyshev’s inequality) ............................................................... 50
22. Exercise (Generalized Markov’s inequality) ................................................. 50
23. Exercise (Independence of functions of independent random variables.) ........... 55
24. Exercise ......................................................................................... 56
25. Exercise ......................................................................................... 58
26. Exercise ......................................................................................... 58
27. Exercise (Independence in conditional expectation) ...................................... 61
28. Exercise (Measurable sections) ................................................................. 64
29. Exercise (Equivalence of conditional independence relationships) .................... 71
30. Exercise (Chain rule for conditional independence) ........................................ 72
31. Exercise (Almost sure version of the continuous mapping theorem) ............... 75
32. Exercise (Applications of Borel–Cantelli Lemma) ........................................ 76
33. Exercise (Convergence in probability and arithmetic) ..................................... 79
34. Exercise (Implications of L^p convergence) .................................................. 80
35. Exercise (Convex increasing functions of submartingales) .............................. 94
1 Motivation

Placeholder for these notes.
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 argument—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.
2 Measurable spaces

Reading: Çinlar, I.1.

Supplemental:

Overview. Recall from last time that defining a probability distribution on an uncountable space can lead to fundamental issues. Ideally, 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 ran into problems when trying to work in an uncountable space—in essence, there were too many points (uncountably many) that we required the probability to “measure” in a self-consistent way. 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.

For the sake of time, we’re skipping over some material that is quite important to measure theory, though we won’t need it later on. In particular, we won’t cover the sections “p-systems and d-systems” and “Monotone class theorem” Çinlar [Çin11]. I encourage you to read them on your own.

2.1 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 $A \supseteq E$.

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 “master” 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” as combining sets of outcomes via logical operations (e.g., “It’s raining today.” and “It rained yesterday.”). 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

$$\bigcap_{i \in I} A_i , \bigcup_{i \in I} A_i$$

(2.1)

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

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

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

(2.2)

(2.3)

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

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

---

1Thanks to Trevor Campbell for the analogy.
Example 2.1. Limits of monotone sequences of sets. Show that $A_n \nearrow \cup_m A_m$ for any monotone increasing sequence $(A_n)$.

Evaluating (2.2) in both cases yields the answer. In particular, consider $\limsup_n A_n$ term-by-term:

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

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

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

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, $A_n \nearrow \cup_m A_m$.

\textbf{Exercise 1} (Limits of monotone decreasing sequences of sets):

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

### 2.2 \(\sigma\)-algebras

\textbf{Warning!} \(\sigma\)-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). A \(\sigma\)-algebra is the mathematical formalism of what we can possibly query.

(Most authors write “\(\sigma\)-algebra”, but you may see “sigma-alebra”. You may see both within the same text (as in Çınlar) including Greek letters in \LaTeX section headings causes problems for the hyperref package.\footnote{Aside: most of the time this can be dealt with by using the \texttt{texorpdfstring} command.}

\textbf{Definition.} A non-empty collection $\mathcal{E}$ of $E$ is called an \textbf{algebra} if it is closed under \textit{finite} unions and complements. $\mathcal{E}$ is a \textbf{\(\sigma\)-algebra} if it is closed under \textit{countable} unions and complements. In math:

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

Observe that a \(\sigma\)-algebra is also closed under countable intersections:

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

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 \( E \) (closed under complements).

This type of recursive closure suggests an algorithm to create a \( \sigma \)-algebra from any collection of sets.\(^3\)

Conversely (and somewhat less intuitively), for a given \( \sigma \)-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 \( \sigma \)-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 \( \sigma \)-algebra also contains the limits of all monotone sequences of its sets. In fact, this is how Doob [Doo94] defines \( \sigma \)-algebra.

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

**Exercise 2** (Countable unions and limits of monotone increasing sequences of sets):
Show that a collection of sets \( E \) 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 intersections and limits of monotone decreasing sequences of sets):
Show that a collection of sets \( E \) 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.

**Exercise 4** (Intersection of \( \sigma \)-algebras):

a) Let \( \mathcal{E}_1 \) and \( \mathcal{E}_2 \) be \( \sigma \)-algebras on \( E \). Show that \( \mathcal{E}_1 \cap \mathcal{E}_2 \) is a valid \( \sigma \)-algebra on \( E \).

b) Let \( (\mathcal{E}_n)_{n \geq 1} \) be a (countable) collection of \( \sigma \)-algebras on \( E \). Show that their intersection is a valid \( \sigma \)-algebra on \( E \).

### 2.3 Examples of \( \sigma \)-algebras

**Trivial \( \sigma \)-algebra.** Every \( \sigma \)-algebra on \( E \) contains at least \( \emptyset \) and \( E \). \( E = \{\emptyset, E\} \) is called the **trivial** \( \sigma \)-algebra.

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

\(^3\)This is what Rohlin had in mind.
**Generated σ-algebra.** Fix an arbitrary collection \( C \) of subsets of \( E \), and consider all the σ-algebras that contain \( C \). At the very least, \( 2^E \) contains \( C \). Take the intersection of all of those σ-algebras and call the result the σ-algebra **generated** by \( C \), denoted \( \sigma C \) or \( \sigma(\mathcal{C}) \).

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

Observe (convince yourself) that every member of \( \sigma C \) can be obtained by at most a countable number of set operations (complement, union, intersection) applied to at most a countable number of sets from \( C \).

σ-algebras generated by random variables are used extensively in probability.

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

**Proposition 2.1.** Let \( C \) and \( D \) be two collections of subsets of \( E \). Then:

(a) If \( C \subset D \) then \( \sigma C \subset \sigma D \).

(b) If \( C \subset \sigma D \) then \( \sigma C \subset \sigma D \).

(c) If \( C \subset \sigma D \) and \( D \subset \sigma C \), then \( \sigma C = \sigma D \).

(d) If \( C \subset D \subset \sigma C \), then \( \sigma C = \sigma D \).

**Proof.**

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

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

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

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

We will see an example of how to use these relationships in Example 2.2.

**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].

A **topology** \( \tau_S \) on a set \( S \) is a collection of subsets of \( S \) satisfying:

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

---

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.)
A non-empty set $S$ equipped with a topology $\tau_{S}$ is called a topological space, denoted $(S, \tau_{S})$. Members of $\tau_{S}$ 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 $\tau_{S}$.) The complement of an open set is a closed set.

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 $\epsilon$-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.

**Borel $\sigma$-algebra.** If $(E, \tau_{E})$ is a topological space then the $\sigma$-algebra generated by the collection of all open subsets of $E$ (i.e., $\tau_{E}$) is called the Borel $\sigma$-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 $\sigma$-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 $\sigma$-algebra) from them.
Example 2.2. Show that $\mathcal{B}(\mathbb{R})$ is generated by the collection of all open intervals.

Let $\mathcal{C}_I$ denote the collection of all open intervals of $\mathbb{R}$, and $\sigma \mathcal{C}_I$ the $\sigma$-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 $\sigma$-algebra (i.e., the Borel $\sigma$-algebra).

Each open interval is an open set, so $\mathcal{C}_I \subset \mathcal{O}$, which by Proposition 2.1 (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, i.e., $A \in \sigma \mathcal{O}$ for all open sets $A \in \mathcal{O}$. Therefore, $\mathcal{O} \subset \sigma \mathcal{C}_I$. Proposition 2.1 (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 $(x, \infty)$.

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

Example 2.3. 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.

2.4 Measurable spaces

A measurable space is a pair $(E, \mathcal{E})$, where $E$ is a set and $\mathcal{E}$ is a $\sigma$-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 $E$.

Exercise 5 (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 $\sigma$-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 $D$.
2.4.1 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

\[
A \times B = \{(x, y) : x \in A, y \in B\}.
\]

(2.5)

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

The product \(\sigma\)-algebra on \(E \times F\) is the collection of all measurable rectangles, 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.

\(^5\)A \times B\) is also called the Cartesian product of \(A\) and \(B\).
3 Measurable functions

Reading: Çinlar, I.2.

Supplemental:

Overview. With Section 2 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.

3.1 Measurability of functions

Let $E$ and $F$ be sets. A function or mapping $f$ 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, is also sometimes written as $f : x \mapsto x^2 + 5$ 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

$$f^{-1}B = \{x \in E : f(x) \in B\}.$$  \hfill (3.1)

Picture on board.

The inverse image is essentially a set operation that “plays well” with other set operations (i.e., commutes, distributes).

**Lemma 3.1.** Let $f$ be a mapping from $E$ into $F$. Then,

$$f^{-1}\emptyset = \emptyset, \quad f^{-1}F = E, \quad f^{-1}(B \setminus C) = (f^{-1}B) \setminus (f^{-1}C)$$  \quad (3.2)

$$f^{-1}\bigcup_i B_i = \bigcup_i f^{-1}B_i, \quad f^{-1}\bigcap_i B_i = \bigcap_i f^{-1}B_i,$$

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

**Exercise 6** (Proof of Lemma 3.1):

---

Be aware that different sub-fields of mathematics may or may not treat these as synonymous (Çinlar does). For example, this Math Stack Exchange thread. I will treat them as synonymous.
Prove Lemma 3.1.

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

As a further simplification, when both \(E\) and \(F\) 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 3.2.** 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 \(E\) and \(\mathcal{F}\) if and only if \(f^{-1}B \in \mathcal{E}\) for every \(B \in \mathcal{F}_0\).

**Exercise 7 (Proof of Proposition 3.2):**

Prove Proposition 3.2.

*Hint:* Use Lemma 3.1.

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

\[
f^{-1} \mathcal{F} = \{ f^{-1}B : B \in \mathcal{F} \} .
\]

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

**Exercise 8 (\(\sigma\)-algebra generated by a function):**

Show that \(f^{-1} \mathcal{F}\) as in Eq. (3.3) is a \(\sigma\)-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 \(\sigma\)-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.

### 3.2 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

\[
(g \circ f)(x) = g(f(x)), \quad x \in E.
\]

In statistics, machine learning, and related fields,\(^7\) we often compose functions and typically do so without any second thoughts about measurability. That’s because measurable functions of measurable functions are measurable.

**Proposition 3.3.** 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 9** (Measurable functions of measurable functions are measurable): Prove Proposition 3.3.

### 3.3 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 \(\bar{\mathbb{R}}\), 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.

Recall from Example 2.2 that various different collections of intervals generate \(\mathcal{B}(\mathbb{R})\). Recall also from Proposition 3.2 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 3.4.** A mapping \(f : E \to \bar{\mathbb{R}}\) is \(\mathcal{E}\)-measurable if and only if, for every \(r \in \mathbb{R}\), \(f^{-1}(-\infty, r] \in \mathcal{E}\).

**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

---

\(^7\)Artificial neural networks, i.e., deep learning, are iterated compositions of functions.
function whose value at $x$ is $f(x) \vee g(x)$. For a measurable space $(E, \mathcal{E})$ and a function $f : E \to \mathbb{R}$, 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. 1.2.9] because we can obtain many results for arbitrary $f$ from the corresponding results for positive functions.

### 3.4 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:
\[ \inf_n f_n, \quad \sup_n f_n, \quad \liminf_n f_n, \quad \limsup_n f_n. \tag{3.5} \]

Each of these are functions defined on $E$ pointwise: fix $x \in E$ and construct the sequence of numbers $(f_1(x), f_2(x), \ldots)$. Sequences of real numbers and taking the inf, sup, liminf, or limsup 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 result. 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 $(f_1(x), f_2(x), \ldots)$.

In general, lim inf is dominated by lim sup. 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, $f_1 \leq f_2 \leq \ldots$ (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 $f_n \searrow f$. Likewise, $f_n \nearrow f$ means that $(f_n)$ is decreasing and has limit $f$.

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

**Theorem 3.5.** Let $(f_n)$ be a sequence of $\mathcal{E}$-measurable numerical functions. Then each of the four functions in (3.5) 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 $r \in \mathbb{R}$,
\[ f^{-1}[-\infty, r] = \{x : f(x) \leq r\} = \cap_n \{x : f_n(x) \leq r\} = \cap_n f_n^{-1}[-\infty, r]. \tag{3.6} \]

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 3.4, $f = \sup_n f_n$ 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. \qed
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.

### 3.5 Simple functions and approximating measurable functions

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

$$1_A(x) = \begin{cases} 
1 & \text{if } x \in A \\
0 & \text{if } x \notin A.
\end{cases} \quad (3.7)$$

$1_A$ 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

$$f = \sum_{i=1}^{n} a_i 1_{A_i}, \quad (3.8)$$

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 $f = \sum_{i=1}^{m} b_i 1_{B_i}$. (Convince yourself that this is true.) This is called the **canonical form** of the simple function $f$.

**Measurability.** Proposition 3.4 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 10** (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 $d_n : \mathbb{R} \rightarrow \mathbb{R}$ as

$$d_n(r) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} 1_{\left(\frac{k-1}{2^n}, \frac{k}{2^n}\right)}(r) + n 1_{[n, \infty)}(r). \quad (3.9)$$
Draw on board for $n = 1, 2$.

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

**Exercise 11 (Plotting the dyadic functions):**

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

**Measurability.** Recall from Example 2.2 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 3.6.** For each $n \in \mathbb{N}$, each $d_n$ is an increasing right-continuous simple function on $\overline{\mathbb{R}}_+$, and $d_n(r)$ increases to $r$ (that is, $d_n(r) \nearrow r$) for each $r \in \overline{\mathbb{R}}_+$ 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 $n \to \infty$, $d_n \circ f \nearrow f$. If $f$ is $\mathcal{E}$-measurable, then so is $d_n \circ f$ for each $n \in \mathbb{N}$, by Proposition 3.3. 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.

**Theorem 3.7.** 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 $\sigma$-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: $f = f^+ - f^-$. This technique will be useful on problem 4 of the first assignment.

**Proof.** If a positive function $f$ is the limit of an increasing sequence of positive simple functions, then by Theorem 3.5, 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 (3.9) 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 3.3). 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 $x \in E$. 

20
3.6 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 \(\sigma\)-algebras.
- If \(E\) is a complete separable metric space (c.s.m.s.), 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\(^9\) 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.

**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:\(^{10}\)

- \(\{1, 2, \ldots, n\}\) and its discrete \(\sigma\)-algebra;
- \(\{1, 2, \ldots\}\) and its discrete \(\sigma\)-algebra;
- \([0, 1]\) and its Borel \(\sigma\)-algebra.

**Notation**

Çinlar [Çin11] uses \(\mathcal{E}\) to denote both the \(\sigma\)-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:

---

8*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), \((E, d)\).

9From Wikipedia: “Polish spaces are so named because they were first extensively studied by Polish topologists and logicians—Sierpiński, Kuratowski, Tarski and others”.

10This 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.
• \( \mathcal{M} \) 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, \( \mathcal{E}_+^{\text{fin}} \) is the class of positive \( \mathcal{E} \)-measurable numerical functions.
4 Measures, probability measures, and probability spaces

Reading: Çinlar, I.3.

Supplemental:

Overview. Sections 2 and 3 followed Çinlar [Çin11] pretty closely. This section will continue to do so, though we will introduce a few ideas from section II.1.

4.1 Measures and measure spaces

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

a) \textit{Zero on the empty set}: \(\mu(\emptyset) = 0\).

b) \textit{Countable additivity}: \(\mu(\bigcup_n A_n) = \sum_n \mu(A_n)\) for every disjointed sequence \((A_n)\) in \(\mathcal{E}\).

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

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

4.1.1 Probability measures and probability spaces

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

The fact that the total mass of \(\Omega\) 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 \(\Omega\) is called the \textit{sample space} and its elements \(\omega\) are called \textit{outcomes}. The \(\sigma\)-algebra \(\mathcal{H}\) is called the \textit{grand history} (maybe a bit less common these days) and its elements are called \textit{events}. \(\Omega\) is often described as containing all possible states of the world/universe, each corresponding to an element \(\omega\), and the random variables we work with are functions whose value changes depending on the state of the world. I never found this analogy helpful, but some people do.

4.2 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 4.1. Dirac measure. Let \(x\) be a fixed point of \(E\). For each \(A \in \mathcal{E}\), define

\[
\delta_x(A) = \begin{cases} 
1 & \text{if } x \in A \\
0 & \text{if } x \notin A
\end{cases} \quad (4.1)
\]

\(\delta_x\) is a measure on \((E, \mathcal{E})\) called the \textit{Dirac measure}.  

23
Example 4.2. 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). $\nu_D$ is a measure on $(E, \mathcal{E})$ called a counting measure. Often, $D$ is taken to be countable, in which case

$$\nu_D(A) = \sum_{x \in D} \delta_x(A), \quad A \in \mathcal{E}. \quad (4.2)$$

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

$$\bar{\nu}_D(A) = \sum_{x \in D} m(x) \delta_x(A), \quad A \in \mathcal{E}. \quad (4.3)$$

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

$$\hat{\nu}_D(A) = \frac{1}{|D|} \sum_{x \in D} \delta_x(A), \quad A \in \mathcal{E}, \quad (4.4)$$

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

Example 4.3. Lebesgue measure. A measure $\lambda$ 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”.

The notation $\lambda$ for Lebesgue measure is fairly standard, though Çinlar uses $\text{Leb}$. I will use $\lambda$.

4.3 Some properties of measures

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

Proposition 4.1. Measures and probability measures have the following properties:
<table>
<thead>
<tr>
<th>Property</th>
<th>Measure space $(E, \mathcal{E}, \mu)$</th>
<th>Probability space $(\Omega, \mathcal{H}, \mathbb{P})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norming:</td>
<td>$\mu(\emptyset) = 0$</td>
<td>$\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$</td>
</tr>
<tr>
<td>Countable &amp; finite additivity: ${H_n}$ disjointed $\Rightarrow$</td>
<td>$\mu(\bigcup_n H_n) = \sum_n \mu(H_n)$</td>
<td>$\mathbb{P}(\bigcup_n H_n) = \sum_n \mathbb{P}(H_n)$</td>
</tr>
<tr>
<td>Monotonicity: $H \subset K$ $\Rightarrow$</td>
<td>$\mu(H) \leq \mu(K)$</td>
<td>$\mathbb{P}(H) \leq \mathbb{P}(K)$</td>
</tr>
<tr>
<td>Sequential continuity $H_n \nearrow H$ $\Rightarrow$</td>
<td>$\mu(H_n) \nearrow \mu(H)$</td>
<td>$\mathbb{P}(H_n) \nearrow \mathbb{P}(H)$</td>
</tr>
<tr>
<td>$H_n \searrow H$ $\Rightarrow$</td>
<td>$\mu(\bigcup_n H_n) \leq \sum_n \mu(H_n)$</td>
<td>$\mathbb{P}(\bigcup_n H_n) \leq \sum_n \mathbb{P}(H_n)$</td>
</tr>
</tbody>
</table>

Boole’s inequality (a.k.a. union bound)

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

**Exercise 12** (Restrictions and traces, Çinlar Ex. I.3.11): Let $(E, \mathcal{E})$ be a measurable space and $\mu$ a measure on it. Fix $D \in \mathcal{E}$.

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

b) Let $\mathcal{D}$ be the trace of $\mathcal{E}$ on $D$ (recall Exercise 5). Define $\nu(A) = \mu(A)$ for $A \in \mathcal{D}$. Show that $\nu$ is a measure on $(D, \mathcal{D})$. It is called the restriction of $\mu$ to $D$.

**Exercise 13** (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 $\mu$ be a measure on $(D, \mathcal{D})$ and define $\nu$ by

$$\nu(A) = \mu(A \cap D), \quad A \in \mathcal{E}. \quad (4.5)$$

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

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

- If $c > 0$ and $\mu$ is a measure on $(E, \mathcal{E})$, then so is $c\mu$.
- If $\mu$ and $\nu$ 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.
Less straightforward is the following uniqueness of measures result. We won’t prove it, but it’s a useful result. See Çinlar [Çin11, Prop. 3.7] if you’re interested in the proof.

**Proposition 4.2.** Let \((E, \mathcal{E})\) be a measurable space. Let \(\mu\) and \(\nu\) be measures on it with \(\mu(E) = \nu(E) < \infty\). Let \(\mathcal{C}\) be a collection of sets that is closed under finite intersections, and such that \(\sigma\mathcal{C} = \mathcal{E}\). If \(\mu\) and \(\nu\) agree on all elements of \(\mathcal{C}\), then \(\mu\) and \(\nu\) are identical.

You may see the term p-system or \(\pi\)-system to describe a collection of sets that is closed under finite intersections. We will not concern ourselves with them, but it’s good to know the term.

**Corollary 4.3.** Let \(\mu\) and \(\nu\) 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 \(r \in \mathbb{R}\).

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 \(\mu\) be a measure on \((E, \mathcal{E})\). A point \(x\) is called an atom of \(\mu\) if \(\mu(\{x\}) > 0\). If \(\mu\) 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, \(\mu = \lambda + \nu\), where \(\lambda\) is a diffuse measure and \(\nu\) is purely atomic [see Çinlar [Çin11, Prop. 3.9]].

### 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 \((\Omega, \mathcal{H}, P)\), a measurable set \(B\) with \(P(B) = 0\) 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 \(\mu\) 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].)

### Almost everywhere/surely

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

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

### Image measures

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

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

(4.6)
This is well-defined because of the measurability of \( f \), and is called the image, or \textbf{image measure}, of \( \nu \) under \( f \). Other notation includes, variously, as \( f \circ \nu \), \( h(\nu) \), \( \nu \circ h \), and \( \nu_h \).

Another common name for the image measure is the \textbf{pushforward} measure.

\textbf{Exercise 14 (Image measure):}

Show that \( \mu = \nu \circ f^{-1} \) is a measure on \((F, \mathcal{F})\).
5 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.

From here on, we will be firmly in the realm of probability, with all of the randomness and uncertainty that comes with it. Keep in mind that measure theory is always in the background, even if we don’t appeal directly to measure theoretic results (though we often will). 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.

