Theory of Probability

Measure theory, classical probability and stochastic analysis

Lecture Notes

by Gordan Žitković

Contents

Preface

These notes were written (and are still being heavily edited) to help students with the graduate courses Theory of Probability I and II offered by the Department of Mathematics, University of Texas at Austin.

Statements, proofs, or entire sections marked by an asterisk (∗) are not a part of the syllabus and can be skipped when preparing for midterm, final and prelim exams. .

GORDAN ŽITKOVIĆ
Austin, TX
December 2010.

Part I

Theory of Probability I

Chapter 1

Measurable spaces

Before we delve into measure theory, let us fix some notation and terminology.

- $\subseteq$ denotes a subset (not necessarily proper).

- A set A is said to be **countable** if there exists an injection (one-to-one mapping) from A into $\mathbb{N}$. Note that finite sets are also countable. Sets which are not countable are called **uncountable**.

- For two functions $f : B \to C$, $g : A \to B$, the **composition** $f \circ g : A \to C$ of f and g is given by $(f \circ g)(x) = f(g(x))$, for all $x \in A$.

- $\{A_n\}_{n\in\mathbb{N}}$ denotes a sequence. More generally, $(A_\gamma)_{\gamma\in\Gamma}$ denotes a collection indexed by the set Γ.

1.1 Families of Sets

Definition 1.1 (Order properties) *A (countable) family $\{A_n\}_{n\in\mathbb{N}}$ of subsets of a non-empty set S is said to be*

1. **increasing** *if $A_n \subseteq A_{n+1}$ for all $n \in \mathbb{N}$,*

2. **decreasing** *if $A_n \supseteq A_{n+1}$ for all $n \in \mathbb{N}$,*

3. **pairwise disjoint** *if $A_n \cap A_m = \emptyset$ for $m \neq n$,*

4. *a* **partition** *of S if $\{A_n\}_{n\in\mathbb{N}}$ is pairwise disjoint and $\cup_n A_n = S$.*

We use the notation $A_n \nearrow A$ to denote that the sequence $\{A_n\}_{n\in\mathbb{N}}$ is increasing and $A = \cup_n A_n$. Similarly, $A_n \searrow A$ means that $\{A_n\}_{n\in\mathbb{N}}$ is decreasing and $A = \cap_n A_n$.

Here is a list of some properties that a family $\mathcal{S}$ of subsets of a nonempty set S can have:

(A1) $\emptyset \in \mathcal{S}$,

(A2) $S \in \mathcal{S}$,

(A3) $A \in \mathcal{S} \Rightarrow A^c \in \mathcal{S}$,

(A4) $A, B \in \mathcal{S} \Rightarrow A \cup B \in \mathcal{S}$,

(A5) $A, B \in \mathcal{S}, A \subseteq B \Rightarrow B \setminus A \in \mathcal{S}$,

(A6) $A, B \in \mathcal{S} \Rightarrow A \cap B \in \mathcal{S}$,

(A7) $A_n \in \mathcal{S}$ for all $n \in \mathbb{N} \Rightarrow \cup_n A_n \in \mathcal{S}$,

(A8) $A_n \in \mathcal{S}$, for all $n \in \mathbb{N}$ and $A_n \nearrow A$ implies $A \in \mathcal{S}$,

(A9) $A_n \in \mathcal{S}$, for all $n \in \mathbb{N}$ and $\{A_n\}_{n \in \mathbb{N}}$ is pairwise disjoint implies $\cup_n A_n \in \mathcal{S}$,

---

Definition 1.2 (Families of sets) *A family $\mathcal{S}$ of subsets of a non-empty set S is called an*

1. **algebra** *if it satisfies (A1),(A3) and (A4),*

2. σ-**algebra** *if it satisfies (A1), (A3) and (A7)*

3. π-**system** *if it satisfies (A6),*

4. λ-**system** *if it satisfies (A2), (A5) and (A8).*

Problem 1.3 Show that:

1. Every σ-algebra is an algebra.

2. Each algebra is a π-system and each σ-algebra is an algebra and a λ-system.

3. A family $\mathcal{S}$ is a σ-algebra if and only if it satisfies (A1), (A3), (A6) and (A9).

4. A λ-system which is a π-system is also a σ-algebra.

5. There are π-systems which are not algebras.

6. There are algebras which are not σ-algebras (*Hint:* Pick all finite subsets of an infinite set. That is not an algebra yet, but sets can be added to it so as to become an algebra which is not a σ-algebra.)

7. There are λ-systems which are not π-systems.

Definition 1.4 (Generated σ-algebras) *For a family $\mathcal{A}$ of subsets of a non-empty set S, the intersection of all σ-algebras on S that contain $\mathcal{A}$ is denoted by $\sigma(\mathcal{A})$ and is called the σ-**algebra generated by** $\mathcal{A}$.*

Remark 1.5 Since the family 2^S of *all* subsets of S is a σ-algebra, the concept of a generated σ-algebra is well defined: there is always at least one σ-algebra containing $\mathcal{A}$ - namely 2^S. $\sigma(\mathcal{A})$ is itself a σ-algebra (why?) and it is the smallest (in the sense of set inclusion) σ-algebra that contains $\mathcal{A}$. In the same vein, one can define the algebra, the π-system and the λ-system generated by $\mathcal{A}$. The only important property is that intersections of σ-algebras, π-systems and λ-systems are themselves σ-algebras, π-systems and λ-systems.

Problem 1.6 Show, by means of an example, that the *union* of a family of algebras (on the same S) does not need to be an algebra. Repeat for σ-algebras, π-systems and λ-systems.

Definition 1.7 (Topology) *A **topology** on a set S is a family τ of subsets of S which contains $\emptyset$ and S and is closed under finite intersections and arbitrary (countable or uncountable!) unions. The elements of τ are often called the **open sets**. A set S on which a topology is chosen (i.e., a pair (S, τ) of a set and a topology on it) is called a **topological space**.*

Remark 1.8 Almost all topologies in these notes will be generated by a metric, i.e., a set $A \subset S$ will be open if and only if for each $x \in A$ there exists $\varepsilon > 0$ such that $\{y \in S : d(x, y) < \varepsilon\} \subseteq A$. The prime example is $\mathbb{R}$ where a set is declared open if it can be represented as a union of open intervals.

Definition 1.9 (Borel σ-algebras) *If (S, τ) is a topological space, then the σ-algebra $\sigma(\tau)$, generated by all open sets, is called the **Borel σ-algebra** on (S, τ).*

Remark 1.10 We often abuse terminology and call S itself a topological space, if the topology τ on it is clear from the context. In the same vein, we often speak of the Borel σ-algebra on a set S.

Example 1.11 Some important σ-algebras. Let S be a non-empty set:

1. The set $\mathcal{S} = 2^S$ (also denoted by $\mathcal{P}(S)$) consisting of all subsets of $\mathcal{S}$ is a σ-algebra.

2. At the other extreme, the family $\mathcal{S} = \{\emptyset, S\}$ is the smallest σ-algebra on S. It is called the **trivial σ-algebra** on $\mathcal{S}$.

3. The set $\mathcal{S}$ of all subsets of S which are either countable or whose complements are countable is a σ-algebra. It is called the **countable-cocountable σ-algebra** and is the smallest σ-algebra on S which contains all singletons, i.e., for which $\{x\} \in \mathcal{S}$ for all $x \in S$.

4. The Borel σ-algebra on $\mathbb{R}$ (generated by all open sets as defined by the Euclidean metric on $\mathbb{R}$), is denoted by $\mathcal{B}(\mathbb{R})$.

Problem 1.12 Show that the $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{A})$, for any of the following choices of the family $\mathcal{A}$:

1. $\mathcal{A} = \{$all open subsets of $\mathbb{R}$ $\}$,

2. $\mathcal{A} = \{$all closed subsets of $\mathbb{R}$ $\}$,

3. $\mathcal{A} = \{$all open intervals in $\mathbb{R}\}$,

4. $\mathcal{A} = \{$all closed intervals in $\mathbb{R}\}$,

5. $\mathcal{A} = \{$all left-closed right-open intervals in $\mathbb{R}\}$,

6. $\mathcal{A} = \{$all left-open right-closed intervals in $\mathbb{R}\}$, and

7. $\mathcal{A} = \{$all open intervals in $\mathbb{R}$ with rational end-points$\}$

8. $\mathcal{A} = \{$all intervals of the form $(-\infty, r]$, where r is rational$\}$.

(*Hint:* An arbitrary open interval $I = (a, b)$ in $\mathbb{R}$ can be written as $I = \cup_{n \in \mathbb{N}}[a + n^{-1}, b - n^{-1}]$.)

1.2 Measurable mappings

Definition 1.13 (Measurable spaces) *A pair $(S, \mathcal{S})$ consisting of a non-empty set S and a σ-algebra $\mathcal{S}$ of its subsets is called a* **measurable space**.

If $(S, \mathcal{S})$ is a measurable space, and $A \in \mathcal{S}$, we often say that A **is measurable in** $\mathcal{S}$.

Definition 1.14 (Pull-backs and push-forwards) *For a function $f : S \to T$ and subsets $A \subseteq S$, $B \subseteq T$, we define the*

1. **push-forward** $f(A)$ *of $A \subseteq S$ as*

$$f(A) = \{f(x) \, : \, x \in A\} \subseteq T,$$

2. **pull-back** $f^{-1}(B)$ *of $B \subseteq T$ as*

$$f^{-1}(B) = \{x \in S \, : \, f(x) \in B\} \subseteq S.$$

It is often the case that the notation is abused and the pull-back of B under f is denoted simply by $\{f \in B\}$. This notation presupposes, however, that the domain of f is clear from the context.

Problem 1.15 Show that the pull-back operation preserves the elementary set operations, i.e., for $f : S \to T$, and $B, \{B_n\}_{n \in \mathbb{N}} \in T$,

1. $f^{-1}(T) = S$, $f^{-1}(\emptyset) = \emptyset$,

2. $f^{-1}(\cup_n B_n) = \cup_n f^{-1}(B_n)$,

3. $f^{-1}(\cap_n B_n) = \cap_n f^{-1}(B_n)$, and

4. $f^{-1}(B^c) = [f^{-1}(B)]^c$.

Give examples showing that the push-forward analogues of the statements (1), (3) and (4) above are <u>not true</u>.

(*Note:* The assumption that the families in (2) and (3) above are countable is not necessary. Uncountable unions or intersections commute with the pull-back, too.)

Definition 1.16 (Measurability) *A mapping $f : S \to T$, where $(S, \mathcal{S})$ and $(T, \mathcal{T})$ are measurable spaces, is said to be $(\mathcal{S}, \mathcal{T})$-***measurable*** if $f^{-1}(B) \in \mathcal{S}$ for each $B \in \mathcal{T}$.*

Remark 1.17 When $T = \mathbb{R}$, we tacitly assume that the Borel σ-algebra is defined on T, and we simply call f *measurable*. In particular, a function $f : \mathbb{R} \to \mathbb{R}$, which is measurable with respect to the pair of the Borel σ-algebras is often called a **Borel function**.

Proposition 1.18 (A measurability criterion) *Let $(S, \mathcal{S})$ and $(T, \mathcal{T})$ be two measurable spaces, and let $\mathcal{C}$ be a subset of $\mathcal{T}$ such that $\mathcal{T} = \sigma(\mathcal{C})$. If $f : S \to T$ is a mapping with the property that $f^{-1}(C) \in \mathcal{S}$, for any $C \in \mathcal{C}$, then f is $(\mathcal{S}, \mathcal{T})$-measurable.*

PROOF Let $\mathcal{D}$ be the family of subsets of T defined by

$$\mathcal{D} = \{B \subset T \, : \, f^{-1}(B) \in \mathcal{S}\}.$$

By the assumptions of the proposition, we have $\mathcal{C} \subseteq \mathcal{D}$. On the other hand, by Problem 1.15, the family $\mathcal{D}$ has the structure of the σ-algebra, i.e., $\mathcal{D}$ is a σ-algebra that contains $\mathcal{C}$. Remembering that $\mathcal{T} = \sigma(\mathcal{C})$ is the *smallest* σ-algebra that contains $\mathcal{C}$, we conclude that $\mathcal{T} \subseteq \mathcal{D}$. Consequently, $f^{-1}(B) \in \mathcal{S}$ for all $B \in \mathcal{T}$. ∎

Problem 1.19 Let $(S, \mathcal{S})$ and $(T, \mathcal{T})$ be measurable spaces.

1. Suppose that S and T are topological spaces, and that $\mathcal{S}$ and $\mathcal{T}$ are the corresponding Borel σ-algebras. Show that each continuous function $f : S \to T$ is $(\mathcal{S}, \mathcal{T})$-measurable. (*Hint:* Remember that the function f is continuous if the pull-backs of open sets are open.)

2. Let $f : S \to \mathbb{R}$ be a function. Show that f is measurable if and only if

$$\{x \in S \, : \, f(x) \leq q\} \in \mathcal{S}, \text{ for all rational } q.$$

3. Find an example of $(S, \mathcal{S})$, $(T, \mathcal{T})$ and a measurable function $f : S \to T$ such that $f(A) = \{f(x) \, : \, x \in A\} \notin \mathcal{T}$ for all nonempty $A \in \mathcal{S}$.

Proposition 1.20 (Compositions of measurable maps) *Let $(S, \mathcal{S})$, $(T, \mathcal{T})$ and $(U, \mathcal{U})$ be measurable spaces, and let $f : S \to T$ and $g : T \to U$ be measurable functions. Then the composition $h = g \circ f : S \to U$, given by $h(x) = g(f(x))$ is $(\mathcal{S}, \mathcal{U})$-measurable.*

PROOF It is enough to observe that $h^{-1}(B) = f^{-1}(g^{-1}(B))$, for any $B \subseteq U$. ∎

Corollary 1.21 (Compositions with a continuous maps) *Let $(S, \mathcal{S})$ be a measurable space, T be a topological space and $\mathcal{T}$ the Borel σ-algebra on T. Let $g : T \to \mathbb{R}$ be a continuous function. Then the map $g \circ f : S \to \mathbb{R}$ is measurable for each measurable function $f : S \to T$.*

Definition 1.22 (Generation by a function) *Let $f : S \to T$ be a map from the set S into a measurable space $(T, \mathcal{T})$. The σ-algebra generated by f, denoted by $\sigma(f)$, is the intersection of all σ-algebras $\mathcal{S}$ on S which make f $(\mathcal{S}, \mathcal{T})$-measurable.*

The letter Γ will typically be used to denote an abstract index set - we only assume that it is nonempty, but make no other assumptions about its cardinality.

Definition 1.23 (Generation by several functions) *Let $(f_\gamma)_{\gamma \in \Gamma}$ be a family of maps from a set S into a measurable space $(T, \mathcal{T})$. The σ-algebra generated by $(f_\gamma)_{\gamma \in \Gamma}$, denoted by $\sigma\left((f_\gamma)_{\gamma \in \Gamma}\right)$, is the intersection of all σ-algebras on S which make each f_γ, $\gamma \in \Gamma$, measurable.*

Problem 1.24 In the setting of Definitions 1.22 and 1.23, show that

1. for $f : S \to T$, we have

   (1.1) $$\sigma(f) = \{f^{-1}(B) \: : \: B \in \mathcal{T}\}.$$

2. for a familty $f_\gamma : S \to T$, $\gamma \in \Gamma$, we have

   (1.2) $$\sigma\left((f_\gamma)_{\gamma \in \Gamma}\right) = \sigma\left(\bigcup_{\gamma \in \Gamma} f_\gamma^{-1}(\mathcal{T})\right),$$

   where $f_\gamma^{-1}(\mathcal{T}) = \{f_\gamma^{-1}(B) \: : \: B \in \mathcal{T}\}$.

(*Note:* Note how the right-hand sides differ in (1.1) and (1.2).)

1.3 Products of measurable spaces

Definition 1.25 (Products, choice functions) *Let $(S_\gamma)_{\gamma \in \Gamma}$ be a family of sets, parametrized by some (possibly uncountable) index set Γ. The **product** $\prod_{\gamma \in \Gamma} S_\gamma$ is the set of all functions $s : \Gamma \to \cup_\gamma S_\gamma$ (called **choice functions**) with the property that $s(\gamma) \in S_\gamma$.*

Remark 1.26

1. When Γ is finite, each function $\boldsymbol{s} : \Gamma \to \cup_{\gamma \in \Gamma} S_\gamma$ can be identified with an ordered "tuple" $(\boldsymbol{s}(\gamma_1), \ldots, \boldsymbol{s}(\gamma_n))$, where n is the cardinality (number of elements) of Γ, and $\gamma_1, \ldots, \gamma_n$ is some ordering of its elements. With this identification, it is clear that our definition of a product coincides with the well-known definition in the finite case.

2. The celebrated *Axiom of Choice* in set theory postulates that no matter what the family $(S_\gamma)_{\gamma \in \Gamma}$ is, there exists at least one choice function. In other words, axiom of choice simply asserts that products of sets are non-empty.

Definition 1.27 (Natural projections) *For $\gamma_0 \in \Gamma$, the function $\pi_{\gamma_0} : \prod_{\gamma \in \Gamma} S_\gamma \to S_{\gamma_0}$ defined by*

$$\pi_{\gamma_0}(\boldsymbol{s}) = \boldsymbol{s}(\gamma_0), \textit{ for } \boldsymbol{s} \in \prod_{\gamma \in \Gamma} S_\gamma,$$

is called the **(natural) projection onto the coordinate** γ_0.

Definition 1.28 (Products of measurable spaces) *Let $\{(S_\gamma, \mathcal{S}_\gamma)\}_{\gamma \in \Gamma}$ be a family of measurable spaces. The* **product** $\otimes_{\gamma \in \Gamma}(S_\gamma, \mathcal{S}_\gamma)$ *is a measurable space $(\prod_{\gamma \in \Gamma} S_\gamma, \otimes_{\gamma \in \Gamma} \mathcal{S}_\gamma)$, where $\otimes_{\gamma \in \Gamma} \mathcal{S}_\gamma$ is the smallest σ-algebra that makes all natural projections $(\pi_\gamma)_{\gamma \in \Gamma}$ measurable.*

Example 1.29 When Γ is finite, the above definition can be made more intuitive. Suppose, just for simplicity, that $\Gamma = \{1, 2\}$, so that $(S_1, \mathcal{S}_1) \otimes (S_2, \mathcal{S}_2)$ is a measurable space of the form $(S_1 \times S_2, \mathcal{S}_1 \otimes \mathcal{S}_2)$, where $\mathcal{S}_1 \otimes \mathcal{S}_2$ is the smallest σ-algebra on $S_1 \times S_2$ which makes both π_1 and π_2 measurable. The pull-backs under π_1 of sets in $\mathcal{S}_1$ are given by

$$\pi^{-1}(B_1) = \{(x, y) \in S_1 \times S_2 \; : \; x \in B_1\} = B_1 \times S_2, \text{ for } B_1 \in \mathcal{S}_1.$$

Similarly

$$\pi^{-1}(B_2) = S_1 \times B_2, \text{ for } B_2 \in \mathcal{S}_2.$$

Therefore, by Problem 1.24,

$$\mathcal{S}_1 \otimes \mathcal{S}_2 = \sigma\Big(\{B_1 \times S_2, S_1 \times B_2 \; : \; B_1 \in \mathcal{S}_1, B_2 \in \mathcal{S}_2\}\Big).$$

Equivalently (why?)

$$\mathcal{S}_1 \otimes \mathcal{S}_2 = \sigma\Big(\{B_1 \times B_2 \; : \; B_1 \in \mathcal{S}_1, B_2 \in \mathcal{S}_2\}\Big).$$

In a completely analogous fashion, we can show that, for finitely many measurable spaces $(S_1, \mathcal{S}_1), \ldots, (S_n, \mathcal{S}_n)$, we have

$$\bigotimes_{i=1}^{n} \mathcal{S}_i = \sigma\Big(\{B_1 \times B_2 \times \cdots \times B_n \; : \; B_1 \in \mathcal{S}_1, B_2 \in \mathcal{S}_2, \ldots, B_n \in \mathcal{S}_n\}\Big)$$

The same goes for countable products. Uncountable products, however, behave very differently.

Problem 1.30 We know that the Borel σ-algebra (based on the usual Euclidean topology) can be constructed on each $\mathbb{R}^n$. A σ-algebra on $\mathbb{R}^n$ (for $n > 1$), can also be constructed as a product σ-algebra $\otimes_{i=1}^n \mathcal{B}(\mathbb{R})$. A third possibility is to consider the mixed case where $1 < m < n$ is picked and the σ-algebra $\mathcal{B}(\mathbb{R}^m) \otimes \mathcal{B}(\mathbb{R}^{n-m})$ is constructed on $\mathbb{R}^n$ (which is now interpreted as a product of $\mathbb{R}^m$ and $\mathbb{R}^{n-m}$). Show that we get the same σ-algebra in all three cases.

Problem 1.31 Let $(P, \mathcal{P})$, $\{(S_\gamma, \mathcal{S}_\gamma)\}_{\gamma \in \Gamma}$ be measurable spaces and set $S = \prod_{\gamma \in \Gamma} S_\gamma$, $\mathcal{S} = \otimes_{\gamma \in \Gamma} \mathcal{S}_\gamma$. Prove that a map $f : P \to S$ is $(\mathcal{P}, \mathcal{S})$-measurable if and only if the composition $\pi_\gamma \circ f : P \to S_\gamma$ is $(\mathcal{P}, \mathcal{S}_\gamma)$ measurable for each $\gamma \in \Gamma$. (*Note:* Loosely speaking, this result states that a "vector"-valued mapping is measurable if and only if all of its components are measurable.)

Definition 1.32 (Cylinder sets) *Let $\{(S_\gamma, \mathcal{S}_\gamma)\}_{\gamma \in \Gamma}$ be a family of measurable spaces, and let $(\prod_{\gamma \in \Gamma} S_\gamma, \otimes_{\gamma \in \Gamma} \mathcal{S}_\gamma)$ be its product. A subset $C \subseteq \prod_{\gamma \in \Gamma} S_\gamma$ is called a* **cylinder set** *if there exist a finite subset $\{\gamma_1, \ldots, \gamma_n\}$ of Γ, as well as a measurable set $B \in \mathcal{S}_{\gamma_1} \otimes \mathcal{S}_{\gamma_2} \otimes \cdots \otimes \mathcal{S}_{\gamma_n}$ such that*

$$C = \{s \in \prod_{\gamma \in \Gamma} S_\gamma \ : \ (s(\gamma_1), \ldots, s(\gamma_n)) \in B\}.$$

A cylinder set for which the set B can be chosen of the form $B = B_1 \times \cdots \times B_n$, for some $B_1 \in \mathcal{S}_1, \ldots, B_n \in \mathcal{S}_n$ is called a **product cylinder set**. *In that case*

$$C = \{s \in \prod_{\gamma \in \Gamma} S_\gamma \ : \ (s(\gamma_1) \in B_1, s(\gamma_2) \in B_2, \ldots, s(\gamma_n) \in B_n\}.$$

Problem 1.33

1. Show that the family of product cylinder sets generates the product σ-algebra.

2. Show that (not-necessarily-product) cylinders are measurable in the product σ-algebra.

3. Which of the 4 families of sets from Definition 1.2 does the collection of all product cylinders belong to in general? How about (not-necessarily-product) cylinders?

Example 1.34 The following example will play a major role in probability theory. Hence the name **coin-toss space**. Here $\Gamma = \mathbb{N}$ and for $i \in \mathbb{N}$, $(S_i, \mathcal{S}_i)$ is the discrete two-element space $S_i = \{-1, 1\}$, $\mathcal{S}_i = 2^{S_1}$. The product $\prod_{i \in \mathbb{N}} S_i = \{-1, 1\}^{\mathbb{N}}$ can be identified with the set of all sequences $s = (s_1, s_2, \ldots)$, where $s_i \in \{-1, 1\}$, $i \in \mathbb{N}$. For each cylinder set C, there exists (why?) $n \in \mathbb{N}$ and a subset B of $\{-1, 1\}^n$ such that

$$C = \{s = (s_1, \ldots, s_n, s_{n+1}, \ldots) \in \{-1, 1\}^{\mathbb{N}} \ : \ (s_1, \ldots, s_n) \in B\}.$$

The product cylinders are even simpler - they are always of the form $C = \{-1, 1\}^{\mathbb{N}}$ or $C = C_{n_1, \ldots, n_k; b_1, \ldots, b_k}$, where

(1.3) $$C_{n_1, \ldots, n_k; b_1, \ldots, b_k} = \left\{s = (s_1, s_2, \ldots) \in \{-1, 1\}^{\mathbb{N}} \ : \ s_{n_1} = b_1, \ldots, s_{n_k} = b_k\right\},$$

for some $k \in \mathbb{N}$, $1 \le n_1 < n_2 < \cdots < n_k \in \mathbb{N}$ and $b_1, b_2, \ldots, b_k \in \{-1, 1\}$.

We know that the σ-algebra $\mathcal{S} = \otimes_{i \in \mathbb{N}} \mathcal{S}_i$ is generated by all projections $\pi_i : \{-1, 1\}^{\mathbb{N}} \to \{-1, 1\}$, $i \in \mathbb{N}$, where $\pi_i(\boldsymbol{s}) = s_i$. Equivalently, by Problem 1.33, $\mathcal{S}$ is generated by the collection of all cylinder sets.

Problem 1.35 One can obtain the product σ-algebra $\mathcal{S}$ on $\{-1, 1\}^{\mathbb{N}}$ as the Borel σ-algebra corresponding to a particular topology which makes $\{-1, 1\}^{\mathbb{N}}$ compact. Here is how. Start by defining a mapping $d : \{-1, 1\}^{\mathbb{N}} \times \{-1, 1\}^{\mathbb{N}} \to [0, \infty)$ by

(1.4) $$d(\boldsymbol{s}^1, \boldsymbol{s}^2) = 2^{-i(\boldsymbol{s}^1, \boldsymbol{s}^2)}, \text{ where } i(\boldsymbol{s}^1, \boldsymbol{s}^2) = \inf\{i \in \mathbb{N} : s_i^1 \neq s_i^2\},$$

for $\boldsymbol{s}^j = (s_1^j, s_2^j, \dots)$, $j = 1, 2$.

1. Show that d is a metric on $\{-1, 1\}^{\mathbb{N}}$.

2. Show that $\{-1, 1\}^{\mathbb{N}}$ is compact under d. (*Hint:* Use the diagonal argument.)

3. Show that each cylinder of $\{-1, 1\}^{\mathbb{N}}$ is both open and closed under d.

4. Show that each open ball is a cylinder.

5. Show that $\{-1, 1\}^{\mathbb{N}}$ is separable, i.e., it admits a countable dense subset.

6. Conclude that $\mathcal{S}$ coincides with the Borel σ-algebra on $\{-1, 1\}^{\mathbb{N}}$ under the metric d.

1.4 Real-valued measurable functions

Let $\mathcal{L}^0(S, \mathcal{S}; \mathbb{R})$ (or, simply, $\mathcal{L}^0(S; \mathbb{R})$ or $\mathcal{L}^0(\mathbb{R})$ or $\mathcal{L}^0$ when the domain $(S, \mathcal{S})$ or the co-domain $\mathbb{R}$ are clear from the context) be the set of all $\mathcal{S}$-measurable functions $f : S \to \mathbb{R}$. The set of non-negative measurable functions is denoted by $\mathcal{L}^0_+$ or $\mathcal{L}^0([0, \infty))$.

Proposition 1.36 (Measurable functions form a vector space) $\mathcal{L}^0$ *is a vector space, i.e.*

$$\alpha f + \beta g \in \mathcal{L}^0, \text{ whenever } \alpha, \beta \in \mathbb{R}, \, f, g \in \mathcal{L}^0.$$

PROOF Let us define a mapping $F : S \to \mathbb{R}^2$ by $F(x) = (f(x), g(x))$. By Problem 1.30, the Borel σ-algebra on $\mathbb{R}^2$ is the same as the product σ-algebra when we interpret $\mathbb{R}^2$ as a product of two copies of $\mathbb{R}$. Therefore, since its compositions with the coordinate projections are precisely the functions f and g, Problem 1.31 implies that F is $(\mathcal{S}, \mathcal{B}(\mathbb{R}^2))$-measurable.

Consider the function $\phi : \mathbb{R}^2 \to \mathbb{R}$ given by $\phi(x, y) = \alpha x + \beta y$. It is linear, and, therefore, continuous. By Corollary 1.21, the composition $\phi \circ F : \mathcal{S} \to \mathbb{R}$ is $(\mathcal{S}, \mathcal{B}(\mathbb{R}))$-measurable, and it only remains to note that

$$(\phi \circ F)(x) = \phi(F(x)) = \alpha f(x) + \beta g(x), \text{ i.e., } \phi \circ F = \alpha f + \beta g. \qquad \blacksquare$$

In a similar manner (the functions $(x, y) \mapsto \max(x, y)$ and $(x, y) \to xy$ are continuous from $\mathbb{R}^2$ to $\mathbb{R}$ - why?) one can prove the following proposition.

Proposition 1.37 (Products and maxima preserve measurability) *Let f, g be in $\mathcal{L}^0$. Then*

1. $fg \in \mathcal{L}^0$,

2. $\max(f, g)$ *and* $\min(f, g) \in \mathcal{L}^0$,

Even though the map $x \mapsto 1/x$ is not defined on the whole $\mathbb{R}$, the following problem is not too hard:

Problem 1.38 Suppose that $f \in \mathcal{L}^0$ has the property that $f(x) \neq 0$ for all $x \in S$. Then the function $1/f$ is also in $\mathcal{L}^0$.

For $A \subseteq S$, the **indicator function** $\mathbf{1}_A$ is defined by

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

Despite their simplicity, indicators will be extremely useful throughout these notes.

Problem 1.39 Show that for $A \subset S$, we have $A \in \mathcal{S}$ if and only if $\mathbf{1}_A \in \mathcal{L}^0$.

Remark 1.40 Since it contains the products of pairs of its elements, the set $\mathcal{L}^0$ has the structure of an *algebra* (not to be confused with the algebra of sets defined above). It is true, however, that any algebra $\mathcal{A}$ of subsets of a non-empty set S, together with the operations of union, intersection and complement forms a *Boolean algebra*. Alternatively, it can be given the (algebraic) structure of a *commutative ring with a unit*. Indeed, under the operation $\triangle$ of symmetric difference, $\mathcal{A}$ is an Abelian group (prove that!). If, in addition, the operation of intersection is introduced in lieu of multiplication, the resulting structure is, indeed, the one of a commutative ring.

Additionally, a natural partial order given by $f \preceq g$ if $f(x) \leq g(x)$, for all $x \in S$, can be introduced on $\mathcal{L}^0$. This order is compatible with the operations of addition and multiplication and has the additional property that each pair $\{f, g\} \subseteq \mathcal{L}^0$ admits a *least upper bound*, i.e., the element $h \in \mathcal{L}^0$ such that $f \preceq h$, $g \preceq h$ and $h \preceq k$, for any other k with the property that $f, g \preceq k$. Indeed, we simply take $h(x) = \max(f(x), g(x))$. A similar statement can be made for a *greatest lower bound*. A vector space with a partial order which satisfies the above properties is called a *vector lattice*.

Since a limit of a sequence of real numbers does not necessarily belong to $\mathbb{R}$, it is often necessary to consider functions which are allowed to take the values ∞ and $-\infty$. The set $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty, -\infty\}$ is called the **extended set of real numbers**. Most (but not all) of the algebraic and topological structure from $\mathbb{R}$ can be lifted to $\bar{\mathbb{R}}$. In some cases there is no unique way to do that, so we choose one of them as a matter of convention.

1. **Arithmetic operations.** For $x, y \in \bar{\mathbb{R}}$, all the arithmetic operations are defined in the usual way when $x, y \in \mathbb{R}$. When one or both are in $\{\infty, -\infty\}$, we use the following convention, where $\oplus \in \{+, -, *, /\}$:

   We define $x \oplus y = z$ if all pairs of sequences $\{x_n\}_{n \in \mathbb{N}}$, $\{y_n\}_{n \in \mathbb{N}}$ in $\mathbb{R}$ such that $x = \lim_n x_n$, $y = \lim_n y_n$ and $x_n \oplus y_n$ is well-defined for all $n \in \mathbb{N}$, we have

   $$z = \lim_n (x_n \oplus y_n).$$

Otherwise, $x \oplus y$ is not defined. This basically means that all intuitively obvious conventions (such as $\infty + \infty = \infty$ and $\frac{a}{0} = \infty$ for $a > 0$ hold). In measure theory, however, we do make one important *exception* to the above rule. We set

$$0 \times \infty = \infty \times 0 = 0 \times (-\infty) = (-\infty) \times 0 = 0.$$

2. **Order.** $-\infty < x < \infty$, for all $x \in \mathbb{R}$. Also, *each* non-empty subset of $\bar{\mathbb{R}}$ admits a supremum and an infimum in $\bar{\mathbb{R}}$ in an obvious way.