5.1 Probability spaces as models of reality

The inherent uncertainty 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, (Ω, ℱ, 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 ω ∈ Ω are called outcomes.
- A special collection of subsets, a σ-algebra, of Ω, denoted ℱ. The sets in ℱ are called events.\footnote{Other common notation for the σ-algebra on Ω is A or F.}
- A probability measure, ℙ, satisfying:
  - ℙ(Ω) = 1 (this implies that ℙ(∅) = 0).
  - For any countable disjoint collection (An) ∈ ℱ, ℙ(∪n An) = ∑n ℙ(An) (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:

\footnote{Çinlar [Çin11] says that ℱ 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.}
Ω₁ = \{(H,H), (H,T), (T,H), (T,T)\}.

• \(\mathcal{H}\) is all possible subsets of \(Ω₁\), \(2^{Ω₁}\). 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 \(Ω₁\).

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

Note that because \(Ω₁\) 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\(^{13}\) that, upon stopping, reports the fraction \(p ∈ (0, 1)\) of \(360°\) that it rotated (modulo \(360°\));

• a machine that could make biased coins such that the probability of coming up heads is equal to a user-specified \(p ∈ (0, 1)\).

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 \(Ω₂ = (0,1) × Ω₁\) and \(\mathcal{H} = \mathcal{B}((0,1)) ⊗ 2^{Ω₁}\). How might we construct a probability measure on the following \(\mathcal{H}\)-measurable events?

• \(\{0.1\} \cap E_{\text{tails}}\)

• \([0.1, 0.5] \cap E_{\text{tails}}\)

• \([0.1, 0.1 + \delta] \cap E_{\text{tails}}, \text{for any } 0 < \delta < 0.9\)

• \(E_{\text{tails}}\)

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 \(\sigma\)-algebra, \(\mathcal{B}((0,1)) ⊗ 2^{Ω₁}\) (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.

5.2 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 \((Ω, \mathcal{H})\),

\[
P = \{\mathbb{P}_θ : \mathcal{H} \to [0, 1], \ θ ∈ Θ\}\.
\]

Here, \(Θ\) is an index set for the model. If \(Θ\) is finite-dimensional (e.g., \(Θ = \mathbb{R}^d\)) 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 \(Θ_p × Θ_n\), where \(Θ_p\) has finite dimension and \(Θ_n\) does not.

\(^{13}\)Something like this.
Example 5.1. Normal family. Let $\mathcal{N}_{\theta, \sigma^2}, \theta \in \mathbb{R}$, denote the normal distribution with mean $\theta$ and variance $\sigma^2$. A statistical model might be

$$\mathcal{P} = \{\mathcal{N}_{\theta, 1} : \theta \in \mathbb{R}\}.$$ 

Clearly, this model is parametric.

Example 5.2. 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 $S_K$. The sequence might be infinite, in which case it lies in $S_\infty$.

Define the mixture of normals distribution as

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

A statistical model using this might be

$$\mathcal{P} = \{(p \circ \mathcal{N})_\theta : \theta \in \mathbb{R} \times \mathbb{R}_+ \times S_K\}.$$ 

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 inference. Frequentist inference 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; hence the central importance of parameter estimation in frequentist inference.

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.
6 Random variables

Reading: Çinlar [Çin11], II.1

Supplemental:

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

\[
X^{-1}A := \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{H} \quad A \in \mathcal{E}.
\]  

(6.1)

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 \(\mathcal{H}\): \(1_H\) 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.

6.1 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 \(X\),

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

(6.2)

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 \(\Omega\), and then determine the image measure \(\mathbb{P} \circ X^{-1}\). Rather, we define the distribution \(\mu_X\) 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 \(\mu_X\) being a probability measure.
Proposition 6.1. 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 \(\sigma\)-algebra on whatever space in which they take values. A common interpretation in such models is that \(\Omega\) 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
\[
\mathbb{P}\{\omega \in \Omega : X(\omega) = Y(\omega)\} = \mathbb{P}\{X = Y\} = 1.
\]
(6.3)

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

\(X\) and \(Y\) are said to be **equal in distribution** if
\[
\mathbb{P}\{X \in A\} = \mathbb{P}\{Y \in A\}, \quad A \in \mathcal{E},
\]
(6.4)
denoted \(X \overset{d}{=} Y\). Note that in general, this is a much weaker form of equality: almost-sure equality implies distributional equality.

**Distributions on \(\mathbb{R}\).** In light of the result Corollary 4.3 on the uniqueness of probability measures on \(\mathbb{R}\), 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**
\[
F(x) = \mu([-\infty, x]) = \mathbb{P}\{X \leq x\}, \quad x \in \mathbb{R}.
\]
(6.5)

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. You will be asked to prove some of their properties on Assignment 2.

**Exercise 15:**

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
\[
F(x) = \int_{-\infty}^{x} f(x) \lambda(dx),
\]
(6.6)
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); \(f = \frac{dF}{dx}\) 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 \(dx\).
6.2 Examples of random variables

**Example 6.1. Poisson distribution.** Let $X$ be a random variable taking values in $\mathbb{N} = \{0, 1, 2, \ldots\}$, with the discrete $\sigma$-algebra $2^\mathbb{N}$. $X$ has **Poisson distribution** with mean $c > 0$ is

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

(6.7)

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

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

(6.8)

**Example 6.2. 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

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

(6.9)

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

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

(6.10)

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 $\chi^2$-**distribution** with $n$ degrees of freedom.
Example 6.3. 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 $V$ of vertices and the set $E \subseteq V \times V$ 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 ($V = \{1, 2, \ldots, n\} := [n]$) and records a value in $\{0, 1\}$ for each possible edge in $[n] \times [n]^2$.

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 $\sigma$-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 $V(G_{n-1})$.
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 $A \subseteq [n] \times \{0, 1\}^{[n] \times [n]}$, 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.)

---

*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 6.4. Function of a random variable. Let \( X \) be a random variable taking values in a measurable space \((\mathcal{E}, \mathcal{E})\), and \((\mathcal{F}, \mathcal{F})\) another measurable space such that the mapping \( f : \mathcal{E} \rightarrow \mathcal{F} \) 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 \((\mathcal{F}, \mathcal{F})\). This follows from the measurability of the composition of measurable functions (Proposition 3.3). If \( \mu \) is the distribution of \( X \), then the distribution \( \nu \) 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 6.5. 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
\[
X \overset{d}{=} -\ln(U)/b .
\]
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 \( \text{rand()} \) 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 16 (Random variable from undergraduate probability):

Give the full definition (i.e., set of values, \( \sigma \)-algebra) of a random variable you studied in under-
graduate probability. State at least one distributional identity involving that random variable.

Exercise 17 (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.

6.3 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 $\sigma$-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 $\pi$ on the product space, called the **joint distribution** of $X$ and $Y$. In order to specify $\pi$ it is sufficient to specify

$$\pi(A \times B) = \mathbb{P}\{X \in A, Y \in B\} = \mathbb{P}\{\{X \in A\} \cap \{Y \in B\}\}, \quad A \in \mathcal{E}, \; B \in \mathcal{F}. \quad (6.11)$$

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

$$\mu(A) = \mathbb{P}\{X \in A\} = \pi(A \times F), \quad A \in \mathcal{E} \quad \text{and} \quad \nu(B) = \mathbb{P}\{Y \in B\} = \pi(E \times B), \quad B \in \mathcal{F}. \quad (6.12)$$

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:

$$\pi(A \times B) = \mathbb{P}\{X \in A, Y \in B\} = \mathbb{P}\{X \in A\}\mathbb{P}\{Y \in B\} = \mu(A)\nu(B), \quad A \in \mathcal{E}, \; B \in \mathcal{F}. \quad (6.13)$$

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.

6.4 Information and determinability

**Theorem 6.2.** Let $X$ be a random variable taking values in some measurable space $(E, \mathcal{E})$. A mapping $V : \Omega \to \mathbb{R}$ belongs to $\sigma X$ (is $\sigma X/\mathcal{B}(\mathbb{R})$-measurable) if and only if $V = f \circ X$ for some deterministic $\mathcal{E}/\mathcal{B}(\mathbb{R})$-measurable function $f$.

Question 4 of Assignment 1 was a slightly less general version of this. One interpretation is that $\sigma X$ is the collection of all $\mathbb{R}$-valued random variables that are measurable functions of $X$ (and $X$ only). One can think of $\sigma X$ as a body of information; measurability with respect to it is In this sense, $X$ **determines** all $\sigma X$-measurable random variables. A random variable $Y$ that is not $\sigma X$-measurable is not determined by $X$, i.e., there is no function $f$ such that $Y = f(X)$—there is leftover randomness. When we study conditional expectations, we will see that conditioning on $\sigma X$ yields a “best estimate” of $Y$ given the information in $\sigma X$. 36
7 Integration

Reading: Çinlar [Çin11], I.4

Supplemental:

7.1 Definition and desiderata

Let \((E, E, \mu)\) be a measure space. Recall that \(E_{\text{fin}}\) denotes the collection of all \(E/B(\mathbb{R})\)-measurable functions, and \(E_{\text{fin}}^+\) 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 \(\mu\) by (these are just different ways of writing the same thing)

\[
\mu f = \mu(f) = \int_E \mu(dx)f(x) = \int_E f \, d\mu.
\]

(7.1)

The notation \(\mu f\) suggests a kind of multiplication, and indeed integration behaves much like multiplication via the properties (which we will prove) for all \(a, b \in \mathbb{R}_+\) and \(f, g, f_n \in E_{\text{fin}}:\)

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 \(f_n \uparrow f\) then \(\mu f_n \uparrow \mu f\).

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 \(f = \sum_{i=1}^n a_i 1_{A_i}\). Then the integral is defined as

\[
\mu f = \sum_{i=1}^n a_i \mu(A_i).
\]

(7.2)

b) **Positive functions:** Let \(f \in E_{\text{fin}}\), and put \(f_n = d_n \circ f\) with \(d_n\) the dyadic function defined in (3.9) (with properties given in Lemma 3.6). Then each \(f_n\) is simple and positive, and \(f_n \uparrow f\). The integral \(\mu f_n\) is defined in part a) above; the sequence of numbers \(\mu f_n\) is increasing so the limit exists (we will establish this below). We define

\[
\mu f = \lim_n \mu f_n.
\]

(7.3)

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

\[
\mu f = \mu f^+ - \mu f^-.
\]

(7.4)

provided at least one of the terms on the RHS is finite. Otherwise, if both terms are infinite, we define \(\mu f\) 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.

Some observations and immediate implications for positive simple functions. Let \( f, g \) be simple and positive. Then the following are true:

- The formula (7.2) for \( \mu f \) still holds even if \( f \) is not in canonical form. (This is true due to the finite additivity of \( \mu \).)
- Linearity of the integral follows from the finite additivity of \( \mu \).
- The integral is monotonic: if \( f \leq g \) then \( \mu f \leq \mu g \). This follows from the linearity property applied to the functions \( f \) and \( g - f \):
  \[
  \mu f \leq \mu f + \mu(g - f) = \mu(f + g - f) = \mu g .
  \]
- By monotonicity, 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 \(+\infty\)).

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 \( f \)):

\[
(d_n \circ f)(x) = \sum_{k=1}^{2^n} \frac{1}{2^n} [\frac{k-1}{2^n}, \frac{k}{2^n}) (f(x)) + n1_{[n,\infty)}(f(x)) .
\] (7.5)

Applied to the sum, the general 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 \( \delta_{x_0} \). The integral as defined above yields \( \delta_{x_0} f = f(x_0) \) for every \( f \in E^\text{fn} \). 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 \( f \in F^\text{fn} \). 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 \( \mu \) is the restriction of the Lebesgue measure on \( \mathbb{R}^d \) to \( (E, \mathcal{E}) \). For \( f \in E^\text{fn} \), 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.

7.3 Basic properties of integration

Integrability. A function \( f \in E^\text{fn} \) is said to be integrable if \( \mu f \) 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 \( f \in E^\text{fn} \) and \( A \in \mathcal{E} \). Then \( f1_A \in E^\text{fn} \) and the integral of \( f \) over \( A \) is defined to be the integral of \( f1_A \), denoted

\[
\mu(f1_A) = \int_A \mu(dx) f(x) = \int_A f \ d\mu .
\]

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

\[
\mu(f1_A) + \mu(f1_B) = \mu(f1_C) .
\]
Proof. If $f$ is simple, the lemma follows from the linearity property established for simple functions:

$$
\mu(f1_A) + \mu(f1_B) = \mu(f1_A + f1_B) = \mu(f(1_A + 1_B)) = \mu(f1_{A\cup B}) = \mu(f1_C).
$$

For arbitrary $f \in \mathcal{E}_+^{fn}$, putting $f_n = d_n \circ f$, we get

$$
\mu(d_n \circ (f1_A)) + \mu(d_n \circ (f1_B)) = \mu(d_n \circ (f1_C))
$$

for all $n$ since the $f_n$ are simple. Observing that $f_n1_A = d_n \circ f(1_A)$ (convince yourself of this), and likewise for $B$ and $C$, we get

$$
\mu(d_n \circ (f1_A)) + \mu(d_n \circ (f1_B)) = \mu(d_n \circ (f1_C))
$$

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

Positivity and monotonicity.

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

**Proof.** Positivity of the integral for $f \in \mathcal{E}_+^{fn}$ 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. 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 $\mu f_n$ 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, $f \in \mathcal{E}_+^{fn}$ so $\mu f$ is well-defined. Since $(f_n)$ is increasing, Proposition 7.2 on monotonicity implies that the integrals $(\mu f_n)$ form an increasing sequence of numbers. Hence, $\lim_n \mu f_n$ exists. We want to show that it is equal to $\mu f$.

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 \( \mu \) (Proposition 4.1),

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

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, \( f_n \geq f_n \sum_{i=1}^m \mathbf{1}_{B_i} \), and therefore

\[
\begin{align*}
\lim_n \mu f_n & \geq \lim_n \mu(f_n \sum_{i=1}^m \mathbf{1}_{B_i}) \quad \text{(monotonicity)} \\
& = \lim_n \sum_{i=1}^m \mu(f_n \mathbf{1}_{B_i}) \quad \text{(finite additivity)} \\
& = \sum_{i=1}^m \lim_n \mu(f_n \mathbf{1}_{B_i}) \quad \text{(interchange limit with finite sum)} \\
& \geq \sum_{i=1}^m b_i \mu(B_i) \quad \text{(by (7.8))} \\
& = \mu g \quad \text{(definition of integral for simple functions)} .
\end{align*}
\]
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 $f, g \in \mathcal{E}_+^\infty$ and $a, b \in \mathbb{R}$,

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

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

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

Suppose that $f, g \in \mathcal{E}_+^\infty$ 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,

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

$$\mu (a \lim f_n + b \lim g_n) = a\mu (\lim f_n) + b\mu (\lim g_n)$$

$$\mu (af + bg) = a\mu f + b\mu g .$$

### 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 $\mu (f 1_A) = 0$ for every $f \in \mathcal{E}^\infty$.

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

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

**Exercise 18** (Insensitivity of integration):

Prove Proposition 7.5.
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 \(\mathcal{E}^\text{fin}_+\) into \(\bar{\mathbb{R}}^+\). Then there exists a unique measure \(\mu\) on \((E, \mathcal{E})\) such that \(L(f) = \mu f\) for every \(f \in \mathcal{E}^\text{fin}_+\) if and only if:

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

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

c) \((f_n) \subset \mathcal{E}^\text{fin}_+\) 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 two necessary intermediate results on interchanging lim inf/lim sup and integration.

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 \(\mu\) is finite, then we can take \(g = b\) in the DCT.
Theorem 7.8 (Bounded Convergence Theorem). Let \((f_n)\) be a sequence of \(E\)-measurable functions, bounded by \(b \in \mathbb{R}_+\). Suppose that \(\mu\) 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 4.6) that for a measurable function \(h : F \to E\) and a measure \(\nu\) on \((F, \mathcal{F})\), the image of \(\nu\) 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 \(\nu\) on \((F, \mathcal{F})\).

**Theorem 7.9.** For every \(f \in \mathcal{E}_+^\infty\), \((\nu \circ h^{-1}) f = \nu(f \circ h)\).

**Proof.** Sketch of proof: Define \(L : \mathcal{E}_+^\infty \to \mathbb{R}_+\) 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 \(\mu\) on \((E, \mathcal{E})\). Then, show that \(\mu\) and \(\nu \circ h^{-1}\) agree for all \(B \in \mathcal{E}\). \(\square\)

**Exercise 19 (Proof of Theorem 7.9):**

Complete the proof of Theorem 7.9.

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 \(\mu\) and \(\nu\) expressed in terms of the Lebesgue measure on \(\mathbb{R}^d\) and the Jacobian of the transformation \(h\).
Example 7.3. Beta-gamma magic. Let $X$ and $Y$ be independent random gamma variables (as in Example 6.2), 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,

\[ P(U \in du) = \beta_{a,b}(du) = du \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1}(1-u)^{b-1}, \quad 0 < u < 1. \quad (7.11) \]

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

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

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

45
8 Expectation

Reading: Çinlar [Çin11], I.4, II.2

Supplemental:

Random variables are measurable functions; expectations of random variables are integrals of measurable functions. Hence, we’ve already developed most of the necessary theory. 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

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

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 \(\mathbb{E}_{X\sim \mathbb{P}}[f(X)]\).

\(\mathbb{E}\) 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 \(\mathbb{E}X1_H\).

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 \(\mu\) is the distribution of \(X\), then

\[
\mathbb{E}f \circ X = \mu f
\]

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

**Proof.** The first statement is true because of what we know about integration with respect to image measures (Theorem 7.9): \(\mu = \mathbb{P} \circ X^{-1}\), so \(\mu f = \mathbb{P}(f \circ X) = \mathbb{E} f \circ X\) at least for \(f \in \mathcal{E}^\mu\). Conversely, if (8.2) holds for all \(f \in \mathcal{E}^\mu\), taking \(f = 1_A\) we see that

\[
\mu(A) = \mu 1_A = \mathbb{E} 1_A \circ X = \mathbb{P}\{X \in A\} ,
\]

which is the distribution of \(X\). \(\square\)

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

\[
\nu(A) = \mu(p1_A) = \int_A \mu(dx)p(x) , \quad A \in \mathcal{E} .
\] (8.3)

\(\nu\) is a measure on \((E, \mathcal{E})\). (Can you prove this?) It is called the indefinite integral of \(p\) with respect to \(\mu\). Like the image measure, we can choose to integrate with \(\mu\) or with \(\nu\).

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

The proof has the same structure as that of Theorem 7.9.

In more explicit notation, (8.3) is

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

Informally, we write

\[
\nu(dx) = \mu(dx)p(x) , \quad x \in E .
\] (8.5)

This aligns with ideas we may have taken for granted in introductory probability. Heuristically, \(\mu(dx)\) is the amount of mass \(\mu\) 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 \(\nu\) with respect to \(\mu\). The function \(p\) is called the density function of \(\nu\) relative to \(\mu\), with the following notation:

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

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

\[
p = \frac{\mathbb{P}(X \in dx)}{\mathbb{P}(X' \in dx)} , \quad x \in E .
\] (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 the same \(p\) 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 now, though we may prove it later in the course.

At a minimum, \(\mu\) and \(\nu\) need to agree on non-null sets. In particular, a measure \(\mu\) on \((E, \mathcal{E})\) is absolutely continuous with respect to \(\mu\) (also on \((E, \mathcal{E})\)) if, for every set \(A \in \mathcal{E}\),

\[
\mu(A) = 0 \Rightarrow \nu(A) = 0 .
\]
This is denoted $\nu \ll \mu$.

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

For the purposes of stating the theorem, we need the following definition. A measure $\mu$ on $(E, \mathcal{E})$ is $\sigma$-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 $\sigma$-finite.

In this course, we will only work with probability measures, which are finite by definition.

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

$$\int_E \nu(dx) f(x) = \int_E \mu(dx) p(x) f(x), \quad f \in \mathcal{E}_+^{fn}.$$  \hspace{1cm} (8.8)

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

We won’t prove this now, but we may prove it later in the course.

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

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

is a cumulative distribution function consisting purely of jumps. $F$ corresponds to some discrete probability measure $\mu$ 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 $\nu$ 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 $\mu$ or $F$ with respect to $\nu$. 
Example 8.2. c.d.f. and p.d.f. Let $F$ be a distribution function, and assume that $F$ is differentiable (in the same sense as in calculus). Then its derivative $f$ satisfies

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

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

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

$f$ is the probability density function of $\mu$ or $F$ with respect to the Lebesgue measure $\lambda$.

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

Proposition 8.4. Let $\mu$ be a $\sigma$-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 $f \in E_{\infty}$, 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

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

iii) If $\eta$ is a measure, $\nu$ is a $\sigma$-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 $\mu$ and $\nu$ 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 $\mu_1, \nu_1$ be $\sigma$-finite measures on $(E, \mathcal{E})$ and $\mu_2, \nu_2$ $\sigma$-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

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

See Appendix A.1 for an application.
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 $\mathbb{R}$, with distribution $\mu$. Then $E[X^n]$ is called the $n$th moment of $X$. For $n = 1$, $E[X]$ is the mean. Assuming the mean is finite, e.g., $E[X] = a < \infty$, the $n$th moment of $X - a$ is called the $n$th centered moment of $X$. A familiar example is the variance of $X$, $\text{Var}[X] = E[(X - a)^2]$.

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 $c > 0$,

$$P\{X > c\} \leq \frac{1}{c}E[X]. \quad (8.9)$$

**Exercise 20** (Proof of Markov’s inequality):

Prove Markov’s inequality (8.9).

*Hint:* Use the fact that $X \geq c1_{X>c}$.

**Exercise 21** (Chebyshev’s inequality):

Assume that $E[X]$ is finite. Apply Markov’s inequality to $(X - E[X])^2$ to show that

$$P\{|X - E[X]| > c\} \leq \frac{1}{c^2}\text{Var}[X]. \quad (8.10)$$

**Exercise 22** (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 $c \in \mathbb{R}$,

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

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

$$\varphi(E[X]) \leq E[\varphi(X)]. \quad (8.12)$$
The measure-theoretic version of this is (with $\mu$ the distribution of $X$)

$$\varphi\left(\int_{\mathbb{R}} x\mu(dx)\right) \leq \int_{\mathbb{R}} \varphi(x)\mu(dx).$$

**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 $\nu$ and $\mu$ is

$$\text{KL}(\nu||\mu) = \int_{E} \ln\left(\frac{d\nu}{d\mu}\right) d\nu. \quad (8.13)$$

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,

$$\text{KL}(\nu||\mu) = -\int_{E} \ln\left(\frac{d\mu}{d\nu}\right) d\nu \geq \ln\left(\int_{E} \frac{d\mu}{d\nu} d\nu\right) $$

$$= \ln\left(\int_{E} d\mu\right) = 0.$$

Hence, we have shown that when $\nu$ and $\mu$ are mutually absolutely continuous, $\text{KL}(\nu||\mu) \geq 0$. (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 $\mu$. Then for $r \in \mathbb{R}^+$, the random variable $e^{-rX}$ takes values in $[0, 1]$, and

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

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 $\mu$ (also of $X$). It can be shown that the Laplace transform determines the distribution: if $\mu$ and $\nu$ are distributions on $\mathbb{R}^+$, 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)

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

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 $\mu$, $e^{irX} = \cos rX + i \sin rX$ is a complex-valued random variable, and the notion of expectation extends
naturally:

$$\hat{\mu}_r = \mathbb{E}[e^{i r X}] = \int_{\mathbb{R}} e^{i r x} \mu(dx) = \int_{\mathbb{R}} \mu(dx) \cos rx + i \int_{\mathbb{R}} \mu(dx) \sin rx.$$  \hspace{1cm} (8.16)

The Fourier transform determines the distribution.

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

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

52
9 Independence

Reading: Çınlar [Çin11], II.5

Supplemental:

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 $\sigma$-algebras.

A $\sigma$-algebra $\mathcal{G}$ is a sub-$\sigma$-algebra of $\mathcal{H}$ if $\mathcal{G} \subset \mathcal{H}$. Çınlar advocates for thinking of the sub-$\sigma$-algebra $\mathcal{G}$ both as a collection of events, 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 $\sigma$-algebra generated by a function, defined in Section 3.1, (3.3)), because of the following theorem—a version of which you proved in Assignment 1.

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 $\sigma$-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 $\sigma X$, denoted $V \in \sigma X$, if it is $\sigma X/\mathcal{B}(\mathbb{R})$-measurable.

Recall from Theorem 6.2 that a mapping $V : \Omega \to \bar{\mathbb{R}}$ belongs to $\sigma X$ if and only if

$$V = f \circ X,$$

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

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

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

(9.1)

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

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

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

(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-$\sigma$-algebra $\mathcal{F}_t$ of $\mathcal{H}$ for each $t \in T$, the collection $\{\mathcal{F}_t : t \in T\}$ is an independency of 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: $\sigma X$ and $\sigma Y$ are independent.
The following test for independence of \( \sigma \)-algebras echoes the test for measurability of functions in Proposition 3.2: 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-\( \sigma \)-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

\[
P(H_1 \cap \cdots \cap H_n) = P(H_1) \cdots P(H_n),
\]

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

See Çinlar [Çin11, Prop. II.5.2] for a proof.

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 \( \{ T_1, T_2, \ldots \} \) 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 9.2.** Every partition of an independency is an independency.

### 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. 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

\[
E[f_1(X_1) \cdots f_n(X_n)] = E[f_1(X_1)] \cdots E[f_n(X_n)],
\]

for all \( f_1 \in \mathcal{E}^{\text{fin}}_{1,+}, \ldots, f_n \in \mathcal{E}^{\text{fin}}_{n,+} \).

**Proof.** We need to show that (9.2) holds for all positive \( V_i \in \sigma X_1, \ldots, V_n \in \sigma X_n \) if and only if (9.3) holds for all positive \( f_1 \in \mathcal{E}^{\text{fin}}_{1,+}, \ldots, f_n \in \mathcal{E}^{\text{fin}}_{n,+} \). This is immediate from Theorem 6.2: \( V_i \in \sigma X_i \) if and only if \( V_i = f_i(X_i) \) for some \( f_i \in \mathcal{E}^{\text{fin}}_i \). \( \square \)

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

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

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 23 (Independence of functions of independent random variables.): Prove Proposition 9.5.

Hint: Let \( Y_i = f_i(X_i) \), \( i = 1, \ldots, n \). Then \( \sigma Y_i \subset \sigma X_i \).

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 \( \mu \) and \( \nu \), respectively. Then, the distribution of \( (X,Y) \) is the product measure \( \mu \times \nu \), and the distribution of \( X + Y \), \( \mu \ast \nu \), is given by

\[
(\mu \ast \nu)f = \mathbb{E}[f(X + Y)] = \int_{\mathbb{R}} \mu(dx) \int_{\mathbb{R}} \nu(dy) f(x + y) .
\]  

(9.5)

The distribution \( \mu \ast \nu \) is called the convolution of \( \mu \) and \( \nu \). Of course, because \( X + Y = Y + X \), \( \mu \ast \nu = \nu \ast \mu \).

In many cases, there are easier ways to identify the distribution of \( X + Y \) than to compute (9.5). The following example gives one.

Example 9.1. 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 \( Z \approx X + Y \).

9.3 Tail fields and Kolmogorov’s 0-1 law

Let \( (\mathcal{G}_n) \) be a sequence of sub-\( \sigma \)-algebras of \( \mathcal{H} \). For the purposes of this section, we think of \( \mathcal{G}_n \) as the information revealed by the \( n \)th trial of an experiment. Then the information about the future after \( n \) is \( \mathcal{T}_n = \bigvee_{m>n} \mathcal{G}_m \). The \( \sigma \)-algebra consisting of events whose occurrences are unaffected by anything in finite time is the tail \( \sigma \)-algebra, \( \mathcal{T} = \bigcap_n \mathcal{T}_n \).
Example 9.2. Tail $\sigma$-algebra. Let $X_1, X_2, \ldots$ 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 $\omega$ 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 9.6 (Kolmogorov’s 0-1 law). Let $\mathcal{G}_1, \mathcal{G}_2, \ldots$ be independent. Then $\mathbb{P}(H)$ is either 0 or 1 for every event $H$ in the tail $\sigma$-algebra $\mathcal{T}$.

Proof. By Proposition 9.2 on partition independencies, $\{\mathcal{G}_1, \ldots, \mathcal{G}_n, \mathcal{T}_n\}$ is an independency for each $n$, which implies that so is $\{\mathcal{G}_1, \ldots, \mathcal{G}_n, \mathcal{T}\}$ for each $n$, since $\mathcal{T} \subset \mathcal{T}_n$. Thus, by definition $\{\mathcal{T}, \mathcal{G}_1, \mathcal{G}_2, \ldots\}$ is an independency, and so is $\{\mathcal{T}, \mathcal{T}_0\}$ by Proposition 9.2 (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-$\sigma$-algebra (for example, $\limsup_n \frac{1}{n} S_n$) is almost surely constant.

Corollary 9.7. Let $\mathcal{G}_1, \mathcal{G}_2, \ldots$ 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 $V \equiv c$.

Exercise 24:

Prove Corollary 9.7.
10 Conditioning and disintegration

**Reading:** Çinlar [Çin11], Chapter IV

**Supplemental:**

Conditioning is one of the most important concepts in modern probability, especially for applications in statistics and machine learning. Martingales and Markov processes cannot even be defined 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.

### 10.1 Conditional expectations

For two $\mathbb{R}$-valued random variables $X$ and $Y$, a heuristic notion of the conditional expectation $E[X \mid Y = y]$ 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-$\sigma$-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 $\mathcal{F}$-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 $H$,

$$
E[X \mid H] = \frac{1}{\mathbb{P}(H)} \int_H \mathbb{P}(d\omega)X(\omega) = \frac{1}{\mathbb{P}(H)} E[X1_H] .
$$

The quantity $E[X \mid H]$ 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., $\sigma$-algebras). Conditioning on a random variable $Y$ is really short-hand for conditioning on $\sigma Y$.

Now suppose that the sub-$\sigma$-algebra $\mathcal{F}$ is generated by a measurable partition $(H_n)$ of $\Omega$. Now for fixed $\omega$, consider our estimate of $X(\omega)$ given the information $\mathcal{F}$: we can tell which of the events $H_1, H_2, \ldots$ includes $\omega$, and therefore our estimate

$$
\hat{X}(\omega) = E[X \mid \mathcal{F}] = \sum_n E[X \mid H_n]1_{H_n}(\omega) .
$$

(10.1)

If we do this for each $\omega \in \Omega$, we have defined a random variable $\hat{X} : \Omega \to \mathbb{R}_+$. This is the conditional expectation of $X$ given $\mathcal{F}$.

Observe that because $\hat{X} = E[X \mid \mathcal{F}]$ 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, $\hat{X}$ is $\mathcal{F}$-measurable; it is determined by the information in $\mathcal{F}$. Second, $E[VX] = E[V\hat{X}]$ for every $V$ that belongs to $\mathcal{F}_+$. (These are worth checking.) These two properties will be used to define conditional expectation.
Exercise 25:
Show that \( \bar{X} \) in Eq. (10.1) belongs to \( \mathcal{F} \).

Exercise 26:
Show that for \( \bar{X} \) in Eq. (10.1), \( \mathbb{E}[VX] = \mathbb{E}[V \bar{X}] \) for every \( V \) that belongs to \( \mathcal{F}_+ \).

\[ \text{Hint: First, fix } n \text{ and let } V = \mathbf{1}_{H_n} \text{ and show the desired equality. Then, extend that to arbitrary } \]
\[ V \in \mathcal{F}_+ \text{ using the MCT, observing that for this particular } \mathcal{F}, \text{ all such } V = \sum_n a_n \mathbf{1}_{H_n}. \]

Definition. Let \( \mathcal{F} \) be a sub-\( \sigma \)-algebra of \( \mathcal{H} \). The conditional expectation of \( X \) given \( \mathcal{F} \), denoted \( \mathbb{E}[X | \mathcal{F}] \), is defined as follows:

i) For \( X \) in \( \mathcal{H}_+ \), \( \mathbb{E}[X | \mathcal{F}] \) 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 \( \mathcal{F}_+ \).\(^{14}\)

Then we write \( \mathbb{E}[X | \mathcal{F}] = \bar{X} \) and call \( \bar{X} \) a version of \( \mathbb{E}[X | \mathcal{F}] \).

ii) For arbitrary \( X \in \mathcal{H} \), if \( \mathbb{E}[X] \) exists, then we define

\[ \mathbb{E}[X | \mathcal{F}] = \mathbb{E}[X^+ | \mathcal{F}] - \mathbb{E}[X^- | \mathcal{F}] \, . \]

Otherwise, if \( \mathbb{E}[X^+] = \mathbb{E}[X^-] = +\infty \), then \( \mathbb{E}[X | \mathcal{F}] \) 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 \( Y \equiv Z \). 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, \( Y \equiv Z \). We use this fact to establish uniqueness of conditional expectation.

Proposition 10.1. Let \( X \) be a random variable belonging to \( \mathcal{H}_+ \). Then the conditional expectation defined above is unique up to almost-sure equivalence.

\(^{14}\)“Projection” is intentional: an alternative definition of conditional expectation is via projection in Hilbert spaces. See Çinlar [Çin11].
Proof. Consider two versions of $E[X \mid F]$ for $X \geq 0$, $\bar{X}$ and $\bar{X}'$. Then both versions belong to $F_+$ and $E[VX] = E[V\bar{X}] = E[V\bar{X}']$ for all $V \in F_+$. Therefore, $\bar{X} \equiv \bar{X}'$.

Conversely, if $E[X \mid F] = \bar{X}$ and if $\bar{X}' \in F_+$ and $\bar{X}' \equiv \bar{X}$, then $\bar{X}'$ satisfies the measurability and projection properties in the definition and hence is a version of $E[X \mid F]$.

The uniqueness extends to arbitrary $X$ for which $E[X]$ exists via an integrability argument (see Çinlar, IV.1.6f).

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 $E[X \mid F] = \bar{X}$ almost surely.

Integrability. With $V = 1$, if $X \in H_+$, then $E[X] = E[E[X \mid F]]$. Therefore, if $X$ is integrable then so is $E[X \mid F]$.

For general $X$ belonging to $H$, the previous argument applies separately to $X^+$ and $X^-$. Hence, if $X$ is integrable then so is $E[X \mid F] = \bar{X}$, and the projection property can be expressed as

$$E[V(X - \bar{X})] = 0 \quad \text{for every bounded } V \in F.$$ 

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 10.2.** Let $X \in H_+$. Let $F$ be a sub-$\sigma$-algebra of $H$. Then $E[X \mid F]$ exists and is unique up to equivalence.

**Proof.** For each event $H$ in $F$, define

$$P(H) = \mathbb{P}(H) \quad \text{and} \quad Q(H) = \int_H \mathbb{P}(d\omega)X(\omega). \quad (10.2)$$

That is, $P$ is the restriction of $\mathbb{P}$ to $F$, and $Q$ is obtained by averaging $X$ over the sets in $F$. On the measurable space $(\Omega, F)$, $P$ 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 $F_+$ such that for every $H \in F$,

$$E[X1_H] = \int_{\Omega} \mathbb{P}(d\omega)X(\omega)1_H(\omega) \quad (\text{definition})$$

$$= \int_{\Omega} Q(d\omega)1_H(\omega) \quad (\text{Equation (10.2)})$$

$$= \int_{\Omega} P(d\omega)\bar{X}(\omega)1_H(\omega) \quad (\text{Radon–Nikodym theorem})$$

$$= \int_{\Omega} \mathbb{P}(d\omega)\bar{X}(\omega)1_H(\omega) = E[\bar{X}1_H] \quad (\text{Equation (10.2)}) .$$

Recall (from the remark above) that this is equivalent to $E[1_H] = E[\bar{X}1_H]$ for all $V$ belonging to $F_+$. This shows that $\bar{X}$ is a version of $E[X \mid F]$. Uniqueness was established in Proposition 10.1 above. 

59
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 Çınlar [Çin11, Remark IV.1.9] for a detailed explanation.

**Proposition 10.3.** Let $\mathcal{F}$ be a sub-$\sigma$-algebra of $\mathcal{H}$, $X, Y, (X_n)_{n\geq 1}, (Y_n)_{n\geq 1}$ be $\mathbb{R}$-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:** $X_n \geq 0, X_n \not\in \mathcal{F} \Rightarrow \mathbb{E}[X_n \mid \mathcal{F}] \not\in \mathbb{E}[X \mid \mathcal{F}]$.
- **Dominated Convergence Theorem:** $X_n \rightarrow X, |X_n| \leq Y$ with $Y$ integrable $\Rightarrow \mathbb{E}[X_n \mid \mathcal{F}] \rightarrow \mathbb{E}[X \mid \mathcal{F}]$.
- **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-$\sigma$-algebras of $\mathcal{H}$. At a high level, $\mathcal{F}$ contains less information than $\mathcal{G}$. Çınlar 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 10.4.** Let $\mathcal{F}$ and $\mathcal{G}$ be sub-$\sigma$-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:** $\mathcal{F} \subset \mathcal{G} \Rightarrow \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{F}] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{F}] \mid \mathcal{G}] = \mathbb{E}[X \mid \mathcal{F}]$.

**Proof.** We give the proofs for positive $W$ and $X$; the general case follows from the usual arguments.

a) Suppose $X \in \mathcal{H}_+$ and $W \in \mathcal{F}_+$. Then $\tilde{X} = \mathbb{E}[X \mid \mathcal{F}] \in \mathcal{F}_+$ and

$$\mathbb{E}[V \cdot (WX)] = \mathbb{E}[(VW) \cdot X] = \mathbb{E}[(VW) \cdot \tilde{X}] = \mathbb{E}[V \cdot (W\tilde{X})],$$

for every $V \in \mathcal{F}_+$, by the projection property and the fact that $VW \in \mathcal{F}_+$. Hence, $W\tilde{X} = W\mathbb{E}[X \mid \mathcal{F}]$ is a version of $\mathbb{E}[WX \mid \mathcal{F}]$.

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.
Independence. If \( X \) and \( Y \) are independent random variables taking values in some measurable space \((\mathcal{H}, \mathcal{F})\), then conditioning on \( \sigma Y \) tells us nothing about \( X \).

**Proposition 10.6.** Suppose that \( X \) and \( Y \) are independent random variables taking values in \((\mathcal{D}, \mathcal{D})\) and \((\mathcal{E}, \mathcal{E})\), respectively. If \( f \) belongs to \( \mathcal{D}_+ \) and \( g \) to \( \mathcal{E}_+ \), then
\[
\mathbb{E}[f \circ X \mid g \circ Y \mid \sigma Y] = g \circ Y \mathbb{E}[f \circ X].
\]

**Exercise 27** (Independence in conditional expectation):
Prove Proposition 10.6.

### 10.2 Conditional probabilities and distributions

Conditional expectations are (almost) all we need to define conditional probabilities. For a sub-\( \sigma \)-algebra \( \mathcal{F} \) of \( \mathcal{H} \), for each event \( H \in \mathcal{H} \),
\[
\mathbb{P}[H \mid \mathcal{F}] = \mathbb{E}[1_H \mid \mathcal{F}], \tag{10.4}
\]
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
\[
\mathbb{P}(G \cap H) = \mathbb{P}(G)\mathbb{P}(H \mid G). \tag{10.5}
\]
When \( \mathbb{P}(G) > 0 \), this is unique.

However, a more general version will define conditional probability in terms of \( \sigma \)-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.
10.3 Interjection: kernels and product spaces

We need a bit of mathematical background to study conditional probability. In particular, we need some tools for moving between measurable spaces. This material comes from Çinlar [Çin11, Ch. I.6].

**Transition kernels.** Let \((E, \mathcal{E})\) and \((F, \mathcal{F})\) be two measurable spaces, and let \(K\) be a mapping from \(E \times F\) into \(\mathbb{R}_+\). 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 \(E = F = \{1, \ldots, m\}\) is equipped with its discrete \(\sigma\)-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

\[
K(x, B) = \sum_{y \in B} K(x, \{y\}) = \sum_{y \in B} P_{x,y}, \quad B \subset \{1, \ldots, m\}.
\]

This special case (of discrete \(E, F\)) informs the choice of notation \(Kf\) and \(\mu K\) 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 \(\nu\) on \((F, \mathcal{F})\):

\[
K(x, B) = \int_B \nu(dy)k(x, y), \quad x \in E, \quad B \in \mathcal{F}.
\]

**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 10.7.** Let \(K\) be a transition kernel from \((E, \mathcal{E})\) into \((F, \mathcal{F})\). Then:

i)

\[
Kf(x) = \int_F K(x, dy)f(y), \quad x \in E,
\]

defines a function \(Kf\) that belongs to \(\mathcal{E}_+\) for every function \(f\) belonging to \(\mathcal{F}_+\).

ii)

\[
\mu K(B) = \int_E \mu(dx)K(x, B), \quad B \in \mathcal{F},
\]

defines a measure on \((F, \mathcal{F})\) for each measure \(\mu\) on \((E, \mathcal{E})\).
iii)

\[ (\mu K)f = \mu(Kf) = \int_E \mu(dx) \int_F K(x, dy)f(y) \]  \hspace{1cm} (10.9)

for every measure \( \mu \) on \((E, \mathcal{E})\) and function \( f \) belonging to \( \mathcal{F}_+ \).

**Proof sketch.** The proof of i) relies on the fact that \( K(x, B) \) is a measure on \((F, \mathcal{F})\), following the usual simple function/limit of simple functions proof technique.

ii) and iii) are proven at the same time, by using the characterization theorem for the integral, Theorem 7.6. Define \( L : \mathcal{F}_+ \to \mathbb{R}_+ \) 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 \( \mu \). Similarly, the MCT applied to the measures \( B \mapsto K(x, B) \) and to \( \mu \) characterize \( L(f) \) as the integral \( \nu f \), for some unique \( \nu \). Taking \( f = 1_B \) shows that \( \mu K = \nu \).

See Çinlar [Çin11, Theorem I.6.3] for the full proof. \( \square \)

**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,

\[ K^0 = I, \quad K^1 = K, \quad K^2 = KK, \quad K^3 = KK^2, \ldots, \]

where \( I \) is the identity kernel on \((E, \mathcal{E})\) defined by \( I(x, A) = 1_A(x), \quad x \in E, A \in \mathcal{E} \).

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 \( \sigma \)-finite if the measure \( B \mapsto K(x, B) \) is \( \sigma \)-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.** Note: I skipped this section in lecture. The two most important results are Theorems 10.10 and 10.11 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 a 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 $G \subseteq \mathbb{R}$ and $G = B(G)$, in which case we use our usual notation $f \in \mathcal{E} \otimes \mathcal{F}$ to say that $f$ is $\mathcal{E} \otimes \mathcal{F}/B(\mathbb{R})$-measurable.

**Lemma 10.8 (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 28 (Measurable sections).**

Prove Lemma 10.8. (See Cinlar [Cin11, 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 Cin11, Exercise 1.6.28].

For functions defined on a product space, the following result generalizes the operation $K : \mathcal{F}^{\text{fin}} \to \mathcal{E}^{\text{fin}}$, $f \mapsto Kf$ of Theorem 10.7.

**Proposition 10.9.** 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 $T : (\mathcal{E} \otimes \mathcal{F})^{\text{fin}}_+ \to \mathcal{E}^{\text{fin}}_+$ is linear and continuous under increasing limits. That is,

a) $T(af + bg) = aTf + bTg$ for positive $f$ and $g$ in $(\mathcal{E} \otimes \mathcal{F})^{\text{fin}}_+$ and $a, b \in \mathbb{R}_+$; and

b) $Tf_n \nearrow Tf$ for every positive sequence $(f_n) \subseteq (\mathcal{E} \otimes \mathcal{F})^{\text{fin}}_+$ 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 Cinlar [Cin11, Proposition 6.9].