3. **Convergence.** It is impossible to extend the usual (Euclidean) metric from $\mathbb{R}$ to $\bar{\mathbb{R}}$, but a metric $d' : \mathbb{R} \times \mathbb{R} \to [0, 1)$ given by

$$d'(x, y) = |\arctan(y) - \arctan(x)|,$$

extends readily to a metric on $\bar{\mathbb{R}}$ if we set $\arctan(\infty) = \pi/2$ and $\arctan(-\infty) = -\pi/2$. We define convergence (and topology) on $\mathbb{R}$ using d'. For example, a sequence $\{x_n\}_{n \in \mathbb{N}}$ in $\bar{\mathbb{R}}$ converges to $+\infty$ if

   a) It contains only a finite number of terms equal to $-\infty$,

   b) Every subsequence of $\{x_n\}_{n \in \mathbb{N}}$ whose elements are in $\mathbb{R}$ converges to $+\infty$ (in the usual sense).

We define the notions of **limit superior** and **limit inferior** on $\bar{\mathbb{R}}$ for a sequence $\{x_n\}_{n \in \mathbb{N}}$ in the following manner:

$$\limsup_n x_n = \inf_n S_n, \text{ where } S_n = \sup_{k \geq n} x_k,$$

and

$$\liminf_n x_n = \sup_n I_n, \text{ where } I_n = \inf_{k \geq n} x_k.$$

If you have forgotten how to manipulate limits inferior and superior, here is an exercise to remind you:

Problem 1.41 Let $\{x_n\}_{n \in \mathbb{N}}$ be a sequence in $\bar{\mathbb{R}}$. Prove the following statements:

1. $a \in \bar{\mathbb{R}}$ satisfies $a \geq \limsup_n x_n$ if and only if for any $\varepsilon \in (0, \infty)$ there exists $n_\varepsilon \in \mathbb{N}$ such that $x_n \leq a + \varepsilon$ for $n \geq n_\varepsilon$.

2. $\liminf_n x_n \leq \limsup_n x_n$.

3. Define
$$A = \{\lim_k x_{n_k} : x_{n_k} \text{ is a convergent (in } \bar{\mathbb{R}}) \text{ subsequence of } \{x_n\}_{n \in \mathbb{N}}\}.$$

   Show that
   $$\{\liminf_n x_n, \limsup_n x_n\} \subseteq A \subseteq [\liminf_n x_n, \limsup_n x_n].$$

Having introduced a topology on $\bar{\mathbb{R}}$ we immediately have the σ-algebra $\mathcal{B}(\bar{\mathbb{R}})$ of Borel sets there and the notion of measurability for functions mapping a measurable space $(S, \mathcal{S})$ into $\bar{\mathbb{R}}$.

Problem 1.42 Show that a subset $A \subseteq \bar{\mathbb{R}}$ is in $\mathcal{B}(\bar{\mathbb{R}})$ if and only if $A \setminus \{\infty, -\infty\}$ is Borel in $\mathbb{R}$. Show that a function $f : S \to \bar{\mathbb{R}}$ is measurable in the pair $(S, \mathcal{B}(\bar{\mathbb{R}}))$ if and only if the sets $f^{-1}(\{\infty\})$, $f^{-1}(\{-\infty\})$ and $f^{-1}(A)$ are in $\mathcal{S}$ for all $A \in \mathcal{B}(\mathbb{R})$ (equivalently and more succinctly, $f \in \mathcal{L}^0(\bar{\mathbb{R}})$ iff $\{f = \infty\}, \{f = -\infty\} \in \mathcal{S}$ and $f\mathbf{1}_{\{f \in \mathbb{R}\}} \in \mathcal{L}^0$).

The set of all measurable functions $f : S \to \bar{\mathbb{R}}$ is denoted by $\mathcal{L}^0(S, \mathcal{S}; \bar{\mathbb{R}})$, and, as always we leave out S and $\mathcal{S}$ when no confusion can arise. The set of extended non-negative measurable functions often plays a role, so we denote it by $\mathcal{L}^0([0, \infty])$ or $\mathcal{L}^0_+(\bar{\mathbb{R}})$. Unlike $\mathcal{L}^0(\mathbb{R})$, $\mathcal{L}^0(\bar{\mathbb{R}})$ is not a vector space, but it retains all the order structure. Moreover, it is particularly useful because, unlike $\mathcal{L}^0(\mathbb{R})$, it is closed with respect to the limiting operations. More precisely, for a sequence $\{f_n\}_{n \in \mathbb{N}}$ in $\mathcal{L}^0(\bar{\mathbb{R}})$, we define the functions $\limsup_n f_n : S \to [-\infty, \infty]$ and $\liminf_n f_n : S \to [-\infty, \infty]$ by

$$(\limsup_n f_n)(x) = \limsup_n f_n(x) = \inf_n \left(\sup_{k \geq n} f_k(x) \right),$$

and

$$(\liminf_n f_n)(x) = \liminf_n f_n(x) = \sup_n \left(\inf_{k \geq n} f_k(x) \right).$$

Then, we have the following result, where the supremum and infimum of a sequence of functions are defined pointwise (just like the limits superior and inferior).

Proposition 1.43 (Limiting operations preserve measurability) *Let $\{f_n\}_{n \in \mathbb{N}}$ be a sequence in $\mathcal{L}^0(\bar{\mathbb{R}})$. Then*

1. *$\sup_n f_n, \inf_n f_n \in \mathcal{L}^0(\bar{\mathbb{R}})$,*

2. *$\limsup_n f_n, \liminf_n f_n \in \mathcal{L}^0(\bar{\mathbb{R}})$,*

3. *if $f(x) = \lim_n f_n(x)$ exists in $\bar{\mathbb{R}}$ for each $x \in S$, then $f \in \mathcal{L}^0(\bar{\mathbb{R}})$, and*

4. *the set $A = \{\lim_n f_n$ exists in $\bar{\mathbb{R}}\}$ is in $\mathcal{S}$.*

PROOF

1. We show only the statement for the supremum. It is clear that it is enough to show that the set $\{\sup_n f_n \leq a\}$ is in $\mathcal{S}$ for all $a \in (-\infty, \infty]$ (why?). This follows, however, directly from the simple identity
$$\{\sup_n f_n \leq a\} = \cap_n \{f_n \leq a\},$$
and the fact that σ-algebras are closed with respect to countable intersections.

2. Define $g_n = \sup_{k \geq n} f_k$ and use part 1. above to conclude that $g_n \in \mathcal{L}^0(\bar{\mathbb{R}})$ for each $n \in \mathbb{N}$. Another appeal to part 1. yields that $\limsup_n f_n = \inf_n g_n$ is in $\mathcal{L}^0(\bar{\mathbb{R}})$. The statement about the limit inferior follows in the same manner.

3. If the limit $f(x) = \lim_n f_n(x)$ exists for all $x \in S$, then $f = \liminf_n f_n$ which is measurable by part 2. above.

4. The statement follows from the fact that $A = f^{-1}(\{0\})$, where
$$f(x) = \arctan\left(\limsup_n f_n(s)\right) - \arctan\left(\liminf_n f_n(x)\right).$$

(*Note:* The unexpected use of the function arctan is really noting to be puzzled by. The only property needed is its measurability (it is continuous) and monotonicity+bijectivity from

$[-\infty, \infty]$ to $[-\pi/2, \pi/2]$. We compose the limits superior and inferior with it so that we don't run into problems while trying to subtract $+\infty$ from itself.) ∎

1.5 Additional Problems

Problem 1.44 Which of the following are σ-algebras on $\mathbb{R}$?

1. $\mathcal{S} = \{A \subseteq \mathbb{R} \,:\, 0 \in A\}$.

2. $\mathcal{S} = \{A \subseteq \mathbb{R} \,:\, A \text{ is finite}\}$.

3. $\mathcal{S} = \{A \subseteq \mathbb{R} \,:\, A \text{ is finite, or } A^c \text{ is finite}\}$.

4. $\mathcal{S} = \{A \subseteq \mathbb{R} \,:\, A \text{ is countable or } A^c \text{ is countable}\}$.

5. $\mathcal{S} = \{A \subseteq \mathbb{R} \,:\, A \text{ is open}\}$.

6. $\mathcal{S} = \{A \subset \mathbb{R} \,:\, A \text{ is open or } A \text{ is closed}\}$.

Problem 1.45 A **partition** S is a family $\mathcal{P}$ of non-empty subsets of S with the property that each $\omega \in S$ belongs to exactly one $A \in \mathcal{P}$.

1. Show that the number of different algebras on a finite set S is equal to the number of different partitions of S. (*Note:* This number for $S_n = \{1, 2, \ldots, n\}$ is called the nth **Bell number** B_n, and no nice closed-form expression for it is known. See below, though.)

2. How many algebras are there on the set $S = \{1, 2, 3\}$?

3. Does there exist an algebra with 754 elements?

4. For $N \in \mathbb{N}$, let a_n be the number of different algebras on the set $\{1, 2, \ldots, n\}$. Show that $a_1 = 1$, $a_2 = 2$, $a_3 = 5$, and that the following recursion holds (where $a_0 = 1$ by definition),

$$a_{n+1} = \sum_{k=0}^{n} \binom{n}{k} a_k.$$

5. Show that the exponential generating function for the sequence $\{a_n\}_{n \in \mathbb{N}}$ is $f(x) = e^{e^x - 1}$, i.e., that

$$\sum_{n=0}^{\infty} a_n \frac{x^n}{n!} = e^{e^x - 1} \text{ or, equivalently, } a_n = \left(\tfrac{d^n}{dx^n} e^{e^x - 1}\right)\Big|_{x=0}.$$

Problem 1.46 Let $(S, \mathcal{S})$ be a measurable space. For $f, g \in \mathcal{L}^0$ show that the sets $\{f = g\} = \{x \in S \,:\, f(x) = g(x)\}$, $\{f < g\} = \{x \in S \,:\, f(x) < g(x)\}$ are in $\mathcal{S}$.

Problem 1.47 Show that all

1. monotone,

2. convex

functions $f : \mathbb{R} \to \mathbb{R}$ are measurable.

Problem 1.48 Let $(S, \mathcal{S})$ be a measurable space and let $f : S \to \mathbb{R}$ be a Borel-measurable function. Show that the graph

$$G_f = \{(x, y) \in S \times \mathbb{R} \; : \; f(x) = y\},$$

of f is a measurable subset in the product space $(S \times \mathbb{R}, \mathcal{S} \otimes \mathcal{B}(\mathbb{R}))$.

Chapter 2

Measures

Definition 2.1 (Measure) *Let $(S, \mathcal{S})$ be a measurable space. A mapping $\mu : \mathcal{S} \to [0, \infty]$ is called a* **(positive) measure** *if*

1. *$\mu(\emptyset) = 0$, and*

2. *$\mu(\cup_n A_n) = \sum_{n \in \mathbb{N}} \mu(A_n)$, for all* pairwise disjoint *sequences $\{A_n\}_{n \in \mathbb{N}}$ in $\mathcal{S}$.*

A triple $(S, \mathcal{S}, \mu)$ consisting of a non-empty set, a σ-algebra $\mathcal{S}$ on it and a measure μ on $\mathcal{S}$ is called a **measure space***.*

Remark 2.2

1. A mapping whose domain is some nonempty set $\mathcal{A}$ of subsets of some set S is sometimes called a **set function**.

2. If the requirement 2. in the definition of the measure is weakened so that it is only required that $\mu(A_1 \cup \cdots \cup A_n) = \mu(A_1) + \cdots + \mu(A_n)$, for $n \in \mathbb{N}$, and pairwise disjoint $A_1, \ldots, A_n$, we say that the mapping μ is a **finitely-additive measure**. If we want to stress that a mapping μ satisfies the original requirement 2. for *sequences* of sets, we say that μ is σ**-additive** (**countably additive**).

Definition 2.3 (Terminology) *A measure μ on the measurable space $(S, \mathcal{S})$ is called*

1. *a* **probability measure***, if $\mu(S) = 1$,*

2. *a* **finite measure***, if $\mu(S) < \infty$,*

3. *a σ-**finite measure**, if there exists a sequence $\{A_n\}_{n \in \mathbb{N}}$ in $\mathcal{S}$ such that $\cup_n A_n = S$ and $\mu(A_n) < \infty$,*

4. **measure** *or* **atom-free***, if* $\mu(\{x\}) = 0$*, whenever* $x \in S$ *and* $\{x\} \in \mathcal{S}$*.*

A set $N \in \mathcal{S}$ *is said to be* **null** *if* $\mu(N) = 0$*.*

Example 2.4 (Examples of measures) Let S be a non-empty set, and let $\mathcal{S}$ be a σ-algebra on S.

1. **Measures on countable sets.** Suppose that S is a finite or countable set. Then each measure μ on $\mathcal{S} = 2^S$ is of the form

$$\mu(A) = \sum_{x \in A} p(x),$$

for some function $p : S \to [0, \infty]$ (why?). In particular, for a finite set S with N elements, if $p(x) = 1/N$ then μ is a probability measure called the **uniform measure on** S. It has the property that $\mu(A) = \frac{\#A}{\#S}$, where $\#$ denotes the cardinality (number of elements).

2. **Dirac measure.** For $x \in S$, we define the set function δ_x on $\mathcal{S}$ by

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

It is easy to check that δ_x is indeed a measure on $\mathcal{S}$. Alternatively, δ_x is called the **point mass at** x (or an **atom on** x, or the **Dirac function**, even though it is not really a function). Moreover, δ_x is a probability measure and, therefore, a finite and a σ-finite measure. It is atom free only if $\{x\} \notin \mathcal{S}$.

3. **Counting Measure.** Define a set function $\mu : \mathcal{S} \to [0, \infty]$ by

$$\mu(A) = \begin{cases} \#A, & A \text{ is finite}, \\ \infty, & A \text{ is infinite}, \end{cases} \quad \text{for } A \in \mathcal{S},$$

where, as above, $\#A$ denotes the number of elements in the set A. Again, it is not hard to check that μ is a measure - it is called the **counting measure**. Clearly, μ is a finite measure if and only is S is a finite set. μ could be σ-finite, though, even without S being finite. Simply take $S = \mathbb{N}, \mathcal{S} = 2^{\mathbb{N}}$. In that case $\mu(S) = \infty$, but for $A_n = \{n\}, n \in \mathbb{N}$, we have $\mu(A_n) = 1$, and $S = \cup_n A_n$. Finally, μ is never atom-free and it is a probability measure only if $\text{card } S = 1$.

Example 2.5 (A finitely-additive set function which is not a measure) Let $S = \mathbb{N}$, and $\mathcal{S} = 2^S$. For $A \in S$ define $\mu(A) = 0$ if A is finite and $\mu(A) = \infty$, otherwise. For $A_1, \ldots, A_n \subseteq S$, we have the following two possibilities:

1. A_i is finite, for each $i = 1, \ldots, n$. Then $\cup_{i=1}^n A_i$ is also finite and so $0 = \mu(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i)$.

2. at least one A_i is infinite. Then $\cup_{i=1}^n A_i$ is also infinite and so $\infty = \mu(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i)$, because $\mu(A_i) = \infty$.

Therefore, μ is finitely additive.

On the other hand, take $A_i = \{i\}$, for $i \in \mathbb{N}$. Then $\mu(A_i) = 0$, for each $i \in \mathbb{N}$, and, so, $\sum_{i \in \mathbb{N}} \mu(A_i) = 0$, but $\mu(\cup_i A_i) = \mu(\mathbb{N}) = \infty$.

(*Note:* It is possible to construct very simple-looking finite-additive measures which are not σ-additive. For example, there exist $\{0,1\}$-valued finitely-additive measures on all subsets of $\mathbb{N}$, which are not σ-additive. Such objects are called **ultrafilters** and their existence is equivalent to a certain version of the Axiom of Choice.)

Proposition 2.6 (First properties of measures) *Let $(S, \mathcal{S}, \mu)$ be a measure space.*

1. *For $A_1, \ldots, A_n \in \mathcal{S}$ with $A_i \cap A_j = \emptyset$, for $i \neq j$, we have*

$$\sum_{i=1}^{n} \mu(A_i) = \mu(\cup_{i=1}^{n} A_i).$$

(Finite additivity)

2. *If $A, B \in \mathcal{S}$, $A \subseteq B$, then*
$$\mu(A) \leq \mu(B).$$

(Monotonicity of measures)

3. *If $\{A_n\}_{n \in \mathbb{N}}$ in $\mathcal{S}$ is increasing, then*

$$\mu(\cup_n A_n) = \lim_n \mu(A_n) = \sup_n \mu(A_n).$$

(Continuity with respect to increasing sequences)

4. *If $\{A_n\}_{n \in \mathbb{N}}$ in $\mathcal{S}$ is decreasing and $\mu(A_1) < \infty$, then*

$$\mu(\cap_n A_n) = \lim_n \mu(A_n) = \inf_n \mu(A_n).$$

(Continuity with respect to decreasing sequences)

5. *For a sequence $\{A_n\}_{n \in \mathbb{N}}$ in $\mathcal{S}$, we have*

$$\mu(\cup_n A_n) \leq \sum_{n \in \mathbb{N}} \mu(A_n).$$

(Subadditivity w.r.t. general sequences)

PROOF

1. Note that the sequence $A_1, A_2, \ldots, A_n, \emptyset, \emptyset, \ldots$ is pairwise disjoint, and so, by σ-additivity,

$$\mu(\cup_{i=1}^{n} A_i) = \mu(\cup_{i \in \mathbb{N}} A_i) = \sum_{i \in \mathbb{N}} \mu(A_i) = \sum_{i=1}^{n} \mu(A_i) + \sum_{i=n+1}^{\infty} \mu(\emptyset) = \sum_{i=1}^{n} \mu(A_i).$$

2. Write B as a disjoint union $A \cup (B \setminus A)$ of elements of $\mathcal{S}$. By (1) above,

$$\mu(B) = \mu(A) + \mu(B \setminus A) \geq \mu(A).$$

3. Define $B_1 = A_1$, $B_n = A_n \setminus A_{n-1}$ for $n > 1$. Then $\{B_n\}_{n \in \mathbb{N}}$ is a pairwise disjoint sequence in $\mathcal{S}$ with $\cup_{k=1}^{n} B_k = A_n$ for each $n \in \mathbb{N}$ (why?). By σ-additivity we have

$$\mu(\cup_n A_n) = \mu(\cup_n B_n) = \sum_{n \in \mathbb{N}} \mu(B_n) = \lim_n \sum_{k=1}^{n} \mu(B_k) = \lim_n \mu(\cup_{k=1}^{n} B_k) = \lim_n \mu(A_n).$$

4. Consider the increasing sequence $\{B_n\}_{n \in \mathbb{N}}$ in $\mathcal{S}$ given by $B_n = A_1 \setminus A_n$. By De Morgan laws, finiteness of $\mu(A_1)$ and (3) above, we have

$$\mu(A_1) - \mu(\cap_n A_n) = \mu(A_1 \setminus (\cap_n A_n)) = \mu(\cup_n B_n) = \lim_n \mu(B_n) = \lim_n \mu(A_1 \setminus A_n)$$
$$= \mu(A_1) - \lim_n \mu(A_n).$$

Subtracting both sides from $\mu(A_1) < \infty$ produces the statement.

5. We start from the observation that for $A_1, A_1 \in \mathcal{S}$ the set $A_1 \cup A_2$ can be written as a disjoint union

$$A_1 \cup A_2 = (A_1 \setminus A_2) \cup (A_2 \setminus A_1) \cup (A_1 \cap A_2),$$

so that

$$\mu(A_1 \cup A_2) = \mu(A_1 \setminus A_2) + \mu(A_2 \setminus A_1) + \mu(A_1 \cap A_2).$$

On the other hand,

$$\mu(A_1) + \mu(A_2) = (\mu(A_1 \setminus A_2) + \mu(A_1 \cap A_2)) + \Big(\mu(A_2 \setminus A_1) + \mu(A_1 \cap A_2)\Big)$$
$$= \mu(A_1 \setminus A_2) + \mu(A_2 \setminus A_1) + 2\mu(A_1 \cap A_2),$$

and so

$$\mu(A_1) + \mu(A_2) - \mu(A_1 \cup A_2) = \mu(A_1 \cap A_2) \geq 0.$$

Induction can be used to show that

$$\mu(A_1 \cup \cdots \cup A_n) \leq \sum_{k=1}^{n} \mu(A_k).$$

Since all $\mu(A_n)$ are nonnegative, we now have

$$\mu(A_1 \cup \cdots \cup A_n) \leq \alpha, \text{ for each } n \in \mathbb{N}, \text{ where } \alpha = \sum_{n \in \mathbb{N}} \mu(A_n).$$

The sequence $\{B_n\}_{n \in \mathbb{N}}$ given by $B_n = \cup_{k=1}^{n} A_k$ is increasing, so the continuity of measure with respect to increasing sequences implies that

$$\mu(\cup_n A_n) = \mu(\cup_n B_n) = \lim_n \mu(B_n) = \lim_n \mu(A_1 \cup \cdots \cup A_n) \leq \alpha. \qquad \blacksquare$$

Remark 2.7 The condition $\mu(A_1) < \infty$ in the part (4) of Proposition 2.6 cannot be significantly relaxed. Indeed, let μ be the counting measure on $\mathbb{N}$, and let $A_n = \{n, n+1, \dots\}$. Then $\mu(A_n) = \infty$ and, so $\lim_n \mu(A_n) = \infty$. On the other hand, $\cap A_n = \emptyset$, so $\mu(\cap_n A_n) = 0$.

In addition to unions and intersections, one can produce other important new sets from sequences of old ones. More specifically, let $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of subsets of S. The subset $\liminf_n A_n$ of S, defined by

$$\liminf_n A_n = \cup_n B_n, \text{ where } B_n = \cap_{k \geq n} A_k,$$

is called the **limit inferior** of the sequence A_n. It is also denoted by $\underline{\lim}_n A_n$ or $\{A_n, \text{ ev.}\}$ (*ev.* stands for *eventually*). The reason for this last notation is the following: $\liminf_n A_n$ is the set of all $x \in S$ which belong to A_n for *all but finitely many values* of the index n.

Similarly, the subset $\limsup_n A_n$ of S, defined by

$$\limsup_n A_n = \cap_n B_n, \text{ where } B_n = \cup_{k \geq n} A_k,$$

is called the **limit superior** of the sequence A_n. It is also denoted by $\overline{\lim}_n A_n$ or $\{A_n, \text{ i.o.}\}$ (*i.o.* stands for *infinitely often*). In words, $\limsup_n A_n$ is the set of all $x \in S$ which belong A_n for *infinitely many values* of n. Clearly, we have

$$\liminf_n A_n \subseteq \limsup_n A_n.$$

Problem 2.8 Let $(S, \mathcal{S}, \mu)$ be a finite measure space. Show that

$$\mu(\liminf A_n) \leq \liminf \mu(A_n) \leq \limsup \mu(A_n) \leq \mu(\limsup A_n),$$

for any sequence $\{A_n\}_{n \in \mathbb{N}}$ in $\mathcal{S}$. Give an example of a (single) sequence $\{A_n\}_{n \in \mathbb{N}}$ for which all inequalities above are strict.

(*Hint:* For the second part, a measure space with finite (and small) S will do.)

Proposition 2.9 (Borel-Cantelli Lemma I) *Let $(S, \mathcal{S}, \mu)$ be a measure space, and let $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of sets in $\mathcal{S}$ with the property that $\sum_{n \in \mathbb{N}} \mu(A_n) < \infty$. Then*

$$\mu(\limsup_n A_n) = 0.$$

PROOF Set $B_n = \cup_{k \geq n} A_k$, so that $\{B_n\}_{n \in \mathbb{N}}$ is a decreasing sequence of sets in $\mathcal{S}$ with $\limsup_n A_n = \cap_n B_n$. Using the subadditivity of measures of Proposition 2.6, part 5., we get

$$(2.1) \qquad\qquad \mu(B_n) \leq \sum_{k=n}^{\infty} \mu(A_n).$$

Since $\sum_{n \in \mathbb{N}} \mu(A_n)$ converges, the right-hand side of (2.1) can be made arbitrarily small by choosing large enough $n \in \mathbb{N}$. Hence $\mu(\limsup_n A_n) = 0$. ∎

2.2 Extensions of measures and the coin-toss space

Example 1.34 of Chapter 1 has introduced a measurable space $(\{-1,1\}^{\mathbb{N}}, \mathcal{S})$, where $\mathcal{S}$ is the product σ-algebra on $\{-1,1\}^{\mathbb{N}}$. The purpose of the present section is to turn $(\{-1,1\}^{\mathbb{N}}, \mathcal{S})$ into a measure space, i.e., to define a suitable measure on it. It is easy to construct just any measure on $\{-1,1\}^{\mathbb{N}}$, but the one we are after is the one which will justify the name *coin-toss space*.

The intuition we have about tossing a fair coin infinitely many times should help us start with the definition of the coin-toss measure - denoted by μ_C - on cylinders. Since the coordinate spaces $\{-1,1\}$ are particularly simple, each product cylinder is of the form $C = \{-1,1\}^{\mathbb{N}}$ or $C = C_{n_1,\dots,n_k;b_1,\dots,b_k}$, as given by (1.3), for a choice $1 \leq n_1 < n_2 < \dots, n_k \in \mathbb{N}$ of coordinates and the corresponding values $b_1, \dots, b_k \in \{-1,1\}$. In the language of elementary probability, each cylinder corresponds to the event when the outcome of the n_i-th coin is $b_i \in \{-1,1\}$, for $k = 1, \dots, n$. The measure (probability) of this event can only be given by

$$\text{(2.2)} \qquad \mu_C(C_{n_1,\dots,n_k;b_1,\dots,b_k}) = \underbrace{\tfrac{1}{2} \times \tfrac{1}{2} \times \cdots \times \tfrac{1}{2}}_{k \text{ times}} = 2^{-k}.$$

The hard part is to extend this definition to *all* elements of $\mathcal{S}$, and not only cylinders. For example, in order to state the law of large numbers later on, we will need to be able to compute the measure of the set

$$\left\{ s \in \{-1,1\}^{\mathbb{N}} : \lim_{n} \tfrac{1}{n} \sum_{k=1}^{n} s_k = \tfrac{1}{2} \right\},$$

which is clearly not a cylinder.

Problem 1.33 states, however, that cylinders form an algebra and generate the σ-algebra $\mathcal{S}$. Luckily, this puts us close to the conditions of the following important theorem of Caratheodory. The proof does not use unfamiliar methodology, but we omit it because it is quite long and tricky.

Theorem 2.10 (Caratheodory's Extension Theorem) *Let S be a non-empty set, let $\mathcal{A}$ be an algebra of its subsets and let $\mu : \mathcal{A} \to [0,\infty]$ be a set-function with the following properties:*

1. *$\mu(\emptyset) = 0$, and*

2. *$\mu(A) = \sum_{n=1}^{\infty} \mu(A_n)$, if $\{A_n\}_{n\in\mathbb{N}}$ is a pairwise-disjoint family in $\mathcal{A}$ and $A = \cup_n A_n \in \mathcal{A}$.*

Then, there exists a measure $\tilde{\mu}$ on $\sigma(\mathcal{A})$ with the property that $\mu(A) = \tilde{\mu}(A)$ for $A \in \mathcal{A}$.

Remark 2.11 In words, a σ-additive measure on an algebra $\mathcal{A}$ can be extended to a σ-additive measure on the σ-algebra generated by $\mathcal{A}$. It is clear that the σ-additivity requirement of Theorem 2.10 is necessary, but it is quite surprising that it is actually sufficient.

In order to apply Theorem 2.10 in our situation, we need to check that μ is indeed a countably-additive measure on the algebra $\mathcal{A}$ of all cylinders. The following problem will help pinpoint the hard part of the argument:

Problem 2.12 Let $\mathcal{A}$ be an algebra on the non-empty set S, and let $\mu : \mathcal{A} \to [0,\infty]$ be a finite ($\mu(S) < \infty$) and finitely-additive set function on $\mathcal{S}$ with the following, additional, property:

$$\text{(2.3)} \qquad \lim_{n} \mu(A_n) = 0, \text{ whenever } A_n \searrow \emptyset.$$

Then μ is σ-additive on $\mathcal{A}$, i.e., it satisfies the conditions of Theorem 2.10.

The part about finite additivity is easy (but messy) and we leave it to the reader:

Problem 2.13 Show that the set-function μ_C, defined by (2.2) on the algebra $\mathcal{A}$ of cylinders, is finitely additive.

Lemma 2.14 (Conditions of Caratheodory's theorem) *The the set-function μ_C, defined by (2.2) on the algebra $\mathcal{A}$ of cylinders, has the property* (2.3).

PROOF By Problem 1.35, cylinders are closed sets, and so $\{A_n\}_{n\in\mathbb{N}}$ is a sequence of closed sets whose intersection is empty. The same problem states that $\{-1,1\}^{\mathbb{N}}$ is compact, so, by the finite-intersection property[1], we have $A_{n_1} \cap \ldots A_{n_k} = \emptyset$, for some finite collection $n_1, \ldots, n_k$ of indices. Since $\{A_n\}_{n\in\mathbb{N}}$ is decreasing, we must have $A_n = \emptyset$, for all $n \geq n_k$, and, consequently, $\lim_n \mu(A_n) = 0$. ∎

Proposition 2.15 (Existence of the coin-toss measure) *There exists a measure μ_C on $(\{-1,1\}^{\mathbb{N}}, \mathcal{S})$ with the property that (2.2) holds for all cylinders.*

PROOF Thanks to Lemma 2.14, Theorem 2.10 can now be used. ∎

In order to prove uniqueness, we will need the celebrated π-λ Theorem of Eugene Dynkin:

Theorem 2.16 (Dynkin's "π-λ" Theorem) *Let $\mathcal{P}$ be a π-system on a non-empty set S, and let Λ be a λ-system which contains $\mathcal{P}$. Then Λ also contains the σ-algebra $\sigma(\mathcal{P})$ generated by $\mathcal{P}$.*

PROOF Using the result of part 4. of Problem 1.3, we only need to prove that $\lambda(\mathcal{P})$ (where $\lambda(\mathcal{P})$ denotes the λ-system generated by $\mathcal{P}$) is a π-system. For $A \subseteq S$, let $\mathcal{G}_A$ denote the family of all subsets of S whose intersections with A are in $\lambda(\mathcal{P})$:

$$\mathcal{G}_A = \{C \subseteq S \,:\, C \cap A \in \lambda(\mathcal{P})\}.$$

Claim 1: $\mathcal{G}_A$ is a λ-system for $A \in \lambda(\mathcal{P})$.

- Since $A \in \lambda(\mathcal{P})$, clearly $S \in \mathcal{G}_A$.

- For an increasing family $\{C_n\}_{n\in\mathbb{N}}$ in $\mathcal{G}_A$ we have $(\cup_n C_n) \cap A = \cup_n(C_n \cap A)$. Each $C_n \cap A$ is in Λ, and the family $\{C_n \cap A\}_{n\in\mathbb{N}}$ is increasing, so $(\cup_n C_n) \cap A \in \Lambda$.

---

[1]The *finite-intersection property* refers to the following fact, familiar from real analysis: If a family of closed sets of a compact topological space has empty intersection, then it admits a *finite* subfamily with an empty intersection.

- Finally, for $C_1, C_2 \in \mathcal{G}$ with $C_1 \subseteq C_2$, we have

$$(C_2 \setminus C_1) \cap A = (C_2 \cap A) \setminus (C_1 \cap A) \in \Lambda,$$

because $C_1 \cap A \subseteq C_2 \cap A$.

Since $\mathcal{P}$ is a π-system, for any $A \in \mathcal{P}$, we have $\mathcal{P} \subseteq \mathcal{G}_A$. Therefore, $\lambda(\mathcal{P}) \subseteq \mathcal{G}_A$, because $\mathcal{G}_A$ is a λ-system. In other words, for $A \in \mathcal{P}$ and $B \in \lambda(\mathcal{P})$, we have $A \cap B \in \lambda(\mathcal{P})$.

That means, however, that $\mathcal{P} \subseteq \mathcal{G}_B$, for any $B \in \lambda(\mathcal{P})$. Using the fact that $\mathcal{G}_B$ is a λ-system we must also have $\lambda(\mathcal{P}) \subseteq \mathcal{G}_B$, for any $B \in \lambda(\mathcal{P})$, i.e., $A \cap B \in \lambda(\mathcal{P})$, for all $A, B \in \lambda(\mathcal{P})$, which shows that $\lambda(\mathcal{P})$ is π-system. ∎

Proposition 2.17 (Measures which agree on a π-system) *Let $(S, \mathcal{S})$ be a measurable space, and let $\mathcal{P}$ be a π-system which generates $\mathcal{S}$. Suppose that μ_1 and μ_2 are two measures on $\mathcal{S}$ with the property that $\mu_1(S) = \mu_2(S) < \infty$ and*

$$\mu_1(A) = \mu_2(A), \text{ for all } A \in \mathcal{P}.$$

Then $\mu_1 = \mu_2$, i.e., $\mu_1(A) = \mu_2(A)$, for all $A \in \mathcal{S}$.

PROOF Let $\mathcal{L}$ be the family of all subsets A of $\mathcal{S}$ for which $\mu_1(A) = \mu_2(A)$. Clearly $\mathcal{P} \subseteq \mathcal{L}$, but $\mathcal{L}$ is, potentially, bigger. In fact, it follows easily from the elementary properties of measures (see Proposition 2.6) and the fact that $\mu_1(S) = \mu_2(S) < \infty$ that it necessarily has the structure of a λ-system. By Theorem 2.16 (the π-λ Theorem), $\mathcal{L}$ contains the σ-algebra generated by $\mathcal{P}$, i.e., $\mathcal{S} \subseteq \mathcal{L}$. On the other hand, by definition, $\mathcal{L} \subseteq \mathcal{S}$ and so $\mu_1 = \mu_2$. ∎

Remark 2.18 It seems that the structure of a λ-system is defined so that it would exactly describe the structure of the family of all sets on which two measures (with the same total mass) agree. The structure of the π-system corresponds to the minimal assumption that allows Proposition 2.17 to hold.