For the purposes of stating the proof, I reverted back to the $(\mathcal{E} \otimes \mathcal{F})^{\text{fin}}_+$ notation, to avoid confusion about what $T$ operates on. I’ll switch back to $(\mathcal{E} \otimes \mathcal{F})^{\text{fin}}_+$ 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 $(\mathcal{E} \otimes \mathcal{F})^{\text{fin}}_+$. By Lemma 10.8, 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 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 $T_{1_{A \times B}}$ for $A \in \mathcal{E}, B \in \mathcal{F}$ are $\mathcal{E}$-measurable. Specifically,

$$T_{1_{A \times B}}(x) = \int_{\mathcal{F}} K(x, dy) 1_{A \times B}(x, y) = 1_A(x) K(x, B),$$

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 $T_{1_{A \times B}}(x)$ 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. Note: I resumed here in lecture. For our purposes, this is the key construction, and something we’ve been working towards for awhile. 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 dominates Bayesian modeling. The following result shows that this very general method is valid.

**Theorem 10.10** (Marginal-conditional-joint construction). Let $\mu$ be a finite measure on $(E, \mathcal{E})$, and $K$ a finite transition kernel from $(E, \mathcal{E})$ into $(F, \mathcal{F})$. Then $\pi$ is the unique (finite) measure on the product space $(E \times F, \mathcal{E} \otimes \mathcal{F})$ satisfying,

$$\pi(A \times B) = \int_A \mu(dx) K(x, B), \quad A \in \mathcal{E}, B \in \mathcal{F}. \quad (10.10)$$

Moreover,

$$\pi f = \int_{E \otimes F} \mu(dx) \int_F K(x, dy) f(x, y), \quad f \in (\mathcal{E} \otimes \mathcal{F})_+. \quad (10.11)$$

**Proof.** Clearly, $\pi$ defined by (10.10) is a measure on the product space that inherits the finiteness of $\mu$ and $K$. To show that it is unique, assume that $\hat{\pi}$ is another measure satisfying (10.10). Then

$$\pi(A \times B) = \hat{\pi}(A \times B), \quad A \in \mathcal{E}, B \in \mathcal{F}.$$

Thus, $\pi$ and $\hat{\pi}$ agree on the set of measurable rectangles $A \times B$, which is closed under unions and which generates $\mathcal{E} \otimes \mathcal{F}$. Thus, by Proposition 4.2, $\pi = \hat{\pi}$.

To prove (10.11), note that by Proposition 10.9 it can be written as

$$\pi f = \int_{E \otimes F} \mu(dx) T f(x) = \mu(T f),$$

with $T f$ belonging to $\mathcal{E}_+$. Now Theorem 7.6 can be used, with $L(f) = \mu(T f)$, to show that the equality holds. □

Note that this theorem holds under more general conditions on $\mu$ 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 $\nu$ on $(F, \mathcal{F})$, then $\pi$ 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 10.11** (Fubini). Let $\mu$ and $\nu$ be finite measures on $(E, \mathcal{E})$ and $(F, \mathcal{F})$, respectively. There exists a finite measure $\pi$ on $(E \times F, \mathcal{E} \otimes \mathcal{F})$ such that, for each $f \in (\mathcal{E} \otimes \mathcal{F})^+$,

$$
\pi f = \int_{E} \mu(dx) \int_{F} \nu(dy) f(x, y) = \int_{F} \nu(dy) \int_{E} \mu(dx) f(x, y).
$$

**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 $h = \pi \circ h^{-1}$ via Proposition 4.2. Let $\hat{f}(y, x) = f(x, y)$. Then

$$
\hat{h} = (\pi \circ h^{-1}) \hat{f} = \pi(\hat{f} \circ h) = \pi f
$$

since $\hat{f} \circ h(x, y) = \hat{f}(y, x) = f(x, y)$.

Kernels and randomization. We know from our study of distribution functions that for a random variable $X$ with distribution function $F$ (and quantile function $Q$), we can generate a realization of $X$ with a uniform random variable $U$: $X = Q(U)$. We end this detour with a result that extends that construction to more general spaces in a way that can be used with conditional distributions.

**Proposition 10.12.** Let $K$ be a probability transition kernel from a measurable space $(E, \mathcal{E})$ into a standard measurable space $(F, \mathcal{F})$. Then there exists a family of $F$-valued random variables $(Y_x)$ with distribution $K(x, \bullet)$ for every $x \in E$. More precisely, there exists a measurable function $f : E \times [0, 1] \to F$ such that if $U$ is a Unif$[0, 1]$ random variable, then we may set $Y_x = f(x, U)$ for every $x \in E$.

**Proof.** First, assume that $F = [0, 1]$ and $\mathcal{F} = \mathcal{B}([0, 1])$. Define

$$
f(x, u) = \inf \{ r \in [0, 1] \cap \mathbb{Q} : K(x, [0, r]) > u \}, \quad x \in E, \; u \in [0, 1],
$$

and note that $f$ is $(\mathcal{E} \otimes \mathcal{B}([0, 1]))$-measurable because the set $\{(x, u) : K(x, [0, u]) < u\}$ is measurable for each $r$ (and by Theorem 3.5, so is their infimum). If $U \sim$ Unif$[0, 1]$, then setting $Y_x = f(x, Y)$, we have

$$
\mathbb{P}\{Y_x \leq r\} = \mathbb{P}\{f(x, U) \leq r\} = \mathbb{P}\{U \leq K(x, [0, r])\} = K(x, [0, r]), \quad r \in [0, 1] \cap \mathbb{Q}.
$$

The intervals $[0, r], \; r \in \mathbb{Q}$, generate $\mathcal{B}([0, 1])$ so $f(x, U)$ has distribution $K(x, \bullet)$, by Corollary 4.3. For the general case of a standard measurable space $(F, \mathcal{F})$, recall (see Section 3.6) that by definition there is an isomorphism $g$ from $F$ into some Borel subset $G \subset [0, 1]$. Now, we can construct a transition probability kernel from $(E, \mathcal{E})$ into $(G, \mathcal{G})$ by composing $K$ with $g$: $K(g^{-1}B)$.

Because $g$ is an isomorphism, it is straightforward to show that $K$ is a transition probability kernel from $(E, \mathcal{E})$ into $(G, \mathcal{G})$. Therefore, the previous part of the proof applies to guarantee the existence
of the random variables \((\hat{Y}_x)\) taking values in \(G\), with distribution \(\hat{K}(x, \cdot)\). Let \(h : G \to F\) be the functional inverse of \(g\) and define \(Y_x = h(\hat{Y}_x)\). Observe that
\[
P\{Y_x \in B\} = P\{\hat{Y}_x \in h^{-1}B\} = \hat{K}(x, h^{-1}B) = K(x, (g^{-1} \circ h^{-1})B) = K(x, B),
\]
which shows that \(Y_x\) has distribution \(K(x, \cdot)\) for each \(x \in E\). □

10.4 Conditional probabilities and distributions, continued

Recall that before our detour into kernels and product spaces, we defined the conditional probability as
\[
P[H \mid \mathcal{F}](\omega) = E[1_H \mid \mathcal{F}](\omega), \quad H \in \mathcal{H}, \ \omega \in \Omega, \quad (10.12)
\]
where \(\mathcal{F}\) is a sub-\(\sigma\)-algebra of \(\mathcal{H}\), and I am making explicit the dependence on \(\omega\). Now let \(Q(H)\) be a version of \(P[H \mid \mathcal{F}]\) 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 \(\omega \mapsto Q(H)\) is \(\mathcal{F}\)-measurable for each \(H \in \mathcal{H}\) (by definition of conditional expectation); and by the monotone convergence property of Proposition 10.3,
\[
Q(\omega)(\bigcup_n H_n) = \sum_n Q(\omega)(H_n), \quad (H_n) \text{ disjointed in } \mathcal{H}, \quad (10.13)
\]
for almost every \(\omega\). The almost is a limitation: we would need to pin down the null sets (and therefore the almost sure event \(\Omega_0\)) for which (10.13) holds, but this generally depends on the sequence \(h = (H_n)\). We might run into some serious technical difficulties in specifying the set of \(\omega\) in \(\Omega\) for which \(H \mapsto Q(\omega)(H)\) is a probability measure. This set is \(\Omega_0 = \cap_k \Omega_k\), where the intersection is taken over all disjointed sequences \(h \subset \mathcal{H}\). As Çinlar writes, “\(\Omega_0\) 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 \(P[H \mid \mathcal{F}]\) 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 \(P[\, \cdot \, \mid \mathcal{F}]\) 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 10.13.** Suppose that \(P[\, \cdot \, \mid \mathcal{F}]\) has a regular version \(Q\). Then
\[
QX : \omega \mapsto Q(\omega)(X) = \int_\Omega Q(\omega)(d\omega')X(\omega')
\]
(10.14)
is a version of \(E[X \mid \mathcal{F}]\) for every random variable \(X\) whose expectation exists.

**Proof.** It is sufficient to prove this for positive \(X\) belonging to \(\mathcal{H}\). By Theorem 10.7 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 \(V \in \mathcal{F}_+\),
\[
E[VX] = E[VQX].
\]

67
Fix $V$. For $X = 1_H$, this follows from the definition of $Q(H)$ as a version of $E[H \mid F]$. This extends to simple and then arbitrary positive random variables by the linearity and monotone convergence properties of the operators $X \mapsto QX$ (Proposition 10.3) and $Z \mapsto E[Z]$ (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-$\sigma$-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

$$P[Y \in B \mid \mathcal{F}](\omega) = L_\omega(B), \quad \omega \in \Omega, \ B \in \mathcal{E}. \quad (10.15)$$

If $P[\cdot \mid \mathcal{F}]$ has a regular version $Q$, then

$$L_\omega(B) = Q_\omega\{Y \in B\}, \quad \omega \in \Omega, \ B \in \mathcal{E}, \quad (10.16)$$

defines a version $L$ of the conditional distribution of $Y$ given $\mathcal{F}$. In general the problem is to find a regular version of $P[\cdot \mid \mathcal{F}]$ restricted to $\sigma Y$. The following standard existence result does so under fairly general conditions, but requires some regularity conditions on $(E, \mathcal{E})$.

**Theorem 10.14.** 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 $\mathcal{F}$ (10.15). Moreover, $P[\cdot \mid \mathcal{F}]$ 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 Proposition 10.12. 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 $([0,1]), B([0,1])$, and then construct the kernel from the distribution function. See Cinlar [Cin11, Theorem IV.2.10].

The second claim, that $P[\cdot \mid \mathcal{F}]$ 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 $P[\cdot \mid \mathcal{F}]$ given by (10.15).

**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 $\sigma$-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,

$$P[\cdot \mid X = x] = K(x, \cdot). \quad (10.17)$$

**Theorem 10.15.** 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 $\mathcal{F} = \sigma X$.

68
Proof. Applying Theorem 10.14 with $\mathcal{F} = \sigma X$ implies the existence of a regular version

$$L_\omega(B) = Q_\omega \{ Y \in B \} = P[Y \in B \mid \mathcal{F}](\omega) = \mathbb{E}[\mathbb{1}_{\{ Y \in B \}} \mid \mathcal{F}](\omega), \quad \omega \in \Omega, \ B \in \mathcal{H},$$

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 6.2 that a mapping $V : \Omega \to \bar{\mathbb{R}}$ belongs to $\sigma X$ 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 $Q_\omega$, $B \mapsto K(X(\omega), B)$ is almost surely a probability measure on $(E, \mathcal{E})$, and by the previous arguments $\omega \mapsto K(X(\omega), B)$ belongs to $\sigma X$.

Disintegration. In the previous section, we constructed a joint measure $\pi$ from a measure $\mu$ and a transition kernel $K$, with (10.11)

$$\pi f = \int_D \mu(dx) \int_E K(x, dy)f(x, y), \quad f \in (D \otimes \mathcal{E})^+, \quad (10.18)$$

and $\pi$ satisfying

$$\pi(A \times B) = \int_A \mu(dx)K(x, B), \quad A \in D, \ B \in \mathcal{E}. \quad (10.19)$$

A natural question is whether a measure $\pi$ on $(D \times E, D \otimes \mathcal{E})$ has a disintegration into components, in which case we would write (informally),

$$\pi(dx, dy) = \mu(dx)K(x, dy), \quad x \in D, \ y \in E. \quad (10.20)$$

The probabilistic interpretation is as follows: if $\pi$ 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 $\pi$ is the converse of the construction of $\pi$ in Theorem 10.10. The following result is the exact converse of that theorem, except that we also require here that $(E, \mathcal{E})$ be standard.

**Theorem 10.16** (Disintegration). Let $\pi$ be a probability measure on the product space $(D \times E, D \otimes \mathcal{E})$, and suppose that $(E, \mathcal{E})$ is standard. Then there exist a probability measure $\mu$ on $(D, D)$ and a transition probability kernel $K$ from $(D, D)$ into $(E, \mathcal{E})$ such that (10.20) holds.

The theorem has some implications. One is that for every $f$ belonging to $(D \otimes \mathcal{E})^+$,

$$\mathbb{E}[f(X, Y) \mid \sigma X] = \int_E K(X, dy)f(X, y), \quad (10.21)$$

which further implies that

$$\mathbb{E}[f(X, Y)] = \int_D \mu(dx) \int_E K(x, dy)f(x, y). \quad (10.22)$$

69
Proof of Theorem 10.16. This can be cast as a special case of Theorem 10.14: Let $W = D \times E$, $\mathcal{W} = D \otimes \mathcal{E}$, $\mathbb{P} = \pi$. On the probability space $(W, \mathcal{W}, \mathbb{P})$, define the random variable $X(w) = x$ and $Y(w) = y$, for $w = (x, y) \in W$. Let $\mu$ be the distribution of $X$: $\mu(A) = \pi(A \times E)$, $A \in D$. 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 $A \times E$, $A \in D$, and therefore a random variable $V$ belongs to $\mathcal{F}_+$ if and only if $V(w) = V(x, y) = v(x)$, for some mapping $v : D \rightarrow \mathbb{R}_+$. Therefore, using the projection property of $L_w(B)$, and writing $\mathbb{E}_\pi$ for expectation with respect to $\pi$,

$$
\pi(A \times B) = \mathbb{E}_\pi[1_A \circ X \cdot 1_B \circ Y] = \mathbb{E}_\pi[1_A \circ X \cdot K(X, B)] = \int_D \mu(dx)1_A(x)K(x, B).
$$

This establishes (10.20), and also (10.18) for $f = 1_{A \times B}$; it is extended to measurable $f$ by the usual arguments.

10.5 Conditional independence

This is an important generalization of independence (Section 9), which is recovered in the special case of conditioning on a trivial $\sigma$-algebra.

Let $\mathcal{F}, \mathcal{F}_1, \ldots, \mathcal{F}_n$ be sub-$\sigma$-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,

$$
\mathbb{E}[V_1 \cdots V_n \mid \mathcal{F}] = \mathbb{E}[V_1 \mid \mathcal{F}] \cdots \mathbb{E}[V_n \mid \mathcal{F}].
$$

(10.23)

Recall that both sides of this equality are random variables, so there is an implicit “$\mathbb{P}$-almost surely” here.

Comparing the definition (10.23) 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 to 9.4) 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 $\mathcal{F}_1 \perp_{\mathcal{F}_2} \mathcal{F}_2$ denote that $\mathcal{F}_1$ and $\mathcal{F}_2$ are conditionally independent given $\mathcal{F}$. Furthermore, conditioning on multiple $\sigma$-algebras means that we condition on the $\sigma$-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 10.17. The following are equivalent:

a) \( \mathcal{F}_1 \perp \mathcal{F}_2 \).

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,

\[
\mathbb{E}[VV_1 V_2] = \mathbb{E}[\mathbb{E}[VV_1 V_2 \mid \mathcal{F}]] = \mathbb{E}[\mathbb{E}[(V_1 \mathbb{E}[V_2 \mid \mathcal{F}]) \mid \mathcal{F}]] = \mathbb{E}[V_1 \mathbb{E}[V_2 \mid \mathcal{F}]].
\]

On the other hand,

\[
\mathbb{E}[VV_1 V_2] = \mathbb{E}[\mathbb{E}[VV_1 V_2 \mid \mathcal{F}, \mathcal{F}_1]] = \mathbb{E}[\mathbb{E}[V_1 \mathbb{E}[V_2 \mid \mathcal{F}, \mathcal{F}_1]]] = \mathbb{E}[V_1 \mathbb{E}[V_2 \mid \mathcal{F}, \mathcal{F}_1]].
\]

Therefore,

\[
\mathbb{E}[VV_1 \mathbb{E}[V_2 \mid \mathcal{F}, \mathcal{F}_1]] = \mathbb{E}[V_1 \mathbb{E}[V_2 \mid \mathcal{F}]].
\]

Now, random variables of the form \( VV_1 \) 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). \( \square \)

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 10.18. For any \( \sigma \)-algebras \( \mathcal{F}, \mathcal{F}_1, \mathcal{F}_2 \), we have \( \mathcal{F}_1 \perp \mathcal{F}_2 \) if and only if,

\[
\mathbb{P}[H \mid \mathcal{F}, \mathcal{F}_1] = \mathbb{P}[H \mid \mathcal{F}] \quad \text{for each event } H \in \mathcal{F}_2.
\]

Corollary 10.19. For any \( \sigma \)-algebras \( \mathcal{F}, \mathcal{F}_1, \mathcal{F}_2 \), we have \( \mathcal{F}_1 \perp \mathcal{F}_2 \) if and only if \( (\mathcal{F}, \mathcal{F}_1) \perp \mathcal{F}_2 \).

Exercise 29 (Equivalence of conditional independence relationships):

Prove Corollary 10.19.

Proposition 10.20 (Chain rule).

\( \mathcal{F} \perp \mathcal{G}(\mathcal{F}_1, \mathcal{F}_2, \ldots) \) if and only if \( \mathcal{F} \perp (\mathcal{G}, \mathcal{F}_1, \ldots, \mathcal{F}_n) \mathcal{F}_{n+1} \) for \( n \geq 0 \).

In particular,

\[
\mathcal{F} \perp \mathcal{G}(\mathcal{F}_1, \mathcal{F}_2) \quad \text{if and only if} \quad \mathcal{F} \perp \mathcal{G} \mathcal{F}_1 \quad \text{and} \quad \mathcal{F} \perp (\mathcal{G}, \mathcal{F}_1) \mathcal{F}_2. \tag{10.24}
\]
Exercise 30 (Chain rule for conditional independence):
Prove (10.24).

The previous results specified how to check for conditional independence via $\sigma$-algebras. An alternative is to check via an independent randomization. The proof relies on an application of Proposition 10.12.

Proposition 10.21. 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 $X \overset{a.s.}{=} f(Y, U)$ 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 $Z$.

10.6 Statistical sufficiency

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 is a measurable function $S : E \to S$ 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 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 $(E, \mathcal{E})$ into $(F, \mathcal{F})$ such that

$$P[X \in \cdot \mid S = s] = K(s, \cdot), \quad \text{for each } P \in \mathcal{P}.$$  (10.25)

Example 10.1. 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, $K(s, \cdot)$ 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 $\sigma S$, and therefore any measurable function $f$ of $S$ is also a sufficient statistic because $f$ belongs to $\sigma S$. Hence, a sufficient statistic typically is not unique; let $\mathcal{S}_P$ 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 6.2) that a minimal sufficient $\sigma$-algebra can be thought of as $\sigma T = \bigcap_{S \in \mathcal{S}_P} \sigma S$.

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

\footnote{This section borrows heavily from the lecture notes of Peter Orbanz, \url{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,

$$\mathcal{P}(S, K) := \{ P : P[\cdot | S] = K(S, \cdot) \} .$$

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 $K(s, \cdot)$ for some $s \in F$. Therefore, every $P \in \mathcal{P}(S, K)$ has the representation

$$P(\cdot) = \int_F K(s, \cdot) \nu_P(ds) ,$$

for some probability measure $\nu_P$ 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.

### 10.7 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 Extension 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.
11 Convergence

Reading: Çinlar [Çin11], III.2
Supplemental: Çinlar [Çin11], III.1

Proving convergence of sequences of random variables is a core area of probability, and is often applied to statistical problems. The aim of this section is to provide the basic concepts that we will use in our study of martingales and of Markov chains. Çinlar [Çin11, Chapter III] contains much more material, but I will streamline the ideas as much as possible.

There are three primary modes of convergence that we will study: almost sure convergence, convergence in probability, and convergence in distribution (also known as weak convergence). We will also briefly touch on $L^p$-convergence.

11.1 Convergence of real sequences

I will introduce notation and review a few basic ideas about convergence of real sequences. I won’t go over everything we need here—please read Çinlar [Çin11], III.1 (and see the same for proofs of all results in this subsection).

Let $(x_n)$ be a sequence of real numbers, indexed by $\mathbb{N}_+ = \{1, 2, \ldots\}$. Then
\[
\liminf_{n} x_n = \sup_{m} \inf_{n \geq m} x_n, \quad \limsup_{n} x_n = \inf_{m} \sup_{n \geq m} x_n,
\]
(11.1)
are well-defined numbers, possibly infinite. If they are equal to the same number $x$ then $(x_n)$ is said to have limit $x$, written $\lim_{n} x_n = x$ or $x_n \to x$. The sequence is said to be convergent if the limit exists and is a real (finite) number.

Characterization. Çinlar has some nice notation: for $\epsilon \in \mathbb{R}_+$, let $i_\epsilon$ be the indicator on the interval $(\epsilon, \infty)$,
\[
i_\epsilon(x) = 1_{(\epsilon, \infty)}(x) = \begin{cases} 
1 & x > \epsilon \\
0 & x \leq \epsilon
\end{cases}.
\]
(11.2)

Let $(x_n)$ be a sequence of numbers in $\mathbb{R}$. The sequence converges to $x$ if and only if the sequence $|x_n - x|$ converges to 0. When $x$ is known, it is typically easier to work with $|x_n - x|$ because of its positivity. The classical (real analysis) statement of convergence says that a sequence of positive numbers $(x_n)$ converges to 0 if and only if for every $\epsilon > 0$ there exists a $k$ such that $x_n \leq \epsilon$ for all $n \geq k$, i.e.

\[x_n \to 0 \iff \sum_{n} i_\epsilon(x_n) < \infty, \quad \text{for every } \epsilon > 0.\]
(11.3)

Since every term on the right-hand side is either 0 or 1,
\[\sum_{n} i_\epsilon(x_n) < \infty \iff \limsup_{n} i_\epsilon(x_n) = 0 \iff \lim_{n} i_\epsilon(x_n) = 0.\]
(11.4)

This is the basic characterization of convergence. Çinlar [Çin11], III.1 gives a number of useful ways to establish convergence—I will refer to them as necessary.
11.2 Almost sure convergence

As always \((\Omega, \mathcal{H}, \mathbb{P})\) is our basic probability space. \((X_n)\) denotes a sequence of random variables in \(\mathbb{R}\).

A sequence \((X_n)\) is said to be **almost surely convergent** if the numerical sequence \((X_n(\omega))\) is convergent for \((\mathbb{P})\)-almost every \(\omega\). It is said to **converge** to \(X\) if \(X\) is an almost surely \(\mathbb{R}\)-valued random variable and \(\lim_n X_n(\omega) = X(\omega)\) for almost every \(\omega\). This is denoted by \(X_n \xrightarrow{a.s.} X\).

This definition is basically almost sure pointwise convergence of the measurable functions \(X_n: \Omega \to \mathbb{R}\). Since \(\liminf_n X_n\) and \(\limsup_n X_n\) are random variables, \(\Omega_0 = \{ \omega \in \Omega : \liminf_n X_n(\omega) = \limsup_n X_n(\omega) \in \mathbb{R} \}\) is an event in \(\mathcal{H}\). The sequence \((X_n)\) is almost surely convergent if \(\mathbb{P}(\Omega_0) = 1\).

Observe that if \(X_\prime\) is another random variable such that \(X \xrightarrow{a.s.} X_\prime\), then \(X_n \xrightarrow{a.s.} X\) implies that \(X_n \xrightarrow{a.s.} X_\prime\), too.

The following theorem characterizes almost sure convergence.

**Theorem 11.1.** The sequence \((X_n)\) converges to \(X\) almost surely if and only if, for every \(\varepsilon > 0\),

\[
\sum_n \varepsilon \circ |X_n - X| < \infty \quad \text{almost surely.} \tag{11.5}
\]

**Proof. Necessity:** Suppose that \(X_n \xrightarrow{a.s.} X\). Let \(\Omega_0\) be the almost sure set on which \(X_n(\omega) \to X(\omega)\), and let \(Y_n = |X_n - X|\). Then for each \(\omega \in \Omega_0\), \(\sum_n \varepsilon \circ Y_n < \infty\) for every \(\varepsilon > 0\), by (11.3). Thus, for fixed \(\varepsilon\), (11.5) holds on the almost sure set \(\Omega_0\), which does not depend on \(\varepsilon\).

**Sufficiency:** Suppose that, for each \(\varepsilon > 0\), (11.5) holds. Let \((\varepsilon_m)\) be a sequence strictly decreasing to 0, and \(N_m = \sum_n \varepsilon_m \circ Y_n\). We have \(\mathbb{P}\{N_m < \infty\} = 1\) for every \(m\) by assumption. Since \(\varepsilon_{m+1} < \varepsilon_m\), \(\varepsilon_{m+1} \geq \varepsilon_m\) and therefore \(N_{m+1} \geq N_m\). Thus, the events \(\{N_m < \infty\}\) are shrinking to \(\Omega_0 = \bigcap_m \{N_m < \infty\}\).

By sequential continuity of \(\mathbb{P}\), we have \(\mathbb{P}(\Omega_0) = \lim_m \mathbb{P}\{N_m < \infty\} = 1\). Furthermore, again by (11.3), \(X_n(\omega) \to X(\omega)\) for every \(\omega \in \Omega_0\). \(\square\)

**Exercise 31** (Almost sure version of the continuous mapping theorem):

Let \(X\) and \((X_n)\) be random variables taking values in \(\mathbb{R}\), and \(f: \mathbb{R} \to \mathbb{R}\) a function with at most a countable number of points of discontinuity. Show that \(X_n \xrightarrow{a.s.} X\) implies that \(f \circ X_n \xrightarrow{a.s.} f \circ X\).

**Convergence in metric spaces.** Recall that a metric space is a pair \((E, d)\), where \(E\) is a set and \(d\) is a metric on \(E\) (see pp. 10-11). Let \(X\) and \(X_1, X_2, \ldots\) be random variables taking values in \(E\). Then the sequence \((X_n)\) is said to converge to \(X\) almost surely if the \(\mathbb{R}_+\)-valued random variables \(d(X_n, X)\) converge to 0 almost surely.

**Borel–Cantelli lemmas.**
Lemma 11.2 (Borel–Cantelli). Let \((H_n)\) be a sequence of events. Then
\[
\sum_n \mathbb{P}(H_n) < \infty \Rightarrow \sum_n 1_{H_n} < \infty \text{ almost surely.} \tag{11.6}
\]

**Proof.** Let \(N = \sum_n 1_{H_n}\) be the random variable under question. By the MCT, \(E[N] = \sum_n \mathbb{P}(H_n)\). Clearly, \(E[N] < \infty\) implies that \(N < \infty\) almost surely, which is what is claimed. \(\square\)

Another way to interpret this is that if \((H_n)\) is a sequence of events, then \(N\) is the number of times these events occur. So if \(\sum_n \mathbb{P}(H_n) < \infty\), then almost surely only a finite number of events \(H_n\) occur.

The following suite of results follow from Lemma 11.2.

**Proposition 11.3.** \(X_n \xrightarrow{a.s.} X\) if any of the following holds:

i) \(\sum_n \mathbb{P}\{|X_n - X| > \varepsilon\} < \infty\) for every \(\varepsilon > 0\).

ii) There exists a sequence \((\varepsilon_n)\) decreasing to 0 such that \(\sum_n \mathbb{P}\{|X_n - X| > \varepsilon_n\} < \infty\).

iii) There exists a sequence \((\varepsilon_n)\) of strictly positive numbers such that \(\sum_n \varepsilon_n < \infty\) and \(\sum_n \mathbb{P}\{|X_{n+1} - X_n| > \varepsilon_n\} < \infty\).

**Exercise 32 (Applications of Borel–Cantelli Lemma):**

Prove Proposition 11.3.

In general, the converse of Lemma 11.2 is not true, unless the events are independent.

Lemma 11.4 (Second Borel–Cantelli). Let \((H_n)\) be a sequence of independent events. Then
\[
\sum_n \mathbb{P}(H_n) = \infty \Rightarrow \sum_n 1_{H_n} = \infty \text{ almost surely.} \tag{11.7}
\]

**Proof.** Recall that \(\{\sum_n 1_{H_n} = \infty\}\) is also written as \(\{H_n \text{ i.o.}\}\). Recall further from Section 9.3 that the event \(\{H_n \text{ i.o.}\}\) is in the tail-\(\sigma\)-algebra, and therefore
\[
\mathbb{P}\{H_n \text{ i.o.}\} = \mathbb{P}(\cap_m \cup_{n \geq m} H_n) = \lim_m \mathbb{P}(\cup_{n \geq m} H_n).
\]

Suppose that \(\sum_n \mathbb{P}(H_n) = \infty\). Note that this implies \(\sum_{n \geq m} \mathbb{P}(H_n) = \infty\), \(m \geq 1\). Now,
\[
\mathbb{P}(N = \infty) = \lim_m \mathbb{P}(\cup_{n \geq m} H_n) = 1 - \lim_m \mathbb{P}(\cap_{n \geq m} H_n^c).
\]
Now, by independence of \((H_n)\) and the fact that \(1 - x \leq e^{-x}\) for \(x \geq 0\),
\[
\mathbb{P}(N = \infty) = 1 - \prod_{n \geq m} \mathbb{P}(H_n^c) \\
= 1 - \prod_{n \geq m} (1 - \mathbb{P}(H_n)) \\
\geq 1 - \prod_{n \geq m} e^{-\mathbb{P}(H_n)} \\
= 1 - \exp \left( \sum_{n \geq m} \mathbb{P}(H_n) \right) \\
= 1 - e^{-\infty} = 1.
\]

### 11.3 Convergence in probability

A weaker notion than almost sure convergence is convergence in probability. Let \(X_1, X_2, \ldots\) be \(\mathbb{R}\)-valued random variables. The sequence \((X_n)\) is said to converge to \(X\) in probability if, for every \(\epsilon > 0\),
\[
\lim_{n \to \infty} \mathbb{P}\{|X_n - X| > \epsilon\} = 0.
\] (11.8)

We denote this as \(X_n \xrightarrow{p} X\).

For comparison, observe that we can write almost sure convergence similarly: for every \(\epsilon > 0\), \((X_n)\) converges to \(X\) almost surely if
\[
\mathbb{P}\{\lim_n |X_n - X| > \epsilon\} = 0.
\] (11.9)

Convergence in probability is widely used in modern probability—more so than almost sure convergence. This is because almost sure convergence requires satisfying an extremely strong condition: pointwise convergence on a set of probability 1. Convergence in probability is much less stringent. The following example illustrates.

**Example 11.1.** Convergence in probability but not almost surely. Let \(\Omega = (0, 1]\), \(\mathcal{H}\) the Borel \(\sigma\)-algebra on \((0, 1]\), and \(\mathbb{P}\) the Lebesgue measure. Let \(X_1, X_2, X_3, X_4, X_5, X_6, X_7, \ldots\) be the indicators of \((0, 1/2], (1/2, 1], (0, 1/3], (1/3, 2/3], (2/3, 1], (0, 1/4], \ldots\), respectively. Then for arbitrary \(\epsilon \in (0, 1]\), the probabilities \(\mathbb{P}\{X_n > \epsilon\}\) for the sequence \((1, 1/2, 1/2, 1/3, 1/3, 1/3, 1/4, \ldots)\), whose limit is 0. Thus, \(X_n \xrightarrow{p} 0\). But, for every \(\omega \in \Omega\), the sequence \((X_n(\omega))\) consists of zeros and ones without end (i.e., switching back and forth infinitely often). Therefore, \(\liminf_n (X_n(\omega)) = 0\) and \(\limsup_n (X_n(\omega)) = 1\), and the set of \(\omega\) for which \((X_n(\omega))\) converges is empty.

The following theorem summarizes the relationship between almost sure convergence and convergence in probability.

**Theorem 11.5.**

i) If \((X_n)\) converges to \(X\) almost surely, then in converges to \(X\) in probability, as well.

77
ii) If \((X_n)\) converges to \(X\) in probability, then it has a subsequence that converges to \(X\) almost surely.

iii) If every subsequence of \((X_n)\) has a further subsequence that converges to \(X\) almost surely, then \((X_n)\) converges to \(X\) in probability.

Proof. For simplicity, assume that the \((X_n)\) are positive and \(X = 0\); this is the same as replacing \(X_n\) with \(|X_n - X|\). Recall that \(i_\varepsilon\) is the indicator of the interval \((\varepsilon, \infty)\). Define the bounded sequence \((p_n)\) by

\[
p_n(\varepsilon) = E[i_\varepsilon \circ X_n] = P\{X_n > \varepsilon\}.
\]

i) Suppose that \(X_n \overset{a.s.}{\rightarrow} 0\). Fix \(\varepsilon > 0\). Then \(i_\varepsilon \circ X_n \overset{a.s.}{\rightarrow} 0\), which by the bounded convergence theorem implies that \(\lim_n p_n(\varepsilon) = \lim_n E[i_\varepsilon \circ X_n] = E[\lim_n i_\varepsilon \circ X_n] = 0\). Therefore, \(X_n \overset{p}{\rightarrow} 0\).

ii) Suppose that \(X_n \overset{p}{\rightarrow} 0\). Let \(\varepsilon_k = 1/k\), for \(k \geq 1\), and set \(n_0 = 0\). For each \(k \geq 1\), \(p_n(\varepsilon_k) \rightarrow 0\) by assumption (of convergence in probability). Therefore, there exists some \(n_k > n_{k-1}\) such that \(p(\varepsilon_k) \leq 2^{-k}\) for all \(n \geq n_k\). Our subsequence is then \(X_{n_1}, X_{n_2}, \ldots\), and

\[
\sum_{k \geq 1} P\{X_{n_k} > \varepsilon_k\} \leq \sum_{k \geq 1} 2^{-k} = 1.
\]

By Proposition 11.3ii, the subsequence \((X_{n_k})\) converges to 0 almost surely. \(\Box\)

To prove part iii), we need the following lemma characterizing convergence of sequences of real numbers via convergent subsequences.

**Lemma 11.6.** Let \((x_n)\) be a sequence of real numbers. If every subsequence that has a limit has the same value \(x\) for its limit, then the sequence \((x_n)\) tends to \(x\) (infinite values are allowed). If the sequence is bounded, and every convergent subsequence of it has the same limit \(x\), then \((x_n)\) converges to \(x\).

**Proof of Theorem 11.5, continued.**

iii) Assume that every subsequence of \((X_n)\) has a further subsequence that converges to 0 almost surely. Fix some \(\varepsilon > 0\). Consider the bounded sequence of numbers \((p_n(\varepsilon))\). Let \(N\) be a subsequence of \(N\) along which \((p_n(\varepsilon))\) converges; denote its limit by \(p\).\(^{16}\) By assumption, \(N\) has a subsequence \(N'\) such that \(X_{n', \overset{p}{\rightarrow}} 0\), for \(n' \in N'\). By part a) above, we have that \(p_n(\varepsilon) \rightarrow 0\) along \(N'\), which means that the limit \(p = 0\). Since every subsequence of \((X_n)\) has a further subsequence for which this is true, by Lemma 11.6, the original sequence \((p_n(\varepsilon))\) converges to 0 for every \(\varepsilon > 0\), and therefore \(X_n \overset{p}{\rightarrow} 0\). \(\Box\)

\(^{16}\) We know that such a convergent subsequence exists by the Bolzano–Weierstrass theorem, which says that every bounded sequence in \(\mathbb{R}^n\) has a convergent subsequence.
Convergence of continuous functions. An application of Theorem 11.5 is an in probability version of the continuous mapping theorem.

**Proposition 11.7.** Let \( f: \mathbb{R} \to \mathbb{R} \) be continuous. If \( X_n \xrightarrow{p} X \) then \( f \circ X_n \xrightarrow{p} f \circ X \).

The proof demonstrates a useful technique: using almost sure convergence of subsequences to show convergence in probability of the entire sequence.

**Proof.** Suppose that \( X_n \xrightarrow{p} X \). Let \( N' \) be a subsequence of \( N \). Since the original sequence \( (X_n) \) converges in probability to \( X \), so does the subsequence \( (X_{n'}) \), \( n' \in N' \). By Theorem 11.5b, there is a further subsequence \( N'' \) such that \( X_{n''} \xrightarrow{a.s.} X \), \( n'' \in N'' \), which by Exercise 31 (almost sure version of continuous mapping) implies that \( f \circ X_{n''} \xrightarrow{a.s.} f \circ X \). Applying Theorem 11.5c, we have that \( f \circ X_n \xrightarrow{p} f \circ X \) for the original sequence. \( \square \)

Convergence and arithmetic. The previous proof technique can be used to show that convergence in probability is preserved under arithmetical operations.

**Theorem 11.8.** Suppose that \( X_n \xrightarrow{p} X \) and \( Y_n \xrightarrow{p} Y \). Then:

i) \( X_n + Y_n \xrightarrow{p} X + Y \);

ii) \( X_n - Y_n \xrightarrow{p} X - Y \);

iii) \( X_n Y_n \xrightarrow{p} XY \); and

iv) \( X_n/Y_n \xrightarrow{p} X/Y \), provided that, almost surely, \( Y_n \) and \( Y \) are non-zero.

**Exercise 33** (Convergence in probability and arithmetic):

Prove Theorem 11.8i using the method of Proposition 11.7.

11.4 Convergence in \( L^p \)

In order to define this mode of convergence, we need to define \( L^p \) spaces, which we previously skipped. See Činlar [Čin11, Ch. II.3].

**\( L^p \) spaces.** Let \( X \) be a random variables taking values in \( \mathbb{R} \), and for \( p \in [1, \infty] \), define

\[
||X||_p = \begin{cases} 
(\mathbb{E}[|X|^p])^{1/p} & p \in [1, \infty) \\
\inf \{b \in \mathbb{R}_+ : |X| \leq b \text{ almost surely} \} & .
\end{cases}
\]  

(11.10)

It is easy to see that

\[
||X||_p = 0 \Rightarrow X \xrightarrow{a.s.} 0 , \quad \text{and} \quad ||cX||_p = c||X||_p .
\]  

(11.11)

Furthermore, it can be shown that (see Hölder’s inequality, Činlar Theorem II.3.6)

\[
0 \leq ||X||_p \leq ||X||_q \leq +\infty \quad \text{if} \quad 1 \leq p \leq q \leq +\infty .
\]  

(11.12)
For each \( p \in [1, \infty] \), denote by \( L^p \) the collection of all \( \mathbb{R} \)-valued random variables \( X \) with \( ||X||_p < \infty \). For \( X \) in \( L^p \), \( ||X||_p \) is called the \( L^p \)-norm of \( X \). For \( p = \infty \), \( ||X||_\infty \) is also called the essential supremum of \( X \). Clearly, \( X \) is in \( L^p \) if and only if \( |X|^p \) is integrable, and \( X \) is in \( L^\infty \) if and only if \( X \) is almost surely bounded.

These ideas have natural generalizations for measures and functions: for a measure \( \mu \) on some measurable space \( (E, \mathcal{E}) \) and a measurable function \( f : E \to \mathbb{R}, ||f||_p = (\int f(x)^p \mu(dx))^{1/p} \). The \( L^p \) space \( L^p(\mu) \) is the collection of all \( \mathbb{R} \)-valued functions \( f \) such that \( ||f||_p < \infty \). In the probability literature, it is common to see \( L^p(Q) \), etc., for some probability measure \( Q \), to differentiate between the \( L^p \) spaces of different probability measures.

**\( L^p \) convergence.** We will now restrict ourselves to \( p \in [1, \infty) \). A sequence \( (X_n) \) is said to converge in \( L^p \) to \( X \) if

\[
\lim_n \mathbb{E}[|X_n - X|^p] = 0 . \tag{11.13}
\]

We denote this as \( X_n \xrightarrow{L^p} X \).

Recall that \( L^p \) is the collection of all \( \mathbb{R} \)-valued random variables with finite \( L^p \)-norm \( (11.10) \). If we identify as a single random variable all random variables that are almost surely equal to each other, then \( L^p \) is a normed vector space. Clearly, \( (11.13) \) is equivalent to

\[
\lim_n ||X_n - X||_p = 0 , \tag{11.14}
\]

which is the usual notion of convergence in a normed vector space. Hence, \( L^p \) convergence aligns with much of our usual intuition about convergence.

**Proposition 11.9.** Suppose that \( (X_n) \) and \( X \) are in \( L^p \) for some \( p \in [1, \infty) \). Then \( X_n \xrightarrow{L^p} X \) implies the following:

i) \( X_n \xrightarrow{L^q} X \) for each \( q \in [1, p] \); and

ii) \( X_n \xrightarrow{p} X \).

**Exercise 34 (Implications of \( L^p \) convergence):**

Prove Proposition 11.9.

*Hint:* For ii), use Markov’s inequality.

**Example 11.2. Convergence in probability and not \( L^1 \).**

Recall the setup from Example 11.1: \( \Omega = (0,1], \mathcal{H} \) the Borel \( \sigma \)-algebra on \( (0,1] \), and \( \mathbb{P} \) the Lebesgue measure. Let \( X_1, X_2, X_3, X_4, X_5, X_6, X_7, \ldots \) be the indicators of \( (0,1], (0,1/2], (1/2, 1], (0,1/3], (1/3, 2/3], (2/3, 1], (0,1/4], \ldots \), respectively. We showed in that example that \( X_n \xrightarrow{p} 0 \). We have that \( X_n \xrightarrow{L^1} 0 \), as well: \( \mathbb{E}[|X_n|] \to 0 \).

A slight alteration changes that. Let \( (\tilde{X}_n) = (X_1, 2X_2, 3X_3, 3X_4, 3X_5, 3X_6, 4X_7, \ldots) \). Now for \( \epsilon \in (0,1) \), \( \mathbb{P}\{\tilde{X}_n > \epsilon\} = \mathbb{P}\{X_n > \epsilon\} \), so \( \tilde{X}_n \xrightarrow{p} 0 \). However, \( \mathbb{E}[|X_n|] = 1 \) for all \( n \), \( (\tilde{X}_n) \) does not converge to 0 in \( L^1 \). But it cannot converge elsewhere, say some \( a \neq 0 \), otherwise Proposition 11.9 would imply that \( X_n \xrightarrow{a} a \). So \( (\tilde{X}_n) \) does not converge in \( L^1 \).
Convergence in $L^1$ is more common in the literature than any other $p$. In fact, one often sees a proof of $L^p$-convergence (and in particular, $L^2$) in order to show, via Proposition 11.9, $L^1$-convergence. One reason for this is that it allows interchanging limits with expectation: if $X_n \overset{L^1}{\to} X$, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$. This is a corollary to the following more general result.

**Proposition 11.10.** If $X_n \overset{L^1}{\to} X$, then for every bounded random variable $Y$,

$$\lim_n \mathbb{E}[X_n Y] = \mathbb{E}[XY].$$

**Proof.** Suppose that $|Y| \leq b$ almost surely, and that $X_n \overset{L^1}{\to} X$. Then

$$|\mathbb{E}[X_n Y] - \mathbb{E}[XY]| \leq \mathbb{E}[|X_n Y - XY|] \leq b \mathbb{E}|X_n - X| \to 0.$$ 

We will find this particularly useful when working with martingales.