Proposition 2.19 (Uniqueness of the coin-toss measure) *The measure μ_C is the unique measure on $(\{-1, 1\}^{\mathbb{N}}, \mathcal{S})$ with the property that (2.2) holds for all cylinders.*

PROOF The existence is the content of Proposition 2.15. To prove uniqueness, it suffices to note that algebras are π-systems and use Proposition 2.17. ∎

Problem 2.20 Define $D_1, D_2 \subseteq \{-1, 1\}^{\mathbb{N}}$ by

1. $D_1 = \{s \in \{-1, 1\}^{\mathbb{N}} : \limsup_n s_n = 1\}$,

2. $D_2 = \{s \in \{-1, 1\}^{\mathbb{N}} : \exists N \in \mathbb{N}, \ s_N = s_{N+1} = s_{N+2}\}$.

Show that $D_1, D_2 \in \mathcal{S}$ and compute $\mu(D_1), \mu(D_2)$.

Our next task is to probe the structure of the σ-algebra $\mathcal{S}$ on $\{-1,1\}^{\mathbb{N}}$ a little bit more and show that $\mathcal{S} \neq 2^{\{-1,1\}^{\mathbb{N}}}$. It is interesting that such a result (which deals exclusively with the structure of $\mathcal{S}$) requires a use of a measure in its proof.

Example 2.21 ((*) A non-measurable subset of $\{-1,1\}^{\mathbb{N}}$) Since σ-algebras are closed under countable set operations, and since the product σ-algebra $\mathcal{S}$ for the coin-toss space $\{-1,1\}^{\mathbb{N}}$ is generated by sets obtained by restricting finite collections of coordinates, one is tempted to think that $\mathcal{S}$ contains *all* subsets of $\{-1,1\}^{\mathbb{N}}$. That is not the case. We will use the axiom of choice, together with the fact that a measure μ_C can be defined on the whole of $\{-1,1\}^{\mathbb{N}}$, to show to "construct" an example of a non-measurable set.

Let us start by constructing a relation $\sim$ on $\{-1,1\}^{\mathbb{N}}$ in the following way: we set $s^1 \sim s^2$ if and only if there exists $n \in \mathbb{N}$ such that $s_k^1 = s_k^2$, for $k \geq n$ (here, as always, $s^i = (s_1^i, s_2^i, \ldots)$, $i = 1, 2$). In words, s^1 and s^2 are related if they only differ in a finite number of coordinates. It is easy to check that $\sim$ is an equivalence relation and that it splits $\{-1,1\}^{\mathbb{N}}$ into disjoint equivalence classes. One of the many equivalent forms of the axiom of choice states that there exists a subset N of $\{-1,1\}^{\mathbb{N}}$ which contains exactly one element from each of the equivalence classes.

Let us suppose that N is an element in $\mathcal{S}$ and see if we can reach a contradiction. Let F denote the set of all finite subsets of $\mathbb{N}$. For each nonempty $\boldsymbol{n} = \{n_1, \ldots, n_k\} \in F$, let us define the mapping $T_{\boldsymbol{n}} : \{-1,1\}^{\mathbb{N}} \to \{-1,1\}^{\mathbb{N}}$ in the following manner:

$$
(T_{\boldsymbol{n}}(\boldsymbol{s}))_l = \begin{cases} s_l, & l \in \boldsymbol{n}, \\ -s_l, & l \notin \boldsymbol{n}. \end{cases}
$$

In words, $T_{\boldsymbol{n}}$ flips the signs of the elements of its argument on the positions corresponding to $\boldsymbol{n}$. We define $T_{\emptyset} = \mathrm{Id}$, i.e., $T_{\emptyset}(\boldsymbol{s}) = \boldsymbol{s}$.

Since $\boldsymbol{n}$ is finite, T_f preserves the $\sim$-equivalence class of each element. Consequently (and using the fact that N contains exactly one element from each equivalence class) the sets N and $T_{\boldsymbol{n}}(N) = \{T_{\boldsymbol{n}}(\boldsymbol{s}) : \boldsymbol{s} \in N\}$ are disjoint. Similarly and more generally, the sets $T_{\boldsymbol{n}}(N)$ and $T_{\boldsymbol{n}'}(N)$ are also disjoint whenever $\boldsymbol{n} \neq \boldsymbol{n}'$. On the other hand, each $\boldsymbol{s} \in \{-1,1\}^{\mathbb{N}}$ is equivalent to some $\hat{\boldsymbol{s}} \in N$, i.e., it can be obtained from $\hat{\boldsymbol{s}}$ by flipping a finite number of coordinates. Therefore, the family

$$
\mathcal{N} = \{T_{\boldsymbol{n}}(N) : \boldsymbol{n} \in F\}
$$

forms a partition of $\{-1,1\}^{\mathbb{N}}$.

The mapping $T_{\boldsymbol{n}}$ has several other nice properties. First of all, it is immediate that it is involutory, i.e., $T_{\boldsymbol{n}} \circ T_{\boldsymbol{n}} = \mathrm{Id}$. To show that it is $(\mathcal{S}, \mathcal{S})$-measurable, we need to prove that its composition with each projection map $\pi_k : \mathcal{S} \to \{-1,1\}$ is measurable. This follows immediately from the fact that for $k \in \mathbb{N}$

$$
(T_{\boldsymbol{n}} \circ \pi_k)^{-1}(\{1\}) = \begin{cases} C_{k;1}, & k \notin \boldsymbol{n}, \\ C_{k;-1}, & k \in \boldsymbol{n}, \end{cases}
$$

where. for $i \in \{-1,1\}$, $C_{k;i} = \{\boldsymbol{s} \in \{-1,1\}^{\mathbb{N}} : s_k = i\}$ - a cylinder. If we combine the involutivity and measurability of $T_{\boldsymbol{n}}$, we immediately conclude that $T_{\boldsymbol{n}}(A) \in \mathcal{S}$ for each $A \in \mathcal{S}$. In particular, $\mathcal{N} \subseteq \mathcal{S}$.

In addition to preserving measurability, the map $T_{\boldsymbol{n}}$ also preserves the measure[2] the in μ_C, i.e., $\mu_C(T_{\boldsymbol{n}}(A)) = \mu_C(A)$, for all $A \in \mathcal{S}$. To prove that, let us pick $\boldsymbol{n} \in F$ and consider the set-function

---

[2]Actually, we say that a map f from a measure space $(S, \mathcal{S}, \mu_S)$ to the measure space $(T, \mathcal{T}, \mu_T)$ is **measure preserving** if it is measurable and $\mu_S(f^{-1}(A)) = \mu_T(A)$, for all $A \in \mathcal{T}$. The involutivity of the map $T_{\boldsymbol{n}}$ implies that this general definition agrees with our usage in this example.

$\mu_{\boldsymbol{n}} : \mathcal{S} \to [0, 1]$ given by

$$\mu_{\boldsymbol{n}}(A) = \mu_C(T_{\boldsymbol{n}}(A)).$$

It is a simple matter to show that $\mu_{\boldsymbol{n}}$ is, in fact, a measure on $(S, \mathcal{S})$ with $\mu_{\boldsymbol{n}}(S) = 1$. Moreover, thanks to the simple form (2.2) of the action of the measure μ_C on cylinders, it is clear that $\mu_{\boldsymbol{n}} = \mu_C$ on the algebra of all cylinders. It suffices to invoke Proposition 2.17 to conclude that $\mu_{\boldsymbol{n}} = \mu_C$ on the entire $\mathcal{S}$, i.e., that $T_{\boldsymbol{n}}$ preserves μ_C.

The above properties of the maps $T_{\boldsymbol{n}}$, $\boldsymbol{n} \in F$ can imply the following: N is a partition of $\mathcal{S}$ into countably many measurable subsets of equal measure. Such a partition $\{N_1, N_2, \dots\}$ cannot exist, however. Indeed if it did, one of the following two cases would occur:

1. $\mu(N_1) = 0$. In that case $\mu(S) = \mu(\cup_k N_k) = \sum_n \mu(N_k) = \sum_n 0 = 0 \neq 1 = \mu(S)$.

2. $\mu(N_1) = \alpha > 0$. In that case $\mu(S) = \mu(\cup_k N_k) = \sum_n \mu(N_k) = \sum_n \alpha = \infty \neq 1 = \mu(S)$.

Therefore, the set N cannot be measurable in $\mathcal{S}$.

(*Note:* Somewhat heavier set-theoretic machinery can be used to prove that most of the subsets of S are not in $\mathcal{S}$, in the sense that the cardinality of the set $\mathcal{S}$ is strictly smaller than the cardinality of the set 2^S of all subsets of S)

2.3 The Lebesgue measure

As we shall see, the coin-toss space can be used as a sort of a universal measure space in probability theory. We use it here to construct the Lebesgue measure on $[0, 1]$. We start with the notion somewhat dual to the already introduced notion of the pull-back in Definition 1.14. We leave it as an exercise for the reader to show that the set function $f_*(\mu)$ from Definition 2.22 is indeed a measure.

Definition 2.22 (Push-forwards) *Let $(S, \mathcal{S}, \mu)$ be a measure space and let $(T, \mathcal{T})$ be a measurable space. The measure $f_*(\mu)$ on $(T, \mathcal{T})$, defined by*

$$f_*\mu(B) = \mu(f^{-1}(B)), \text{ for } B \in \mathcal{T},$$

is called the **push-forward** *of the measure μ by f.*

Let $f : \mathcal{S} \to [0, 1]$ be the mapping given by

$$f(\boldsymbol{s}) = \sum_{k=1}^{\infty} \left(\tfrac{1+s_k}{2} \right) 2^{-k}, \ \boldsymbol{s} \in \{-1, 1\}^{\mathbb{N}}.$$

The idea is to use f to establish a correspondence between all real numbers in $[0, 1]$ and their expansions in the binary system, with the coding $-1 \mapsto 0$ and $1 \mapsto 1$. It is interesting to note that f is not one-to-one[3] , as it, for example, maps $\boldsymbol{s}_1 = (1, -1, -1, \dots)$ and $\boldsymbol{s}_2 = (-1, 1, 1, \dots)$ into the same value - namely $\frac{1}{2}$. Let us show, first, that the map f is continuous in the metric d

---

[3]The reason for this is, poetically speaking, that $[0, 1]$ is not the Cantor set.

defined by part (1.4) of Problem 1.33. Indeed, we pick s_1 and s_2 in $\{-1, 1\}^{\mathbb{N}}$ and remember that for $d(s_1, s_2) = 2^{-n}$, the first $n - 1$ coordinates of s_1 and s_2 coincide. Therefore,

$$|f(s_1) - f(s_2)| \leq \sum_{k=n}^{\infty} 2^{-k} = 2^{-n+1} = 2d(s_1, s_2).$$

Hence, the map f is Lipschitz and, therefore, continuous.

The continuity of f (together with the fact that $\mathcal{S}$ is the Borel σ-algebra for the topology induced by the metric d) implies that $f : (\{-1, 1\}^{\mathbb{N}}, \mathcal{S}) \to ([0, 1], \mathcal{B}([0, 1]))$ is a measurable mapping. Therefore, the push-forward $\lambda = f_*(\mu)$ is well defined on $([0, 1], \mathcal{B}([0, 1]))$, and we call it the **Lebesgue measure** on $[0, 1]$.

Proposition 2.23 (Intuitive properties of the Lebesgue measure) *The Lebesgue measure λ on $([0, 1], \mathcal{B}([0, 1]))$ satisfies*

(2.4) $$\lambda([a, b)) = b - a, \ \lambda(\{a\}) = 0 \ for \ 0 \leq a < b \leq 1.$$

PROOF

1. Consider a, b of the form $b = \frac{k}{2^n}$ and $b = \frac{k+1}{2^n}$, for $n \in \mathbb{N}$ and $k < 2^n$. For such a, b we have $f^{-1}([a, b)) = C_{1,\ldots,n;c_1,c_2,\ldots,c_n}$, where $\overline{c_1 c_2 \ldots c_n}$ is the base-2 expansion of k (after the "recoding" $-1 \mapsto 0$, $1 \mapsto 1$). By the very definition of λ and the form (2.2) of the action of the coin-toss measure μ_C on cylinders, we have

$$\lambda\Big([a, b)\Big) = \mu_C\Big(f^{-1}([a, b))\Big) = \mu_C(C_{1,\ldots,n;c_1,c_2,\ldots,c_n}) = 2^{-n} = \tfrac{k+1}{2^n} - \tfrac{k}{2^n}.$$

Therefore, (2.4) holds for a, b of the form $b = \frac{k}{2^n}$ and $b = \frac{l}{2^n}$, for $n \in \mathbb{N}$, $k < 2^n$ and $l = k + 1$. Using (finite) additivity of λ, we immediately conclude that (2.4) holds for all k, l, i.e., that it holds for all dyadic rationals. A general $a \in (0, 1]$ can be approximated by an increasing sequence $\{q_n\}_{n \in \mathbb{N}}$ of dyadic rationals from the left, and the continuity of measures with respect to decreasing sequences implies that

$$\lambda\Big([a, p)\Big) = \lambda\Big(\cap_n [q_n, p)\Big) = \lim_n \lambda\Big([q_n, p)\Big) = \lim_n (p - q_n) = (p - a),$$

whenever $a \in (0, 1]$ and p is a dyadic rational. In order to remove the dyadicity requirement from the right limit, we approximate it from the left by a sequence $\{p_n\}_{n \in \mathbb{N}}$ of dyadic rationals with $p_n > a$, and use the continuity with respect to increasing sequences to get, for $a < b \in (0, 1)$,

$$\lambda\Big([a, b)\Big) = \lambda\Big(\cup_n [a, p_n)\Big) = \lim_n \lambda\Big([a, p_n)\Big) = \lim_n (p_n - a) = (b - a). \quad \blacksquare$$

The Lebesgue measure has another important property:

Problem 2.24 Show that the Lebesgue measure is **translation invariant**. More precisely, for $B \in \mathcal{B}([0, 1])$ and $x \in [0, 1)$, we have

1. $B +_1 x = \{b + x \;(\mathrm{mod}\, 1)\; :\; b \in B\}$ is in $\mathcal{B}([0, 1])$ and

2. $\lambda(B +_1 x) = \lambda(B)$,

where, for $a \in [0, 2)$, we define $a \;(\mathrm{mod}\, 1) = \begin{cases} a, & a \leq 1, \\ a - 1, & a > 1 \end{cases}$. Geometrically, the set $x +_1 B$ is obtained from B by translating it to the right by x and then shifting the part that is "sticking out" by 1 to the left.) (*Hint:* Use Proposition 2.17 for the second part.)

Finally, the notion of the Lebesgue measure is just as useful on the entire $\mathbb{R}$, as on its compact subset $[0, 1]$. For a general $B \in \mathcal{B}(\mathbb{R})$, we can define the Lebesgue measure of B by measuring its intersections with all intervals of the form $[n, n + 1)$, and adding them together, i.e.,

$$\lambda(B) = \sum_{n=-\infty}^{\infty} \lambda\Big(\big(B \cap [n, n + 1)\big) - n\Big).$$

Note how we are overloading the notation and using the letter λ for both the Lebesgue measure on $[0, 1]$ and the Lebesgue measure on $\mathbb{R}$.

It is a quite tedious, but does not require any new tools, to show that many of the properties of λ on $[0, 1]$ transfer to λ on $\mathbb{R}$:

Problem 2.25 Let λ be the Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Show that

1. $\lambda([a, b)) = b - a$, $\lambda(\{a\}) = 0$ for $a < b$,

2. λ is σ-finite but not finite,

3. $\lambda(B + x) = \lambda(B)$, for all $B \in \mathcal{B}(\mathbb{R})$ and $x \in \mathbb{R}$, where $B + x = \{b + x \; : \; b \in B\}$.

Remark 2.26 The existence of the Lebesgue measure allows to show quickly that the converse of the implication in the Borel-Cantelli Lemma does not hold without additional conditions, even if μ is a probability measure. Indeed, let $\mu = \lambda$ be the Lebesgue measure on $[0, 1]$.

Set $A_n = (0, \frac{1}{n}]$, for $n \in \mathbb{N}$ so that

$$\limsup_n A_n = \bigcap_n \bigcup_{k \geq n} A_k = \bigcap_n A_n = \emptyset,$$

which implies that $\mu(\limsup_n A_n) = 0$. On the other hand

$$\sum_{n \in \mathbb{N}} \mu(A_n) = \sum_{n \in \mathbb{N}} \frac{1}{n} = \infty.$$

We will see later that the converse does hold if the family of sets $\{A_n\}_{n \in \mathbb{N}}$ satisfy the additional condition of independence.

2.4   Signed measures

In addition to (positive) measures, it is sometimes useful to know a few things about measure-like set functions which take values in $\bar{\mathbb{R}}$ (and not in $[0, \infty]$).

Definition 2.27 (Signed measures) *Let $(S, \mathcal{S})$ be a measurable space. A mapping $\mu : \mathcal{S} \to (-\infty, \infty]$ is called a **signed measure** (or **real measure**) if*

1. *$\mu(\emptyset) = 0$, and*

2. *for any pairwise disjoint sequence $\{A_n\}_{n \in \mathbb{N}}$ in $\mathcal{S}$ the series $\sum_n \mu(A_n)$ is summable and $\mu(\cup_n A_n) = \sum_n \mu(A_n)$.*

The notion of convergence here is applied to sequences that may take the value ∞, so we need to be precise about how it is defined. Remember that $a^+ = \max(a, 0)$ and $a^- = \max(-a, 0)$.

Definition 2.28 (Summability for sequences) *A sequence $\{a_n\}_{n \in \mathbb{N}}$ in $(-\infty, \infty]$ is said to be **summable** if $\sum_{n \in \mathbb{N}} a_n^- < \infty$. In that case, the **sum of the series** $\sum_{n \in \mathbb{N}} a_n$ is the (well-defined) extended real number $\sum_{n \in \mathbb{N}} a_n^+ - \sum_{n \in \mathbb{N}} a_n^- \in (-\infty, \infty]$.*

Remark 2.29

1. Simply put, a series with elements in $(-\infty, \infty]$ is summable if the sum of the sub-series of its negative elements is finite. The problem we are trying to avoid is, of course, the one involving $\infty - \infty$. It is easy to show that in the case of a real-valued series, the notion of summability coincides with the notion of absolute convergence.

2. The fact that we are allowing the signed measure to take the value ∞ and not the value $-\infty$ is entirely arbitrary. A completely parallel theory can be built with the opposite choice. What cannot be dealt with in a really meaningful way is the case when both ∞ and $-\infty$ are allowed.

Definition 2.30 (Finite measurable partitions) *A finite collection $\{B_1, \dots, B_n\}$ of measurable subsets of $A \in \mathcal{S}$ is said to be a **finite measurable partition of A** if*

$$B_i \cap B_j = \emptyset \text{ for } i \neq j, \text{ and } A = \cup_i B_i.$$

The family of all finite measurable partitions of the set A is denoted by $P_{[0,\infty)}(A)$.

Definition 2.31 (Total variation) *For $A \in \mathcal{S}$ and a signed measure μ on $\mathcal{S}$, we define the number $|\mu|(A) \in [0, \infty]$, called the **total variation of μ on** A by*

$$|\mu|(A) = \sup_{\{D_1, D_2, \dots, D_n\} \in P_{[0,\infty)}(A)} \sum_{k=1}^{n} |\mu(D_k)|,$$

where the supremum is taken over all finite measurable partitions $D_1, \ldots, D_n$, $n \in \mathbb{N}$ of S. The number $|\mu|(S) \in [0, \infty]$ is called the **total variation (norm) of** μ*.*

The central result about signed measures is the following:

Theorem 2.32 (Hahn-Jordan decomposition) *Let $(S, \mathcal{S})$ be a measure space, and let μ be a signed measure on $\mathcal{S}$. Then there exist two (positive) measures μ^+ and μ^- such that*

1. *μ^- is finite,*

2. *$\mu(A) = \mu^+(A) - \mu^-(A)$,*

3. *$|\mu|(A) = \mu^+(A) + \mu^-(A)$,*

Measures μ^+ and μ^- with the above properties are unique. Moreover, there exists a set $D \in \mathcal{S}$ such that $\mu^+(A) = \mu(A \cap D^c)$ and $\mu^-(A) = -\mu(A \cap D)$ for all $A \in \mathcal{S}$.

PROOF **(*)** Call a set $B \in \mathcal{S}$ *negative* if $\mu(C) \leq 0$, for all $C \in \mathcal{S}$, $C \subseteq B$. Let $\mathcal{P}$ be the collection of all negative sets - it is nonempty because $\emptyset \in \mathcal{P}$. Set

$$\beta = \inf\{\mu(B) \ : \ B \in \mathcal{P}\},$$

and let $\{B_n\}_{n \in \mathbb{N}}$ be a sequence of negative sets with $\mu(B_n) \to \beta$. We define $D = \cup_n B_n$ and note that D is a negative set with $\mu(D) = \beta$ (why?). In particular, $\beta > -\infty$.

Our first order of business is to show that D^c is a *positive* set, i.e. that $\mu(E) \geq 0$ for all $E \subseteq D^c$. Suppose, to the contrary, that $\mu(B) < 0$, for some $B \in \mathcal{S}$, $B \subseteq D^c$. The set B cannot be a negative set - otherwise $D \cup B$ would be a negative set with $\mu(D \cup B) = \mu(D) + \mu(B) = \beta + \mu(B) < \beta$. Therefore, there exists a measurable subset E_1 of B with $\mu(E_1) > 0$, i.e., the set

$$\mathcal{E}_1 = \{E \subseteq B \ : \ E \in \mathcal{S}, \ \mu(E) > 0\}$$

is non-empty. Pick $k_1 \in \mathbb{N}$ such that

$$\tfrac{1}{k_1} \geq \sup\{\mu(E) \ : \ E \in \mathcal{E}_1\},$$

and an "almost-maximal" set $E_1 \in \mathcal{E}_1$ with

$$\tfrac{1}{k_1} \geq \mu(E_1) > \tfrac{1}{k_1+1}.$$

We set $B_1 = B \setminus E_1$ and observe that, since $0 > \mu(E_1) > -\infty$, we have

$$\mu(B_1) = \mu(B) - \mu(E_1) < 0,$$

and so $\mu(B_1) < 0$. Replacing B by B_1, the above discussion can be repeated and a constant k_2 and the an "almost maximal" E_2 with $\tfrac{1}{k_2} \geq \mu(E_2) > \tfrac{1}{k_2+1}$ can be constructed. Continuing in the same manner, we obtain the sequence $\{E_n\}_{n \in \mathbb{N}}$ of pairwise disjoint subsets of B with $\tfrac{1}{k_n} \geq \mu(E_n) > \tfrac{1}{k_n+1}$.

Given that $\mu(B) < 0$, it cannot have subsets of measure ∞. Therefore, $\mu(\cup_n E_n) < \infty$ and

$$\sum_{n \in \mathbb{N}} \frac{1}{k_n + 1} < \sum_{n \in \mathbb{N}} \mu(E_n) = \mu(\cup_n E_n) < \infty,$$

and so $k_n \to 0$, as $n \to \infty$.

Let $F \in \mathcal{S}$ be a subset of $B \setminus \cup_n E_n$. Then, it a subset of $B \setminus \cup_{k=1}^{n} E_k$, and, therefore, by construction, $\mu(F) \leq \frac{1}{k_n}$. The fact that $k_n \to \infty$ now implies that $\mu(F) \leq 0$, which, in turn, implies that $B \setminus \cup_n E_n$ is a negative set. The set D is, however, the maximal negative set, and so $\mu(B \setminus \cup_n E_n) = 0$. On the other hand,

$$\mu(B_0 \setminus \cup_n E_n) = \mu(B_0) - \sum_n \mu(E_n) > \mu(B_0) > 0,$$

a contradiction. Therefore, D^c is a positive set.

Having split S into a disjoint union of a positive and a negative set, we define

$$\mu^+(A) = \mu(A \cap D^c) \text{ and } \mu^-(A) = -\mu(A \cap D),$$

so that both μ^+ and μ^- are (positive measures) with μ^- finite and $\mu = \mu^+ - \mu^-$.

Finally, we need to show that $|\mu| = \mu^+ + \mu^-$. Take $A \in \mathcal{S}$ and $\{B_1, \ldots, B_n\} \in P_{[0,\infty)}(A)$. Then

$$\sum_{k=1}^{n} |\mu(B)| = \sum_{k=1}^{n} |\mu^+(B_k) - \mu^-(B_k)| \leq \sum_{k=1}^{n} \left(\mu^+(B_k) + \mu^-(B_k) \right) = \mu^+(A) + \mu^-(A).$$

To show that the obtained upper bound is tight, we consider the partition $\{A \cap D, A \cap D^c\}$ of A for which we have

$$|\mu(A \cap D)| + |\mu(A \cap D^c)| = \mu^-(A \cap D) + \mu^+(A \cap D^c) = \mu^+(A) + \mu^-(A). \qquad \blacksquare$$

2.5 Additional Problems

Problem 2.33 (Local separation by constants) Let $(S, \mathcal{S}, \mu)$ be a measure space and let the function $f, g \in \mathcal{L}^0(S, \mathcal{S}, \mu)$ satisfy $\mu(\{x \in S : f(x) < g(x)\}) > 0$. Prove or construct a counterexample for the following statement:

"There exist constants $a, b \in \mathbb{R}$ such that $\mu(\{x \in S : f(x) \leq a < b \leq g(x)\}) > 0$."

Problem 2.34 (A pseudometric on sets) Let $(S, \mathcal{S}, \mu)$ be a finite measure space. For $A, B \in \mathcal{S}$ define

$$d(A, B) = \mu(A \bigtriangleup B),$$

where $\bigtriangleup$ denotes the symmetric difference: $A \bigtriangleup B = (A \setminus B) \cup (B \setminus A)$. Show that d is a pseudometric[4] on $\mathcal{S}$, and for $A \in \mathcal{S}$ describe the set of all $B \in \mathcal{S}$ with $d(A, B) = 0$.

---

[4]Let X be a nonempty set. A function $d : X \times X \to [0, \infty)$ is called a **pseudo metric** if

1. $d(x, y) + d(y, x) \geq d(x, z)$, for all $x, y, z \in X$,

2. $d(x, y) = d(y, x)$, for all $x, y \in X$, and

3. $d(x, x) = 0$, for all $x \in X$.

Note how the only difference between a metric and a pseudometric is that for a metric $d(x, y) = 0$ implies $x = y$, while no such requirement is imposed on a pseudometric.

Problem 2.35 (Complete measure spaces) A measure space $(S, \mathcal{S}, \mu)$ is called **complete** if all subsets of null sets are themselves in $\mathcal{S}$. For a (possibly incomplete) measure space $(S, \mathcal{S}, \mu)$ we define the **completion** $(S, \mathcal{S}^*, \mu^*)$ in the following way:

$$\mathcal{S}^* = \{A \cup N^* \ : \ A \in \mathcal{S} \text{ and } N^* \subseteq N \text{ for some } N \in \mathcal{S} \text{ with } \mu(N) = 0\}.$$

For $B \in \mathcal{S}^*$ with representation $B = A \cup N^*$ we set $\mu^*(B) = \mu(A)$.

1. Show that $\mathcal{S}^*$ is a σ-algebra.

2. Show that the definition $\mu^*(B) = \mu(A)$ above does not depend on the choice of the decomposition $B = A \cup N^*$, i.e., that $\mu(\hat{A}) = \mu(A)$ if $B = \hat{A} \cup \hat{N}^*$ is another decomposition of B into a set $\hat{A}$ in $\mathcal{S}$ and a subset $\hat{N}$ of a null set in $\mathcal{S}$.

3. Show that μ^* is a measure on $(S, \mathcal{S}^*)$ and that $(S, \mathcal{S}^*, \mu^*)$ is a complete measure space with the property that $\mu^*(A) = \mu(A)$, for $A \in \mathcal{S}$.

Problem 2.36 (The Cantor set) The **Cantor set** is defined as the collection of all real numbers x in $[0, 1]$ with the representation

$$x = \sum_{n=1}^{\infty} c_n 3^{-n}, \text{ where } c_n \in \{0, 2\}.$$

Show that it is Borel-measurable and compute its Lebesgue measure.

Problem 2.37 (The uniform measure on a circle) Let S^1 be the unit circle, and let $f : [0, 1) \to S^1$ be the "winding map"
$$f(x) = \Big(\cos(2\pi x), \sin(2\pi x)\Big), \ x \in [0, 1).$$

1. Show that the map f is $(\mathcal{B}([0, 1)), \mathcal{S}^1)$-measurable, where $\mathcal{S}^1$ denotes the Borel σ-algebra on S^1 (with the topology inherited from $\mathbb{R}^2$).

2. For $\alpha \in (0, 2\pi)$, let R_α denote the (counter-clockwise) rotation of $\mathbb{R}^2$ with center $(0, 0)$ and angle α. Show that $R_\alpha(A) = \{R_\alpha(x) \ : \ x \in A\}$ is in $\mathcal{S}^1$ if and only if $A \in \mathcal{S}^1$.

3. Let μ^1 be the push-forward of the Lebesgue measure λ by the map f. Show that μ^1 is rotation-invariant, i.e., that $\mu^1(A) = \mu^1\big(R_\alpha(A)\big)$.

(*Note:* The measure μ^1 is called the **uniform measure** (or the **uniform distribution on** S^1.)

Problem 2.38 (Asymptotic densities) We say that the subset A of $\mathbb{N}$ admits **asymptotic density** if the limit
$$d(A) = \lim_n \frac{\#(A \cap \{1, 2, \dots, n\})}{n},$$
exists (remember that $\#$ denotes the number of elements of a set). Let $\mathcal{D}$ be the collection of all subsets of $\mathbb{N}$ which admit asymptotic density.

1. Is $\mathcal{D}$ an algebra? A σ-algebra?

2. Is the map $A \mapsto d(A)$ finitely-additive on $\mathcal{D}$? A measure?

Problem 2.39 (A subset of the coin-toss space) An element in $\{-1,1\}^{\mathbb{N}}$ (i.e., a sequence s with $s = (s_1, s_2, \dots)$ where $s_n \in \{-1, 1\}$ for all $n \in \mathbb{N}$) is said to be **eventually periodic** if there exists $N_0, K \in \mathbb{N}$ such that $s_n = s_{n+K}$ for all $n \geq N_0$. Let $P \subseteq \{-1, 1\}^{\mathbb{N}}$ be the collection of all eventually-period sequences. Show that P is measurable in the product σ-algebra $\mathcal{S}$ and compute $\mu_C(P)$.

Problem 2.40 (Regular measures) The measure space $(S, \mathcal{S}, \mu)$, where (S, d) is a metric space and $\mathcal{S}$ is a σ-algebra on S which contains the Borel σ-algebra $\mathcal{B}(d)$ on S is called **regular** if for each $A \in \mathcal{S}$ and each $\varepsilon > 0$ there exist a closed set C and an open set O such that $C \subseteq A \subseteq O$ and $\mu(O \setminus C) < \varepsilon$.

1. Suppose that $(S, \mathcal{S}, \mu)$ is a regular measure space, and let $(S, \mathcal{B}(d), \mu|_{\mathcal{B}(d)})$ be the measure space obtained from $(S, \mathcal{S}, \mu)$ by restricting the measure μ onto the σ-algebra of Borel sets. Show that $\mathcal{S} \subseteq \mathcal{B}(d)^*$, where $\big(S, \mathcal{B}(d)^*, (\mu|_{\mathcal{B}(d)})^*\big)$ is the completion of $(S, \mathcal{B}(d), \mu|_{\mathcal{B}(d)})$ (in the sense of Problem 2.35).

2. Suppose that (S, d) is a metric space and that μ is a finite measure on $\mathcal{B}(d)$. Show that $(S, \mathcal{B}(d), \mu)$ is a regular measure space.

   (*Hint:* Consider a collection $\mathcal{A}$ of subsets A of S such that for each $\varepsilon > 0$ there exists a closed set C and an open set O with $C \subseteq A \subseteq O$ and $\mu(O \setminus C) < \varepsilon$. Argue that $\mathcal{A}$ is a σ-algebra. Then show that each closed set can be written as an intersection of open sets; use (but prove, first) the fact that the map

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

   is continuous on S for any nonempty $C \subseteq S$.)

3. Show that $(S, \mathcal{B}(d), \mu)$ is regular if μ is not necessarily finite, but has the property that $\mu(A) < \infty$ whenever $A \in \mathcal{B}(d)$ is bounded, i.e., when $\sup\{d(x, y) \,:\, x, y \in A\} < \infty$. (*Hint:* Pick a point $x_0 \in S$ and, for $n \in \mathbb{N}$, define the family $\{R_n\}_{n \in \mathbb{N}}$ of subsets of S as follows:

   $$R_1 = \{x \in S \,:\, d(x, x_0) < 2\}, \text{ and}$$
   $$R_n = \{x \in S \,:\, n - 1 < d(x, x_0) < n + 1\}, \text{ for } n > 1,$$

   as well as a sequence $\{\mu_n\}_{n \in \mathbb{N}}$ of set functions on $\mathcal{B}(d)$, given by $\mu_n(A) = \mu(A \cap R_n)$, for $A \in \mathcal{B}(d)$. Under the right circumstances, even countable unions of closed sets are closed).