**Uniform integrability.** Convergence in $L^1$ implies convergence in probability. However, the converse is not generally true. To get from convergence in probability to $L^1$, we need a further condition. One might be that $(X_n)$ is dominated, i.e., $|X_n| \leq Y$ almost surely for some integrable random variable $Y$. This is a pretty strong condition, and we might want something weaker.

Revisiting Example 11.2 illustrates what we need to address: even though $(X_n) \overset{P}{\to} 0$, i.e., $\mathbb{P}$ gives less and less mass to non-zero values of $X_n$, which non-zero value has non-zero mass is growing at the same rate as its probability decreases. Non-zero probability mass, however small, is escaping “off to infinity”. If we can control this escape, then we can get convergence in $L^1$. This is known as uniform integrability.

Before giving the definition (which applies to a collection), we illustrate with the simplest case of a single random variable.

**Lemma 11.11.** Let $X$ be a random variable taking values in $\mathbb{R} = (-\infty, +\infty)$. Then, $X$ is integrable if and only if

$$\lim_{b \to \infty} \mathbb{E}[|X| \mathbf{1}_{|X| > b}] = 0. \tag{11.15}$$

**Proof.** Let $Z_b = |X| \mathbf{1}_{|X| > b}$. Note that it is dominated by $|X|$ and goes to 0 as $b \to \infty$, since $X$ takes values in $\mathbb{R}$. If $X$ is integrable, then dominated convergence (Theorem 7.7) implies that $\lim_{b \to \infty} \mathbb{E}[Z_b] = \mathbb{E}[\lim_{b \to \infty} Z_b] = 0$. This is (11.15).

Conversely, if (11.15) holds, choose $b$ large enough to have $\mathbb{E}[Z_b] \leq 1$. Then

$$|X| = |X|\mathbf{1}_{|X| \leq b} + |X|\mathbf{1}_{|X| > b} \leq b + Z_b,$$

which shows that $\mathbb{E}[|X|] \leq b + 1 < \infty$, and thus $X$ is integrable. \qed
We extend this idea to a collection $\mathcal{K}$ of random variables by taking the limit (11.15) \textit{uniformly} in $\mathcal{K}$. Specifically, $\mathcal{K}$ is said to be \textit{uniformly integrable} if,

$$k(b) = \sup_{X \in \mathcal{K}} \mathbb{E}[|X| \mathbf{1}_{|X|>b}],$$

(11.16)
goest to 0 as $b \to \infty$.

If the collection $\mathcal{K}$ is finite and each $X \in \mathcal{K}$ is integrable, then $\mathcal{K}$ is uniformly integrable:

$$\lim_{b \to \infty} \sup_{X \in \mathcal{K}} k(b) = \sup_{X \in \mathcal{K}} \lim_{b \to \infty} k(b) = 0 \quad \text{(Lemma 11.11)}.$$

The harder case is when $\mathcal{K}$ is not finite, and we can’t necessarily interchange $\lim_b$ with $\sup_{\mathcal{K}}$. We need to consider what other conditions would give us uniform integrability. There turn out to be a few; the following three that are relatively easy to work with. (See Çinlar [Çin11, Ch. II.3] for more.)

**Proposition 11.12.** Let $\mathcal{K}$ be a collection of $\mathbb{R}$-valued random variables. If any of the following conditions hold, then $\mathcal{K}$ is uniformly integrable.

1. **$L^p$-boundedness.** $\sup_{X \in \mathcal{K}} \mathbb{E}[|X|^p] < \infty$ for some $p > 1$.
2. **Integrable envelope.** There is a random variable $Y$ with $\mathbb{E}[|Y|] < \infty$ such that each $X \in \mathcal{K}$ satisfies
   $$\int_{\{|X| > b\}} |X| d\mathbb{P} \leq \int_{\{|X| > b\}} |Y| d\mathbb{P}, \quad \text{for all } b > 0.$$
3. There is a (positive) random variable $Y$ with $\mathbb{E}[Y] < \infty$ such that $|X| \leq Y$ almost surely, for each $X \in \mathcal{K}$.

\textbf{Proof.}

1. Assume without loss of generality that $X \geq 0$ for each $X \in \mathcal{K}$. Fix $p > 1$ such that condition i) holds, and let $K$ be some constant such that $\sup_{X \in \mathcal{K}} \mathbb{E}[|X|^p] < k < \infty$. If $X \geq b$, then $X^{1/p} \leq b^{1/p}$, and thus $X \leq X^p b^{1/p}$. Therefore,
   $$\mathbb{E}[|X| \mathbf{1}_{|X|>b}] \leq b^{1-p} \mathbb{E}[|X|^p \mathbf{1}_{|X|>b}] \leq \frac{k}{b^{p-1}},$$
   and hence
   $$\lim_{b \to \infty} \sup_{X \in \mathcal{K}} \mathbb{E}[|X| \mathbf{1}_{|X|>b}] \leq \lim_{b \to \infty} \frac{k}{b^{p-1}} = 0.$$

2. By a similar argument as Lemma 11.11, if condition ii) holds then,
   $$\lim_{b \to \infty} \sup_{X \in \mathcal{K}} \mathbb{E}[|X| \mathbf{1}_{|X|>b}] \leq \lim_{b \to \infty} \mathbb{E}[|Y| \mathbf{1}_{|Y|>b}] \overset{\text{(DCT)}}{=} \mathbb{E} \left[ \lim_{b \to \infty} |Y| \mathbf{1}_{|Y|>b} \right] = 0.$$

3. If $|X| \leq Y$ almost surely then $|X| \mathbf{1}_{|X|>b} \leq |Y| \mathbf{1}_{|Y|>b} \leq Y \mathbf{1}_{Y>b}$ almost surely, and in particular condition ii) holds.
The following result is the main theorem relating $L^1$-convergence to convergence in probability. Its proof can be found as part of Çınlar [Çın11, Thm. III.4.6].

**Theorem 11.13.** Let $(X_n)$ and $X$ be random variables in $L^1$. Then $X_n \xrightarrow{L^1} X$ if and only if: (i) $X_n \xrightarrow{p} X$; and (ii) the collection $(X_n)$ is uniformly integrable.

Using Theorem 11.5, we can prove the following corollary.

**Corollary 11.14.** Let $(X_n)$ and $X$ be random variables in $L^1$, and assume that $(X_n)$ is uniformly integrable. If $X_n \xrightarrow{a.s.} X$ then $X_n \xrightarrow{L^1} X$. Conversely, if $X_n \xrightarrow{L^1} X$ then $(X_n)$ has a subsequence that converges to $X$ almost surely.

### 11.5 Overview of modes of convergence

**Cauchy criterion.** For completeness, an additional characterization of the different modes of convergence is the Cauchy criterion. Recall that a sequence of real numbers $(x_n)$, the Cauchy criterion says that $(x_n)$ converges if and only if, for every $\epsilon > 0$, there is some $k$ such that

$$|x_n - x_m| \leq \epsilon, \quad \text{for all } m \geq k \text{ and } n \geq k.$$  

An equivalent way of writing this is $\lim_{m,n \to \infty} |x_n - x_m| = 0$. Note that it doesn’t say where the sequence converges, but it is useful nonetheless.

The probabilistic versions of this are summarized in the following result.

**Theorem 11.15 (Probabilistic Cauchy criteria).** Let $(X_n)$ be a sequence of $\mathbb{R}$-valued random variables.

i) **Almost sure.** The sequence is convergent almost surely if and only if

$$\lim_{m,n \to \infty} |X_n - X_m| = 0, \quad \text{almost surely.}$$

ii) **In probability.** The sequence converges in probability if and only if, for every $\epsilon > 0$,

$$\lim_{m,n \to \infty} P(|X_n - X_m| > \epsilon) = 0.$$ 

iii) **In $L^1$.** The sequence converges in $L^1$ if and only if

$$\lim_{m,n \to \infty} \mathbb{E}[|X_n - X_m|] = 0.$$ 

### 11.6 Convergence in distribution

The previous modes of convergence that we studied pertained to $\mathbb{R}$-valued random variables (i.e., measurable functions on $\Omega$) converging either:
a) pointwise on \( P \)-almost sure sets (almost sure);
b) on sets with limiting probability is 1 (in probability);
c) in expectation (in \( L^p \)).

**Convergence in distribution**, or **weak convergence**, has to do with the convergence of sequences of probability measures on a topological space. We will work exclusively with convergence in \( \mathbb{R} \), but these ideas can be ported to general topological spaces.

Let \( X_1, X_2, \ldots, X \) be \( \mathbb{R} \)-valued random variables whose distributions are \( \mu_1, \mu_2, \ldots, \mu \) (probability measures on \( \mathbb{R} \)), respectively. Let \( C_b = C_b(\mathbb{R}, \mathbb{R}) \) denote the collection of bounded continuous functions from \( \mathbb{R} \) into \( \mathbb{R} \).

The sequence \( (\mu_n) \) is said to **converge weakly** to \( \mu \) if

\[
\lim_{n \to \infty} \mu_n f = \mu f , \quad \text{for every } f \in C_b .
\]

The sequence \( (X_n) \) is said to **converge in distribution** to \( X \), denoted \( X_n \overset{d}{\to} X \) if \( (\mu_n) \) converges to \( \mu \) weakly, that is, if

\[
\lim_{n \to \infty} \mathbb{E}[f \circ X_n] = \mathbb{E}[f \circ X] , \quad \text{for every } f \in C_b .
\]  \( \text{(11.17)} \)

**Example 11.3. Weak convergence to Lebesgue measure.** Let \( \mu_n \) be the probability measure that puts mass \( 1/n \) on each of the points \( 1/n, 2/n, \ldots, n/n \). Then, for \( f \in C_b \),

\[
\mu_n f = \sum_{k=1}^{n} \frac{1}{n} f \left( \frac{k}{n} \right) \overset{L^1}{\to} \int_0^1 du f(u) = \lambda f .
\]

Alternatively, let \( U_n \) be a random variable with distribution \( \mu_n \) (i.e., the discrete uniform distribution on \( \{1/n, 2/n, \ldots, 1\} \)), and \( U \sim \text{Unif}(0, 1] \). Then \( U_n \overset{d}{\to} U \).

Weak convergence is implied by each of the other modes of convergence that we have seen. The following result illustrates.

**Proposition 11.16.** Let \( (X_n) \) and \( X \) be \( \mathbb{R} \)-valued random variables such that \( X_n \overset{p}{\to} X \). Then \( X_n \overset{d}{\to} X \).

**Proof.** Let \( f \in C_b \). Since \( f \) is continuous, by Proposition 11.7 we know that \( f \circ X_n \overset{p}{\to} f \circ X \). Since \( f \) is also bounded, by Proposition 11.12iii we know that \( (f \circ X_n) \) is uniformly integrable and therefore (by Theorem 11.13) \( f \circ X_n \overset{L^1}{\to} f \circ X \). Therefore, (11.17) holds and \( X_n \overset{d}{\to} X \). \( \square \)

There is a (very weak) partial converse.

**Proposition 11.17.** Let \( (X_n) \) and \( X \) be \( \mathbb{R} \)-valued random variables such that \( X_n \overset{d}{\to} X \). If \( X = x_0 \) for some constant \( x_0 \), then \( X_n \overset{p}{\to} X \), as well.

One might be tempted to think that \( X_n \overset{d}{\to} X \) implies that \( \mathbb{P}\{X_n \in A\} \) converges to \( \mathbb{P}\{X \in A\} \), for all Borel sets \( A \). In reality, this is only true for some special sets. In the case of \( \mathbb{R} \)-valued random variables, this is intimately related to the convergence of distribution functions...
\( F_n(x) = \mu_n(-\infty, x] \). At a high level, weak convergence only implies (and requires) convergence of the distribution functions on the sets of continuity. The proof of this fact is rather technical (e.g., [Çin11, Prop. 5.7]), but we state the result for posterity.

**Theorem 11.18.** The following are equivalent:

i) \( (\mu_n) \) converges weakly to \( \mu \).

ii) The distribution functions \( F_n(x) \to F(x) \) for every continuity point \( x \) of \( F \).

iii) The quantile functions \( Q_n(u) \to Q(u) \) for every continuity point \( u \) of \( Q \).

There is much more that can (and probably should) be said about weak convergence. See Çinlar [Çin11, Ch. III.5]. For the sake of brevity, we devote our attention to two important results that will let us prove the Central Limit Theorem.

**Tightness and Prohorov’s Theorem.** For each \( f \in C_b \), the sequence of numbers \( (\mu_n f) \) is bounded. Therefore, using Lemma 11.6, if every subsequence of \( (\mu_n f) \) has a further subsequence that converges to \( \mu f \), then \( \mu_n f \to \mu f \). In other words, if every subsequence of \( (\mu_n) \) has a further subsequence that converges weakly to \( \mu \), then \( (\mu_n) \) converges weakly to \( \mu \).

In order to make this idea work, we need a regularity condition: the sequence \( (\mu_n) \) is said to be tight if for every \( \epsilon > 0 \), there is a compact set \( K \) such that \( \mu_n(K) > 1 - \epsilon \) for all \( n \).

Recall that in a Euclidean space (such as \( \mathbb{R}^d \)), a set \( K \) is compact if and only if it is closed and bounded. (For a general topological space, see [AB06, Ch. 2].) Therefore, tightness amounts to there being some finite \( r \in \mathbb{R} \) such that

\[
\lim_{r \to \infty} \sup_{n} \mu_n([-r, r]^c) = 0 .
\]

**Theorem 11.19** (Prohorov’s theorem). If \( (\mu_n) \) is tight then every subsequence of it has a further subsequence that is weakly convergent.

**Sketch of proof.** At a high level, the proof goes as follows: Let \( (F_n) \) be the sequence of distribution functions corresponding to \( (\mu_n) \). Helly’s theorem [Çin11, Thm. III.1.9] states that every sequence of distribution functions has a subsequence that converges pointwise to some distribution function \( F \) at every continuity point of \( F \). Using this, we see that every subsequence of \( (F_n) \) has a further subsequence that converges to a distribution function \( F \), corresponding to a measure \( \mu \). Using tightness, we can show that \( \mu \) is a probability measure (i.e., \( F(-\infty) = 0 \) and \( F(\infty) = 1 \)); Theorem 11.18ii shows that \( (\mu_n) \) converges weakly to \( \mu \) along the sub-subsequence. \( \square \)

See Çinlar [Çin11, Thm. III.3.15] for the complete proof.

**Convergence of Fourier transforms.** Recall that the Fourier transform, of a probability distribution \( \mu_n \) is

\[
f_n(r) = \int_{\mathbb{R}} \mu_n(dx) e^{i r x} , \quad r \in \mathbb{R} .
\]

Prohorov’s theorem is the key to showing the close relationship between weak convergence and Fourier transforms.
Theorem 11.20. The sequence \((\mu_n)\) is weakly convergent if and only if

\[
\lim_{n \to \infty} f_n(r) = f(r) \tag{11.19}
\]

exists for every \(r \in \mathbb{R}\) and the function \(f\) is continuous at 0. If this holds, then \(f\) is the Fourier transform of a probability measure \(\mu\) on \(\mathbb{R}\), and \(\mu\) is the weak limit of \((\mu_n)\).

Proof. Necessity. Assume that \((\mu_n)\) converges weakly to a probability measure \(\mu\) on \(\mathbb{R}\). Then since 
\(g : x \mapsto \cos(rx)\) and \(h : x \mapsto \sin(rx)\) are in \(C_b\), \(\mu_n g \to \mu g\) and \(\mu_n h \to \mu h\), and hence

\[f_n(r) = \mu_n g + i\mu_n h \to \mu g + i\mu h = f(r),\]

where \(f\) is the Fourier transform of \(\mu\). Then, \(f\) is continuous at 0 because it is a probability measure on \(\mathbb{R}\).17

Sufficiency. This argument is a bit more complicated. First we will show tightness of \((\mu_n)\), then use Prohorov’s theorem to argue that every subsequence has a further subsequence along which \((\mu_n)\) converges to some probability measure \(\mu'\). By the necessity part of the proof, that implies that \((f_n)\) converges to \(f'\) along that subsequence. By assumption, though, \((f_n)\) converges to \(f\), so \(f' = f\) no matter the subsequence. Therefore, every subsequence of \((\mu_n)\) has a further subsequence that converges weakly to the same probability measure \(\mu\), whose Fourier transform is \(f\), and therefore \((\mu_n)\) converges weakly to \(\mu\).

Suppose that the limits (11.19) exist and \(f\) is continuous at 0. To show tightness, note that by the bounded convergence theorem (Theorem 7.8) we have,

\[\lim_{n \to \infty} \frac{1}{b} \int_{-b}^{b} dr |1 - f_n(r)| = \frac{1}{b} \int_{-b}^{b} dr |1 - f(r)|.\]

By the continuity of \(f\) at 0, as \(b \to 0\), \(f(r) \to f(0) = \lim_n f_n(0) = 1\), so the right-hand side of the previous equation goes to 0 as \(b \to 0\).

Therefore, for every \(\epsilon > 0\) there is some \(b > 0\) such that the right-hand side is less than \(\epsilon/2\), and thus,

\[\frac{1}{b} \int_{-b}^{b} dr |1 - f_n(r)| \leq \epsilon, \tag{11.20}\]

for all but finitely many \(n\).

Now, by Fubini’s theorem (Theorem 10.11),

\[\frac{1}{b} \int_{-b}^{b} dr (1 - f_n(r)) = \int_{\mathbb{R}} \mu_n(dx) \frac{1}{b} \int_{-b}^{b} dr (1 - e^{i\pi x}) .\]

Recall that \(e^{i\pi x} = \cos rx + i \sin rx\). Furthermore, observe that \(\sin\) is an odd function so its integral over a symmetric interval vanishes, and therefore,

\[\frac{1}{b} \int_{-b}^{b} dr (1 - f_n(r)) = 2 \int_{\mathbb{R}} \mu_n(dx) \left( 1 - \frac{\sin bx}{bx} \right).\]

17To see this, note that \(|e^{i\pi x}| = 1\) for all \(r\), and therefore we can apply the bounded convergence theorem (Theorem 7.8) to show that \(f\) is continuous at 0 (and actually that it is uniformly continuous).
Now, $1 - y^{-1} \sin y$ is always positive and exceeds $1/2$ when $|y| > 2$. (See the Fig. 1 for an illustration with $b = 1$.)

Therefore,

$$\frac{1}{b} \int_{-b}^{b} dr (1 - f_n(r)) \geq \int_{\mathbb{R}} \mu_n(dx) \cdot 1_{(2, \infty)}(b|x|) = 1 - \mu_n \left[ \frac{2}{b} - \frac{2}{b} \right].$$

Putting this together with (11.20), for every $\epsilon > 0$ we have that $K = [-2/b, 2/b]$ is such that $\mu(K) \geq 1 - \epsilon$ for some $b > 0$, for all but finitely many $n$. By taking $b$ smaller if necessary, we can ensure that $\mu_n(K) \geq 1 - \epsilon$ for all $n$. Hence $(\mu_n)$ is tight.

Figure 1: The function $\frac{\sin x}{x}$ shown over the interval $(-5, 5)$ (left) and $(-100, 100)$ (right).

We can translate this into convergence in distribution, recalling that the characteristic function of a random variable $X$ is the Fourier transform of its distribution.

**Corollary 11.21.** $X_n \xrightarrow{d} X$ if and only if

$$\lim_{n \to \infty} \mathbb{E}[e^{irX_n}] = \mathbb{E}[e^{irX}], \quad r \in \mathbb{R}.$$ 

### 11.7 Central Limit Theorem

From these results follows one of the most important achievements in probability theory: the Central Limit Theorem (CLT). The basic idea is that if $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables with mean $\mu$ and variance $\sigma^2 < \infty$, and $S_n = \sum_{i=1}^{n} X_i$ is their sum, then the distribution of,

$$Z_n = \frac{(S_n - n\mu)/(\sigma \sqrt{n})}{},$$

for large $n$ is approximately standard normal. The importance of this result is that we haven’t assumed anything about the distribution of $X_i$ except that it has finite variance. The standard proof technique (which we follow) obscures some of the intuition, but the high-level idea is that $Z_n$ has mean 0 and variance 1 for all $n$; for large $n$, all higher-order fluctuations are negligible.

To state the result, let $Z$ denote a standard normal random variable, i.e., $Z \sim \mathcal{N}(0, 1)$. 

87
Theorem 11.22 (Central Limit Theorem). Let \((X_n)\) be a sequence of i.i.d. \(\mathbb{R}\)-valued random variables with mean \(\mu < \infty\) and variance \(\sigma^2 < \infty\). Then \((Z_n)\) (as defined above) converges to \(Z\) in distribution.

Proof. Let \(f\) denote the characteristic function of \(Y_n = (X_n - \mu) / \sigma\) (i.e., the Fourier transform of the distribution of \(Y_n\)). Observe that \(Y_n\) has mean 0 and variance 1. A version of Taylor’s theorem yields

\[
f(r) = f(0) + f'(0)r + \frac{1}{2} f''(0)r^2 (1 + h(r)),
\]

with some function \(h\) such that \(|h(r)| \to 0\) as \(r \to 0\).

Recall that for a random variable \(X\) with distribution \(\mu\) and characteristic function \(f\), if \(E[|X|^m] < \infty\), then

\[
\lim_{r \to \infty} \frac{d^m}{dr^m} f(r) = i^m E[X^m].
\]

For \(Y_n\), this yields \(f(0) = 1\), \(f'(0) = E[Y_n] = 0\), and \(f''(0) = E[Y_n^2] = 1\), and thus

\[
E[e^{i r Z_n}] = \left[ f\left(\frac{r}{\sqrt{n}}\right)^n = \left[ 1 - \frac{r^2/2}{n} (1 + h(r/\sqrt{n})) \right]^n \to e^{-r^2/2}, \quad \text{as } n \to \infty,
\]

because \((1 + c_\alpha/n)^n \to e^c\) if \(c_\alpha \to c\). This is the characteristic function of a standard normal random variable \(Z\): \(E[e^{i r Z}] = e^{-r^2}\).

Therefore, by Corollary 11.21, we have \(Z_n \xrightarrow{d} Z\).

It’s worth squinting to see what actually happened there. For convenience, denote the \(m\)th derivative of \(f\) with respect to \(r\) as \(f^{(m)}\). Let’s take a closer look at \(h(r/\sqrt{n})\) (assuming the higher-order derivatives exist):

\[
h(r/\sqrt{n}) = \frac{2n}{f^{(2)}(0)r^2} \left( \frac{1}{8n^{3/2}} f^{(3)}(0)r^3 + \frac{1}{16n^2} f^{(4)}(0)r^4 + \ldots \right).
\]

For large \(n\), these terms are dominated by the behavior of their dependence on \(n\), i.e., \(n^{-1/2}, n^{-1}\), and so on. They all tend to zero as \(n \to \infty\). Observe that this would be true for any \(h(r/n^\alpha)\), with \(\alpha > 0\).

On the other hand, consider \([f(r/n^\alpha)]^n\) as \(n\) grows large: if \(\alpha > 1/2\), then we would have \([f(r/n^\alpha)]^n \to 1\) because the resulting sequence \((c_n)\) would converge to zero; if \(\alpha < 1/2\), \([f(r/n^\alpha)]^n \to \infty\). The proof works because \(\alpha = 1/2\) is just right. See the discussion below about the relationship between the CLT and the Law of Large Numbers for more on this.

There are stronger versions of the CLT that relax the i.i.d. assumption, and that have weaker assumptions on the moments; they are all describing a similar phenomenon to the basic CLT above. See Çinlar [Çin1, Ch. III.8] for much more on this.
11.8 Laws of large numbers

Like the CLT, the law of large numbers (LLN) is a landmark result from classical probability theory. It is typically presented before the CLT because it says something somewhat less informative (more on this below). The trade-off for the less informative conclusion is that it is a stronger convergence result: the so-called weak law of large numbers (WLLN) states that the sample average converges in probability to its mean; the strong LLN (SLLN) establishes convergence almost surely.

The proof of the almost sure version requires the following technical lemma.

**Lemma 11.23.** Let \((x_n)\) be a sequence of positive numbers and \(\bar{x}_n = (x_1 + \cdots + x_n)/n\) the average of the first \(n\) entries. Let \(N = (n_k)\) be a subsequence of \(\mathbb{N}\) with \(\lim_k n_{k+1}/n_k = r > 0\). If the sequence \((\bar{x}_n)\) converges along \(N\) to \(\mu\), then
\[
x/r \leq \liminf_n \bar{x}_n \leq \limsup_n \bar{x}_n \leq rx.
\]

See [Cin11, Lemma 1.7] for the proof.

**Theorem 11.24.** Suppose that the \(X_n\) are pairwise independent and (marginally) identically distributed with finite mean \(\mu\) and variance \(\sigma^2\). Then, \((\bar{X}_n)\) converges to \(\mu\) in \(L^2\), in probability, and almost surely.

**Proof.** Observe that \(E[S_n] = n\mu\) and \(\text{Var}[S_n] = n\sigma^2\) because of pairwise independence. Thus, \(E[\bar{X}_n] = \mu\) and \(\text{Var}[\bar{X}_n] = \sigma^2/n\). Hence,
\[
E[|\bar{X}_n - \mu|^2] = \text{Var}[\bar{X}_n] = \sigma^2/n \to 0, \quad \text{as } n \to \infty.
\]

In other words, \(\bar{X}_n \overset{L^2}{\to} \mu\). This implies that \(X_n \overset{p}{\to} \mu\), as well.

For almost sure convergence, note that it is enough to show this for positive \(X_n\), because we could show (separately) convergence for the positive and negative parts of general \(X_n\). Assume that \(X_n \geq 0\) for all \(n\). Let \(N = (n_k)\) be defined by \(n_k = k^2\). By Chebyshev’s inequality,
\[
\epsilon^2 \sum_{n \in N} P(|\bar{X}_n - \mu| > \epsilon) \leq \sigma^2 \sum_{k=1}^\infty < \infty.
\]

By Proposition 11.3, this implies that \(\bar{X}_n \overset{a.s.}{\to} \mu\) along \(N\).

Let \(\Omega_0\) be the almost sure set on which this convergence holds. For \(\omega \in \Omega_0\), (the original sequence) \((X_n(\omega))\) is a sequence of positive numbers, and therefore Lemma 11.23 applied to \((\bar{X}_n(\omega))\) with \(r = \lim_k (k+1)^2/k^2 = 1\) yields
\[
\mu \leq \liminf_n \bar{X}_n(\omega) \leq \limsup_n \bar{X}_n(\omega) \leq \mu.
\]

Therefore, for every \(\omega \in \Omega_0\), \((\bar{X}_n(\omega))\) converges to \(\mu\). But \(\Omega_0\) is almost sure, so \(\bar{X}_n \overset{a.s.}{\to} \mu\). \(\square\)
Relationship between CLT and LLN.\footnote{Much of this discussion comes from Jacod and Protter [JP04, Ch. 21].} For simplicity, consider an i.i.d. sequence $(X_n)$, with finite mean $\mu$ and variance $\sigma^2$. Then the SLLN implies that $\bar{X}_n \xrightarrow{n \to \infty} \mu$ (and in $L^2$), i.e.,

$$\lim_{n \to \infty} |\bar{X}_n - \mu| = 0 \quad \text{almost surely and in } L^2.$$  

A natural question often asked in statistics is, “How large must $n$ be so that $\bar{X}_n$ is close to $\mu$?” That is, what is the rate of convergence of $\bar{X}_n$ to $\mu$? For example, is there some $\alpha \in \mathbb{R}$ such that

$$\lim_{n \to \infty} n^{\alpha} |\bar{X}_n - \mu| = c > 0, \quad \text{a.s.}.$$  

It turns out that no such $\alpha$ exists: it cannot be that $n^{\alpha}(\bar{X}_n - \mu)$ converges almost surely to a non-zero constant or a non-zero random variable, even in probability.

However, the CLT says that for $\alpha = 1/2$, $\sqrt{n}(\bar{X}_n - \mu)$ converges in distribution to $N(0, \sigma^2)$. The rate of convergence is, in this weaker sense, $\sqrt{n}$. (This is often quoted as a rule of thumb in statistical applications.)

Convergence of the empirical distribution. Suppose that $(X_n)$ are i.i.d. $\mathbb{R}$-valued random variables in $L^2$. Call their common distribution function $F$, and suppose that we would like to estimate it. Let

$$Y_n(x) = 1_{(-\infty, x)} \circ X_n, \quad x \in \mathbb{R}.$$  

Observe that $(Y_n(x))_n$ are i.i.d. and also in $L^2$. Define

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} Y_n(x), \quad \text{for } x \text{ fixed.} \quad (11.21)$$

As $x$ varies, $F_n$ defines a function on $\mathbb{R}$ called the empirical distribution function. By the SLLN, for each $x \in \mathbb{R}$, we have almost surely that

$$\lim_{n \to \infty} F_n(x) = \mathbb{E}[Y_1(x)] = \mathbb{E}[1_{(-\infty, x)} \circ X_1] = \mathbb{P}\{X_1 \leq x\} = F(x).$$

Therefore, $F_n(x) \xrightarrow{\text{a.s.}} F(x)$ for each $x \in \mathbb{R}$; the empirical distribution function converges to the true distribution function almost surely and in $L^2$.

It is straightforward to see that the above argument works for a general measurable space $(E, \mathcal{E})$, with $1_{(-\infty, x)} \circ X_n$ replaced by $1_A \circ X_n$ for any $A \in \mathcal{E}$, and therefore $F_n(A) \xrightarrow{\text{a.s.}} \mu(A)$, where $\mu$ is the (common) distribution of $X_n$. Specifically,

$$F_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_A \circ X_i = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(A) \xrightarrow{\text{a.s.}} \mu(A), \quad A \in \mathcal{E}.$$  

Using this, it can be shown that

$$\int_E f(x) F_n(dx) = \int_E \frac{1}{n} \sum_{i=1}^{n} f(x) \delta_{X_i}(dx) = \frac{1}{n} \sum_{i=1}^{n} f \circ X_i \xrightarrow{\text{a.s.}} \mu f. \quad (11.22)$$
This is the basis for Monte Carlo methods (based on i.i.d. samples from the unknown distribution). In practice, we often cannot sample i.i.d. from our distribution of interest $\mu$, but we might be able to generate samples from a Markov chain whose stationary distribution is $\mu$. In that case, we need a stronger LLN, known as the ergodic LLN, which says that we can estimate integrals using samples from a Markov chain in the same way as in (11.22). This is the basis for Markov chain Monte Carlo (MCMC) methods. See, for example, Meyn and Tweedie [MT93] or, for a nice overview, Roberts and Rosenthal [RR04].
12 Martingales

Reading: Jacod and Protter [JP04], Ch. 24-27
Supplemental: Çinlar [Çin11], XXX

12.1 Basic definitions

We will consider families of random variables \((X_n)_{n \in \mathbb{N}}\) that have very special properties. The ideas in this section generalize to more general index sets, with \(\mathbb{R}\) an important case.

Filtrations. A filtration is a family \(\mathcal{F} = (\mathcal{F}_n)_{n \in \mathbb{N}}\) of \(\sigma\)-algebras that satisfies

\[ n \leq m \Rightarrow \mathcal{F}_n \subseteq \mathcal{F}_m. \tag{12.1} \]

The monotonicity of the filtration means that there is a uniquely determined, smallest \(\sigma\)-algebra that contains all \(\sigma\)-algebras in \(\mathcal{F}\):

\[ \mathcal{F}_\infty = \sigma \left( \bigcup_{n \in \mathbb{N}} \mathcal{F}_n \right) \]

Let \(\mathcal{F} = (\mathcal{F}_n)_{n \in \mathbb{N}}\) be a filtration. A family of random variables \((X_n)_{n \in \mathbb{N}}\) is called adapted to \(\mathcal{F}\) if \(X_n \in \mathcal{F}_n\) for each \(n \in \mathbb{N}\). Clearly, every sequence \((X_n)\) is adapted to its canonical filtration, defined by

\[ \mathcal{F}_n = \sigma \left( \bigcup_{m \leq n} \sigma(X_m) \right). \]

An adapted family \((X_n, \mathcal{F}_n)_{n \in \mathbb{N}}\) is called a martingale if:

i) \(E[|X_n|] < \infty\) for each \(n\) (that is, each \(X_n\) is integrable); and

ii) \(E[X_m | \mathcal{F}_n] \overset{a.s.}{=} X_n\) for each \(m \geq n\).

We also say that \((X_n)\) is an \(\mathcal{F}\)- or \((\mathcal{F}_n)\)-martingale. Note that for index set \(\mathbb{N}\), it is enough that \(E[X_{n+1} | \mathcal{F}_n]\) for all \(n \geq 0\); more general index sets require the more general condition above.

If the ‘\(\overset{a.s.}{=}\)’ is replaced by ‘\(\leq\)’ a.s., then \((X_n, \mathcal{F}_n)\) is called a supermartingale; if replaced by ‘\(\geq\)’ a.s., it is called a submartingale.

At a high level, a martingale is, on average, constant; a submartingale (supermartingale) increases (decreases) on average. I’ll say more about the origin of the names a bit later.

Examples.
Example 12.1. Martingale: sum of centered independent random variables. Let \((X_n)\) be a sequence of independent integrable random variables, with respective means \((\nu_n)\). Let \(F_n\) be the canonical filtration of \((X_n)\), with \(F_0 = \{\emptyset, \Omega\}\). Let \(S_0 = 0\) and \(S_n = \sum_{i=1}^{n}(X_i - \nu_i)\). Then \((S_n, F_n)_{n \geq 0}\) is a martingale because,

\[
E[S_m | F_n] = E[S_n + (S_m - S_n) | F_n] = S_n + E\left[\sum_{i=n+1}^{m} (X_i - \nu_i) \bigg| F_n \right]
\]

\[
= S_n + \sum_{i=n+1}^{m} (E[X_i] - \nu_i)
\]

\[
= S_n.
\]

Example 12.2. Martingale: product of independent random variables. Let \((R_n)\) be a sequence of independent random variables with mean 1 and finite variances. Let \(X_0 = 1\) and

\[
X_n = X_0 \prod_{i=1}^{n} R_i,
\]

and \(F\) the canonical filtration of \((R_n)\). Then \(E[X_{n+1} | F_n] = X_n E[R_{n+1} | F_n] = X_n\), so \((X_n)\) is a \(F\)-martingale.

In each of the previous examples, the filtrations were comprised of independent \(\sigma\)-algebras; by Kolmogorov’s 0-1 law, the limits must be constant almost surely. Much more interesting objects occur in practice, which give rise to random limits. The following is a simple (canonical) example.

Example 12.3. Pólya’s urn. Imagine an urn (a fancy bucket) that contains \(R_0 = 1\) red ball and \(W_0 = 1\) white ball. At each time step \(n \geq 1\), I draw a ball from the urn, then replace that ball and add another of the same color. Let \(X_n \in \{0, 1\}\) record the color of the ball drawn at step \(n\), with 0 corresponding to red. The proportion of white balls after \(n\) steps is

\[
W'_n = \frac{W_n}{W_n + R_n} = \frac{W_n}{n+2}.
\]

Now, with \(F\) the canonical filtration of \((X_n)\),

\[
E[W'_{n+1} | F_n] = (n + 3)^{-1}(W_n + E[X_n | F_n])
\]

\[
= (n + 3)^{-1}W_n(1 + (n + 2)^{-1}) = (n + 2)^{-1}W_n = W'_n,
\]

and therefore \((W_n)\) is an \(F\)-martingale. It turns out (we won’t show this) that the limiting proportion of white balls, \(W_\infty\) is a random variable with distribution \(\text{Unif}(0, 1)\).

Some basic properties. Martingales are hugely important in probability theory and in applications, in large part because they are disarmingly simple. Once it has been established that something is a martingale (or, in many cases, a submartingale) proving many desirable properties becomes quite simple. The hard part, however, often is finding a martingale lurking around in the system being studied. Except for some simple cases (e.g., the first two examples above), there aren’t a lot of
recipes for constructing a (sub)martingale, unless one is already in hand.

Some useful properties and observations are listed below.

a) The martingale property $E[X_m | F_n] = X_n$ can be expressed equivalently as
$$\int_A X_n dP = \int_A X_m dP, \quad \text{for each } A \in F_n.$$  

b) The martingale property $E[X_m | F_n] = X_n$ can be (and is) expressed equivalently as,
$$E[(X_m - X_n) | F_n] \equiv 0, \quad \text{for } m \geq n.$$  

c) Let $X$ be an $F$-submartingale. For $m \geq n$, the random variable $E[(X_m - X_n) | F_n]$ is non-negative, and therefore it is almost surely zero if and only if its expectation $E[E[(X_m - X_n) | F_n]] = 0$. It follows that the submartingale $X$ is a martingale if and only if $E[X_n] = E[X_0]$ for all $n \geq 0$.

d) If $X$ and $Y$ are $F$-submartingales, then so is $aX + bY$, for $a, b \in \mathbb{R}_+$. This is because, for $m \geq n$,
$$E[aX_m + bY_m | F_n] = aE[X_m | F_n] + bE[Y_m | F_n] \geq aX_n + bY_n.$$  

e) If $X$ and $Y$ are $F$-martingales, then so is $aX + bY$, for $a, b \in \mathbb{R}$. This can be shown as above.

f) Let $f$ be a convex function on $\mathbb{R}$. If $X$ is an $F$-martingale, and if $f \circ X_n$ is integrable for every $n$, then $f \circ X$ is an $F$-submartingale. This follows from Jensen’s inequality (for conditional expectations): for $m \geq n$,
$$E[f \circ X_m | F_n] \geq f \circ E[X_m | F_n] = f \circ X_n.$$  

Two important examples (in the context of martingale theory) are $|x|$ and $x^+ = \max\{0, x\}$.

g) The previous property also applies to $F$-submartingales, if $f$ is non-decreasing.

**Exercise 35** (Convex increasing functions of submartingales): Prove property g) above.

**Stopping times.** The “constant expectation” property of martingales can be extended to hold over certain random times. A **stopping time** is a random variable taking values in $\mathbb{N} \cup \{+\infty\}$, such that the event $\{T \geq n\} \in F_n$, for all $n$.

Constant integer-valued random variables are (uninteresting) stopping times. More interesting stopping times are constructed from sequences of random variables. For example, let
$$T = \begin{cases} 
\inf_{n \geq 0} \{n : X_n \geq 73\} & \text{if } X_n \geq 73 \text{ for some } n \in \mathbb{N} \\
+\infty & \text{otherwise.}
\end{cases}$$

**Theorem 12.1.** Let $T$ be a stopping time bounded be some $c \in \mathbb{R}$ and let $(X_n)$ be a martingale with respect to its canonical filtration. Then $E[X_T] = E[X_0]$.  

94
12.2 Martingale Convergence Theorem

Ultimately, we will prove the following result, which ...

**Theorem 12.2** (Martingale Convergence Theorem). Let \( X = (X_n, \mathcal{F}_n)_{n \in \mathbb{N}} \) be a submartingale.

i) If \( X \) is bounded in \( L^1 \), the sequence \( (X_n)_{n \in \mathbb{N}} \) converges to some \( X_\infty \) almost surely as \( n \to \infty \), and \( X_\infty \in L^1 \).

ii) If \( X \) is even a martingale, and if the collection \( (X_n)_{n \in \mathbb{N}} \) is uniformly integrable, then \( (X_n)_{n \in \mathbb{N}} \) converges almost surely and in \( L^1 \) to some \( X_\infty \) as \( n \to \infty \). Moreover, \( X_n = \mathbb{E}[X_\infty | \mathcal{F}_n] \) for all \( n \in \mathbb{N} \).

iii) Conversely, for any integrable, real-valued random variable \( X \), \((\mathbb{E}[X | \mathcal{F}_n], \mathcal{F}_n)_{n \in \mathbb{N}} \), is a uniformly integrable martingale.

**Remarks.** Observe that in part i) we get almost sure convergence of \( (X_n) \) if it is bounded in \( L^1 \), and the limit is also in \( L^1 \). However, we do not have convergence in \( L^1 \); that is, it may not be the case that \( \mathbb{E}[X_n] \to \mathbb{E}[X_\infty] \). For that, we need the uniform integrability condition of part ii).

The \( L^1 \)-convergence claim in part ii) also holds for submartingales, with the conclusion that \( X_n \leq \mathbb{E}[X_\infty | \mathcal{F}_n] \). This implies that \( X_\infty \) extends \( X \) to a submartingale \( \tilde{X} = (X_n)_{n \in \mathbb{N} \cup \{\infty\}} \) (and therefore also for martingales).

**Upcrossings.** To show that \( (X_n) \) converges almost surely, we need to show that \( \lim_n X_n \) exists (almost surely), which is true if \( \liminf_n X_n = \limsup_n X_n \). We will use the idea of upcrossings to prove by contradiction that this is true: Suppose that \( \liminf_n X_n < \limsup_n X_n \). Then there are numbers \( a < b \) such that \( \liminf_n X_n < a < b < \limsup_n X_n \). Recall that for any \( a > \liminf_n X_n \), \((X_n)\) will fall in the interval \([\liminf_n X_n, a] \) infinitely often; likewise for the interval \([b, \limsup_n X_n] \).

Therefore, \((X_n)\) makes upward crossings, or “upcrossings”, from below \( a \) to above \( b \) infinitely many times. Thus, if we can show that the number of upcrossings is finite almost surely for all \( a < b \), then it must be that \( \liminf_n X_n = \limsup_n X_n \).

Formally, we express upcrossings via stopping times. Define \( T_0 = 0 \), and inductively for \( j \geq 0 \),

\[
\begin{align*}
S_{j+1} &= \min\{n > T_j : X_n \leq a\} \quad \text{first time after } T_j \text{ that } X_n \leq a \\
T_{j+1} &= \min\{n > S_{j+1} : X_n \geq b\} \quad \text{first time after } S_{j+1} \text{ that } X_n \geq b.
\end{align*}
\]

(12.2)

We also use the conventions that \( \min\emptyset = +\infty \) and \( \max\emptyset = 0 \). Observe that all \( S_j \) and \( T_j \) are stopping times. The number of \([a, b]\)-upcrossings up to time \( n \) is defined as

\[
U_n(a, b) = \min\{k : T_k \leq n\} = \sum_{k=1}^{\infty} 1_{(0,n]} \circ T_k.
\]

This is the quantity we want to control (i.e., show finiteness of) as \( n \) grows.

---

\[19\]Note that this is not claiming that \( X_n \xrightarrow{L^1} X_\infty \).
To that end, define the upcrossing indicator process

$$C_n = \sum_{k=1}^{\infty} 1_{(S_k,T_k)}(n) = 1\{\text{an upcrossing is being made at time } n\}.$$  \hfill(12.3)

$(C_n)$ is a predictable process (that is, $C_n \in F_{n-1}$ for all $n$). To see this, note that we can write (12.3) as $C_1 = 1\{X_0 < a\}$ and

$$C_n = 1\{C_{n-1} = 1 \text{ and } X_{n-1} \leq b\} + 1\{C_{n-1} = 0 \text{ and } X_{n-1} < a\}, \quad n \geq 2. \hfill(12.4)$$

We will use $(C_n)$ to accumulate $(X_n)$ during active upcrossings, via

$$Y_n = \sum_{m=1}^{n} C_m (X_m - X_{m-1}), \quad n \geq 2. \hfill(12.5)$$

The upcrossing indicator $(C_n)$ just picks out parts of $(X_n)$ that are traversing from below $a$ to above $b$. Hence,

$$Y_n \geq (b - a)U_n(a,b) - |(X_n - a)^-| \quad \text{almost surely.} \hfill(12.6)$$

Furthermore, observe that because $C_{m+1} \leq 1$ and $E[X_{m+1} - X_m \mid F_m] \geq 0$,

$$E[Y_{m+1} - Y_m \mid F_m] = E[C_{m+1}(X_{m+1} - X_m) \mid F_m] \leq E[C_{m+1}E[X_{m+1} - X_m \mid F_m] \leq E[X_{m+1} - X_m \mid F_m].$$

Now taking expectations of both sides and summing over $m$ yields,

$$E[Y_n] \leq E[(X_n - X_0)]. \hfill(12.7)$$

From this and (12.6) we can deduce the following lemma, which is an important result on its own.