(*Note:* It follows now from the fact that Lebesgue measure of any ball is finite, that the Lebesgue measure on $\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\big)$ is regular.)

Chapter 3

Lebesgue Integration

3.1 The construction of the integral

Unless expressly specified otherwise, we pick and fix a measure space $(S, \mathcal{S}, \mu)$ and assume that all functions under consideration are defined there.

> **Definition 3.1 (Simple functions)** *A function $f \in \mathcal{L}^0(S, \mathcal{S}, \mu)$ is said to be* **simple** *if it takes only a finite number of values.*

The collection of all simple functions is denoted by $\mathcal{L}^{\text{Simp},0}$ (more precisely by $\mathcal{L}^{\text{Simp},0}(S, \mathcal{S}, \mu)$) and the family of non-negative simple functions by $\mathcal{L}_+^{\text{Simp},0}$. Clearly, a simple function $f : S \to \mathbb{R}$ admits a (not necessarily unique) representation

$$(3.1) \qquad\qquad f = \sum_{k=1}^{n} \alpha_k \mathbf{1}_{A_k},$$

for $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and $A_1, \ldots, A_n \in \mathcal{S}$. Such a representation is called the **simple-function representation** of f.

When the sets A_k, $k = 1, \ldots, n$ are intervals in $\mathbb{R}$, the graph of the simple function f looks like a collection of steps (of heights $\alpha_1, \ldots, \alpha_n$). For that reason, the simple functions are sometimes referred to as *step functions*.

The Lebesgue integral is very easy to define for non-negative simple functions and this definition allows for further generalizations. In fact, the progression of events you will see in this section is typical for measure theory: you start with indicator functions, move on to non-negative simple functions, then to general non-negative measurable functions, and finally to (not-necessarily-non-negative) measurable functions. This approach is so common, that it has a name - the **Standard Machine**.

Definition 3.2 (Lebesgue integration for simple functions) *For* $f \in \mathcal{L}_+^{\mathrm{Simp},0}$ *we define the* **(Lebesgue) integral** $\int f \, d\mu$ *of f with respect to μ by*

$$\int f \, d\mu = \sum_{k=1}^{n} \alpha_k \mu(A_k) \in [0, \infty],$$

where $f = \sum_{k=1}^{n} \alpha_k \mathbf{1}_{A_k}$ is a simple-function representation of f,

---

Problem 3.3 Show that the Lebesgue integral is well-defined for simple functions, i.e., that the value of the expression $\sum_{k=1}^{n} \alpha_k \mu(A_k)$ does not depend on the choice of the simple-function representation of f.

Remark 3.4

1. It is important to note that $\int f \, d\mu$ can equal $+\infty$ even if f never takes the value $+\infty$. It is enough to pick $f = \mathbf{1}_A$ where $\mu(A) = +\infty$ - indeed, then $\int f \, d\mu = 1\mu(A) = \infty$, but f only takes values in the set $\{0, 1\}$. This is one of the reasons we start with *non-negative* functions. Otherwise, we would need to deal with the (unsolvable) problem of computing $\infty - \infty$. On the other hand, such examples cannot be constructed when μ is a finite measure. Indeed, it is easy to show that when $\mu(S) < \infty$, we have $\int f \, d\mu < \infty$ for all $f \in \mathcal{L}_+^{\mathrm{Simp},0}$.

2. One can think of the (simple) Lebesgue integral as a generalization of the notion of (finite) additivity of measures. Indeed, if the simple-function representation of f is given by $f = \sum_{k=1}^{n} \mathbf{1}_{A_k}$, for pairwise disjoint $A_1, \ldots, A_n$, then the equality of the values of the integrals for two representations $f = \mathbf{1}_{\cup_{k=1}^{n} A_k}$ and $f = \sum_{k=1}^{n} \mathbf{1}_{A_k}$ is a simple restatement of finite additivity. When $A_1, \ldots, A_n$ are not disjoint, then the finite additivity gives way to finite subadditivity

$$\mu(\cup_{k=1}^{n} A_k) \leq \sum_{k=1}^{n} \mu(A_k),$$

   but the integral $\int f \, d\mu$ "takes into account" those x which are covered by more than one A_k, $k = 1, \ldots, n$. Take, for example, $n = 2$ and $A_1 \cap A_2 = C$. Then

$$f = \mathbf{1}_{A_1} + \mathbf{1}_{A_2} = \mathbf{1}_{A_1 \setminus C} + 2\mathbf{1}_C + \mathbf{1}_{A_2 \setminus C},$$

   and so

$$\int f \, d\mu = \mu(A_1 \setminus C) + \mu(A_2 \setminus C) + 2\mu(C) = \mu(A_1) + \mu(A_2) + \mu(C).$$

It is easy to see that $\mathcal{L}_+^{\mathrm{Simp},0}$ is a **convex cone**, i.e., that it is closed under finite linear combinations with non-negative coefficients. The integral map $f \mapsto \int f \, d\mu$ preserves this structure:

Problem 3.5 For $f_1, f_2 \in \mathcal{L}_+^{\mathrm{Simp},0}$ and $\alpha_1, \alpha_2 \geq 0$ we have

1. if $f_1(x) \leq f_2(x)$ for all $x \in S$ then $\int f_1 \, d\mu \leq \int f_2 \, d\mu$, and

2. $\int (\alpha_1 f_1 + \alpha_2 f_2) \, d\mu = \alpha_1 \int f_1 \, d\mu + \alpha_2 \int f_2 \, d\mu$.

Having defined the integral for $f \in \mathcal{L}_+^{\mathrm{Simp},0}$, we turn to general non-negative measurable functions. In fact, at no extra cost we can consider a slightly larger set consisting of all measurable $[0,\infty]$-valued functions which we denote by $\mathcal{L}_+^0\big([0,\infty]\big)$. While there is no obvious advantage at this point of integrating a function which takes the value $+\infty$, it will become clear soon how convenient it really is.

Definition 3.6 (Lebesgue integral for nonnegative functions) *For a function $f \in \mathcal{L}_+^0\big([0,\infty]\big)$, we define the* **Lebesgue integral** $\int f \, d\mu$ *of f by*

$$\int f \, d\mu = \sup \left\{ \int g \, d\mu \ : \ g \in \mathcal{L}_+^{\mathrm{Simp},0}, \ g(x) \leq f(x), \ \forall \, x \in S \right\} \in [0,\infty].$$

Remark 3.7 While there is no question that the expression above defines uniquely the number $\int f \, d\mu$, one can wonder if it matches the previously given definition of the Lebesgue integral for simple functions. A simple argument based on the monotonicity property of part 1. of Problem 3.5 can be used to show that this is, indeed, the case.

Problem 3.8 Show that $\int f \, d\mu = \infty$ if there exists a measurable set A with $\mu(A) > 0$ such that $f(x) = \infty$ for $x \in A$. On the other hand, show that $\int f \, d\mu = 0$ for f of the form

$$f(x) = \infty \mathbf{1}_A(x) = \begin{cases} \infty, & x \in A, \\ 0, & x \neq A, \end{cases}$$

whenever $\mu(A) = 0$. (*Note:* Relate this to our convention that $\infty \times 0 = 0 \times \infty = 0$.)

Finally, we are ready to define the integral for general measurable functions. Each $f \in \mathcal{L}^0$ can be written as a difference of two functions in $\mathcal{L}_+^0$ in many ways. There exists a decomposition which is, in a sense, minimal. We define

$$f^+ = \max(f, 0), \ f^- = \max(-f, 0),$$

so that $f = f^+ - f^-$ (and both f^+ and f^- are measurable). The minimality we mentioned above is reflected in the fact that for each $x \in S$, at most one of f^+ and f^- is non-zero.

Definition 3.9 (Integrable functions) *A function $f \in \mathcal{L}^0$ is said to be* **integrable** *if*

$$\int f^+ \, d\mu < \infty \ and \ \int f^- \, d\mu < \infty.$$

The collection of all integrable functions in $\mathcal{L}^0$ is denoted by $\mathcal{L}^1$. The family of integrable functions is tailor-made for the following definition:

Definition 3.10 (The Lebesgue integral) *For $f \in \mathcal{L}^1$, we define the* **Lebesgue integral** $\int f \, d\mu$ *of f by*

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

Remark 3.11

1. We have seen so far two cases in which an integral for a function $f \in \mathcal{L}^0$ can be defined: when $f \geq 0$ or when $f \in \mathcal{L}^1$. It is possible to combine the two and define the Lebesgue integral for all functions $f \in \mathcal{L}^0$ with $f^- \in \mathcal{L}^1$. The set of all such functions is denoted by $\mathcal{L}^{0-1}$ and we set

$$\int f \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu \in (-\infty, \infty], \text{ for } f \in \mathcal{L}^{0-1}.$$

   Note that no problems of the form $\infty - \infty$ arise here, and also note that, like $\mathcal{L}^0_+$, $\mathcal{L}^{0-1}$ is only a convex cone, and not a vector space. While the notation $\mathcal{L}^0$ and $\mathcal{L}^1$ is quite standard, the one we use for $\mathcal{L}^{0-1}$ is not.

2. For $A \in \mathcal{S}$ and $f \in \mathcal{L}^{0-1}$ we usually write $\int_A f \, d\mu$ for $\int f \mathbf{1}_A \, d\mu$.

Problem 3.12 Show that the Lebesgue integral remains a monotone operation in $\mathcal{L}^{0-1}$. More precisely, show that if $f \in \mathcal{L}^{0-1}$ and $g \in \mathcal{L}^0$ are such that $g(x) \geq f(x)$, for all $x \in S$, then $g \in \mathcal{L}^{0-1}$ and $\int g \, d\mu \geq \int f \, d\mu$.

3.2 First properties of the integral

The wider the generality to which a definition applies, the harder it is to prove theorems about it. Linearity of the integral is a trivial matter for functions in $\mathcal{L}^{\mathrm{Simp},0}_+$, but you will see how much we need to work to get it for $\mathcal{L}^0_+$. In fact, it seems that the easiest route towards linearity is through two important results: an approximation theorem and a convergence theorem. Before that, we need to pick some low-hanging fruit:

Problem 3.13 Show that for $f_1, f_2 \in \mathcal{L}^0_+\big([0, \infty]\big)$ and $\alpha \in [0, \infty]$ we have

1. if $f_1(x) \leq f_2(x)$ for all $x \in S$ then $\int f_1 \, d\mu \leq \int f_2 \, d\mu$.

2. $\int \alpha f \, d\mu = \alpha \int f \, d\mu$.

Theorem 3.14 (Monotone convergence theorem) *Let $\{f_n\}_{n \in \mathbb{N}}$ be a sequence in $\mathcal{L}^0_+\big([0, \infty]\big)$ with the property that*

$$f_1(x) \leq f_2(x) \leq \ldots \text{ for all } x \in S.$$

Then

$$\lim_n \int f_n \, d\mu = \int f \, d\mu,$$

where $f(x) = \lim_n f_n(x) \in \mathcal{L}^0_+\big([0, \infty]\big)$, for $x \in S$.

PROOF The (monotonicity) property (1) of Problem 3.13 above implies immediately that the sequence $\int f_n \, d\mu$ is non-decreasing and that $\int f_n \, d\mu \leq \int f \, d\mu$. Therefore, $\lim_n \int f_n \, d\mu \leq \int f \, d\mu$. To show the opposite inequality, we deal with the case $\int f \, d\mu < \infty$ and pick $\varepsilon > 0$ and $g \in \mathcal{L}_+^{\text{Simp},0}$ with $g(x) \leq f(x)$, for all $x \in S$ and $\int g \, d\mu \geq \int f \, d\mu - \varepsilon$ (the case $\int f \, d\mu = \infty$ is similar and left to the reader). For $0 < c < 1$, define the (measurable) sets $\{A_n\}_{n \in \mathbb{N}}$ by

$$A_n = \{f_n \geq cg\}, n \in \mathbb{N}.$$

By the increase of the sequence $\{f_n\}_{n \in \mathbb{N}}$, the sets $\{A_n\}_{n \in \mathbb{N}}$ also increase. Moreover, since the function cg satisfies $cg(x) \leq g(x) \leq f(x)$ for all $x \in S$ and $cg(x) < f(x)$ when $f(x) > 0$, the increasing convergence $f_n \to f$ implies that $\cup_n A_n = S$. By non-negativity of f_n and monotonicity,

$$\int f_n \, d\mu \geq \int f_n \mathbf{1}_{A_n} \, d\mu \geq c \int g \mathbf{1}_{A_n} \, d\mu,$$

and so

$$\sup_n \int f_n \, d\mu \geq c \sup_n \int g \mathbf{1}_{A_n} \, d\mu.$$

Let $g = \sum_{i=1}^k \alpha_i \mathbf{1}_{B_i}$ be a simple-function representation of g. Then

$$\int g \mathbf{1}_{A_n} \, d\mu = \int \sum_{i=1}^k \alpha_i \mathbf{1}_{B_i \cap A_n} \, d\mu = \sum_{i=1}^k \alpha_i \mu(B_i \cap A_n).$$

Since $A_n \nearrow S$, we have $A_n \cap B_i \nearrow B_i$, $i = 1, \ldots, k$, and the continuity of measure implies that $\mu(A_n \cap B_i) \nearrow \mu(B_i)$. Therefore,

$$\int g \mathbf{1}_{A_n} \, d\mu \nearrow \sum_{i=1}^k \alpha_i \mu(B_i) = \int g \, d\mu.$$

Consequently,

$$\lim_n \int f_n \, d\mu = \sup_n \int f_n \, d\mu \geq c \int g \, d\mu, \text{ for all } c \in (0,1),$$

and the proof is completed when we let $c \to 1$. ∎

Remark 3.15

1. The monotone convergence theorem is a testament to the incredible robustness of the Lebesgue integral. This stability with respect to limiting operations is one of the reasons why it is a de-facto "industry standard".

2. The "monotonicity" condition in the monotone convergence theorem cannot be dropped. Take, for example $S = [0,1]$, $\mathcal{S} = \mathcal{B}([0,1])$, and $\mu = \lambda$ (the Lebesgue measure), and define

$$f_n = n\mathbf{1}_{(0,n^{-1}]}, \text{ for } n \in \mathbb{N}.$$

Then $f_n(0) = 0$ for all $n \in \mathbb{N}$ and $f_n(x) = 0$ for $n > \frac{1}{x}$ and $x > 0$. In either case $f_n(x) \to 0$. On the other hand

$$\int f_n \, d\lambda = n\lambda\left((0, \tfrac{1}{n}]\right) = 1,$$

so that

$$\lim_n \int f_n \, d\lambda = 1 > 0 = \int \lim_n f_n \, d\lambda.$$

We will see later that the while the equality of the limit of the integrals and the integral of the limit will not hold in general, they will always be ordered in a specific way, if the functions $\{f_n\}_{n \in \mathbb{N}}$ are non-negative (that will be the content of Fatou's lemma below).

Proposition 3.16 (Approximation by simple functions) *For each $f \in \mathcal{L}_+^0 \left([0, \infty]\right)$ there exists a sequence $\{g_n\}_{n \in \mathbb{N}} \in \mathcal{L}_+^{\text{Simp},0}$ such that*

1. *$g_n(x) \leq g_{n+1}(x)$, for all $n \in \mathbb{N}$ and all $x \in S$,*

2. *$g_n(x) \leq f(x)$ for all $x \in S$,*

3. *$f(x) = \lim_n g_n(x)$, for all $x \in S$, and*

4. *the convergence $g_n \to f$ is uniform on each set of the form $\{f \leq M\}$, $M > 0$, and, in particular, on the whole S if f is bounded.*

PROOF For $n \in \mathbb{N}$, let A_k^n, $k = 1, \ldots, n2^n$ be a collection of subsets of S given by

$$A_k^n = \{\tfrac{k-1}{2^n} \leq f < \tfrac{k}{2^n}\} = f^{-1}\left([\tfrac{k-1}{2^n}, \tfrac{k}{2^n})\right), \ k = 1, \ldots, n2^n.$$

Note that the sets A_k^n, $k = 1, \ldots, n2^n$ are disjoint and that the measurability of f implies that $A_k^n \in \mathcal{S}$ for $k = 1, \ldots, n2^n$. Define the function $g_n \in \mathcal{L}_+^{\text{Simp},0}$ by

$$g_n = \sum_{k=1}^{n2^n} \tfrac{k-1}{2^n} \mathbf{1}_{A_k^n} + n \mathbf{1}_{\{f \geq n\}}.$$

The statements 1., 2., and 4. follow immediately from the following three simple observations:

- $g_n(x) \leq f(x)$ for all $x \in S$,

- $g_n(x) = n$ if $f(x) = \infty$, and

- $g_n(x) > f(x) - 2^{-n}$ when $f(x) < n$.

Finally, we leave it to the reader to check the simple fact that $\{g_n\}_{n \in \mathbb{N}}$ is non-decreasing. ∎

Problem 3.17 Show, by means of an example, that the sequence $\{g_n\}_{n \in \mathbb{N}}$ would not necessarily be monotone if we defined it in the following way:

$$g_n = \sum_{k=1}^{n^2} \tfrac{k-1}{n} \mathbf{1}_{\{f \in [\frac{k-1}{n}, \frac{k}{n})\}} + n \mathbf{1}_{\{f \geq n\}}.$$

Proposition 3.18 (Linearity of the integral for non-negative functions) *For* $\alpha_1, \alpha_2 \geq 0$ *and* $f_1, f_2 \in \mathcal{L}^0_+\big([0, \infty]\big)$ *we have*

$$\int (\alpha_1 f_1 + \alpha_2 f_2)\, d\mu = \alpha_1 \int f_1\, d\mu + \alpha_2 \int f_2\, d\mu.$$

PROOF Thanks to Problem 3.13 it is enough to prove the statement for $\alpha_1 = \alpha_2 = 1$. Let $\{g_n^1\}_{n \in \mathbb{N}}$ and $\{g_n^2\}_{n \in \mathbb{N}}$ be sequences in $\mathcal{L}^{\mathrm{Simp},0}_+$ which approximate f^1 and f^2 in the sense of Proposition 3.16. The sequence $\{g_n\}_{n \in \mathbb{N}}$ given by $g_n = g_n^1 + g_n^2$, $n \in \mathbb{N}$, has the following properties:

- $g_n \in \mathcal{L}^{\mathrm{Simp},0}_+$ for $n \in \mathbb{N}$,

- $g_n(x)$ is a nondecreasing sequence for each $x \in S$,

- $g_n(x) \to f_1(x) + f_2(x)$, for all $x \in \mathcal{S}$.

Therefore, we can apply the linearity of integration for the simple functions and the monotone convergence theorem (Theorem 3.14) to conclude that

$$\int (f_1 + f_2)\, d\mu = \lim_n \int (g_n^1 + g_n^2)\, d\mu = \lim_n \left(\int g_n^1\, d\mu + \int g_n^2\, d\mu \right) = \int f_1\, d\mu + \int f_2\, d\mu. \qquad \blacksquare$$

Corollary 3.19 (Countable additivity of the integral) *Let* $\{f_n\}_{n \in \mathbb{N}}$ *be a sequence in* $\mathcal{L}^0_+\big([0, \infty]\big)$. *Then*

$$\int \sum_{n \in \mathbb{N}} f_n\, d\mu = \sum_{n \in \mathbb{N}} \int f_n\, d\mu.$$

PROOF Apply the monotone convergence theorem to the partial sums $g_n = f_1 + \cdots + f_n$, and use linearity of integration. $\qquad \blacksquare$

Once we have established a battery of properties for non-negative functions, an extension to $\mathcal{L}^1$ is not hard. We leave it to the reader to prove all the statements in the following problem:

Problem 3.20 The family $\mathcal{L}^1$ of integrable functions has the following properties:

1. $f \in \mathcal{L}^1$ iff $\int |f|\, d\mu < \infty$,

2. $\mathcal{L}^1$ is a vector space,

3. $\left| \int f\, d\mu \right| \leq \int |f|\, d\mu$, for $f \in \mathcal{L}^1$.

4. $\int |f + g|\, d\mu \leq \int |f|\, d\mu + \int |g|\, d\mu$, for all $f, g \in \mathcal{L}^1$.

We conclude the present section with two results, which, together with the monotone convergence theorem, play the central role in the Lebesgue integration theory.

Theorem 3.21 (Fatou's lemma) *Let $\{f_n\}_{n\in\mathbb{N}}$ be a sequence in $\mathcal{L}_+^0\left([0,\infty]\right)$. Then*

$$\int \liminf_n f_n \, d\mu \le \liminf_n \int f_n \, d\mu.$$

PROOF Set $g_n(x) = \inf_{k\ge n} f_k(x)$, so that $g_n \in \mathcal{L}_+^0\left([0,\infty]\right)$ and $g_n(x)$ is a non-decreasing sequence for each $x \in S$. The monotone convergence theorem and the fact that $\liminf f_n(x) = \sup_n g_n(x) = \lim_n g_n(x)$, for all $x \in S$, imply that

$$\int g_n \, d\mu \nearrow \int \liminf_n \, d\mu.$$

On the other hand, $g_n(x) \le f_k(x)$ for all $k \ge n$, and so

$$\int g_n \, d\mu \le \inf_{k\ge n} \int f_k \, d\mu.$$

Therefore,

$$\lim_n \int g_n \, d\mu \le \lim_n \inf_{k\ge n} \int f_k \, d\mu = \liminf_n \int f_k \, d\mu. \qquad \blacksquare$$

Remark 3.22

1. The inequality in the Fatou's lemma does not have to be equality, even if the limit $\lim_n f_n(x)$ exists for all $x \in S$. You can use the sequence $\{f_n\}_{n\in\mathbb{N}}$ of Remark 3.15 to see that.

2. Like the monotone convergence theorem, Fatou's lemma requires that all function $\{f_n\}_{n\in\mathbb{N}}$ be non-negative. This requirement is necessary - to see that, simply consider the sequence $\{-f_n\}_{n\in\mathbb{N}}$, where $\{f_n\}_{n\in\mathbb{N}}$ is the sequence of Remark 3.15 above.

3. The strength of Fatou's lemma comes from the fact that, apart from non-negativity, it requires no special properties for the sequence $\{f_n\}_{n\in\mathbb{N}}$. Its conclusion is not as strong as that of the monotone convergence theorem, but it proves to be very useful in various settings because it gives an upper bound (namely $\liminf_n \int f_n \, d\mu$) on the integral of the non-negative function $\liminf f_n$.

Theorem 3.23 (Dominated convergence theorem) *Let $\{f_n\}_{n\in\mathbb{N}}$ be a sequence in $\mathcal{L}^0$ with the property that there exists $g \in \mathcal{L}^1$ such that $|f_n(x)| \le g(x)$, for all $x \in X$ and all $n \in \mathbb{N}$. If $f(x) = \lim_n f_n(x)$ for all $x \in S$, then $f \in \mathcal{L}^1$ and*

$$\int f \, d\mu = \lim_n \int f_n \, d\mu.$$

PROOF The condition $|f_n(x)| \leq g(x)$, for all $x \in X$ and all $n \in \mathbb{N}$ implies that $g(x) \geq 0$, for all $x \in S$. Since $f_n^+ \leq g$, $f_n^- \leq g$ and $g \in \mathcal{L}^1$, we immediately have $f_n \in \mathcal{L}^1$, for all $n \in \mathbb{N}$. The limiting function f inherits the same properties $f^+ \leq g$ and $f^- \leq g$ from $\{f_n\}_{n \in \mathbb{N}}$ so $f \in \mathcal{L}^1$, too.

Clearly $g(x) + f_n(x) \geq 0$ for all $n \in \mathbb{N}$ and all $x \in S$, so we can apply Fatou's lemma to get

$$\int g\, d\mu + \liminf_n \int f_n\, d\mu = \liminf_n \int (g + f_n)\, d\mu \geq \int \liminf_n (g + f_n)\, d\mu$$
$$= \int (g + f)\, d\mu = \int g\, d\mu + \int f\, d\mu.$$

In the same way (since $g(x) - f_n(x) \geq 0$, for all $x \in S$, as well), we have

$$\int g\, d\mu - \limsup_n \int f_n\, d\mu = \liminf_n \int (g - f_n)\, d\mu \geq \int \liminf_n (g - f_n)\, d\mu$$
$$= \int (g - f)\, d\mu = \int g\, d\mu - \int f\, d\mu.$$

Therefore

$$\limsup_n \int f_n\, d\mu \leq \int f\, d\mu \leq \liminf_n \int f_n\, d\mu,$$

and, consequently, $\int f\, d\mu = \lim_n \int f_n\, d\mu$. ∎

Remark 3.24 The dominated convergence theorem combines the lack of monotonicity requirements of Fatou's lemma and the strong conclusion of the monotone convergence theorem. The price to be paid is the uniform boundedness requirement. There is a way to relax this requirement a little bit (using the concept of *uniform integrability*), but not too much. Still, it is an unexpectedly useful theorem.

3.3   Null sets

An important property - inherited directly from the underlying measure - is that it is blind to sets of measure zero. To make this statement precise, we need to introduce some language:

Definition 3.25 (Null sets) *Let $(S, \mathcal{S}, \mu)$ be a measure space.*

1. *$N \in \mathcal{S}$ is said to be a **null set** if $\mu(N) = 0$.*

2. *A function $f : S \to \bar{\mathbb{R}}$ is called a **null function** if there exists a null set N such that $f(x) = 0$ for $x \in N^c$.*

3. *Two functions f, g are said to be **equal almost everywhere** - denoted by $f = g$, a.e. - if $f - g$ is a null function, i.e., if there exists a null set N such that $f(x) = g(x)$ for all $x \in N^c$.*

Remark 3.26

1. In addition to almost-everywhere equality, one can talk about the almost-everywhere version of any relation between functions which can be defined on points. For example, we write $f \leq g$, a.e. if $f(x) \leq g(x)$ for all $x \in S$, except, maybe, for x in some null set N.

2. One can also define the a.e. equality of sets: we say that $A = B$, a.e., for $A, B \in \mathcal{S}$ if $\mathbf{1}_A = \mathbf{1}_B$, a.e. It is not hard to show (do it!) that $A = B$ a.e., if and only if $\mu(A \,\triangle\, B) = 0$ (Remember that $\triangle$ denotes the symmetric difference: $A \,\triangle\, B = (A \setminus B) \cup (B \setminus A)$).

3. When a property (equality of functions, e.g.) holds almost everywhere, the set where it fails to hold is not necessarily null. Indeed, there is no guarantee that it is measurable at all. What is true is that it is *contained* in a measurable (and null) set. Any such (measurable) null set is often referred to as the **exceptional set**.

Problem 3.27 Prove the following statements:

1. The almost-everywhere equality is an equivalence relation between functions.

2. The family $\{A \in \mathcal{S} \;:\; \mu(A) = 0 \text{ or } \mu(A^c) = 0\}$ is a σ-algebra (the so-called μ-**trivial σ-algebra**).

The "blindness" property of the Lebesgue integral we referred to above can now be stated formally:

Proposition 3.28 (The "blindness" property of the Lebesgue integral) *Suppose that $f = g$, a.e,. for some $f, g \in \mathcal{L}^0_+$. Then*

$$\int f \, d\mu = \int g \, d\mu.$$

PROOF Let N be an exceptional set for $f = g$, a.e., i.e., $f = g$ on N^c and $\mu(N) = 0$. Then $f\mathbf{1}_{N^c} = g\mathbf{1}_{N^c}$, and so $\int f\mathbf{1}_{N^c} \, d\mu = \int g\mathbf{1}_{N^c} \, d\mu$. On the other hand $f\mathbf{1}_N \leq \infty \mathbf{1}_N$ and $\int \infty \mathbf{1}_N \, d\mu = 0$, so, by monotonicity, $\int f\mathbf{1}_N \, d\mu = 0$. Similarly $\int g\mathbf{1}_N \, d\mu = 0$. It remains to use the additivity of integration to conclude that

$$\int f \, d\mu = \int f\mathbf{1}_{N^c} \, d\mu + \int f\mathbf{1}_N \, d\mu = \int g\mathbf{1}_{N^c} \, d\mu + \int g\mathbf{1}_N \, d\mu = \int g \, d\mu. \qquad \blacksquare$$

A statement which can be seen as a converse of Proposition 3.28 also holds:

Problem 3.29 Let $f \in \mathcal{L}^0_+$ be such that $\int f \, d\mu = 0$. Show that $f = 0$, a.e. (*Hint:* What is the negation of the statement "$f = 0$. a.e." for $f \in \mathcal{L}^0_+$?)

The monotone convergence theorem and the dominated convergence theorem both require the sequence $\{f_n\}_{n \in \mathbb{N}}$ functions to converge for each $x \in S$. A slightly weaker notion of convergence is required, though:

Definition 3.30 (Almost-everywhere convergence) *A sequence of functions $\{f_n\}_{n \in \mathbb{N}}$ is said to* **converge almost everywhere** *to the function f, if there exists a null set N such that*

$$f_n(x) \to f(x) \text{ for all } x \in N^c.$$

Remark 3.31 If we want to emphasize that $f_n(x) \to f(x)$ for *all* $x \in S$, we say that $\{f_n\}_{n \in \mathbb{N}}$ converges to f **everywhere**.

Proposition 3.32 (Monotone (almost-everywhere) convergence theorem) *Let* $\{f_n\}_{n \in \mathbb{N}}$ *be a sequence in* $\mathcal{L}^0_+([0, \infty])$ *with the property that*

$$f_n \leq f_{n+1} \text{ a.e., for all } n \in \mathbb{N}.$$

Then

$$\lim_n \int f_n \, d\mu = \int f \, d\mu,$$

if $f \in \mathcal{L}^0_+$ *and* $f_n \to f$, *a.e.*

PROOF There are "$\infty + 1$ a.e.-statements" we need to deal with: one for each $n \in \mathbb{N}$ in $f_n \leq f_{n+1}$, a.e., and an extra one when we assume that $f_n \to f$, a.e. Each of them comes with an exceptional set; more precisely, let $\{A_n\}_{n \in \mathbb{N}}$ be such that $f_n(x) \leq f_{n+1}(x)$ for $x \in A_n^c$ and let B be such that $f_n(x) \to f(x)$ for $x \in B^c$. Define $A \in \mathcal{S}$ by $A = (\cup_n A_n) \cup B$ and note that A is a null set. Moreover, consider the functions $\tilde{f}, \{\tilde{f}_n\}_{n \in \mathbb{N}}$ defined by $\tilde{f} = f\mathbf{1}_{A^c}$, $\tilde{f}_n = f_n\mathbf{1}_{A^c}$. Thanks to the definition of the set A, $\tilde{f}_n(x) \leq \tilde{f}_{n+1}(x)$, for all $n \in \mathbb{N}$ and $x \in S$; hence $\tilde{f}_n \to \tilde{f}$, everywhere. Therefore, the monotone convergence theorem (Theorem 3.14) can be used to conclude that $\int \tilde{f}_n \, d\mu \to \int \tilde{f} \, d\mu$. Finally, Proposition 3.28 implies that $\int \tilde{f}_n \, d\mu = \int f_n \, d\mu$ for $n \in \mathbb{N}$ and $\int \tilde{f} \, d\mu = \int f \, d\mu$. ∎

Problem 3.33 State and prove a version of the dominated convergence theorem where the almost-everywhere convergence is used. Is it necessary for all $\{f_n\}_{n \in \mathbb{N}}$ to be dominated by g for all $x \in S$, or only almost everywhere?

Remark 3.34 There is a subtlety that needs to be pointed out. If a sequence $\{f_n\}_{n \in \mathbb{N}}$ of measurable functions converges to the function f *everywhere*, then f is necessarily a measurable function (see Proposition 1.43). However, if $f_n \to f$ only almost everywhere, there is no guarantee that f is measurable. There is, however, always a measurable function which is equal to f almost everywhere; you can take $\liminf_n f_n$, for example.

3.4 Additional Problems

Problem 3.35 (The monotone-class theorem) Prove the following result, known as the *monotone-class theorem* (remember that $a_n \nearrow a$ means that a_n is a non-decreasing sequence and $a_n \to a$)

Let $\mathcal{H}$ *be a class of bounded functions from* S *into* $\mathbb{R}$ *satisfying the following conditions*

1. $\mathcal{H}$ *is a vector space,*

2. *the constant function* 1 *is in* $\mathcal{H}$, *and*

3. *if* $\{f_n\}_{n \in \mathbb{N}}$ *is a sequence of non-negative functions in* $\mathcal{H}$ *such that* $f_n(x) \nearrow f(x)$, *for all* $x \in S$ *and* f *is bounded, then* $f \in \mathcal{H}$.

Then, if $\mathcal{H}$ *contains the indicator* $\mathbf{1}_A$ *of every set* A *in some* π-*system* $\mathcal{P}$, *then* $\mathcal{H}$ *necessarily contains every bounded* $\sigma(\mathcal{P})$-*measurable function on* S.

(*Hint:* Use Theorems 3.16 and 2.16)

Problem 3.36 (A form of continuity for Lebesgue integration) Let $(S, \mathcal{S}, \mu)$ be a measure space, and suppose that $f \in \mathcal{L}^1$. Show that for each $\varepsilon > 0$ there exists $\delta > 0$ such that if $A \in \mathcal{S}$ and $\mu(A) < \delta$, then $\left| \int_A f \, d\mu \right| < \varepsilon$.

Problem 3.37 (Sums as integrals) Consider the measurable space $(N, 2^N, \mu)$, where μ is the counting measure.

1. For a function $f : \mathbb{N} \to [0, \infty]$, show that

$$\int f \, d\mu = \sum_{n=1}^{\infty} f(n).$$

2. Use the monotone convergence theorem to show the following special case of Fubini's theorem

$$\sum_{k=1}^{\infty} \sum_{n=1}^{\infty} a_{kn} = \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} a_{kn},$$

whenever $\{a_{kn} \, : \, k, n \in \mathbb{N}\}$ is a double sequence in $[0, \infty]$.

3. Show that $f : \mathbb{N} \to \mathbb{R}$ is in $\mathcal{L}^1$ if and only if the series

$$\sum_{n=1}^{\infty} f(n),$$

converges absolutely.

Problem 3.38 (A criterion for integrability) Let $(S, \mathcal{S}, \mu)$ be a finite measure space. For $f \in \mathcal{L}^0_+$, show that $f \in \mathcal{L}^1$ if and only if

$$\sum_{n \in \mathbb{N}} \mu(\{f \geq n\}) < \infty.$$

Problem 3.39 (A limit of integrals) Let $(S, \mathcal{S}, \mu)$ be a measure space, and suppose $f \in \mathcal{L}^1_+$ is such that $\int f \, d\mu = c > 0$. Show that the limit

$$\lim_{n} \int n \log \left(1 + (f/n)^{\alpha} \right) d\mu$$

exists in $[0, \infty]$ for each $\alpha > 0$ and compute its value.

(*Hint:* Prove and use the inequality $\log(1 + x^{\alpha}) \leq \alpha x$, valid for $x \geq 0$ and $\alpha \geq 1$.)

Problem 3.40 (Integrals converge but the functions don't ...) Construct an sequence $\{f_n\}_{n \in \mathbb{N}}$ of continuous functions $f_n : [0, 1] \to [0, 1]$ such that $\int f_n \, d\mu \to 0$, but the sequence $\{f_n(x)\}_{n \in \mathbb{N}}$ is divergent for each $x \in [0, 1]$.

Problem 3.41 (...or they do, but are not dominated) Construct an sequence $\{f_n\}_{n \in \mathbb{N}}$ of continuous functions $f_n : [0, 1] \to [0, \infty)$ such that $\int f_n \, d\mu \to 0$, and $f_n(x) \to 0$ for all x, but $f \notin \mathcal{L}^1$, where $f(x) = \sup_n f_n(x)$.

Problem 3.42 (Functions measurable in the generated σ-algebra) Let $S \neq \emptyset$ be a set and let $f : S \to \mathbb{R}$ be a function. Prove that a function $g : S \to \mathbb{R}$ is measurable with respect to the pair $(\sigma(f), \mathcal{B}(\mathbb{R}))$ if and only if there exists a Borel function $h : \mathbb{R} \to \mathbb{R}$ such that $g = h \circ f$.

Problem 3.43 (A change-of-variables formula) Let $(S, \mathcal{S}, \mu)$ and $(T, \mathcal{T}, \nu)$ be measurable spaces, and let $F : S \to T$ be a measurable function with the property that $\nu = F_*\mu$ (i.e., ν is the push-forward of μ through F). Show that for every $f \in \mathcal{L}^0_+(T, \mathcal{T})$ or $f \in \mathcal{L}^1(T, \mathcal{T})$, we have

$$\int f \, d\nu = \int (f \circ F) \, d\mu.$$

Problem 3.44 (The Riemann Integral) A finite collection $\Delta = \{t_0, \dots, t_n\}$, where $a = t_0 < t_1 < \cdots < t_n = b$ and $n \in \mathbb{N}$, is called a **partition** of the interval $[a, b]$. The set of all partitions of $[a, b]$ is denoted by $P([a, b])$.

For a bounded function $f : [a, b] \to \mathbb{R}$ and $\Delta = \{t_0, \dots, t_n\} \in P([a, b])$, we define its **upper and lower Darboux sums** $U(f, \Delta)$ and $L(f, \Delta)$ by

$$U(f, \Delta) = \sum_{k=1}^{n} \left(\sup_{t \in (t_{k-1}, t_k]} f(t) \right) (t_k - t_{k-1})$$

and

$$L(f, \Delta) = \sum_{k=1}^{n} \left(\inf_{t \in (t_{k-1}, t_k]} f(t) \right) (t_k - t_{k-1}).$$

A function $f : [a, b] \to \mathbb{R}$ is said to be **Riemann integrable** if it is bounded and

$$\sup_{\Delta \in P([a,b])} L(f, \Delta) = \inf_{\Delta \in P([a,b])} U(f, \Delta).$$

In that case the common value of the supremum and the infimum above is called the **Riemann integral** of the function f - denoted by $(R) \int_a^b f(x) \, dx$.

1. Suppose that a bounded Borel-measurable function $f : [a, b] \to \mathbb{R}$ is Riemann-integrable. Show that

$$\int_{[a,b]} f \, d\lambda = (R) \int_a^b f(x) \, dx.$$

2. Find an example of a bounded an Borel-measurable function $f : [a, b] \to \mathbb{R}$ which is not Riemann-integrable.

3. Show that every continuous function is Riemann integrable.

4. It can be shown that for a bounded Borel-measurable function $f : [a, b] \to \mathbb{R}$ the following criterion holds (and you can use it without proof):

   f is Riemann integrable if and only if there exists a Borel set $D \subseteq [a, b]$ with $\lambda(D) = 0$ such that f is continuous at x, for each $x \in [a, b] \setminus D$. Show that

   - all monotone functions are Riemann-integrable,
   - $f \circ g$ is Riemann integrable if $f : [c, d] \to \mathbb{R}$ is Riemann integrable and $g : [a, b] \to [c, d]$ is continuous,

- products of Riemann-integrable functions are Riemann-integrable.

5. Let $([a,b], \mathcal{B}([a,b])^*, \lambda^*)$ be the completion of $([a,b], \mathcal{B}([a,b]), \lambda)$. Show that each Riemann-integrable function on $[a,b]$ is $\mathcal{B}([a,b])^*$-measurable.

   (*Hint:* Pick a sequence $\{\Delta_n\}_{n\in\mathbb{N}}$ in $P([a,b])$ so that $\Delta_n \subseteq \Delta_{n+1}$ and $U(f, \Delta_n) - L(f, \Delta_n) \to 0$. Using those partitions and the function f, define two sequences of Borel-measurable functions $\{\overline{f}_n\}_{n\in\mathbb{N}}$ and $\{\underline{f}_n\}_{n\in\mathbb{N}}$ so that $\underline{f}_n \nearrow \underline{f}$, $\overline{f}_n \searrow \overline{f}$, $\underline{f} \le f \le \overline{f}$, and $\int (\overline{f} - \underline{f})\, d\lambda = 0$. Conclude that f agrees with a Borel measurable function on a complement of a subset of the set $\{\underline{f} \ne \overline{f}\}$ which has Lebesgue measure 0.)

Chapter 4

Lebesgue Spaces and Inequalities

We have seen how the family of all functions $f \in \mathcal{L}^1$ forms a vector space and how the map $f \mapsto ||f||_{\mathcal{L}^1}$, from $\mathcal{L}^1$ to $[0, \infty)$ defined by $||f||_{\mathcal{L}^1} = \int |f| \, d\mu$ has the following properties

1. $f = 0$ implies $||f||_{\mathcal{L}^1} = 0$, for $f \in \mathcal{L}^1$,

2. $||f + g||_{\mathcal{L}^1} \leq ||f||_{\mathcal{L}^1} + ||g||_{\mathcal{L}^1}$, for $f, g \in \mathcal{L}^1$,

3. $||\alpha f||_{\mathcal{L}^1} = |\alpha| \, ||f||_{\mathcal{L}^1}$, for $\alpha \in \mathbb{R}$ and $f \in \mathcal{L}^1$.

Any map from a vector space into $[0, \infty)$ with the properties 1., 2., and 3. above is called a **pseudo norm**. A pair $(V, || \cdot ||)$ where V is a vector space and $|| \cdot ||$ is a pseudo norm on V is called a **pseudo-normed space**.

   If a pseudo norm happens to satisfy the (stronger) axiom

1'. $f = 0$ if and only if $||f||_{\mathcal{L}^1} = 0$, for $f \in \mathcal{L}^1$,

instead of 1., it is called a **norm**, and the pair $(V, || \cdot ||)$ is called a **normed space**.

   The pseudo-norm $|| \cdot ||_{\mathcal{L}^1}$ is, in general, not a norm. Indeed, by Problem 3.29, we have $||f||_{\mathcal{L}^1} = 0$ iff $f = 0$, a.e., and unless $\emptyset$ is the only null-set, there are functions different from the constant function 0 with this property.

Remark 4.1 There is a relatively simple procedure one can use to turn a pseudo-normed space $(V, || \cdot ||)$ into a normed one. Declare two elements x, y in V *equivalent* (denoted by $x \sim y$) if $||y - x|| = 0$, and let $\tilde{V}$ be the quotient space $V/\sim$ (the set of all equivalence classes). It is easy to show that $||x|| = ||y||$ whenever $x \sim y$, so the pseudo-norm $|| \cdot ||$ can be seen as defined on $\tilde{V}$. Moreover, it follows directly from the properties of the pseudo norm that $(\tilde{V}, || \cdot ||)$ is, in fact a normed space. Idea is, of course, bundle together the elements of V which differ by such a "small amount" that $|| \cdot ||$ cannot detect it.

   This construction can be applied to the case of the pseudo-norm $|| \cdot ||_{\mathcal{L}^1}$ on $\mathcal{L}^1$, and the resulting normed space is denoted by $\mathbb{L}^1$. The normed space $\mathbb{L}^1$ has properties similar to those of $\mathcal{L}^1$, but its elements are not functions anymore - they are equivalence classes of measurable functions. Such a point of view is very useful in analysis, but it sometimes leads to confusion in probability (especially when one works with stochastic processes with infinite time-index sets). Therefore, we will stick to $\mathcal{L}^1$ and deal with the fact that it is only a pseudo-normed space.

A pseudo-norm $|| \cdot ||$ on a vector space can be used to define a pseudo metric (pseudo-distance function) on V by the following simple prescription:

$$d(x, y) = ||y - x||, \ x, y \in V.$$

Just like a pseudo norm, a pseudo metric has most of the properties of a metric

1. $d(x, y) \in [0, \infty)$, for $x, y \in V$,

2. $d(x, y) + d(y, z) \geq d(x, z)$, for $x, y, z \in V$,

3. $d(x, y) = d(y, x)$, $x, y \in V$,

4. $x = y$ implies $d(x, y) = 0$, for $x, y \in V$.

The missing axiom is the stronger version of 4. given by

4'. $x = y$ if and only if $d(x, y) = 0$, for $x, y \in V$.

Luckily, a pseudo metric is sufficient for the notion of convergence, where we say that a sequence $\{x_n\}_{n \in \mathbb{N}}$ in V converges towards $x \in V$ if $d(x_n, x) \to 0$, as $n \to \infty$. If we apply it to our original example $(\mathcal{L}^1, || \cdot ||_{\mathcal{L}^1})$, we have the following definition:

Definition 4.2 (Convergence in $\mathcal{L}^1$) *For a sequence $\{f_n\}_{n \in \mathbb{N}}$ in $\mathcal{L}^1$, we say that $\{f_n\}_{n \in \mathbb{N}}$ converges to f in $\mathcal{L}^1$ if*

$$||f_n - f||_{\mathcal{L}^1} \to 0.$$

To get some intuition about convergence in $\mathcal{L}^1$, here is a problem:

Problem 4.3 Show that the conclusion of the dominated convergence theorem (Theorem 3.23) can be replaced by "$f_n \to f$ in $\mathcal{L}^1$". Does the original conclusion follow from the new one?

The only problem that arises when one defines convergence using a pseudo metric (as opposed to a bona-fide metric) is that limits are not unique. This is, however, merely an inconvenience and one gets used to it quite readily:

Problem 4.4 Suppose that $\{f_n\}_{n \in \mathbb{N}}$ converges to f in $\mathcal{L}^1$. Show that $\{f_n\}_{n \in \mathbb{N}}$ also converges to $g \in \mathcal{L}^1$ if and only if $f = g$, a.e.

In addition to the space $\mathcal{L}^1$, one can introduce many other vector spaces of similar flavor. For $p \in [1, \infty)$, let $\mathcal{L}^p$ denote the family of all functions $f \in \mathcal{L}^0$ such that $|f|^p \in \mathcal{L}^1$.

Problem 4.5 Show that there exists a constant $C > 0$ (depending on p, but independent of a, b) such that $(a + b)^p \leq C(a^p + b^p)$, $p \in (0, \infty)$ and for all $a, b \geq 0$. Deduce that $\mathcal{L}^p$ is a vector space for all $p \in (0, \infty)$.

We will see soon that the map $|| \cdot ||_{\mathcal{L}^p}$, defined by

$$||f||_{\mathcal{L}^p} = \left(\int |f|^p \, d\mu \right)^{1/p}, \ f \in \mathcal{L}^p,$$

is a pseudo norm on $\mathcal{L}^p$. The hard part of the proof - showing that $||f + g||_{\mathcal{L}^p} \leq ||f||_{\mathcal{L}^p} + ||g||_{\mathcal{L}^p}$ will be a direct consequence of an important inequality of Minkowski which will be proved below.

Finally, there is a nice way to extend the definition of $\mathcal{L}^p$ to $p = \infty$.

Definition 4.6 (Essential supremum) *A number $a \in \bar{\mathbb{R}}$ is called an* **essential supremum** *of the function $f \in \mathcal{L}^0$ - and is denoted by $a = \mathrm{esssup}\, f$ - if*

1. $\mu(\{f > a\}) = 0$

2. $\mu(\{f > b\}) > 0$ *for any $b < a$.*

A function $f \in \mathcal{L}^0$ with $\mathrm{esssup}\, f < \infty$ is said to be **essentially bounded from above**. *When $\mathrm{esssup}\,|f| < \infty$, we say that f is* **essentially bounded**.

Remark 4.7 Even though the function f may take values larger than a, it does so only on a null set. It can happen that a function is unbounded, but that its essential supremum exists in $\mathbb{R}$. Indeed, take $(S, \mathcal{S}, \mu) = ([0, 1], \mathcal{B}([0, 1]), \lambda)$, and define

$$f(x) = \begin{cases} n, & x = \frac{1}{n} \text{ for some } n \in \mathbb{N}, \\ 0, & \text{otherwise.} \end{cases}$$

Then $\mathrm{esssup}\, f = 0$, since $\lambda(\{f > 0\}) = \lambda(\{1, 1/2, 1/3, \dots\}) = 0$, but $\sup_{x \in [0,1]} f(x) = \infty$.

Let $\mathcal{L}^\infty$ denote the family of all essentially bounded functions in $\mathcal{L}^0$. Define $||f||_{\mathcal{L}^\infty} = \mathrm{esssup}\,|f|$, for $f \in \mathcal{L}^\infty$.

Problem 4.8 Show that $\mathcal{L}^\infty$ is a vector space, and that $|| \cdot ||_{\mathcal{L}^\infty}$ is a pseudo-norm on $\mathcal{L}^\infty$.

The convergence in $\mathcal{L}^p$ for $p > 1$ is defined similarly to the $\mathcal{L}^1$-convergence:

Definition 4.9 (Convergence in $\mathcal{L}^p$) *Let $p \in [1, \infty]$. We say that a sequence $\{f_n\}_{n \in \mathbb{N}}$ in $\mathcal{L}^p$* **converges in $\mathcal{L}^p$** *to $f \in \mathcal{L}^p$ if*
$$||f_n - f||_{\mathcal{L}^p} \to 0, \text{ as } n \to \infty.$$

Problem 4.10 Show that $\{f_n\}_{n \in \mathbb{N}} \in \mathcal{L}^\infty$ converges to $f \in \mathcal{L}^\infty$ in $\mathcal{L}^\infty$ if and only if there exist functions $\{\tilde{f}_n\}_{n \in \mathbb{N}}, \tilde{f}$ in $\mathcal{L}^0$ such that

1. $\tilde{f}_n = f_n$, a.e., and $\tilde{f} = f$, a.e, and

2. $\tilde{f}_n \to \tilde{f}$ uniformly (we say that $g_n \to g$ uniformly if $\sup_x |g_n(x) - g(x)| \to 0$, as $n \to \infty$).

Definition 4.11 (Conjugate exponents) *We say that $p, q \in [1, \infty]$ are* **conjugate exponents** *if* $\frac{1}{p} + \frac{1}{q} = 1$.

Lemma 4.12 (Young's inequality) *For all $x, y \geq 0$ and conjugate exponents $p, q \in [1, \infty)$ we have*

(4.1)
$$\frac{x^p}{p} + \frac{y^q}{q} \geq xy.$$

The equality holds if and only if $x^p = y^q$.

PROOF If $x = 0$ or $y = 0$, the inequality trivially holds so we assume that $x > 0$ and $y > 0$. The function $\log$ is strictly concave on $(0, \infty)$ and $\frac{1}{p} + \frac{1}{q} = 1$, so

$$\log(\tfrac{1}{p}\xi + \tfrac{1}{q}\eta) \geq \tfrac{1}{p}\log(\xi) + \tfrac{1}{q}\log(\eta),$$

for all $\xi, \eta > 0$, with equality if and only if $\xi = \eta$. If we substitute $\xi = x^p$ and $\eta = y^q$, and exponentiate both sides, we get

$$\frac{x^p}{p} + \frac{y^q}{q} \geq \exp(\tfrac{1}{p}\log(x^p) + \tfrac{1}{q}\log(y^q)) = xy,$$

with equality if and only if $x^p = y^q$. ∎

Remark 4.13 If you do not want to be fancy, you can prove Young's inequality by locating the maximum of the function $x \mapsto xy - \frac{1}{p}x^p$ using nothing more than elementary calculus.

Proposition 4.14 (Hölder's inequality) *Let $p, q \in [0, \infty]$ be conjugate exponents. For $f \in \mathcal{L}^p$ and $g \in \mathcal{L}^q$, we have*

(4.2)
$$\int |fg| \, d\mu \leq ||f||_{\mathcal{L}^p} ||g||_{\mathcal{L}^q}.$$

The equality holds if and only if there exist constants $\alpha, \beta \geq 0$ with $\alpha + \beta > 0$ such that $\alpha |f|^p = \beta |g|^q$, a.e.

PROOF We assume that $1 < p, g < \infty$ and leave the (easier) extreme cases to the reader. Clearly, we can also assume that $||f||_{\mathcal{L}^p} > 0$ and $||q||_{\mathcal{L}^q} > 0$ - otherwise, the inequality is trivially satisfied. We define $\tilde{f} = |f| / ||f||_{\mathcal{L}^p}$ and $\tilde{g} = |g| / ||g||_{\mathcal{L}^q}$, so that $||\tilde{f}||_{\mathcal{L}^p} = ||\tilde{g}||_{\mathcal{L}^q} = 1$.

Plugging $\tilde{f}$ for x and $\tilde{g}$ for y in Young's inequality (Lemma 4.12 above) and integrating, we get

(4.3)
$$\frac{1}{p} \int \tilde{f}^p \, d\mu + \frac{1}{q} \int \tilde{g}^q \, d\mu \geq \int \tilde{f}\tilde{g} \, d\mu,$$

and consequently,

$$(4.4) \qquad\qquad \int \tilde{f}\tilde{g}\,d\mu \leq 1,$$

because $\int \tilde{f}^p\,d\mu = ||\tilde{f}||_{\mathcal{L}^p}^p = 1$, and $\int \tilde{g}^q\,d\mu = ||\tilde{g}||_{\mathcal{L}^p} = 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. Hölder's inequality (4.2) now follows by multiplying both sides of (4.4) by $||f||_{\mathcal{L}^p}||g||_{\mathcal{L}^q}$.

If the equality in (4.2) holds, then it also holds a.e. in the Young's inequality (4.3). Therefore, the equality will hold if and only if $||g||_{\mathcal{L}^q}^q\,|f|^p = ||f||_{\mathcal{L}^p}^p\,|g|^q$, a.e. The reader will check that if a pair of constants α, β as in the statement exists, then $(||g||_{\mathcal{L}^q}^q, ||f||_{\mathcal{L}^p}^p)$ must be proportional to it. ∎

For $p = q = 2$ we get the following well-known special case:

Corollary 4.15 (Cauchy-Schwarz inequality) *For $f, g \in \mathcal{L}^2$, we have*

$$\int |fg|\,d\mu \leq ||f||_{\mathcal{L}^2}||g||_{\mathcal{L}^2}.$$

Corollary 4.16 (Minkowski's inequality) *For $f, g \in \mathcal{L}^p$, $p \in [1, \infty]$, we have*

$$(4.5) \qquad\qquad ||f + g||_{\mathcal{L}^p} \leq ||f||_{\mathcal{L}^p} + ||g||_{\mathcal{L}^p}.$$

PROOF Like above, we assume $p < \infty$ and leave the case $p = \infty$ to the reader. Moreover, we assume that $||f + g||_{\mathcal{L}^p} > 0$ - otherwise, the inequality trivially holds. Note, first that for conjugate exponents p, q we have $q(p - 1) = p$. Therefore, Hölder's inequality implies that

$$\int |f|\,|f+g|^{p-1}\,d\mu \leq ||f||_{\mathcal{L}^p}||(f+g)^{p-1}||_{\mathcal{L}^q} = ||f||_{\mathcal{L}^p}\left(\int |f+g|^{q(p-1)}\,d\mu\right)^{1/q}$$

$$= ||f||_{\mathcal{L}^p}||f+g||_{\mathcal{L}^p}^{p/q},$$

and, similarly,

$$\int |g|\,|f+g|^{p-1}\,d\mu \leq ||g||_{\mathcal{L}^p}||f+g||_{\mathcal{L}^p}^{p/q}.$$

Therefore,

$$||f+g||_{\mathcal{L}^p}^p = \int |f+g|^p\,d\mu \leq \int |f|\,|f+g|^{p-1}\,d\mu + \int |g|\,|f+g|^{p-1}\,d\mu$$

$$\leq \left(||f||_{\mathcal{L}^p} + ||g||_{\mathcal{L}^p}\right)||f+g||_{\mathcal{L}^p}^{p-1},$$

and if we divide through by $||f + g||_{\mathcal{L}^p}^{p-1} > 0$, we get (4.5). ∎

Corollary 4.17 ($\mathcal{L}^p$ is pseudo-normed) *($\mathcal{L}^p, || \cdot ||_{\mathcal{L}^p}$) is a pseudo-normed space for each $p \in [1, \infty]$.*

A pseudo-metric space (X, d) is said to be **complete** if each Cauchy sequence converges. A sequence $\{x_n\}_{n \in \mathbb{N}}$ is called a Cauchy sequence if

$$\forall \varepsilon > 0, \ \exists N \in \mathbb{N}, \ m, n \geq N \ \Rightarrow \ d(x_n, x_m) < \varepsilon.$$

A pseudo-normed space $(V, || \cdot ||)$ is called a **pseudo-Banach space** if is it complete for the metric induced by $|| \cdot ||$. If $|| \cdot ||$ is, additionally, a norm, $(V, || \cdot ||)$ is said to be a **Banach space**.

Problem 4.18 Let $\{x_n\}_{n \in \mathbb{N}}$ be Cauchy sequence in a pseudo-metric space (X, d), and let $\{x_{n_k}\}_{k \in \mathbb{N}}$ be a subsequence of $\{x_n\}_{n \in \mathbb{N}}$ which converges to $x \in X$. Show that $x_n \to x$.

Proposition 4.19 ($\mathcal{L}^p$ is pseudo-Banach) *$\mathcal{L}^p$ is a pseudo-Banach space, for $p \in [1, \infty]$.*

PROOF We assume $p \in [1, \infty)$ and leave the case $p = \infty$ to the reader. Let $\{f_n\}_{n \in \mathbb{N}}$ be a Cauchy sequence in $\mathcal{L}^p$. Thanks to the Cauchy property, there exists a subsequence of $\{f_{n_k}\}_{k \in \mathbb{N}}$ such that

$$||f_{n_{k+1}} - f_{n_k}||_{\mathcal{L}^p} < 2^{-k}, \text{ for all } k \in \mathbb{N}.$$

We define the sequence $\{g_k\}_{k \in \mathbb{N}}$ in $\mathcal{L}^0_+$ by $g_k = |f_{n_1}| + \sum_{i=1}^{k-1} |f_{n_{i+1}} - f_{n_i}|$, as well as the function $g = \lim_k g_k \in \mathcal{L}^0([0, \infty])$. The monotone-convergence theorem implies that

$$\int g^p \, d\mu = \lim_n \int g_n^p \, d\mu,$$

and, by Minkowski's inequality, we have

$$\int g_k^p \, d\mu = ||g_k||_{\mathcal{L}^p}^p \leq \left(||f_{n_1}||_{\mathcal{L}^p} + \sum_{i=1}^{k-1} ||f_{n_{k+1}} - f_{n_k}||_{\mathcal{L}^p} \right)^p \leq (||f_{n_1}||_{\mathcal{L}^p} + 1)^p, \ \forall k \in \mathbb{N}.$$

Therefore, $\int g^p \, d\mu \leq (1 + ||f_{n_1}||_{\mathcal{L}^p})^p < \infty$, and, in particular, $g \in \mathcal{L}^p$ and $g < \infty$, a.e. It follows immediately from the absolute convergence of the series $\sum_{i=1}^{\infty} |f_{n_{k+1}} - f_{n_k}|$ that

$$f_{n_k}(x) = f_{n_1}(x) + \sum_{i=1}^{k-1} (f_{n_{k+1}}(x) - f_{n_k}(x)),$$

converges in $\mathbb{R}$, for almost all x. Hence, the function $f = \liminf_k f_{n_k}$ is in $\mathcal{L}^p$ since $|f| \leq g$, a.e.

Since $|f| \leq g$ and $|f_{n_k}| \leq g$, for all $k \in \mathbb{N}$, we have $|f - f_{n_k}|^p \leq 2 |g|^p \in \mathbb{L}^1$, so the dominated convergence theorem implies that $\int |f_{n_k} - f|^p \, d\mu \to 0$, i.e., $f_{n_k} \to f$ in $\mathcal{L}^p$. Finally, we invoke the result of Problem 4.18 to conclude that $f_n \to f$ in $\mathcal{L}^p$. ∎

The following result is a simple consequence of the (proof of) Proposition 4.19.

Corollary 4.20 ($\mathcal{L}^p$-convergent $\Rightarrow$ a.e.-convergent, after passing to subsequence) *For $p \in [1, \infty]$, let $\{f_n\}_{n\in\mathbb{N}}$ be a sequence in $\mathcal{L}^p$ such that $f_n \to f$ in $\mathcal{L}^p$. Then there exists a subsequence $\{f_{n_k}\}_{k\in\mathbb{N}}$ of $\{f_n\}_{n\in\mathbb{N}}$ such that $f_{n_k} \to f$, a.e.*

We have seen above how the concavity of the function $\log$ was used in the proof of Young's inequality (Lemma 4.12). A generalization of the definition of convexity, called *Jensen's inequality*, is one of the most powerful tools in measure theory. Recall that $\mathcal{L}^{0-1}$ denotes the set of all $f \in \mathcal{L}^0$ with $f^- \in \mathcal{L}^1$.

Proposition 4.21 (Jensen's inequality) *Suppose that $\mu(S) = 1$ (μ is a probability measure) and that $\varphi : \mathbb{R} \to \mathbb{R}$ is a convex function. For a function $f \in \mathcal{L}^1$ we have $\varphi(f) \in \mathcal{L}^{0-1}$ and*

$$\int \varphi(f) \, d\mu \geq \varphi(\int f \, d\mu).$$

Before we give a proof, we need a lemma about convex functions:

Lemma 4.22 (Convex functions as suprema of sequences of affine functions) *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function. Then, there exists two sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ of real numbers such that*

$$\varphi(x) = \sup_{n\in\mathbb{N}}(a_n x + b_n).$$

PROOF **(*)** For $x \in \mathbb{R}$, we define the left and right derivative $\frac{\partial}{\partial x}\varphi^-$ and $\frac{\partial}{\partial x}\varphi^+$ of φ at x by

$$\frac{\partial}{\partial x}\varphi^-(x) = \sup_{\varepsilon>0} \tfrac{1}{\varepsilon}(\varphi(x) - \varphi(x - \varepsilon)),$$

and

$$\frac{\partial}{\partial x}\varphi^+(x) = \inf_{\varepsilon>0} \tfrac{1}{\varepsilon}(\varphi(x + \varepsilon) - \varphi(x)).$$

Convexity of the function φ implies that the difference quotient

$$\varepsilon \mapsto \tfrac{1}{\varepsilon}(\varphi(x + \varepsilon) - \varphi(x)), \ \varepsilon \neq 0,$$

is a non-decreasing function (why?), and so both $\frac{\partial}{\partial x}\varphi^-$ and $\frac{\partial}{\partial x}\varphi^+$ are, in fact, limits as $\varepsilon > 0$. Moreover, we always have

(4.6) $$\tfrac{1}{\varepsilon'}(\varphi(x) - \varphi(x - \varepsilon')) \leq \frac{\partial}{\partial x}\varphi^-(x) \leq \frac{\partial}{\partial x}\varphi^+(x) \leq \tfrac{1}{\varepsilon}(\varphi(x + \varepsilon) - \varphi(x)),$$

for all x and all $\varepsilon, \varepsilon' > 0$.

Let $\{q_n\}_{n\in\mathbb{N}}$ be an enumeration of rational numbers in $\mathbb{R}$. For each $n \in \mathbb{N}$ we pick $a_n \in [\frac{\partial}{\partial x}\varphi^-(q_n), \frac{\partial}{\partial x}\varphi^+(q_n)]$ and set $b_n = \varphi(q_n) - a_n q_n$, so that the line $x \mapsto a_n x + b_n$ passes through $(q_n, \varphi(q_n))$ and has a slope which is between the left and the right derivative of φ at q_n.

Let us first show that $\varphi(x) \geq a_n x + b_n$ for all $x \in \mathbb{R}$ and all $n \in \mathbb{N}$. We pick $x \in \mathbb{R}$ and assume that $x \geq q_n$ (the case $x < q_n$ is analogous). If $x = q_n$ then $\varphi(x) = a_n x + b_n$ by construction, and we are done. When $\varepsilon = x - q_n > 0$, relation (4.6) implies that

$$a_n x + b_n = a_n(x - q_n) + \varphi(q_n) \leq \left(\tfrac{\varphi(x) - \varphi(q_n)}{x - q_n}\right)(x - q_n) + \varphi(q_n) = \varphi(x).$$

Conversely, suppose that $\varphi(x) > \sup_n(a_n x + b_n)$, for some $x \in \mathbb{R}$. Both functions φ and ψ, where $\psi(x) = \sup_n(a_n x + b_n)$ are convex (why?), and, therefore, continuous. It follows from $\varphi(x) > \psi(x)$, that there exists a rational number q_n such that $\varphi(q_n) > \psi(q_n)$. This is a contradiction, though, since $\psi(q_n) \geq a_n q_n + b_n = \varphi(q_n)$. ∎

PROOF (PROOF OF PROPOSITION 4.21) Let us first show that $(\varphi(f))^- \in \mathcal{L}^1$. By Lemma 4.22, there exists sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ such that $\varphi(x) = \sup_{n\in\mathbb{N}}(a_n x + b_n)$. In particular, $\varphi(f(x)) \geq a_1 f(x) + b_1$, for all $x \in S$. Therefore,

$$\big(\varphi(f(x))\big)^- \leq (a_1 f(x) + b_1)^- \leq |a_1|\,|f(x)| + |b_1| \in \mathcal{L}^1.$$

Next, we have $\int \varphi(f)\,d\mu \geq \int a_n f + b_n\,d\mu = a_n \int f\,d\mu + b_n$, for all $n \in \mathbb{N}$. Therefore,

$$\int \varphi(f)\,d\mu \geq \sup_n\left(a_n \int f\,d\mu + b_n\right) = \varphi\left(\int f\,d\mu\right).$$

∎

Problem 4.23 State and prove a generalization of Jensen's inequality when φ is defined only on an interval I of $\mathbb{R}$, but $\mu(\{f \notin I\}) = 0$.

Problem 4.24 Use Jensen's inequality on an appropriately chosen measure space to prove the **arithmetic-geometric inequality**

$$\tfrac{a_1 + \cdots + a_n}{n} \geq \sqrt[n]{a_1 \ldots a_n}, \text{ for } a_1, \ldots, a_n \geq 0.$$

The following inequality is know as **Markov's inequality** in probability theory, but not much wider than that. In analysis it is known as **Chebyshev's inequality**.

Proposition 4.25 (Markov's inequality) *For $f \in \mathcal{L}^0_+$ and $\alpha > 0$ we have*

$$\mu(\{f \geq \alpha\}) \leq \tfrac{1}{\alpha}\int f\,d\mu.$$

PROOF Consider the function $g \in \mathcal{L}^0_+$ defined by

$$g(x) = \alpha\mathbf{1}_{\{f \geq \alpha\}} = \begin{cases} \alpha, & f(x) \in [\alpha, \infty) \\ 0, & f(x) \in [0, \alpha). \end{cases}$$

Then $f(x) \geq g(x)$ for all $x \in S$, and so

$$\int f\,d\mu \geq \int g\,d\mu = \alpha\mu(\{f \geq \alpha\}).$$

∎

4.3 Additional problems

Problem 4.26 (Projections onto a convex set) A subset K of a vector space is said to be **convex** if $\alpha x + (1 - \alpha)y \in K$, whenever $x, y \in K$ and $\alpha \in [0, 1]$. Let K be a closed and convex subset of $\mathcal{L}^2$, and let g be an element of its complement $\mathcal{L}^2 \setminus K$. Prove that

1. There exists an element $f^* \in K$ such that $||g - f^*||_{\mathcal{L}^2} \leq ||g - f||_{\mathcal{L}^2}$, for all $f \in K$.