**Lemma 12.3** (Doob’s Upcrossing Inequality). Let $(X_n)$ be a submartingale, let $a < b$ and let $U_n(a,b)$ be the number of upcrossings of $[a,b]$ by time $n$. Then

$$E[U_n(a,b)] \leq \frac{E[(X_n - a)^+ - (X_0 - a)^+]}{b - a}, \quad \text{for all } n \in \mathbb{N}. \hfill(12.8)$$

*Proof.* Let $M_n = (X_n - a)^+$. Since $(X_n)$ is a submartingale, $(M_n)$ is also a submartingale (since $x \mapsto (x - a)^+$ is convex and non-decreasing). Furthermore, an upcrossing of $[a,b]$ by $(X_n)$ is the same as an upcrossing of $[0, b - a]$ by $(M_n)$. Now re-defining $Y_n$ (12.5) in terms of $M_n$, (12.6) becomes (after rearrangement)

$$U_n(a,b) \leq \frac{Y_n}{b - a},$$

because $(M_n)$ never falls below 0 by construction and therefore $|M_n^-| = 0$. Taking expectations and putting in (12.7) yields

$$E[U_n(a,b)] \leq \frac{E[Y_n]}{b - a} \leq \frac{E[M_n - M_0]}{b - a} = \frac{E[(X_n - a)^+ - (X_0 - a)^+]}{b - a}. \hfill(12.9)$$

$\square$
Proof of Theorem 12.2i. First, observe that being bounded in $L^1$ means that for the submartingale $(X_n)$,

$$\mathbb{E}[X_n^+] \leq \mathbb{E}[|X_n|] = 2\mathbb{E}[X_n^+] - \mathbb{E}[X_n] \leq 2\mathbb{E}[X_n^+] - \mathbb{E}[X_0],$$

and therefore

$$\sup_n \mathbb{E}[X_n^+] < \infty \iff \sup_n \mathbb{E}[|X_n|] < \infty.$$  \hspace{1cm} (12.8)$$

As outlined above, we will prove by contradiction that $(X_n)$ has an almost sure limit. Suppose the contrary, and in particular that $\liminf_n X_n < a < b < \limsup_n X_n$ for some $a < b$. Let $U_n(a,b)$ be the number of upcrossings of $[a,b]$ by time $n$, which is non-decreasing and therefore $U_\infty(a,b) = \lim_n U_n(a,b)$ exists. By the Monotone Convergence Theorem, using the fact that $(x-a)^+ \leq x^+ + |a|$ for all $a,x \in \mathbb{R}$,

$$\mathbb{E}[U_\infty(a,b)] = \lim_n \mathbb{E}[U_n(a,b)] \leq \sup_n \frac{\mathbb{E}[(X_n-a)^+]}{b-a} \leq \sup_n \frac{\mathbb{E}[X_n^+] + |a|}{b-a} < \infty.$$  \hspace{1cm} (12.9)$$

Since $\mathbb{E}[U_\infty(a,b)] < \infty$, $P\{U_\infty(a,b) < \infty\} = 1$. Therefore, $(X_n)$ upcrosses $[a,b]$ only finitely many times almost surely. If we let

$$A = \bigcup_{a<b, a,b \in \mathbb{Q}} \{\liminf_n X_n < a < b < \limsup_n X_n\},$$

then because all pairs of rational numbers $a,b \in \mathbb{Q}$ are countable,

$$P(A) \leq \sum_{a<b, a,b \in \mathbb{Q}} P\{\liminf_n X_n < a < b < \limsup_n X_n\} = 0.$$

Hence, $\liminf_n X_n = \limsup_n X_n = \lim_n X_n = X_\infty$ almost surely.

$X_\infty$ may be infinite, so we still need to show that it is in $L^1$. To that end, by Fatou’s Lemma and (12.8),

$$\mathbb{E}[|X_\infty|] = \mathbb{E}[\liminf_n |X_n|] \leq \liminf_n \mathbb{E}[|X_n|] \leq 2\sup_n \mathbb{E}[X_n^+] - \mathbb{E}[X_0] < \infty.$$

Therefore, $X_\infty \in L^1$. \qed

Proof of Theorem 12.2ii. By part i), $X_n \overset{a.s.}{\rightarrow} X_\infty$. Assuming uniform integrability of $(X_n)$ means that $X_n \overset{L^1}{\rightarrow} X_\infty$, as well, by Corollary 11.14. In particular,

$$|\mathbb{E}[X_n] - \mathbb{E}[X_\infty]| \leq \mathbb{E}[|X_n - X_\infty|] \to 0, \text{ as } n \to \infty.$$  \hspace{1cm} (12.10)$$

Showing that $X_n = \mathbb{E}[X_\infty \mid \mathcal{F}_n]$ is equivalent to showing that $\mathbb{E}[^1_H(X_\infty - X_n)] = 0$ for all $H \in \mathcal{F}_n$ (see the ‘Uniqueness and versions’ discussion just before Proposition 10.1). To that end, fix $n$, and let $H \in \mathcal{F}_n$ and $m \geq n$. Then by the martingale property,

$$\mathbb{E}[1_H X_n] = \mathbb{E}[1_H X_m], \text{ } m \geq n.$$
Now, Proposition 11.10 says that if \( X_n \overset{L}{\rightarrow} X \), then for every bounded random variable \( Y \),
\[
\lim_{n \to \infty} E[X_n Y] = E[XY].
\]
Hence, since \( X_m - X_n \overset{L}{\rightarrow} X_\infty - X_n \), with \( Y = 1_H \) we have
\[
E[1_H(X_\infty - X_n)] = \lim_{m \to \infty} E[1_H(X_m - X_n)] = 0.
\]
Since \( H \in \mathcal{F}_n \) was arbitrary, this implies \( X_n = E[X_\infty \mid \mathcal{F}_n] \).

Proof of Theorem 12.2iii. The martingale property follows from the properties of repeated conditioning: for each \( m \leq n \), \( F_m \subset \mathcal{F}_n \), and therefore (almost surely)
\[
E[X_n \mid F_m] = E[E[X \mid \mathcal{F}_n] \mid F_m] = E[X \mid F_m] = X_m,
\]
so \((X_n, \mathcal{F}_n)\) is a martingale.

To show uniform integrability, observe that by Jensen’s inequality (Eq. (8.12)),
\[
|X_n| = |E[X \mid \mathcal{F}_n]| \leq E[|X| \mid \mathcal{F}_n] \text{ almost surely.}
\]
Hence, for every \( A \in \mathcal{F}_n \),
\[
\int_A |X_n| dP \leq \int_A |X| dP,
\]
which holds in particular for \( A = \{|X_n| > b\} \), \( b \in \mathbb{R} \). By Proposition 11.12, since \( X \) is integrable, \((X_n)\) is uniformly integrable.

12.3 Martingale inequalities

A tail bound for an \( \mathbb{R} \)-valued random variable \( X \) is an inequality that upper-bounds the probability that \( X \) takes values far from its mean. Markov’s inequality (8.9) is an example:
\[
P(|X| > c) \leq \frac{1}{c} E[|X|].
\]
Even this elementary example is extremely useful; we’ve used it (and its variants like Chebyshev’s inequality) numerous times in our proofs.

Maximal inequalities. Instead of a single random variable, suppose we have a sequence of random variables \((X_n)_{n \in \mathbb{N}}\). A natural (and useful) question is whether similar tail bounds hold uniformly (i.e., simultaneously) for all random variables in the sequence. Such bounds are called maximal inequalities. For example, when the \( X_n \)’s are independent, much can be said about the sequence of sums \( (S_n) \); Çinlar [Çin11, Ch. III.7] is devoted the topic. The following result is a commonly used example.

Theorem 12.4 (Kolmogorov’s maximal inequality). Suppose that \((X_n)\) is a sequence of independent \( \mathbb{R} \)-valued random variables, each with mean zero. Then, for every \( c \in (0, \infty) \),
\[
P(\max_{k \leq n} |S_k| > c) \leq \frac{1}{c^2} \text{Var}[S_n].
\]
We saw in Example 12.1 that \((S_n)\) is a martingale with respect to the canonical filtration of \((X_n)\); we might ask whether such an inequality holds only for the particular martingale \((S_n)\), or whether it (or something similar) holds for more general martingales. In fact, it does, and Kolmogorov’s inequality is recovered as a special case.

**Theorem 12.5** (Doob’s martingale inequality). Let \(X = (X_n, \mathcal{F}_n)_{n \in \mathbb{N}}\) be a martingale in \(L^p\) for some \(p \in [1, \infty)\). Then, for \(c > 0\),

\[
P\{\max_{k \leq n} |X_k| > c\} \leq \frac{1}{c^p} \mathbb{E}[|X_n|^p].
\]

This result actually follows of a corollary from a more general theorem whose proof is interesting—it involves many of the key ideas related to martingales and stopping times. See Çinlar [Çin11, Thm. V.3.21].

A simpler result that also has a independent-variables counterpart, is **Azuma’s inequality**, which is the martingale generalization of **Hoeffding’s inequality** for independent random variables. (Therefore, it is often called the Azuma–Hoeffding inequality.) Azuma’s inequality provides a powerful tail bound for a martingale whose increments \(\Delta_n = X_n - X_{n-1}\) are bounded almost surely.

**Theorem 12.6** (Azuma–Hoeffding inequality). Let \(X = (X_n, \mathcal{F}_n)_{n \in \mathbb{N}}\) be a martingale with \(\mathbb{E}[X_0] = \mu\). If there are a sequence of constants \(c_n \geq 0\) such that \(|X_1 - \mu| = |\Delta_1| \leq c_1\) and \(|\Delta_n| \leq c_n\) almost surely, then for any \(n \geq 1\) and any \(b > 0\),

\[
P\{|X_n - \mu| > b\} \leq 2 \exp\left(-\frac{b^2}{2 \sum_{m=1}^{n} c_m^2}\right).
\]

The proof relies on the following approximation.

**Lemma 12.7.** Fix some constant \(d > 0\). Then for every \(x\) such that \(|x/d| \leq 1\) and every \(t > 0\),

\[
e^{tx} \leq e^{t^2d^2/2} + \frac{e^{td} - e^{-td}}{2} x.
\]

A proof of this can be found in, e.g., the lecture notes here.

**Proof of Theorem 12.6.** A martingale with its mean subtracted is a mean-zero martingale, so without loss of generality, we may assume that \(\mathbb{E}[X_0] = 0\). By Markov’s inequality, for any \(b > 0\) and \(t > 0\),

\[
P\{X_n > b\} = P\{e^{tX_n} > e^{tb}\} \leq e^{-tb} E[e^{tX_n}] = e^{-tb} E[e^{t\sum_{m=1}^{n} \Delta_m}].
\]

Now, via repeated conditioning and noting that \(\Delta_m \in \mathcal{F}_m\) for \(m \leq m'\),

\[
E\left[e^{t\sum_{m=1}^{m'} \Delta_m}\right] = E\left[e^{t\Delta_1} \prod_{m=2}^{m'} E[e^{t\Delta_m} | \mathcal{F}_{m-1}]\right].
\]
Applying Lemma 12.7 to the conditional expectation with \( x = \Delta_n \) and \( d = c_n \), and using the martingale property \( \mathbb{E}[\Delta_n | \mathcal{F}_{n-1}] = 0 \), we have

\[
\mathbb{E}[e^{t\Delta_n} | \mathcal{F}_{n-1}] \leq e^{t^2c_n^2/2} + \frac{e^{td} - e^{-td}}{2} \mathbb{E}[\Delta_n | \mathcal{F}_{n-1}] = e^{t^2c_n^2/2}.
\] (12.11)

We can continue iterating in this way to obtain

\[
P\{X_n > b\} \leq e^{-tb} \mathbb{E}[e^{t^2 \sum_{m=1}^{n} \Delta_m}] \leq e^{-tb} e^{t^2 \sum_{m=1}^{n} c_m^2/2}.
\]

Optimizing over \( t \) to minimize the bound, we set \( t = b / \sum_{m=1}^{n} c_m^2 \) to obtain

\[
P\{X_n > b\} \leq \exp \left( - \frac{b^2}{2 \sum_{m=1}^{n} c_m^2} \right).
\]

A similar argument for the negative tail of \( X_n \) yields

\[
P\{X_n < -b\} \leq \exp \left( - \frac{b^2}{2 \sum_{m=1}^{n} c_m^2} \right).
\]

Combining the two with a union bound yields the result.

\[\square\]

**Remark: Bounds for sub- and supermartingales.** Observe that the inequality \( \mathbb{E}[e^{t\Delta_n} | \mathcal{F}_{n-1}] \leq e^{tc_n^2} \) from (12.11) holds even for supermartingales, and therefore the one-sided bound,

\[
P\{X_n - \mu \geq b\} \leq \exp \left( - \frac{b^2}{2 \sum_{m=1}^{n} c_m^2} \right),
\]

holds for supermartingales. Since the negative of a supermartingale is a submartingale, the other one-sided bound,

\[
P\{X_n - \mu \leq -b\} \leq \exp \left( - \frac{b^2}{2 \sum_{m=1}^{n} c_m^2} \right),
\]

holds for submartingales.
A Additional results

A.1 An example of the use of Radon–Nikodym derivatives

A simple application of change of measure using Radon–Nikodym derivatives is the following result.

**Proposition A.1.** Let $X$ be a random variable taking values in $(E, \mathcal{E})$, and $\mu$ its distribution. If $\mathbb{E}[f(X)] = 0$ for any $f \in \mathcal{E}_+^n$, then $f(X) = 0$ $\mu$-a.s. Furthermore, if $Y$ is a random variable taking values in $(E, \mathcal{E})$, with distribution $\nu \ll \mu$, then $f(Y) = 0$ $\nu$-a.s.

**Proof.** 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 $f(X) = 0$ $\mu$-a.s. Another application of Proposition 7.5 iii) implies that $f(Y) = 0$ $\nu$-a.s. \qed
References


## Index

*K*-dimensional simplex, 30  
χ²-distribution, 33  
L¹, 80  
Lᵖ, 80  
χ²-distribution, 33  
Lᵖ-norm, 80  
E measurable, 16, 17  
E/F-measurable, 16  
µ-almost everywhere, 26  
π-system, 26  
σ-algebra, 9  
σ-algebra generated by \( f \), 16  
σ-finite, 48, 63  
nth centered moment, 50  
nth moment, 50

absolutely continuous, 32, 47  
adapted, 92  
algebra, 9  
amost everywhere, 26, 32  
amost surely, 26, 32  
amost surely convergent, 75  
atom, 24, 26  
Azuma’s inequality, 99

Borel, 17  
Borel σ-algebra, 12  
Borel sets, 12, 13  
Bounded Convergence Theorem, 44

canonical filtration, 92  
canonical form, 19  
Cauchy criterion, 83  
cdf, 32  
characteristic function, 51, 87  
closed, 8  
closed set, 12  
closed under countable intersections, 8  
collection, 8  
complement, 7  
Complete, 21  
complete, 26  
complete separable metric space, 21  
completion, 26  
composition, 17

Conditional determinism, 60  
conditional determinism, 60  
conditional distribution, 68  
conditional expectation, 57  
conditional expectation of \( X \) given \( F \), 58  
conditional probability, 61  
conditionally independent, 70  
converge, 75  
converge in L¹, 80  
converge in distribution, 84  
converge weakly, 84  
Convergence in distribution, 84  
convergence in probability, 77  
convergent, 74  
convex, 50  
convolution, 55  
counting measure, 24  
cumulative distribution function, 32  
density function, 32, 47  
determines, 36  
diffuse, 26  
Dirac measure, 23  
discrete, 10, 31  
discrete measure, 24  
disintegration, 69  
disjointed, 8  
distribution, 31  
distribution function, 32  
distributional identity, 35  
dominated, 43  
Dominated Convergence Theorem, 43, 60  
dyadic function, 19  
dyadic intervals, 19  
edges, 34  
empirical distribution function, 90  
empirical measure, 24  
equal in distribution, 32  
equivalent, 49  
esential supremum, 80  
event notation, 31  
events, 23, 28
<table>
<thead>
<tr>
<th>Term</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polish space</td>
<td>21</td>
</tr>
<tr>
<td>positive</td>
<td>17</td>
</tr>
<tr>
<td>positive part</td>
<td>18</td>
</tr>
<tr>
<td>Positivity</td>
<td>37</td>
</tr>
<tr>
<td>posterior</td>
<td>30</td>
</tr>
<tr>
<td>preferential attachment model</td>
<td>34</td>
</tr>
<tr>
<td>prior</td>
<td>30</td>
</tr>
<tr>
<td>probabilistic programming</td>
<td>35</td>
</tr>
<tr>
<td>probability density function</td>
<td>49</td>
</tr>
<tr>
<td>probability generating function</td>
<td>52</td>
</tr>
<tr>
<td>probability mass function</td>
<td>48</td>
</tr>
<tr>
<td>probability measure</td>
<td>23, 28</td>
</tr>
<tr>
<td>probability model</td>
<td>28</td>
</tr>
<tr>
<td>probability space</td>
<td>23</td>
</tr>
<tr>
<td>probability transition kernel</td>
<td>62</td>
</tr>
<tr>
<td>product</td>
<td>14</td>
</tr>
<tr>
<td>product $\sigma$-algebra</td>
<td>14</td>
</tr>
<tr>
<td>product measure</td>
<td>66</td>
</tr>
<tr>
<td>product transition kernel</td>
<td>63</td>
</tr>
<tr>
<td>purely atomic</td>
<td>26</td>
</tr>
<tr>
<td>pushforward</td>
<td>27</td>
</tr>
<tr>
<td>Radon–Nikodym derivative</td>
<td>32</td>
</tr>
<tr>
<td>Radon–Nikodym theorem</td>
<td>47</td>
</tr>
<tr>
<td>random element</td>
<td>31</td>
</tr>
<tr>
<td>random experiments</td>
<td>28</td>
</tr>
<tr>
<td>random trees</td>
<td>34</td>
</tr>
<tr>
<td>random variable</td>
<td>31</td>
</tr>
<tr>
<td>Reading</td>
<td>7, 15, 23, 28, 31, 37, 46, 53, 57, 74, 92</td>
</tr>
<tr>
<td>regular conditional probability</td>
<td>67</td>
</tr>
<tr>
<td>regular version</td>
<td>67</td>
</tr>
<tr>
<td>Repeated conditioning</td>
<td>60</td>
</tr>
<tr>
<td>repeated conditioning</td>
<td>60</td>
</tr>
<tr>
<td>reservoir of randomness</td>
<td>32</td>
</tr>
<tr>
<td>restriction</td>
<td>25</td>
</tr>
<tr>
<td>sample space</td>
<td>23, 28</td>
</tr>
<tr>
<td>section</td>
<td>64</td>
</tr>
<tr>
<td>sections</td>
<td>64</td>
</tr>
<tr>
<td>semi-parametric</td>
<td>29</td>
</tr>
<tr>
<td>Separable</td>
<td>21</td>
</tr>
<tr>
<td>simple</td>
<td>19, 31</td>
</tr>
<tr>
<td>simple graph</td>
<td>34</td>
</tr>
<tr>
<td>standard</td>
<td>21</td>
</tr>
<tr>
<td>statistic</td>
<td>72</td>
</tr>
<tr>
<td>statistical model</td>
<td>29</td>
</tr>
<tr>
<td>stopping time</td>
<td>94</td>
</tr>
<tr>
<td>sub-$\sigma$-algebra</td>
<td>53</td>
</tr>
<tr>
<td>sub-Markov kernel</td>
<td>63</td>
</tr>
<tr>
<td>submartingale</td>
<td>92</td>
</tr>
<tr>
<td>subset</td>
<td>7</td>
</tr>
<tr>
<td>sufficient</td>
<td>72</td>
</tr>
<tr>
<td>supermartingale</td>
<td>92</td>
</tr>
<tr>
<td>Supplemental</td>
<td>7, 15, 23, 28, 31, 37, 46, 53, 57, 74, 92</td>
</tr>
<tr>
<td>tail $\sigma$-algebra</td>
<td>55</td>
</tr>
<tr>
<td>tail bound</td>
<td>98</td>
</tr>
<tr>
<td>the background probability space</td>
<td>31</td>
</tr>
<tr>
<td>tight</td>
<td>85</td>
</tr>
<tr>
<td>topological space</td>
<td>12</td>
</tr>
<tr>
<td>topology</td>
<td>11</td>
</tr>
<tr>
<td>trace</td>
<td>13, 25</td>
</tr>
<tr>
<td>transition kernel</td>
<td>62, 63</td>
</tr>
<tr>
<td>trivial</td>
<td>10</td>
</tr>
<tr>
<td>uniformly integrable</td>
<td>82</td>
</tr>
<tr>
<td>union</td>
<td>17</td>
</tr>
<tr>
<td>uniqueness of measures</td>
<td>26</td>
</tr>
<tr>
<td>upcrossings</td>
<td>95</td>
</tr>
<tr>
<td>variance</td>
<td>50</td>
</tr>
<tr>
<td>version</td>
<td>58</td>
</tr>
<tr>
<td>vertices</td>
<td>34</td>
</tr>
<tr>
<td>weak convergence</td>
<td>84</td>
</tr>
</tbody>
</table>