2. $\int (f - f^*)(g - f^*)\, d\mu \leq 0$, for all $f \in K$.

(*Hint:* Pick a sequence $\{f_n\}_{n \in \mathbb{N}}$ in K with $||f_n - g||_{\mathcal{L}^2} \to \inf_{f \in K} ||f - g||_{\mathcal{L}^2}$ and show that it is Cauchy. Use (but prove first) the **parallelogram identity** $2||h||_{\mathcal{L}^2}^2 + 2||k||_{\mathcal{L}^2}^2 = ||h + k||_{\mathcal{L}^2}^2 + ||h - k||_{\mathcal{L}^2}^2$, for $h, k \in \mathcal{L}^2$.)

Problem 4.27 (Egorov's theorem) Suppose that μ is a finite measure, and let $\{f_n\}_{n \in \mathbb{N}}$ be a sequence in $\mathcal{L}^0$ which converges a.e. to $f \in \mathcal{L}^0$. Prove that for each $\varepsilon > 0$ there exists $E \in \mathcal{S}$ with $\mu(E^c) < \varepsilon$ such that
$$\lim_{n \to \infty} \operatorname{esssup} |f_n \mathbf{1}_E - f \mathbf{1}_E| = 0.$$

(*Hint:* Define $A_{n,k} = \cup_{m \geq n}\{|f_m - f| \geq \frac{1}{k}\}$, show that for each $k \in \mathbb{N}$, there exists $n_k \in \mathbb{N}$ such that $\mu(A_{n_k,k}) < \varepsilon/2^k$, and set $E = \cap_k A_{n_k,k}^c$.)

Problem 4.28 (Relationships between different $\mathcal{L}^p$ spaces)

1. Show that for $p, q \in [1, \infty)$, we have
$$||f||_{\mathcal{L}^p} \leq ||f||_{\mathcal{L}^q} \mu(S)^r,$$
where $r = 1/p - 1/q$. Conclude that $\mathcal{L}^q \subseteq \mathcal{L}^p$, for $p \leq q$ if $\mu(S) < \infty$.

2. For $p_0 \in [1, \infty)$, construct an example of a measure space $(S, \mathcal{S}, \mu)$ and a function $f \in \mathcal{L}^0$ such that $f \in \mathcal{L}^p$ if and only if $p = p_0$.

3. Suppose that $f \in \mathcal{L}^r \cap \mathcal{L}^\infty$, for some $r \in [1, \infty)$. Show that $f \in \mathcal{L}^p$ for all $p \in [r, \infty)$ and
$$||f||_{\mathcal{L}^\infty} = \lim_{p \to \infty} ||f||_{\mathcal{L}^p}.$$

Problem 4.29 (Convergence in measure) A sequence $\{f_n\}_{n \in \mathbb{N}}$ in $\mathcal{L}^0$ is said to **converge in measure** toward $f \in \mathcal{L}^0$ if
$$\forall \varepsilon > 0, \ \mu\Big(\big\{|f_n - f| \geq \varepsilon\big\}\Big) \to 0 \text{ as } n \to \infty.$$
Assume that $\mu(S) < \infty$ (parts marked by (†) are true without this assumption).

1. Show that the mapping
$$d(f, g) = \int \frac{|f - g|}{1 + |f - g|}\, d\mu, \ f, g \in \mathcal{L}^0,$$
defines a pseudo metric on $\mathcal{L}^0$ and that convergence in d is equivalent to convergence in measure.

2. Show that $f_n \to f$, a.e., implies that $f_n \to f$ in measure. Give an example which shows that the assumption $\mu(S) < \infty$ is necessary.

3. Give an example of a sequence which converges in measure, but not a.e.

4. (†) For $f \in \mathcal{L}^0$ and a sequence $\{f_n\}_{n \in \mathbb{N}}$ in $\mathcal{L}^0$, suppose that

$$\sum_{n \in \mathbb{N}} \mu\left(\left\{|f_n - f| \geq \varepsilon\right\}\right) < \infty, \text{ for all } \varepsilon > 0.$$

Show that $f_n \to f$, a.e.

5. (†) Show that each sequence convergent in measure has a subsequence which converges a.e.

6. (†) Show that each sequence convergent in $\mathcal{L}^p$, $p \in [1, \infty)$ converges in measure.

7. For $p \in [1, \infty)$, find an example of a sequence which converges in measure, but not in $\mathcal{L}^p$.

8. Let $\{f_n\}_{n \in \mathbb{N}}$ be a sequence in $\mathcal{L}^0$ with the property that any of its subsequences admits a (further) subsequence which converges a.e. to $f \in \mathcal{L}^0$. Show that $f_n \to f$ in measure.

9. Let $\Phi : \mathbb{R}^2 \to \mathbb{R}$ be a continuous function, and let $\{f_n\}_{n \in \mathbb{N}}$ and $\{g_n\}_{n \in \mathbb{N}}$ be two sequences in $\mathcal{L}^0$. If $f, g \in \mathcal{L}^0$ are such that $f_n \to f$ and $g_n \to g$ in measure, then

$$\Phi(f_n, g_n) \to \Phi(f, g) \text{ in measure.}$$

(*Note:* The useful examples include $\Phi(x, y) = x + y$, $\Phi(x, y) = xy$, etc.)

Chapter 5

Theorems of Fubini-Tonelli and Radon-Nikodym

5.1 Products of measure spaces

We have seen in Chapter 2 that it is possible to define products of arbitrary collections of measurable spaces - one generates the σ-algebra on the product by all finite-dimensional cylinders. The purpose of the present section is to extend that construction to products of measure spaces, i.e., to define products of measures.

Let us first consider the case of two measure spaces $(S, \mathcal{S}, \mu_S)$ and $(T, \mathcal{T}, \mu_T)$. If the measures are stripped, the product $S \times T$ is endowed with the product σ-algebra $\mathcal{S} \otimes \mathcal{T} = \sigma(\{A \times B : A \in \mathcal{S}, B \in \mathcal{T}\})$. The family $\mathcal{P} = \{A \times B : A \in \mathcal{S}, B \in \mathcal{T}\}$ serves as a good starting point towards the creation of the product measure $\mu_S \otimes \mu_T$. Indeed, if we interpret of the elements in $\mathcal{P}$ as rectangles of sorts, it is natural to define

$$(\mu_S \otimes \mu_T)(A \times B) = \mu_S(A)\mu_T(B).$$

The family $\mathcal{P}$ is a π-system (why?), but not necessarily an algebra, so we cannot use Theorem 2.10 (Caratheodory's extension theorem) to define an extension of $\mu_S \otimes \mu_T$ to the whole $\mathcal{S} \otimes \mathcal{T}$. It it not hard, however, to enlarge $\mathcal{P}$ a little bit, so that the resulting set is an algebra, but that the measure $\mu_S \otimes \mu_T$ can still be defined there in a natural way. Indeed, consider the smallest algebra that contains $\mathcal{P}$. It is easy to see that it must contain the family $\mathcal{S}$ defined by

$$\mathcal{A} = \{\cup_{k=1}^{n} A_k \times B_k : n \in \mathbb{N}, A_k \in \mathcal{S}, B_k \in \mathcal{T}, k = 1, \ldots, n\}.$$

Problem 5.1 Show that $\mathcal{A}$ is, in fact, an algebra and that each element $C \in \mathcal{A}$ can be written in the form

$$C = \cup_{k=1}^{n} A_k \times B_k,$$

for $n \in \mathbb{N}, A_k \in \mathcal{S}, B_k \in \mathcal{T}, k = 1, \ldots, n$, such that $A_1 \times B_1, \ldots, A_n \times B_n$ are *pairwise disjoint*.

The problem above allows us to extend the definition of the set function $\mu_S \otimes \mu_T$ to the entire $\mathcal{A}$ by

$$(\mu_S \otimes \mu_T)(C) = \sum_{k=1}^{n} \mu_S(A_k)\mu_T(B_k),$$

where $C = \cup_{k=1}^{n} A_k \times B_k$ for $n \in \mathbb{N}, A_k \in \mathcal{S}, B_k \in \mathcal{T}, k = 1, \ldots, n$ is a representation of C with pairwise disjoint $A_1 \times B_1, \ldots, A_n \times B_n$.

At this point, we could attempt to show that the so-defined set function is σ-additive on $\mathcal{A}$ and extend it using the Caratheodory extension theorem. This is indeed possible - under the additional assumption of σ-finiteness - but we will establish the existence of product measures as a side-effect in the proof of Fubini's theorem below.

Lemma 5.2 (Sections of measurable sets are measurable) *Let C be an $\mathcal{S} \otimes \mathcal{T}$-measurable subset of $S \times T$. For each $x \in S$ the **section** $C_x = \{y \in T \, : \, (x, y) \in C\}$ is measurable in $\mathcal{T}$.*

PROOF In the spirit of most of the measurability arguments seen so far in these notes, let $\mathcal{C}$ denote the family of all $C \in \mathcal{S} \times \mathcal{T}$ such that C_x is $\mathcal{T}$-measurable for each $x \in S$. Clearly, the "rectangles" $A \times B$, $A \in \mathcal{S}$, $B \in \mathcal{T}$ are in $\mathcal{A}$ because their sections are either equal to $\emptyset$ or B, for each $x \in S$. Remember that the set of all rectangles generates $\mathcal{S} \otimes \mathcal{T}$. The proof of the theorem will, therefore, be complete once it is established that $\mathcal{C}$ is a σ-algebra. This easy exercise is left to the reader. ∎

Problem 5.3 Show that an analogous result holds for measurable functions, i.e., show that if $f : S \times T \to \bar{\mathbb{R}}$ is a $\mathcal{S} \otimes \mathcal{T}$-measurable function, then the function $x \mapsto f(x, y_0)$ is $\mathcal{S}$-measurable for each $y_0 \in T$, and the function $y \mapsto f(x_0, y)$ is $\mathcal{T}$-measurable for each $y_0 \in T$.

Proposition 5.4 (A simple Cavallieri's principle) *Let μ_S and μ_T be finite measures. For $C \in \mathcal{S} \otimes \mathcal{T}$, define the functions $\varphi_C : T \to [0, \infty)$ and $\psi_C : S \to [0, \infty)$ by*

$$\varphi_C(y) = \mu_S(C_y), \ \psi_C(x) = \mu_T(C_x).$$

Then,

1. *$\varphi_C \in \mathcal{L}_+^0(\mathcal{T})$,*

2. *$\psi_C \in \mathcal{L}_+^0(\mathcal{S})$,*

3. *$\int \varphi_C \, d\mu_T = \int \psi_C \, d\mu_S$.*

PROOF Note that, by Problem 5.3, the function $x \mapsto \mathbf{1}_C(x, y)$ is $\mathcal{S}$-measurable for each $y \in T$. Therefore,

(5.1) $$\int \mathbf{1}_C(\cdot, y) \, d\mu_S = \mu_S(C_y) = \varphi_C(y),$$

and the function φ_C is well-defined.

Let $\mathcal{C}$ denote the family of all sets in $\mathcal{S} \otimes \mathcal{T}$ such that (1), (2) and (3) in the statement of the proposition hold. First, observe that $\mathcal{C}$ contains all rectangles $A \times B$, $A \in \mathcal{S}$, $B \in \mathcal{T}$, i.e., it contains a π-system which generates $\mathcal{S} \otimes \mathcal{T}$. So, by the π-λ Theorem (Theorem 2.16), it will be enough to show that $\mathcal{C}$ is a λ-system. We leave the details to the reader (*Hint:* Use representation (5.1) and the monotone convergence theorem. Where is the finiteness of the measures used?) ∎

Proposition 5.5 (Simple Cavallieri holds for σ-finite measures) *The conclusion of Proposition 5.4 continues to hold if we assume that μ_S and μ_T are only σ-finite.*

PROOF **(*)** Thanks to σ-finiteness, there exists pairwise disjoint sequences $\{A_n\}_{n\in\mathbb{N}}$ and $\{B_n\}_{n\in\mathbb{N}}$ in $\mathcal{S}$ and $\mathcal{T}$, respectively, such that $\cup_n A_n = S$, $\cup_m B_m = T$ and $\mu_S(A_n) < \infty$ and $\mu_S(B_m) < \infty$, for all $m, n \in \mathbb{N}$.

For $m, n \in \mathbb{N}$, define the set-functions μ_S^n and μ_T^m on $\mathcal{S}$ and $\mathcal{T}$ respectively by

$$\mu_S^n(A) = \mu_S(A_n \cap A), \ \mu_T^m(B) = \mu_T(B_m \cap B).$$

It is easy to check that all μ_S^n and μ_T^m, $m, n \in \mathbb{N}$ are finite measures on $\mathcal{S}$ and $\mathcal{T}$, respectively. Moreover, $\mu_S(A) = \sum_{n=1}^{\infty} \mu_S^n(A)$, $\mu_T(B) = \sum_{m=1}^{\infty} \mu_T^m(B)$. In particular, if we set $\varphi_C^n(y) = \mu_S^n(C_y)$ and $\psi_C^m(x) = \mu_T^m(C_x)$, for all $x \in S$ and $y \in S$, we have

$$\varphi_C(y) = \mu_S(C_y) = \sum_{n=1}^{\infty} \mu_S^n(C_y) = \sum_{n=1}^{\infty} \varphi_C^n(y), \text{ and}$$

$$\psi_C(x) = \mu_T(C_x) = \sum_{m=1}^{\infty} \mu_T^m(C_x) = \sum_{m=1}^{\infty} \psi_C^m(x),$$

for all $x \in S$, $y \in T$.

We can apply the conclusion of Proposition 5.4 to all pairs $(S, \mathcal{S}, \mu_S^n)$ and $(T, \mathcal{T}, \mu_T^m)$, $m, n \in \mathbb{N}$, of finite measure spaces to conclude that all elements of the sums above are measurable functions and that so are φ_C and ψ_C.

Similarly, the sequences of non-negative functions $\sum_{i=1}^{n} \varphi_C^n(y)$ and $\sum_{i=1}^{m} \psi_C^i(x)$ are non-decreasing and converge to φ_C and ψ_C. Therefore, by the monotone convergence theorem,

$$\int \varphi_C \, d\mu_T = \lim_n \sum_{i=1}^{n} \int \varphi_C^i \, d\mu_T, \text{ and } \int \psi_C \, d\mu_S = \lim_n \sum_{i=1}^{n} \int \psi_C^i \, d\mu_S.$$

On the other hand, we have $\int \varphi_C^n d\mu_T^m = \int \psi_C^m \, d\mu_S^n$, by Proposition 5.4. Summing over all $n \in \mathbb{N}$ we have

$$\int \varphi_C \, d\mu_T^m = \sum_{n\in\mathbb{N}} \int \psi_C^m \, d\mu_S^n = \int \psi_C^m \, d\mu_S,$$

where the last equality follows from the fact (see Problem 5.7 below) that

$$\sum_{n\in\mathbb{N}} \int f \, d\mu_S^n = \int f \, d\mu_S,$$

for all $f \in \mathcal{L}_+^0$. Another summation - this time over $m \in \mathbb{N}$ - completes the proof. ∎

Remark 5.6 The argument of the proof above uncovers the fact that integration is a bilinear operation, i.e., that the mapping

$$(f, \mu) \to \int f \, d\mu,$$

is linear in both arguments.

Problem 5.7 Let $\{A_n\}_{n\in\mathbb{N}}$ be a measurable partition of S, and let the measure μ^n be defined by $\mu^n(A) = \mu(A \cap A_n)$ for all $A \in \mathcal{S}$. Show that for $f \in \mathcal{L}_+^0$, we have

$$\int f \, d\mu = \sum_{n\in\mathbb{N}} \int f \, d\mu^n.$$

Proposition 5.8 (Finite products of measure spaces) *Let $(S_i, \mathcal{S}_i, \mu_i)$, $i = 1, \dots, n$ be finite measure spaces. There exists a unique measure - denoted by $\mu_1 \otimes \cdots \otimes \mu_n$ - on the product space $(S_1 \times \cdots \times S_n, \mathcal{S}_1 \otimes \cdots \otimes \mathcal{S}_n)$ with the property that*

$$(\mu_1 \otimes \cdots \otimes \mu_n)(A_1 \times \cdots \times A_n) = \mu_1(A_1) \dots \mu_n(A_n),$$

for all $A_i \in \mathcal{S}_i$, $i = 1, \dots, n$. Such a measure is necessarily σ-finite.

PROOF To simplify the notation, we assume that $n = 2$ - the general case is very similar. For $C \in \mathcal{S}_1 \otimes \mathcal{S}_2$, we define

$$(\mu_1 \otimes \mu_2)(C) = \int_{S_2} \varphi_C \, d\mu_2, \text{ where } \varphi_C(y) = \mu_1(C_y) \text{ and } C_y = \{x \in S_1 \,:\, (x, y) \in C\}.$$

It follows from Proposition 5.5 that $\mu_1 \otimes \mu_2$ is well-defined as a map from $\mathcal{S}_1 \otimes \mathcal{S}_2$ to $[0, \infty]$. Also, it is clear that $(\mu_1 \otimes \mu_2)(A \times B) = \mu_1(A)\mu_2(B)$, for all $A \in \mathcal{S}_1$, $B \in \mathcal{S}_2$. It remains to show that $\mu_1 \otimes \mu_2$ is a measure. We start with a pairwise disjoint sequence $\{C_n\}_{n\in\mathbb{N}}$ in $\mathcal{S}_1 \otimes \mathcal{S}_2$. For $y \in S_2$, the sequence $\{(C_n)_y\}_{n\in\mathbb{N}}$ is also pairwise disjoint, and so, with $C = \cup_n C_n$, we have

$$\varphi_C(y) = \mu_1(C_y) = \sum_{n\in\mathbb{N}} \mu_2\Big((C_n)_y\Big) = \sum_{n\in\mathbb{N}} \varphi_{C_n}(y), \ \forall \, y \in S_2.$$

Therefore, by the monotone convergence theorem (see Problem 3.37 for details) we have

$$(\mu_1 \otimes \mu_2)(C) = \int_{S_2} \varphi_C \, d\mu_2 = \sum_{n\in\mathbb{N}} \int_{S_2} \varphi_{C_n} \, d\mu = \sum_{n\in\mathbb{N}} (\mu_1 \otimes \mu_2)(C_n).$$

Finally, let $\{A_n\}_{n\in\mathbb{N}}$, $\{B_n\}_{n\in\mathbb{N}}$ be sequences in $\mathcal{S}_1$ and $\mathcal{S}_2$ (respectively) such that $\mu_1(A_n) < \infty$ and $\mu_2(B_n) < \infty$ for all $n \in \mathbb{N}$ and $\cup_n A_n = S_1$, $\cup_n B_n = S_2$. Define $\{C_n\}_{n\in\mathbb{N}}$ as an enumeration of the countable family $\{A_i \times B_j \,:\, i, j \in \mathbb{N}\}$ in $\mathcal{S}_1 \otimes \mathcal{S}_2$. Then $(\mu_1 \otimes \mu_2)(C_n) < \infty$ and all $n \in \mathbb{N}$ and $\cup_n C_n = S_1 \times S_2$. Therefore, $\mu_1 \otimes \mu_2$ is σ-finite. ∎

The measure $\mu_1 \otimes \cdots \otimes \mu_n$ is called the **product measure**, and the measure space $(S_1 \times \cdots \times S_n, \mathcal{S}_1 \otimes \cdots \otimes \mathcal{S}_n, \mu_1 \otimes \cdots \otimes \mu_n)$ the **product (measure space)** of measure spaces $(S_1, \mathcal{S}_1, \mu_1), \dots, (S_n, \mathcal{S}_n, \mu_n)$.

Now that we know that product measures exist, we can state and prove the important theorem which, when applied to integrable functions bears the name of Fubini, and when applied to non-negative functions, of Tonelli. We state it for both cases simultaneously (i.e., on $\mathcal{L}^{0-1}$) in the case of a product of two measure spaces. An analogous theorem for finite products can be readily derived from it. When the variable or the underlying measure space of integration needs to be specified, we write $\int_S f(x) \, \mu(dx)$ for the Lebesgue integral $\int f \, d\mu$.

Theorem 5.9 (Fubini, Tonelli) *Let $(S, \mathcal{S}, \mu_S)$ and $(T, \mathcal{T}, \mu_T)$ be two σ-finite measure spaces. For $f \in \mathcal{L}^{0-1}(S \times T)$ we have*

(5.2)
$$\int_S \left(\int_T f(x, y) \, \mu_T(dy) \right) \mu_S(dx) = \int_T \left(\int_S f(x, y) \, \mu_S(dx) \right) \mu_T(dy)$$
$$= \int f \, d(\mu_S \otimes \mu_T).$$

PROOF All the hard work has already been done. We simply need to crank the Standard Machine. Let $\mathcal{H}$ denote the family of all functions in $\mathcal{L}^0_+(S \times T)$ with the property that (5.2) holds. Proposition 5.5 implies that $\mathcal{H}$ contains the indicators of all elements of $\mathcal{S} \otimes \mathcal{T}$. Linearity of all components of (5.2) implies that $\mathcal{H}$ contains all simple functions in $\mathcal{L}^0_+$, and the approximation theorem 3.16 implies that the whole $\mathcal{L}^0_+$ is in $\mathcal{H}$. Finally, the extension to $\mathcal{L}^{0-1}$ follows by additivity. ∎

Since f^- is always in $\mathcal{L}^{0-1}$, we have the following corollary

Corollary 5.10 (An integrability criterion) *For $f \in \mathcal{L}^0(S \times T)$, we have*

$$f \in \mathcal{L}^{0-1}(S \times T) \text{ if and only if } \int_S \left(\int_T f^-(x, y) \, \mu_T(dy) \right) \mu_S(dx) < \infty.$$

Example 5.11 (σ-finiteness cannot be left out ...) The assumption of σ-finiteness cannot be left out of the statement of Theorem 5.9. Indeed, let $(S, \mathcal{S}, \mu) = ([0, 1], \mathcal{B}([0, 1]), \lambda)$ and $(T, \mathcal{T}, \nu) = ([0, 1], 2^{[0,1]}, \gamma)$, where γ is the counting measure on $2^{[0,1]}$, so that $(T, \mathcal{T}, \nu)$ fails the σ-finite property. Define $f \in \mathcal{L}^0(S \times T)$ (why it is product-measurable?) by

$$f(x, y) = \begin{cases} 1, & x = y, \\ 0, & x \neq y. \end{cases}$$

Then
$$\int_S f(x, y) \, \mu(dx) = \lambda(\{y\}) = 0,$$

and so
$$\int_T \int_S f(x, y) \, \mu(dx) \, \nu(dy) = \int_{[0,1]} 0 \, \gamma(dy) = 0.$$

On the other hand,
$$\int_T f(x, y) \, \nu(dy) = \gamma(\{x\}) = 1,$$

and so
$$\int_S \int_T f(x, y) \, \nu(dy) \, \mu(dx) = \int_{[0,1]} 1 \, \lambda(dx) = 1.$$

Example 5.12 (... and neither can product-integrability) The integrability of either f^+ or f^- for $f \in \mathcal{L}^0(S \times T)$ is (essentially) necessary for validity of Fubini's theorem, even if all iterated integrals exist. Here is what can go wrong. Let $(S, \mathcal{S}, \mu) = (T, \mathcal{T}, \nu) = (\mathbb{N}, 2^{\mathbb{N}}, \gamma)$, where γ is the counting measure. Define the function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ by

$$ f(n, m) = \begin{cases} 1, & m = n, \\ -1, & m = n + 1, \\ 0, & \text{otherwise} \end{cases} $$

Then

$$ \int_T f(n, m) \, \gamma(dm) = \sum_{m \in \mathbb{N}} f(n, m) = 0 + \cdots + 0 + 1 + (-1) + 0 + \cdots = 0, $$

and so

$$ \int_S \int_T f(n, m) \, \gamma(dm) \, \gamma(dn) = 0. $$

On the other hand,

$$ \int_S f(n, m) \, \gamma(dn) = \sum_{n \in \mathbb{N}} f(n, m) = \begin{cases} 1 + 0 + \cdots = 1, & m = 1 \\ 0 + \cdots + 0 + (-1) + 1 + 0 + \ldots, & m > 1, \end{cases} $$

i.e.,

$$ \int_S f(n, m) \, \gamma(dn) = \mathbf{1}_{\{m=1\}}. $$

Therefore,

$$ \int_T \int_S f(n, m) \, \gamma(dn) \, \gamma(dm) = \int_T \mathbf{1}_{\{m=1\}} \, \gamma(dm) = 1. $$

If you think that using the counting measure is cheating, convince yourself that it is not hard to transfer this example to the setup where $(S, \mathcal{S}, \mu) = (T, \mathcal{T}, \nu) = ([0, 1], \mathcal{B}([0, 1]), \lambda)$.

The existence of the product measure gives us an easy access to the Lebesgue measure on higher-dimensional Euclidean spaces. Just as λ on $\mathbb{R}$ measures the "length" of sets, the Lebesgue measure on $\mathbb{R}^2$ will measure "area", the one on $\mathbb{R}^3$ "volume", etc. Its properties are collected in the following problem:

Problem 5.13 For $n \in \mathbb{N}$, show the following statements:

1. There exists a unique measure λ (note the notation overload) on $\mathcal{B}(\mathbb{R}^n)$ with the property that

$$ \lambda\Big([a_1, b_1) \times \cdots \times [a_n, b_n)\Big) = (b_1 - a_1) \ldots (b_n - a_n), $$

for all $a_1 < b_1, \ldots, a_n < b_n$ in $\mathbb{R}$.

2. The measure λ on $\mathbb{R}^n$ is invariant with respect to all isometries[1] of $\mathbb{R}^n$.

Note: Feel free to use the following two facts without proof:

a) A function $f : \mathbb{R}^n \to \mathbb{R}^n$ is an isometry if and only if there exists $x_0 \in \mathbb{R}^n$ and an orthogonal linear transformation $O : \mathbb{R}^n \to \mathbb{R}^n$ such that $f(x) = x_0 + Ox$.

---

[1]An **isometry** of $\mathbb{R}^n$ is a map $f : \mathbb{R}^n \to \mathbb{R}^n$ with the property that $d(x, y) = d(f(x), f(y))$ for all $x, y \in \mathbb{R}^n$. It can be shown that the isometries of $\mathbb{R}^3$ are precisely translations, rotations, reflections and compositions thereof.

b) Let O be an orthogonal linear transformation. Then R_1 and OR_1 have the same Lebesgue measure, where R_1 denotes the unit rectangle $[0, 1) \times \cdots \times [0, 1)$. (the least painful way to prove this fact is by using the change-of-variable formula for the Riemann integral).

5.2   The Radon-Nikodym Theorem

We start the discussion of the Radon-Nikodym theorem with a simple observation:

Problem 5.14 (Integral as a measure) For a function $f \in \mathcal{L}^0([0, \infty])$, we define the set-function $\nu : \mathcal{S} \to [0, \infty]$ by

(5.3)
$$\nu(A) = \int_A f \, d\mu.$$

1. Show that ν is a measure.

2. Show that $\mu(A) = 0$ implies $\nu(A) = 0$, for all $A \in \mathcal{S}$.

3. Show that the following two properties are equivalent

   - $\mu(A) = 0$ if and only if $\nu(A) = 0$, $A \in \mathcal{S}$, and
   - $f > 0$, a.e.

Definition 5.15 (Absolute continuity, etc.) *Let μ, ν be measures on the measurable space $(S, \mathcal{S})$. We say that*

1. *ν is **absolutely continuous** with respect to μ - denoted by $\nu \ll \mu$ - if $\nu(A) = 0$, whenever $\mu(A) = 0$, $A \in \mathcal{S}$.*

2. *μ and ν are **equivalent** if $\nu \ll \mu$ and $\mu \ll \nu$, i.e., if $\mu(A) = 0 \Leftrightarrow \nu(A) = 0$, for all $A \in \mathcal{S}$,*

3. *μ and ν are **(mutually) singular** - denoted by $\mu \perp \nu$ - if there exists $D \in \mathcal{S}$ such that $\mu(D) = 0$ and $\nu(D^c) = 0$.*

Problem 5.16 Let μ and ν be measures with ν finite and $\nu \ll \mu$. Show that for each $\varepsilon > 0$ there exists $\delta > 0$ such that for each $A \in \mathcal{S}$, we have $\mu(A) \leq \delta \Rightarrow \nu(A) \leq \varepsilon$. Show that the assumption that ν is finite is necessary.

Problem 5.14 states that the prescription (5.3) defines a measure on $\mathcal{S}$ which is absolutely continuous with respect to μ. What is surprising is that the converse also holds under the assumption of σ-finiteness: all absolutely continuous measures on $\mathcal{S}$ are of that form. That statement (and more) is the topic of this section. Since there is more than one measure in circulation, we use the convention that *a.e.* always uses the notion of the null set as defined by the measure μ.

Theorem 5.17 (The Lebesgue decomposition) *Let $(S, \mathcal{S})$ be a measurable space and let μ and ν be two σ-finite measures on $\mathcal{S}$. Then there exists a unique decomposition $\nu = \nu_a + \nu_s$, where*

---

1. $\nu_a \ll \mu$,

2. $\nu_s \perp \mu$.

Furthermore, there exists an a.e.-unique function $f \in \mathcal{L}^0_+$ such that

$$\nu_a(A) = \int_A f \, d\mu.$$

PROOF (*)

   - *Uniqueness.* Suppose that $\nu^1_a + \nu^1_s = \nu = \nu^2_a + \nu^2_s$ are two decompositions satisfying (1) and (2) in the statement. Let D^1 and D^2 be as in the definition of mutual singularity applied to the pairs μ, ν^1_s and μ, ν^2_s, respectively. Set $D = D^1 \cup D^2$, and note that $\mu(D) = 0$ and $\nu^1_s(D^c) = \nu^2_s(D^c) = 0$. For any $A \in \mathcal{S}$, we have $\mu(A \cap D) = 0$ and so, thanks to absolute continuity,

$$\nu^1_a(A \cap D) = \nu^2_a(A \cap D) = 0 \text{ and, consequently, } \nu^1_s(A \cap D) = \nu^2_s(A \cap D) = \nu(A \cap D).$$

By singularity,

$$\nu^1_s(A \cap D^c) = \nu^2_s(A \cap D^c) = 0, \text{ and, consequently, } \nu^1_a(A \cap D^c) = \nu^2_a(A \cap D^c) = \nu(A \cap D^c).$$

Finally,
$$\nu^1_a(A) = \nu^1_a(A \cap D) + \nu^1_a(A \cap D^c) = \nu^2_a(A \cap D) + \nu^2_a(A \cap D^c) = \nu^2(A),$$

and, similarly, $\nu^1_s = \nu^2_s$.

   To establish the uniqueness of the function f with the property that $\nu_a(A) = \int_A f \, d\mu$ for all $A \in \mathcal{S}$, we assume that there are two such functions, f_1 and f_2, say. Define the sequence $\{B_n\}_{n \in \mathbb{N}}$ by

$$B_n = \{f_1 \geq f_2\} \cap C_n,$$

where $\{C_n\}_{n \in \mathbb{N}}$ is a pairwise-disjoint sequence in $\mathcal{S}$ with the property that $\nu(C_n) < \infty$, for all $n \in \mathbb{N}$ and $\cup_n C_n = S$. Then, with $g_n = f_1 \mathbf{1}_{B_n} - f_2 \mathbf{1}_{B_n} \in \mathcal{L}^1_+$ we have

$$\int g_n \, d\mu = \int_{B_n} f_1 \, d\mu - \int_{B_n} f_2 \, d\mu = \nu_a(B_n) - \nu_a(B_n) = 0.$$

By Problem 3.29, we have $g_n = 0$, a.e., i.e., $f_1 = f_2$, a.e., on B_n, for all $n \in \mathbb{N}$, and so $f_1 = f_2$, a.e., on $\{f_1 \geq f_2\}$. A similar argument can be used to show that $f_1 = f_2$, a.e., on $\{f_1 < f_2\}$, as well.

   - *Existence.* By decomposing S into a countable measurable partition whose elements have finite μ and ν measures, we may (and do) assume that both μ and ν are finite. Let $\mathcal{R}$ denote the set of all functions $f \in \mathcal{L}^0([0, \infty])$ with the property that $\nu(A) \geq \int_A f \, d\mu$ for all $A \in \mathcal{S}$. The reader will have no difficulty showing that

1. if $f_1, f_2 \in \mathcal{R}$, then $f = \max(f_1, f_2) \in \mathcal{R}$ (*Hint:* $\int_A f \, d\mu = \int_{A \cap \{f_1 \geq f_2\}} f_1 \, d\mu + \int_{A \cap \{f_1 < f_2\}} f_2 \, d\mu$.)

2. if $\{f_n\}_{n \in \mathbb{N}} \in \mathcal{R}$ and $f_n \nearrow f$, then $f \in \mathcal{R}$.

Let $\{f_n\}_{n \in \mathbb{N}}$ be a sequence in $\mathcal{R}$ such that

$$\int f_n \, d\mu \nearrow \sup \left\{ \int f \, d\mu : f \in \mathcal{R} \right\}.$$

Thanks to (1) above, we can replace f_n by $\max(f_1, \ldots, f_n)$, $n \in \mathbb{N}$, and, thus, assume that the sequence $\{f_n(x)\}_{n\in\mathbb{N}}$ is non-decreasing for each $x \in X$. Part (2), on the other hand, implies that for $f = \lim_n f_n$, we have $f \in \mathcal{R}$ so that

$$(5.4) \qquad \int f \, d\mu \geq \int g \, d\mu \text{ for all } g \in R.$$

In words, f is a maximal "candidate-Radon-Nikodym derivative". We define the measure ν_a by $\nu_a(A) = \int_A f \, d\mu$, so that $\nu_a \ll \nu$. Set $\nu_s = \nu - \nu_a$ and note that ν_s takes values in $[0, \infty]$ thanks to the fact that $f \in \mathcal{R}$.

In order to show that $\mu \perp \nu_s$, we define a sequence $\{\nu_n\}_{n\in\mathbb{N}}$ of set functions $\nu_n : \mathcal{S} \to (-\infty, \infty]$ by $\nu_n = \nu_s - \frac{1}{n}\mu$. Thanks to the assumption $\mu(S) < \infty$, ν_n is a well-defined signed measure, for each $n \in \mathbb{N}$. Therefore, according to the Hahn-Jordan decomposition (see Theorem 2.32), there exists a sequence $\{D_n\}_{n\in\mathbb{N}}$ in $\mathcal{S}$ such that

$$(5.5) \qquad \nu_s(A \cap D_n) \geq \tfrac{1}{n}\mu(A \cap D_n) \text{ and } \nu_s(A \cap D_n^c) \leq \tfrac{1}{n}\mu(A \cap D_n^c),$$

for all $A \in \mathcal{S}$ and $n \in \mathbb{N}$. Moreover, with $D = \cup_n D_n$, we have

$$\nu_s(D^c) \leq \nu_s(D_n^c) \leq \tfrac{1}{n}\mu(D_n^c) \leq \tfrac{1}{n}\mu(S),$$

for all $n \in \mathbb{N}$. Consequently, $\nu_s(D^c) = 0$, so it will suffice to show that $\mu(D) = 0$. For that, we define a sequence $\{f_n\}_{n\in\mathbb{N}}$ in $\mathcal{L}_+^0$ by $f_n = f + \frac{1}{n}\mathbf{1}_{D_n}$ and note that

$$\int_A f_n \, d\mu = \nu_a(A) + \tfrac{1}{n}\mu(D_n \cap A) \leq \nu_a(A) + \nu_s(D_n \cap A) \leq \nu(A).$$

Therefore, $f_n \in \mathcal{R}$ and thanks to the "maximal property" (5.4) of f, we conclude that $f_n = f$, a.e., i.e. $\mu(D_n) = 0$, and, immediately, $\mu(D) = 0$, as required. $\blacksquare$

Corollary 5.18 (Radon-Nikodym) *Let μ and ν be σ-finite measures on $(S, \mathcal{S})$ with $\nu \ll \mu$. Then there exists $f \in \mathcal{L}_+^0$ such that*

$$(5.6) \qquad \nu(A) = \int_A f \, d\mu, \text{ for all } A \in \mathcal{S}.$$

For any other $g \in \mathcal{L}_+^0$ with the same property, we have $f = g$, a.e.

Any function f for which (5.6) holds is called the **Radon-Nikodym derivative** of ν with respect to μ and is denoted by $f = \frac{d\nu}{d\mu}$, a.e. The Radon-Nikodym derivative $f = \frac{d\nu}{d\mu}$ is defined only up to a.e.-equivalence, and there is no canonical way of picking a representative defined for all $x \in S$. For that reason, we usually say that a function $f \in \mathcal{L}_+^0$ is a **version** of the Radon-Nikodym derivative of ν with respect to μ if (5.6) holds. Moreover, to stress the fact that we are talking about a whole class of functions instead of just one, we usually write

$$\frac{d\nu}{d\mu} \in \mathbb{L}_+^0 \text{ and not } \frac{d\nu}{d\mu} \in \mathcal{L}_+^0.$$

We often neglect this fact notationally, and write statements such as "*If $f \in \mathcal{L}^0_+$ and $f = \frac{d\mu}{d\nu}$ then* ...". What we really mean is that the statement holds regardless of the particular *representative f* of the Radon-Nikodym derivative we choose. Also, when we write $\frac{d\nu}{d\mu} = \frac{d\rho}{d\mu}$, we mean that they are equal as elements of $\mathbb{L}^0_+$, i.e., that there exists $f \in \mathcal{L}^0_+$, which is both a version of $\frac{d\mu}{d\nu}$ and a version of $\frac{d\rho}{d\mu}$.

Problem 5.19 Let μ, ν and ρ be σ-finite measures on $(S, \mathcal{S})$. Show that

1. If $\nu \ll \mu$ and $\rho \ll \mu$, then $\nu + \rho \ll \mu$ and

$$\frac{d\nu}{d\mu} + \frac{d\rho}{d\mu} = \frac{d(\nu + \rho)}{d\mu}.$$

2. If $\nu \ll \mu$ and $f \in \mathcal{L}^0_+$, then

$$\int f \, d\nu = \int g \, d\mu \text{ where } g = f \frac{d\nu}{d\mu}.$$

3. If $\nu \ll \mu$ and $\rho \ll \nu$, then $\rho \ll \mu$ and

$$\frac{d\nu}{d\mu} \frac{d\rho}{d\nu} = \frac{d\rho}{d\mu}.$$

   (*Note:* Make sure to pay attention to the fact that different measure give rise to different families of null sets, and, hence, to different notions of *almost everywhere*.)

4. If $\mu \sim \nu$, then

$$\frac{d\mu}{d\nu} = \left(\frac{d\nu}{d\mu} \right)^{-1}.$$

Problem 5.20 Let $\mu_1, \mu_2, \nu_1, \nu_2$ be σ-finite measures with μ_1 and ν_1, as well as μ_2 and ν_2, defined on the same measurable space. If $\nu_1 \ll \mu_1$ and $\nu_2 \ll \mu_2$, show that $\nu_1 \otimes \nu_2 \ll \mu_1 \otimes \mu_2$.

Example 5.21 Just like in the statement of Fubini's theorem, the assumption of σ-finiteness cannot be omitted. Indeed, take $(S, \mathcal{S}) = ([0, 1], \mathcal{B}([0, 1]))$ and consider the Lebesgue measure λ and the counting measure γ on $(S, \mathcal{S})$. Clearly, $\lambda \ll \gamma$, but there is no $f \in \mathcal{L}^0_+$ such that $\lambda(A) = \int_A f \, d\gamma$. Indeed, suppose that such f exists and set $D_n = \{x \in S \ : \ f(x) > 1/n\}$, for $n \in \mathbb{N}$, so that $D_n \nearrow \{f > 0\} = \{f \neq 0\}$. Then

$$1 \geq \lambda(D_n) = \int_{D_n} f \, d\gamma \geq \int_{D_n} \tfrac{1}{n} \, d\gamma = \tfrac{1}{n} \# D_n,$$

and so $\# D_n \leq n$. Consequently, the set $\{f > 0\} = \cup_n D_n$ is countable. This leads to a contradiction since the Lebesgue measure does not "charge" countable sets, and so

$$1 = \lambda([0, 1]) = \int f \, d\gamma = \int_{\{f > 0\}} f \, d\gamma = \lambda(\{f > 0\}) = 0.$$

5.3 Additional Problems

Problem 5.22 (Area under the graph of a function) For $f \in \mathcal{L}^0_+$, let $H = \{(x, r) \in S \times [0, \infty) : f(x) \geq r\}$ be the "region under the graph" of f. Show that $\int f \, d\mu = (\mu \otimes \lambda)(H)$.

(*Note:* This equality is consistent with our intuition that the value of the integral $\int f \, d\mu$ corresponds to the "area of the region under the graph of f".)

Problem 5.23 (A layered representation) Let ν be a measure on $\mathcal{B}([0, \infty))$ such that $N(u) = \nu([0, u)) < \infty$, for all $u \in \mathbb{R}$. Let $(S, \mathcal{S}, \mu)$ be a σ-finite measure space. For $f \in \mathcal{L}^0_+(S)$, show that

1. $\int N \circ f \, d\mu = \int_{[0, \infty)} \mu(\{f > u\}) \, \nu(du)$.

2. for $p > 0$, we have $\int f^p \, d\mu = p \int_{[0, \infty)} u^{p-1} \mu(\{f > u\}) \, \lambda(du)$, where λ is the Lebesgue measure.

Problem 5.24 (A useful integral)

1. Show that $\int_0^\infty \left| \frac{\sin x}{x} \right| dx = \infty$. (*Hint:* Find a function below $\left| \frac{\sin x}{x} \right|$ which is easier to integrate.)

2. For $a > 0$, let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x, y) = \begin{cases} e^{-xy} \sin(x), & 0 \leq x \leq a, \ 0 \leq y, \\ 0, & \text{otherwise.} \end{cases}$

   Show that $f \in \mathcal{L}^1(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), \lambda)$, where λ denotes the Lebesgue measure on $\mathbb{R}^2$.

3. Establish the equality $\int_0^a \frac{\sin x}{x} \, dx = \frac{\pi}{2} - \cos(a) \int_0^\infty \frac{e^{-ay}}{1+y^2} \, dy - \sin(a) \int_0^\infty \frac{y e^{-ay}}{1+y^2} \, dy$.

4. Conclude that for $a > 0$, $\left| \int_0^a \frac{\sin(x)}{x} \, dx - \frac{\pi}{2} \right| \leq \frac{2}{a}$, so that $\lim_{a \to \infty} \int_0^a \frac{\sin(x)}{x} \, dx = \frac{\pi}{2}$.

Problem 5.25 (The Cantor measure) Let $(\{-1, 1\}^{\mathbb{N}}, \mathcal{B}(\{-1, 1\}^{\mathbb{N}}), \mu_C)$ be the coin-toss space. Define the mapping $f : \{-1, 1\}^{\mathbb{N}} \to [0, 1]$ by

$$f(\boldsymbol{s}) = \sum_{n \in \mathbb{N}} (1 + s_n) 3^{-n}, \text{ for } \boldsymbol{s} = (s_1, s_2, \dots).$$

Let δ be the push-forward of μ_C by the map f. It is called the **Cantor measure**.

1. Let d be the metric on $\{-1, 1\}^{\mathbb{N}}$ (as given by the equation (1.4) in Lecture 1). Show that for $\alpha = \log_3(2)$ and $\boldsymbol{s}^1, \boldsymbol{s}^2 \in \{-1, 1\}^{\mathbb{N}}$, we have

$$d(\boldsymbol{s}^1, \boldsymbol{s}^2)^\alpha \leq \left| f(\boldsymbol{s}^2) - f(\boldsymbol{s}^1) \right| \leq 3d(\boldsymbol{s}^1, \boldsymbol{s}^2)^\alpha.$$

2. Show that δ is atom-free, i.e., that $\delta(\{x\}) = 0$, for all $x \in [0, 1]$,

3. For a measure μ on the σ-algebra of Borel sets of a topological space X, the **support** of μ is collection of all $x \in X$ with the property that $\mu(O) > 0$ for each open set O with $x \in O$. Describe the support of δ. (*Hint:* Guess what it is and prove that your guess is correct. Use the result in (1).)

4. Prove that $\delta \perp \lambda$.

(*Note:* The Cantor measure is an example of a **singular measure**. It has no atoms, but is still singular with respect to the Lebesgue measure.)

Problem 5.26 (Joint measurability)

1. Give an example of a function $f : [0, 1] \times [0, 1] \to [0, 1]$ such that $x \mapsto f(x, y)$ and $y \mapsto f(x, y)$ are $\mathcal{B}([0, 1])$-measurable functions for each $y \in [0, 1]$ and $x \in [0, 1]$, respectively, but that f is not $\mathcal{B}([0, 1] \times [0, 1])$-measurable.

   (*Hint:* You can use the fact that there exists a subset of $[0, 1]$ which is not Borel measurable.)

2. Let $(S, \mathcal{S})$ be a measurable space. A function $f : S \times \mathbb{R} \to \mathbb{R}$ is called a **Caratheodory function** if

   - $x \mapsto f(x, y)$ is $\mathcal{S}$-measurable for each $y \in \mathbb{R}$, and
   - $y \mapsto f(x, y)$ is continuous for each $x \in \mathbb{R}$.

   Show that Caratheodory functions are $\mathcal{S} \otimes \mathcal{B}(\mathbb{R})$-measurable. (*Hint:* Express a Caratheodory function as limit of a sequence of the form $f_n = \sum_{k \in \mathbb{Z}} g_{n,k}(x) h_{n,k}(r)$, $n \in \mathbb{N}$.)

Chapter 6

Basic Notions of Probability

A mathematical setup behind a probabilistic model consists of a **sample space** Ω, a family of **events** and a **probability** $\mathbb{P}$. One thinks of Ω as being the set of all possible outcomes of a given random phenomenon, and the occurrence of a particular **elementary outcome** $\omega \in \Omega$ as depending on factors not fully known to the modeler. The family $\mathcal{F}$ is taken to be some collection of subsets of Ω, and for each $A \in \mathcal{F}$, the number $\mathbb{P}[A]$ is interpreted as the likelihood that some $\omega \in A$ occurs. Using the basic intuition that $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$, whenever A and B are disjoint (**mutually exclusive**) events, we conclude $\mathbb{P}$ should have all the properties of a finitely-additive measure. Moreover, a natural choice of normalization dictates that the likelihood of the **certain event** Ω be equal to 1. For reasons that are outside the scope of these notes, a leap of faith is often made and $\mathbb{P}$ is required to be σ-additive. All in all, we can single out probability spaces as a sub-class of measure spaces:

Definition 6.1 (Probability space) *A* **probability space** *is a triple* $(\Omega, \mathcal{F}, \mathbb{P})$, *where* Ω *is a non-empty set,* $\mathcal{F}$ *is a* σ-*algebra on* Ω *and* $\mathbb{P}$ *is a probability measure on* $\mathcal{F}$.

In many (but certainly not all) aspects, probability theory is a part of measure theory. For historical reasons and because of a different interpretation, some of the terminology/notation changes when one talks about measure-theoretic concepts in probability. Here is a list of what is different, and what stays the same:

1. We will always assume - often without explicit mention - that a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is given and fixed.

2. Continuity of measure is called **continuity of probability**, and, unlike the general case, does not require and additional assumptions in the case of a decreasing sequence (that is, of course, because $\mathbb{P}[\Omega] = 1 < \infty$.)

3. A measurable function $f : \Omega \to \mathbb{R}$ is called a **random variable**. Typically, the sample space Ω is too large and clumsy for analysis, so we often focus our attention to real-valued functions X on Ω (random variables are usually denoted by capital letters such as X, Y Z, etc.).

If $\omega \in \Omega$ contains the information about the state of all parts of the model, $X(\omega)$ will typically correspond to a single aspect of it. Therefore $X^{-1}([a,b])$ is the set of all elementary outcomes $\omega \in \Omega$ with for which $X(\omega) \in [a,b]$. If we want to be able to compute the probability $\mathbb{P}[X^{-1}([a,b])]$, the set $X^{-1}([a,b])$ better be an event, i.e., $X^{-1}([a,b]) \in \mathcal{F}$. Hence the measurability requirement.

Sometimes, it will be more convenient for random variables to take values in the extended set $\bar{\mathbb{R}}$ of real numbers. In that case we talk about **extended random variables** or $\bar{\mathbb{R}}$**-valued random variables**.

4. We use the measure-theoretic notation $\mathcal{L}^0, \mathcal{L}^0_+, \mathcal{L}^0(\bar{\mathbb{R}})$, etc. to denote the set of all random variables, non-negative random variables, extended random variables, etc.

5. Let $(S, \mathcal{S})$ be a measurable space. An $(\mathcal{F}, \mathcal{S})$-measurable map $X : \Omega \to S$ is called a **random element (of** S**)**.

   Random variables are random elements, but there are other important examples. If $(S, \mathcal{S}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, we talk about **random vectors**. More generally, if $S = \mathbb{R}^{\mathbb{N}}$ and $\mathcal{S} = \prod_n \mathcal{B}(\mathbb{R})$, the map $X : \Omega \to S$ is called a **discrete-time stochastic process**. Sometimes, the object of interest is a set (the area covered by a wildfire, e.g.) and then S is a collection of subsets of $\mathbb{R}^n$. There are many more examples.

6. The class of null-sets in $\mathcal{F}$ still plays the same role as it did in measure theory, but now we use the acronym **a.s.** (which stands for *almost surely*) instead of the measure-theoretic a.e.

7. The Lebesgue integral with respect to the probability $\mathbb{P}$ is now called **expectation** and is denoted by $\mathbb{E}$, so that we write

$$\mathbb{E}[X] \text{ instead of } \int X \, d\mathbb{P}, \text{ or } \int_\Omega X(\omega) \, \mathbb{P}[d\omega].$$

   For $p \in [1, \infty]$, the $\mathcal{L}^p$ spaces are defined just like before, and have the property that $\mathcal{L}^q \subseteq \mathcal{L}^p$, when $p \leq q$.

8. The notion of a.e.-convergence is now re-baptized as **a.s. convergence**, while convergence in measure is now called **convergence in probability**. We write $X_n \overset{a.s.}{\to} X$ if the sequence $\{X_n\}_{n \in \mathbb{N}}$ of random variables converges to a random variable X, a.s. Similarly, $X_n \overset{\mathbb{P}}{\to} X$ refers to convergence in probability. The notion of convergence in $\mathcal{L}^p$, for $p \in [1, \infty]$ is exactly the same as before. We write $X_n \overset{\mathcal{L}^p}{\to} X$ if $\{X_n\}_{n \in \mathbb{N}}$ converges to X in $\mathcal{L}^p$.

9. Since the constant random variable $X(\omega) = 1$, for $\omega \in \Omega$ is integrable, a special case of the dominated convergence theorem, known as the **bounded convergence theorem** holds in probability spaces:

Theorem 6.2 (Bounded convergence) *Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of random variables such that there exists $M \geq 0$ such that $|X_n| \leq M$, a.s., and $X_n \to X$, a.s., then*

$$\mathbb{E}[X_n] \to \mathbb{E}[X].$$

---

10. The relationship between various forms of convergence can now be represented diagramatically as

$$
\begin{array}{ccc}
& \mathcal{L}^\infty & \\
\swarrow & \downarrow & \searrow \searrow \\
\text{a.s.} & & \mathcal{L}^p \longleftarrow \mathcal{L}^q \\
\searrow & \downarrow & \swarrow \\
& \mathbb{P} &
\end{array}
$$
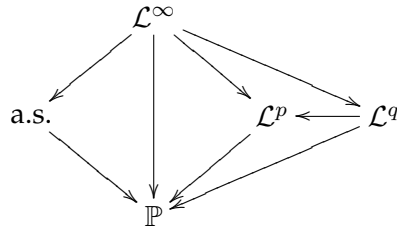
where $1 \le p \le q < \infty$ and an arrow $A \to B$ means that A implies B, but that B *does not imply* A in general.

6.2 Distributions of random variables, vectors and elements

As we have already mentioned, Ω typically too big to be of direct use. Luckily, if we are only interested in a single random variable, all the useful probabilistic information about it is contained in the probabilities of the form $\mathbb{P}[X \in B]$, for $B \in \mathcal{B}(\mathbb{R})$. Btw, it is standard to write $\mathbb{P}[X \in B]$ instead of the more precise $\mathbb{P}[\{X \in B\}]$ or $\mathbb{P}[\{\omega \in \Omega : X(\omega) \in B\}]$. Similarly, we will write $\mathbb{P}[X_n \in B_n, \text{ i.o}]$ instead of $\mathbb{P}[\{X_n \in B_n\} \text{ i.o.}]$ and $\mathbb{P}[X_n \in B_n, \text{ ev.}]$ instead of $\mathbb{P}[\{X_n \in B_n\} \text{ ev.}]$

The map $B \mapsto \mathbb{P}[X \in B]$ is, however, nothing but the push-forward of the measure $\mathbb{P}$ by the map X onto $\mathcal{B}(\mathbb{R})$:

Definition 6.3 (Distribution of a random variable) *The* **distribution** *of the random variable X is the probability measure μ_X on $\mathcal{B}(\mathbb{R})$, defined by*

$$
\mu(B) = \mathbb{P}[X^{-1}(B)],
$$

that is the push-forward of the measure $\mathbb{P}$ by the map X.

In addition to be able to recover the information about various probabilities related to X from μ_X, one can evaluate any possible integral involving a function of X by integrating that function against μ_X (compare the statement to Problem 5.23):

Problem 6.4 Let $g : \mathbb{R} \to \mathbb{R}$ be a Borel function. Then $g \circ X \in \mathcal{L}^{0-1}(\Omega, \mathcal{F}, \mathbb{P})$ if and only if $g \in \mathcal{L}^{0-1}(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_X)$ and, in that case,

$$
\mathbb{E}[g(X)] = \int g \, d\mu_X.
$$

In particular,

$$
\mathbb{E}[X] = \int_{\mathbb{R}} x \mu_X(dx).
$$

Taken in isolation from everything else, two random variables X and Y for which $\mu_X = \mu_Y$ are the same from the probabilistic point of view. In that case we say that X and Y are **equally distributed random variables** and write $X \overset{(d)}{=} Y$. On the other hand, if we are interested in their relationship with a third random variable Z, it can happen that X and Y have the same distribution, but that

their relationship to Z is very different. It is the notion of **joint distribution** that comes to rescue. For a random vector $\boldsymbol{X} = (X_1, \dots, X_n)$, the measure $\mu_{\boldsymbol{X}}$ on $\mathcal{B}(\mathbb{R}^n)$ given by

$$\mu_{\boldsymbol{X}}(B) = \mathbb{P}[\boldsymbol{X} \in B],$$

is called the **distribution** of the random vector $\boldsymbol{X}$. Clearly, the measure $\mu_{\boldsymbol{X}}$ contains the information about the distributions of the individual components $X_1, \dots, X_n$, because

$$\mu_{X_1}(A) = \mathbb{P}[X_1 \in A] = \mathbb{P}[X_1 \in A, X_2 \in \mathbb{R}, \dots, X_n \in \mathbb{R}] = \mu_{\boldsymbol{X}}(A \times \mathbb{R} \times \cdots \times \mathbb{R}).$$

When $X_1, \dots, X_n$ are viewed as components in the random vector $\boldsymbol{X}$, their distributions are sometimes referred to as **marginal distributions**.

Example 6.5 Let $\Omega = \{1, 2, 3, 4\}$, $\mathcal{F} = 2^\Omega$, with $\mathbb{P}$ characterized by $\mathbb{P}[\{\omega\}] = \frac{1}{4}$, for $\omega = 1, \dots, 4$. The map $X : \Omega \to \mathbb{R}$, given by $X(1) = X(3) = 0$, $X(2) = X(4) = 1$, is a random variable and its distribution is the measure $\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ on $\mathcal{B}(\mathbb{R})$ (check that formally!), where δ_a denotes the Dirac measure on $\mathcal{B}(\mathbb{R})$, concentrated on $\{a\}$.

Similarly, the maps $Y : \Omega \to \mathbb{R}$ and $Z : \Omega \to \mathbb{R}$, given by $Y(1) = Y(2) = 0$, $Y(3) = Y(4) = 1$, and $Z(\omega) = 1 - X(\omega)$ are random variables with the same distribution as X. The joint distributions of the random vectors (X, Y) and (X, Z) are very different, though. The pair (X, Y) takes 4 different values $(0, 0), (0, 1), (1, 0), (1, 1)$, each with probability $\frac{1}{4}$, so that the distribution of (X, Y) is given by

$$\mu_{(X,Y)} = \frac{1}{4}\Big(\delta_{(0,0)} + \delta_{(0,1)} + \delta_{(1,0)} + \delta_{(1,1)}\Big).$$

On the other hand, it is impossible for X and Z to take the same value at the same time. In fact, there are only two values that the pair (X, Z) can take - $(0, 1)$ and $(1, 0)$. They happen with probability $\frac{1}{2}$ each, so

$$\mu_{(X,Z)} = \frac{1}{2}\Big(\delta_{(0,1)} + \delta_{(1,0)}\Big).$$

We will see later that the difference between (X, Y) and (X, Z) is best understood if we analyze the way the component random variables depend on each other. In the first case, even if the value of X is revealed, Y can still take the values 0 or 1 with equal probabilities. In the second case, as soon as we know X, we know Z.

More generally, if $X : \Omega \to S$, is a random element with values in the measurable space $(S, \mathcal{S})$, the **distribution of** X is the measure μ_X on $\mathcal{S}$, defined by $\mu_X(B) = \mathbb{P}[X \in B] = \mathbb{P}[X^{-1}(B)]$, for $B \in \mathcal{S}$.

Sometimes it is easier to work with a real-valued function F_X defined by

$$F_X(x) = \mathbb{P}[X \le x],$$

which we call the **(cumulative) distribution function** (**cdf** for short), of the random variable X. The following properties of F_X are easily derived by using continuity of probability from above and from below:

Proposition 6.6 (Preperties of the cdf) *Let X be a random variable, and let F_X be its distribution function. Then,*

1. F_X is non-decreasing and takes values in $[0, 1]$,

> 2. F_X is right continuous,
>
> 3. $\lim_{x\to\infty} F_X(x) = 1$ *and* $\lim_{x\to-\infty} F_X(x) = 0$.

Remark 6.7 A notion of a (cumulative) distribution function can be defined for random vectors, too, but it is not used as often as the single-component case, so we do not write about it here.

The case when μ_X is absolutely continuous with respect to the Lebesgue measure is especially important:

Definition 6.8 (Absolute continuity and pdfs) *A random variable X with the property that $\mu_X \ll \lambda$, where λ is the Lebesgue measure on $\mathcal{B}(\mathbb{R})$, is said to be* **absolutely continuous***. In that case, any Radon-Nikodym derivative $\frac{d\mu_X}{d\lambda}$ is called the* **probability density function (pdf)** *of X, and is denoted by f_X. Similarly, a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is said to be* **absolutely continuous** *if $\mu_{\mathbf{X}} \ll \lambda$, where λ is the Lebesgue measure on $\mathcal{B}(\mathbb{R}^n)$, and the Radon-Nikodym derivative $\frac{d\mu_{\mathbf{X}}}{d\lambda}$, denoted by $f_{\mathbf{X}}$ is called the probability density function (pdf) of $\mathbf{X}$.*

Problem 6.9

1. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an absolutely-continuous random vector. Show that X_k is also absolutely continuous, and that its pdf is given by

$$f_{X_k}(x) = \underbrace{\int_{\mathbb{R}} \ldots \int_{\mathbb{R}}}_{n-1 \text{ integrals}} f(\xi_1, \ldots, \xi_{k-1}, x, \xi_{k+1}, \ldots, \xi_n)\, d\xi_1 \ldots d\xi_{k-1}\, d\xi_{k+1} \ldots d\xi_n.$$

   (*Note:* Note the $f_{X_k}(x)$ is defined only for almost all $x \in \mathbb{R}$; that is because $f_{\mathbf{X}}$ is defined only up to null sets in $\mathcal{B}(\mathbb{R}^n)$.)

2. Let X be an absolutely-continuous random variable. Show that the random vector $(X, -X)$ is *not* absolutely continuous, even though both of its components are .

Problem 6.10 Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an absolutely-continuous random vector with density $f_{\mathbf{X}}$, and let $g : \mathbb{R}^n \to \mathbb{R}$ be a Borel-measurable function with $gf_{\mathbf{X}} \in \mathcal{L}^{0-1}(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \lambda)$. Show that $g(\mathbf{X}) \in \mathcal{L}^{0-1}(\Omega, \mathcal{F}, \mathbb{P})$ and that

$$\mathbb{E}[g(\mathbf{X})] = \int gf_{\mathbf{X}}\, d\lambda = \int_{\mathbb{R}} \ldots \int_{\mathbb{R}} g(\xi_1, \ldots, \xi_n) f_{\mathbf{X}}(\xi_1, \ldots, \xi_n)\, d\xi_1 \ldots d\xi_n.$$

Definition 6.11 (Discrete random variables) *A random variable X is said to be* **discrete** *if there exists a countable set $B \in \mathcal{B}(\mathbb{R})$ such that $\mu_X(B) = 1$.*

Problem 6.12 Show that a sum of two discrete random variables is discrete, but that a sum of two absolutely-continuous random variables does not need to be absolutely continuous.

Definition 6.13 (Singular distributions) *A distribution which has no atoms and is singular with respect to the Lebesgue measure is called* **singular**.

Example 6.14 (A measure which is neither absolutely continuous nor discrete) By Problem 5.25, there exists a measure μ on $[0, 1]$, with the following properties

1. μ has no atoms, i.e., $\mu(\{x\}) = 0$, for all $x \in [0, 1]$,

2. μ and λ (the Lebesgue measure) are mutually singular

3. μ is supported by the Cantor set.

We set $(\Omega, \mathcal{F}, \mathbb{P}) = ([0, 1], \mathcal{B}([0, 1]), \mu)$, and define the random variable $X : \Omega \to \mathbb{R}$, by $X(\omega) = \omega$. It is clear that the distribution μ_X of X has the property that

$$\mu_X(B) = \mu(B \cap [0, 1]),$$

so that X is an example of a random variable with a singular distribution.
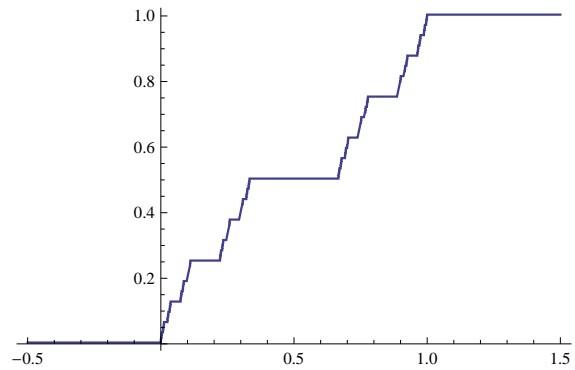
Cdf of the Cantor distribution.

6.3 Independence

The point at which probability departs from measure theory is when independence is introduced. As seen in Example 6.5, two random variables can "depend" on each other in different ways. One extreme (the case of X and Y) corresponds to the case when the dependence is very weak - the distribution of Y stays the same when the value of X is revealed:

Definition 6.15 (Independence of two random variables) *Two random variables X and Y are said to be* **independent** *if*

$$\mathbb{P}[\{X \in A\} \cap \{Y \in B\}] = \mathbb{P}[X \in A] \times \mathbb{P}[Y \in B] \text{ for all } A, B \in \mathcal{B}(\mathbb{R}).$$

It turns out that independence of random variables is a special case of the more-general notion of independence between families of sets.

Definition 6.16 (Independence of families of sets) *Families* $\mathcal{A}_1, \ldots, \mathcal{A}_n$ *of elements in* $\mathcal{F}$ *are said to be*

   1. **independent** *if*

     (6.1) $\qquad\qquad \mathbb{P}[A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}] = \mathbb{P}[A_{i_1}] \times \mathbb{P}[A_{i_2}] \times \cdots \times \mathbb{P}[A_{i_k}],$

     for all $k = 1, \ldots, n$, $1 \le i_1 < i_2 < \cdots < i_k \le n$, *and all* $A_{i_l} \in \mathcal{A}_{i_l}$, $l = 1, \ldots, k$,

   2. **pairwise independent** *if*

$$\mathbb{P}[A_{i_1} \cap A_{i_2}] = \mathbb{P}[A_{i_1}] \times \mathbb{P}[A_{i_2}],$$

     for all $1 \le i_1 < i_2 \le n$, *and all* $A_{i_1} \in \mathcal{A}_{i_1}$, $A_{i_2} \in \mathcal{A}_{i_2}$.

Problem 6.17

   1. Show, by means of an example, that the notion of independence would change if we asked for the product condition (6.1) to hold only for $k = n$ and $i_1 = 1, \ldots, i_k = n$.

   2. Show that, however, if $\Omega \in \mathcal{A}_i$, for all $i = 1, \ldots, n$, it is enough to test (6.1) for $k = n$ and $i_1 = 1, \ldots, i_k = n$ to conclude independence of $\mathcal{A}_i$, $i = 1, \ldots, n$.

Problem 6.18 Show that random variables X and Y are independent if and only if the σ-algebras $\sigma(X)$ and $\sigma(Y)$ are independent.

Definition 6.19 (Independence of random variables and events) *Random variables* $X_1, \ldots, X_n$ *are said to be* **independent** *if the* σ-*algebras* $\sigma(X_1), \ldots, \sigma(X_n)$ *are independent.*

   Events $A_1, \ldots, A_n$ *are called* **independent** *if the families* $\mathcal{A}_i = \{A_i\}$, $i = 1, \ldots, n$, *are independent.*

When only two families of sets are compared, there is no difference between pairwise independence and independence. For 3 or more, the difference is non-trivial:

Example 6.20 (Pairwise independence without independence) Let X_1, X_2 and X_3 be independent random variables, each with the **coin-toss** distribution, i.e., $\mathbb{P}[X_i = 1] = \mathbb{P}[X_i = -1] = \frac{1}{2}$, for $i = 1, 2, 3$. It is not hard to construct a probability space where such random variables may be defined explicitly: let $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8\}$, $\mathcal{F} = 2^{\Omega}$, and let $\mathbb{P}$ be characterized by $\mathbb{P}[\{\omega\}] = \frac{1}{8}$, for all $\omega \in \Omega$. Define

$$X_i(\omega) = \begin{cases} 1, & \omega \in \Omega_i \\ -1, & \text{otherwise} \end{cases}$$

where $\Omega_1 = \{1, 3, 5, 7\}$, $\Omega_2 = \{2, 3, 6, 7\}$ and $\Omega_3 = \{5, 6, 7, 8\}$. It is easy to check that X_1, X_2 and X_3 are independent (X_i is the "i-th bit" in the binary representation of ω).

With X_1, X_2 and X_3 defined, we set

$$Y_1 = X_2 X_3, \; Y_2 = X_1 X_3 \text{ and } Y_3 = X_1 X_2,$$

so that Y_i has a coin-toss distribution, for each $i = 1, 2, 3$. Let us show that Y_1 and Y_2 (and then, by symmetry, Y_1 and Y_3, as well as Y_2 and Y_2) are independent:

$$\begin{aligned}
\mathbb{P}[Y_1 = 1, Y_2 = 1] &= \mathbb{P}[X_2 = X_3, X_1 = X_3] = \mathbb{P}[X_1 = X_2 = X_3] \\
&= \mathbb{P}[X_1 = X_2 = X_3 = 1] + \mathbb{P}[X_1 = X_2 = X_3 = -1] = \tfrac{1}{8} + \tfrac{1}{8} = \tfrac{1}{4} \\
&= \mathbb{P}[Y_1 = 1] \times \mathbb{P}[Y_2 = 1].
\end{aligned}$$

We don't need to check the other possibilities, such as $Y_1 = 1$, $Y_2 = -1$, to conclude that Y_1 and Y_2 are independent (see Problem 6.21 below).

On the other hand, Y_1, Y_2 and Y_3 are not independent:

$$\begin{aligned}
\mathbb{P}[Y_1 = 1, Y_2 = 1, Y_3 = 1] &= \mathbb{P}[X_2 = X_3, X_1 = X_3, X_1 = X_2] = \mathbb{P}[X_1 = X_2 = X_3] \\
&= \tfrac{1}{4} \neq \tfrac{1}{8} = \mathbb{P}[Y_1 = 1] \times \mathbb{P}[Y_2 = 1] \times \mathbb{P}[Y_3 = 1].
\end{aligned}$$

Problem 6.21 Show that if $A_1, \ldots, A_n$ are independent, then so are the families $\mathcal{A}_i = \{A_i, A_i^c\}$, $i = 1, \ldots, n$.

A more general statement is also true (and very useful):

Proposition 6.22 (Independent π-systems imply independent generated σ-algebras) *Let $\mathcal{P}_i$, $i = 1, \ldots, n$ be independent π-systems. Then the σ-algebras $\sigma(\mathcal{P}_i)$, $i = 1, \ldots, n$ are also independent.*

PROOF Let $\mathcal{F}_1$ denote the set of all $C \in \mathcal{F}$ such that

$$\mathbb{P}[C \cap A_{i_2} \cap \cdots \cap A_{i_k}] = \mathbb{P}[C] \times \mathbb{P}[A_{i_2}] \times \cdots \times \mathbb{P}[A_{i_k}],$$

for all $k = 2, \ldots, n$, $1 < i_2 < \cdots < i_k \leq n$, and all $A_{i_l} \in \mathcal{P}_{i_l}$, $l = 2, \ldots, k$. It is easy to see that $\mathcal{F}_i$ is a λ-system which contains the π-system $\mathcal{P}_i$, and so, by the π-λ Theorem, it also contains $\sigma(\mathcal{P}_1)$. Consequently $\sigma(\mathcal{P}_1), \mathcal{P}_2, \ldots \mathcal{P}_n$ are independent families.

A re-play of whole procedure with the families $\mathcal{P}_2, \sigma(\mathcal{P}_1), \mathcal{P}_3, \ldots \mathcal{P}_n$, yields that the families $\sigma(\mathcal{P}_1), \sigma(\mathcal{P}_2), \mathcal{P}_3, \ldots \mathcal{P}_n$ are also independent. Following the same pattern allows us to conclude after n steps that $\sigma(\mathcal{P}_1), \sigma(\mathcal{P}_2), \ldots \sigma(\mathcal{P}_n)$ are independent. ∎

Remark 6.23 All notions of independence above extend to infinite families of objects (random variables, families of sets) by requiring that every finite sub-family be independent.

The result of Proposition 6.22 can be used to help us check independence of random variables:

Problem 6.24 Let $X_1, \ldots, X_n$ be random variables.

1. Show that $X_1, \ldots, X_n$ are independent if and only if

$$\mu_{\boldsymbol{X}} = \mu_{X_1} \otimes \cdots \otimes \mu_{X_n},$$

where $\boldsymbol{X} = (X_1, \ldots, X_n)$.

2. Show that $X_1, \ldots, X_n$ are independent if and only if

$$\mathbb{P}[X_1 \leq x_1, \ldots, X_n \leq x_n] = \mathbb{P}[X_1 \leq x_1] \times \cdots \times \mathbb{P}[X_n \leq x_n],$$

   for all $x_1, \ldots, x_n \in \mathbb{R}$. (*Hint:* Note that the family $\{\{X_i \leq x\} : x \in \mathbb{R}\}$ does not include Ω, so that part (2) of Problem 6.17 cannot be applied directly.)

3. Suppose that the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is absolutely continuous. Then $X_1, \ldots, X_n$ are independent if and only if

$$f_{\boldsymbol{X}}(x_1, \ldots, x_n) = f_{X_1}(x_1) \times \cdots \times f_{X_n}(x_n), \ \lambda\text{-a.e.,}$$

   where λ denotes the Lebesgue measure on $\mathcal{B}(\mathbb{R}^n)$.

4. Suppose that $X_1, \ldots X_n$ are discrete with $\mathbb{P}[X_k \in C_k] = 1$, for countable subsets $C_1, \ldots, C_k$ of $\mathbb{R}$. Show that $X_1, \ldots, X_n$ are independent if and only if

$$\mathbb{P}[X_1 = x_1, \ldots, X_n = x_n] = \mathbb{P}[X_1 = x_1] \times \cdots \times \mathbb{P}[X_n = x_n],$$

   for all $x_i \in C_i$, $i = 1, \ldots, n$.

Problem 6.25 Let $X_1, \ldots, X_n$ be independent random variables. Show that the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is absolutely continuous if and only if each X_i, $i = 1, \ldots, n$ is an absolutely-continuous random variable.

The usefulness of Proposition 6.22 is not exhausted, yet.

Problem 6.26

1. Let $\mathcal{F}_{ij}$ $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, be an independent collection of σ-algebras on Ω. Show that the σ-algebras $\mathcal{G}_1, \ldots, \mathcal{G}_n$, where $\mathcal{G}_i = \sigma(\mathcal{F}_{i1}, \ldots, \mathcal{F}_{im_i})$, are independent.

   (*Hint:* $\cup_{j=1, \ldots, m_i} \mathcal{F}_{ij}$ is a π-system which generates $\mathcal{G}_i$.)

2. Let X_{ij} $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, be an independent random variables, and let $f_i : \mathbb{R}^{m_i} \to \mathbb{R}$, $i = 1, \ldots, n$, be Borel functions. Then the random variables $Y_i = f_i(X_1, \ldots, X_{m_i})$, $i = 1, \ldots, n$ are independent.

Problem 6.27

1. Let $X_1, \ldots, X_n$ be random variables. Show that $X_1, \ldots, X_n$ are independent if and only if

$$\prod_{i=1}^{n} \mathbb{E}[f_i(X_i)] = \mathbb{E}[\prod_{i=1}^{n} f_i(X_i)],$$

   for all n-tuples $(f_1, \ldots, f_n)$ of bounded continuous real functions. (*Hint:* Approximate!)

2. Let $\{X_n^i\}_{n \in \mathbb{N}}$, $i = 1, \ldots, m$ be sequences of random variables such that $X_n^1, \ldots, X_n^m$ are independent for each $n \in \mathbb{N}$. If $X_n^i \overset{a.s.}{\to} X^i$, $i = 1, \ldots, m$, for some $X^1, \ldots, X^m \in \mathcal{L}^0$, show that $X^1, \ldots, X^m$ are independent.

The idea "independent means multiply" applies not only to probabilities, but also to random variables:

Proposition 6.28 (Expectation of a function of independent components) *Let X, Y be independent random variables, and let $h : \mathbb{R}^2 \to [0, \infty)$ be a measurable function. Then*

$$\mathbb{E}[h(X,Y)] = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(x,y) \, \mu_X(dx) \right) \mu_Y(dy).$$

PROOF By independence and part (1) of Problem 6.24, the distribution of the random vector (X, Y) is given by $\mu_X \otimes \mu_Y$, where μ_X is the distribution of X and μ_Y is the distribution of μ_Y. Using Fubini's theorem, we get

$$\mathbb{E}[h(X,Y)] = \int h \, d\mu_{(X,Y)} = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(x,y) \, \mu_X(dx) \right) \mu_Y(dy). \qquad \blacksquare$$

Proposition 6.29 (Independent means multiply) *Let $X_1, X_2, \ldots, X_n$ be independent random variables with $X_i \in \mathcal{L}^1$, for $i = 1, \ldots, n$. Then*

1. *$\prod_{i=1}^n X_i = X_1 \cdots X_n \in \mathcal{L}^1$, and*

2. *$\mathbb{E}[X_1 \cdots X_n] = \mathbb{E}[X_1] \ldots \mathbb{E}[X_n]$.*

The product formula 2. remains true if we assume that $X_i \in \mathcal{L}^0_+$ (instead of $\mathcal{L}^1$), for $i = 1, \ldots, n$.

PROOF Using the fact that X_1 and $X_2 \cdots X_n$ are independent random variables (use part (2) of Problem 6.26), we can assume without loss of generality that $n = 2$.

Focusing first on the case $X_1, X_2 \in \mathcal{L}^0_+$, we apply Proposition 6.28 with $h(x,y) = xy$ to conclude that

$$\mathbb{E}[X_1 X_2] = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} x_1 x_2 \, \mu_{X_1}(dx_1) \right) \mu_{X_2}(dx_2)$$

$$= \int_{\mathbb{R}} x_2 \mathbb{E}[X_1] \, \mu_{X_2}(dx_2) = \mathbb{E}[X_1] \mathbb{E}[X_2].$$

For the case $X_1, X_2 \in \mathcal{L}^1$, we split $X_1 X_2 = X_1^+ X_2^+ - X_1^+ X_2^- - X_1^- X_2^+ + X_1^- X_2^-$ and apply the above conclusion to the 4 pairs $X_1^+ X_2^+$, $X_1^+ X_2^-$, $X_1^- X_2^+$ and $X_1^- X_2^-$. $\qquad \blacksquare$

Problem 6.30 (Conditions for "independent-means-multiply") Proposition 6.29 in states that for independent X and Y, we have

(6.2) $$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y],$$

whenever $X, Y \in \mathcal{L}^1$ or $X, Y \in \mathcal{L}^0_+$. Give an example which shows that (6.2) is no longer necessarily true if $X \in \mathcal{L}^0_+$ and $Y \in \mathcal{L}^1$. (*Hint:* Build your example so that $\mathbb{E}[(XY)^+] = \mathbb{E}[(XY)^-] = \infty$. Use $([0,1], \mathcal{B}([0,1]), \lambda)$ and take $Y(\omega) = \mathbf{1}_{[0,1/2]}(\omega) - \mathbf{1}_{(1/2,0]}(\omega)$. Then show that any random variable X with the property that $X(\omega) = X(1 - \omega)$ is independent of Y.)

Problem 6.31 Two random variables X, Y are said to be **uncorrelated**, if $X, Y \in \mathcal{L}^2$ and $\text{Cov}(X, Y) = 0$, where $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$.

1. Show that for $X, Y \in \mathcal{L}^2$, the expression for $\text{Cov}(X, Y)$ is well defined.

2. Show that independent random variables in $\mathcal{L}^2$ are uncorrelated.

3. Show that there exist uncorrelated random variables which are not independent.

6.4   Sums of independent random variables and convolution

Proposition 6.32 (Convolution as the distribution of a sum) *Let X and Y be independent random variables, and let $Z = X + Y$ be their sum. Then the distribution μ_Z of Z has the following representation:*

$$\mu_Z(B) = \int_{\mathbb{R}} \mu_X(B - y)\, \mu_Y(dy), \text{ for } B \in \mathcal{B}(\mathbb{R}),$$

where $B - y = \{b - y \ : \ b \in B\}$.

PROOF We can view Z as a function $f(x, y) = x + y$ applied to the random vector (X, Y), and so, we have $\mathbb{E}[g(Z)] = \mathbb{E}[h(X, Y)]$, where $h(x, y) = g(x + y)$. In particular, for $g(z) = \mathbf{1}_B(z)$, Proposition 6.28 implies that

$$\mu_Z(B) = \mathbb{E}[g(Z)] = \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbf{1}_{\{x+y \in B\}}\, \mu_X(dx)\, \mu_Y(dy)$$

$$= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathbf{1}_{\{x \in B - y\}} \mu_X(dx) \right)\, \mu_Y(dy) = \int_{\mathbb{R}} \mu_X(B - y)\, \mu_Y(dy).$$

One often sees the expression

$$\int_{\mathbb{R}} f(x)\, dF(x),$$

as notation for the integral $\int f\, d\mu$, where $F(x) = \mu((-\infty, x])$. The reason for this is that such integrals - called the **Lebesgue-Stieltjes** integrals - have a theory parallel to that of the Riemann integral and the correspondence between $dF(x)$ and $d\mu$ is parallel to the correspondence between dx and $d\lambda$.

Corollary 6.33 (Cdf of a sum as a convolution) *Let X, Y be independent random variables, and let Z be their sum. Then*

$$F_Z(\xi) = \int_{\mathbb{R}} F_X(\xi - y)\, dF_Y(y).$$

Definition 6.34 (Convolution of probability measures) *Let μ_1 and μ_2 be two probability measures on $\mathcal{B}(\mathbb{R})$. The* **convolution** *of μ_1 and μ_2 is the probability measure $\mu_1 * \mu_2$ on $\mathcal{B}(\mathbb{R})$, given by*

$$(\mu_1 * \mu_2)(B) = \int_{\mathbb{R}} \mu_1(B - \xi)\, \mu_2(d\xi),\ for\ B \in \mathcal{B}(\mathbb{R}),$$

where $B - \xi = \{x - \xi\ :\ x \in B\} \in \mathcal{B}(\mathbb{R})$.

Problem 6.35 Show that the $*$ is a commutative and associative operation on the set of all probability measures on $\mathcal{B}(\mathbb{R})$. (*Hint:* Use Proposition 6.32).

It is interesting to see how convolution mixes with absolute continuity. To simplify the notation, we write $\int_A f(x)\, dx$ instead of more precise $\int_A f(x)\, \lambda(dx)$ for the (Lebesgue) integral with respect to the Lebesgue measure on $\mathbb{R}$. When $A = [a, b] \in \bar{\mathbb{R}}$, we write $\int_a^b f(x)\, dx$.

Proposition 6.36 (Convolution inherets absolute continuity from either component) *Let X and Y be independent random variables, and suppose that X is absolutely continuous. Then their sum $Z = X + Y$ is also absolutely continuous and its density f_Z is given by*

$$f_Z(z) = \int_{\mathbb{R}} f_X(z - y)\, \mu_Y(dy).$$

PROOF Define $f(z) = \int_{\mathbb{R}} f_X(z - y)\, \mu_Y(dy)$, for some density f_X of X (remember, the density function is defined only λ-a.e.). The function f is measurable (why?) so it will be enough (why?) to show that

$$\mathbb{P}\Big[Z \in [a, b]\Big] = \int_{[a,b]} f(z)\, dz,\ \text{for all}\ -\infty < a < b < \infty.$$

We start with the right-hand side of (6.3) and use Fubini's theorem to get

(6.3)
$$\int_{[a,b]} f(z)\, dz = \int_{\mathbb{R}} \mathbf{1}_{[a,b]}(z) \left(\int_{\mathbb{R}} f_X(z - y)\, \mu_Y(dy) \right) dz$$
$$= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathbf{1}_{[a,b]}(z) f_X(z - y)\, dz \right) \mu_Y(dy)$$

By the translation-invariance property of the Lebesgue measure, we have

$$\int_{\mathbb{R}} \mathbf{1}_{[a,b]}(z) f_X(z - y)\, dz = \int_{\mathbb{R}} \mathbf{1}_{[a-y,b-y]}(z) f_X(z)\, dz = \mathbb{P}\Big[Z \in [a - y, b - y]\Big] = \mu_X\Big([a, b] - y\Big).$$

Therefore, by (6.3) and Proposition 6.32, we have

$$\int_{[a,b]} f(z)\, dz = \int_{\mathbb{R}} \mu_X\Big([a, b] - y\Big)\, \mu_Y(dy) = \mu_Z\Big([a, b]\Big) = \mathbb{P}\Big[Z \in [a, b]\Big]. \qquad \blacksquare$$

Definition 6.37 (Convolution of functions in $\mathcal{L}^1$) *The* **convolution** *of functions f and g in* $\mathcal{L}^1(\mathbb{R})$ *is the function $f * g \in \mathcal{L}^1(\mathbb{R})$ given by*

$$(f * g)(z) = \int_{\mathbb{R}} f(z - x)g(x)\, dx.$$

Problem 6.38

1. Use the reasoning from the proof of Proposition 6.36 to show that the convolution is well-defined operation on $\mathcal{L}^1(\mathbb{R})$.

2. Show that if X and Y are independent absolutely-continuous random variables, then $X + Y$ is also absolutely continuous with density which is the convolution of densities of X and Y.

6.5 Do independent random variables exist?

We leave the most basic of the questions about independence for last: do independent random variable exist? We need a definition and two auxiliary results, first.

Definition 6.39 (Uniform distribution on (a, b)) *A random variable X is said to be* **uniformly distributed on** (a, b), *for $a < b \in \mathbb{R}$, if it is absolutely continuous with density*

$$f_X(x) = \tfrac{1}{b-a}\mathbf{1}_{(a,b)}(x).$$

Our first result states a uniform random variable on $(0, 1)$ can be transformed deterministically into any a random variable of prescribed distribution.

Proposition 6.40 (Transforming a uniform distribution) *Let μ be a measure on $\mathcal{B}(\mathbb{R})$ with $\mu(\mathbb{R}) = 1$. Then, there exists a function $H_\mu : (0, 1) \to \mathbb{R}$ such that the distribution of the random variable $X = H_\mu(U)$ is μ, whenever U is a uniform random variable on $(0, 1)$.*

PROOF Let F be the cdf corresponding to μ, i.e.,

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

The function F is non-decreasing, so it "almost" has an inverse: define

$$H_\mu(y) = \inf\{x \in \mathbb{R} \,:\, F(x) \geq y\}.$$

Since $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$, $H_\mu(y)$ is well-defined and finite for all $y \in (0, 1)$. Moreover, thanks to right-continuity and non-decrease of F, we have

$$H_\mu(y) \leq x \Leftrightarrow y \leq F(x), \text{ for all } x \in \mathbb{R}, y \in (0, 1).$$

Therefore

$$\mathbb{P}[H_\mu(U) \leq x] = \mathbb{P}[U \leq F(x)] = F(x), \text{ for all } x \in \mathbb{R},$$

and the statement of the Proposition follows. ∎

Remark 6.41 Proposition 6.40 is a basis for a technique used to simulate random variables. There are efficient algorithms for producing simulated values which resemble the uniform distribution in $(0, 1)$ (so-called **pseudo-random numbers**). If a simulated value drawn with distribution μ is needed, one can simply apply the function H_μ to a pseudo-random number.

Our next auxiliary result tells us how to construct a sequence of independent uniforms:

Proposition 6.42 (A large-enough probability space exists) *There exists a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$, *and on it a sequence* $\{X_n\}_{n \in \mathbb{N}}$ *of random variables such that*

1. X_n *has the uniform distribution on* $(0, 1)$, *for each* $n \in \mathbb{N}$, *and*

2. *the sequence* $\{X_n\}_{n \in \mathbb{N}}$ *is independent.*

PROOF Set $(\Omega, \mathcal{F}, \mathbb{P}) = (\{-1, 1\}^{\mathbb{N}}, \mathcal{S}, \mu_C)$ - the coin-toss space with the product σ-algebra and the coin-toss measure. Let $a : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ be a bijection, i.e., $(a_{ij})_{i,j \in \mathbb{N}}$ is an arrangement of all natural numbers into a double array. For $i, j \in \mathbb{N}$, we define the map $\xi_{ij} : \Omega \to \{-1, 1\}$, by

$$\xi_{ij}(\boldsymbol{s}) = s_{a_{ij}},$$

i.e., ξ_{ij} is the natural projection onto the a_{ij}-th coordinate. It is straightforward to show that, under $\mathbb{P}$, the collection $(\xi_{ij})_{i,j \in \mathbb{N}}$ is independent; indeed, it is enough to check the equality

$$\mathbb{P}[\xi_{i_1 j_1} = 1, \ldots, \xi_{i_n j_n} = 1] = \mathbb{P}[\xi_{i_1 j_1} = 1] \times \cdots \times \mathbb{P}[\xi_{i_n j_n} = 1],$$

for all $n \in \mathbb{N}$ and all different $(i_1, j_1), \ldots, (i_n, j_n) \in \mathbb{N} \times \mathbb{N}$.

At this point, we use the construction from Section 2.3 of Lecture 2, to construct an independent copy of a uniformly-distributed random variable from each row of $(\xi_{ij})_{i,j \in \mathbb{N}}$. We set

(6.4) $$X_i = \sum_{j=1}^{\infty} \left(\tfrac{1+\xi_{ij}}{2} \right) 2^{-j}, \ i \in \mathbb{N}.$$

By second parts of Problems 6.26 and 6.27, we conclude that the sequence $\{X_i\}_{i \in \mathbb{N}}$ is independent. Moreover, thanks to (6.4), X_i is uniform on $(0, 1)$, for each $i \in \mathbb{N}$. ∎

Proposition 6.43 (Arbitrary independent sequences exist) *Let* $\{\mu_n\}_{n \in \mathbb{N}}$ *be a sequence of probability measures on* $\mathcal{B}(\mathbb{R})$. *Then, there exists a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$, *and a sequence* $\{X_n\}_{n \in \mathbb{N}}$ *of random variables defined there such that*

1. $\mu_{X_n} = \mu_n$, *and*

2. $\{X_n\}_{n \in \mathbb{N}}$ *is independent.*

PROOF Start with the sequence of Proposition 6.42 and apply the function H_{μ_n} to X_n for each $n \in \mathbb{N}$, where H_{μ_n} is as in the proof of Proposition 6.40. ∎

An important special case covered by Proposition 6.43 is the following:

> **Definition 6.44 (Independent and identically distributed sequences)** *A sequence $\{X_n\}_{n \in \mathbb{N}}$ of random variables is said to be **independent and identically distributed (iid)** if $\{X_n\}_{n \in \mathbb{N}}$ is independent and all X_n have the same distribution.*

> **Corollary 6.45 (Iid sequences exist)** *Given a probability measure μ on $\mathbb{R}$, there exist a probability space supporting an iid sequence $\{X_n\}_{n \in \mathbb{N}}$ such that $\mu_{X_n} = \mu$.*

6.6 Additional Problems

Problem 6.46 (The standard normal distribution) An absolutely continuous random variable X is said to have the **standard normal distribution** - denoted by $X \sim N(0,1)$ - if it admits a density of the form

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2), \ x \in \mathbb{R}$$

For a r.v. with such a distribution we write $X \sim N(0,1)$.

1. Show that $\int_{\mathbb{R}} f(x) \, dx = 1$. (*Hint:* Consider the double integral $\int_{\mathbb{R}^2} f(x)f(y) \, dx \, dy$ and pass to polar coordinates.)

2. For $X \sim N(0,1)$, show that $\mathbb{E}[|X|^n] < \infty$ for all $n \in \mathbb{N}$. Then compute the nth moment $\mathbb{E}[X^n]$, for $n \in \mathbb{N}$.

3. A random variable with the same distribution as X^2, where $X \sim N(0,1)$, is said to have the χ^2**-distribution**. Find an explicit expression for the density of the χ^2-distribution.

4. Let Y have the χ^2-distribution. Show that there exists a constant $\lambda_0 > 0$ such that $\mathbb{E}[\exp(\lambda Y)] < \infty$ for $\lambda < \lambda_0$ and $\mathbb{E}[\exp(\lambda Y)] = +\infty$ for $\lambda \geq \lambda_0$. (*Note:* For a random variable $Y \in \mathcal{L}_+^0$, the quantity $\mathbb{E}[\exp(\lambda Y)]$ is called the **exponential moment of order** λ.)

5. Let $\alpha_0 > 0$ be a fixed, but arbitrary constant. Find an example of a random variable $X \geq 0$ with the property that $\mathbb{E}[X^\alpha] < \infty$ for $\alpha \leq \alpha_0$ and $\mathbb{E}[X^\alpha] = +\infty$ for $\alpha > \alpha_0$. (*Hint:* This is not the same situation as in (4) - this time the critical case α_0 is included in a different alternative. Try $X = \exp(Y)$, where $\mathbb{P}[Y \in \mathbb{N}] = 1$.)

Problem 6.47 (The "memory-less" property of the exponential distribution) A random variable is said to have **exponential distribution** with parameter $\lambda > 0$ - denoted by $X \sim \text{Exp}(\lambda)$ - if its distribution function F_X is given by

$$F_X(x) = 0 \text{ for } x < 0, \text{ and } F_X(x) = 1 - \exp(-\lambda x), \text{ for } x \geq 0.$$

1. Compute $\mathbb{E}[X^\alpha]$, for $\alpha \in (-1, \infty)$. Combine your result with the result of part (3) of Problem 6.46 to show that
$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi},$$
where Γ is the Gamma function.

2. Remember that the conditional probability $\mathbb{P}[A|B]$ of A, given B, for $A, B \in \mathcal{F}$, $\mathbb{P}[B] > 0$ is given by
$$\mathbb{P}[A|B] = \mathbb{P}[A \cap B]/\mathbb{P}[B].$$
Compute $\mathbb{P}[X \geq x_2 | X \geq x_1]$, for $x_2 > x_1 > 0$ and compare it to $\mathbb{P}[X \geq (x_2 - x_1)]$.

(*Note:* This can be interpreted as follows: the knowledge that the bulb stayed functional until x_1 does not change the probability that it will not explode in the next $x_2 - x_1$ units of time; bulbs have no memory.)

Conversely, suppose that Y is a random variable with the property that $\mathbb{P}[Y > 0] = 1$ and $\mathbb{P}[Y > y] > 0$ for all $y > 0$. Assume further that

(6.5) $$\mathbb{P}[Y \geq y_2 | Y \geq y_1] = \mathbb{P}[Y \geq y_2 - y_1], \text{ for all } y_2 > y_1 > 0.$$

Show that $Y \sim \text{Exp}(\lambda)$ for some $\lambda > 0$.

(*Hint:* You can use the following fact: let $\phi : (0, \infty) \to \mathbb{R}$ be a Borel-measurable function such that $\phi(y) + \phi(z) = \phi(y + z)$ for all $y, z > 0$. Then there exists a constant $\mu \in \mathbb{R}$ such that $\phi(y) = \mu y$ for all $y > 0$.)

Problem 6.48 (The second Borel-Cantelli Lemma)

1. Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence in $\mathcal{L}^0_+$. Show that there exists a sequence of positive constants $\{c_n\}_{n\in\mathbb{N}}$ with the property that
$$\frac{X_n}{c_n} \to 0, \text{ a.s.}$$
(*Hint:* Use the Borel-Cantelli lemma.)

2. (The first) Borel-Cantelli lemma states that $\sum_{n\in\mathbb{N}} \mathbb{P}[A_n] < \infty$ implies $\mathbb{P}[A_n, \text{ i.o.}] = 0$. There are simple examples showing that the converse does not hold in general. Show that it *does* hold if the events $\{A_n\}_{n\in\mathbb{N}}$ are assumed to be independent. More precisely, show that, for an independent sequence $\{A_n\}_{n\in\mathbb{N}}$, we have
$$\sum_{n\in\mathbb{N}} \mathbb{P}[A_n] = \infty \text{ implies } \mathbb{P}[A_n, \text{ i.o.}] = 1.$$
This is often known as the **second Borel-Cantelli lemma**. (*Hint:* Use the inequality $(1-x) \leq e^{-x}$, $x \in \mathbb{R}$.)

3. Let $\{X_n\}_{n\in\mathbb{N}}$ be an iid (independent and identically distributed) sequence of coin tosses, i.e., independent random variables with $\mathbb{P}[X_n = T] = \mathbb{P}[X_n = H] = 1/2$ for all $n \in \mathbb{N}$ (if you are uncomfortable with T and H, feel free to replace them with 1 and -1). A **tail-run of size** k is a finite sequence of at least k consecutive Ts starting from some index $n \in \mathbb{N}$. Show that for almost every $\omega \in \Omega$ (i.e., almost surely) the sequence $\{X_n(\omega)\}_{n\in\mathbb{N}}$ will contain infinitely many tail runs of size k. Conclude that, for almost every ω, the sequence $X_n(\omega)$ will contain infinitely many tail runs of every length.

4. Let $\{X_n\}_{n\in\mathbb{N}}$ be an iid sequence in $\mathcal{L}^0$. Show that

$$\mathbb{E}[|X_1|] < \infty \text{ if and only if } \mathbb{P}[|X_n| \geq n, \text{ i.o.}] = 0.$$

(*Hint:* Use (but first prove) the fact that $1 + \sum_{n\geq 1}\mathbb{P}[X \geq n] \geq \mathbb{E}[X] \geq \sum_{n\geq 1}\mathbb{P}[X \geq n]$, for $X \in \mathcal{L}^0_+$.)

Chapter 7

Weak Convergence and Characteristic Functions

In addition to the modes of convergence we introduced so far (a.s.-convergence, convergence in probability and $\mathcal{L}^p$-convergence), there is another one, called **convergence in distribution**. Unlike the other three, whether a sequence of random variables (elements) converges in distribution or not depends only on their *distributions*. In addition to its intrinsic mathematical interest, convergence in distribution (or, equivalently, the weak convergence) is precisely the kind of convergence we encounter in the central limit theorem.

   We take the abstraction level up a notch and consider sequences of probability measures on $(S, \mathcal{S})$, where (S, d) is a metric space and $\mathcal{S} = \mathcal{B}(d)$ is the Borel σ-algebra there. In fact, it will always be assumed that S is a metric space and $\mathcal{S}$ is the Borel σ-algebra on it, throughout this chapter.

Definition 7.1 (Weak convergence of probability measures) *Let $\{\mu_n\}_{n\in\mathbb{N}}$ be a sequence of probability measures on $(S, \mathcal{S})$. We say that μ_n converges* **weakly** *to a probability measure μ on $(S, \mathcal{S})$ - and write $\mu_n \overset{w}{\to} \mu$ - if*

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

for all $f \in C_b(S)$, where $C_b(S)$ denotes the set of all continuous and bounded functions $f : S \to \mathbb{R}$.

Remark 7.2 It would be more in tune with standard mathematical terminology to use the term *weak-$*$* convergence instead of weak convergence. For historical reasons, however, we omit the $*$.

Definition 7.3 (Convergence in distribution) *A sequence $\{X_n\}_{n\in\mathbb{N}}$ of random variables (elements) is said to* **converge in distribution** *to the random variable (element) X, denoted by $X_n \overset{\mathcal{D}}{\to} X$, if $\mu_{X_n} \overset{w}{\to} \mu_X$.*

CHAPTER 7. WEAK CONVERGENCE AND CHARACTERISTIC FUNCTIONS

Our first result states that weak limits are unique and for that we need a simple result, first:

Problem 7.4 Let F be a closed set in S. Show that for any $\varepsilon > 0$ there exists a Lipschitz and bounded function $f_{F;\varepsilon} : S \to \mathbb{R}$ such that

1. $0 \leq f_{F;\varepsilon}(x) \leq 1$, for all $x \in \mathbb{R}$,

2. $f_{F;\varepsilon}(x) = 1$ for $x \in F$, and

3. $f_{F;\varepsilon}(x) = 0$ for $d(x, F) \geq \varepsilon$, where $d(x, F) = \inf\{d(x, y) : y \in F\}$.

(*Hint*: Show first that the function $x \mapsto d(x, F)$ is Lipschitz. Then argue that $f_{F;\varepsilon}(x) = h(d(x, F))$ has the required properties for a well-chosen function $h : [0, \infty) \to [0, 1]$.)

Proposition 7.5 (Weak limits are unique) *Suppose that $\{\mu_n\}_{n \in \mathbb{N}}$ is a sequence of probability measures on $(S, \mathcal{S})$ such that $\mu_n \overset{w}{\to} \mu$ and $\mu_n \overset{w}{\to} \mu'$. Then $\mu = \mu'$.*

PROOF By the very definition of weak convergence, we have

$$(7.1) \qquad \int f\, d\mu = \lim_n \int f\, d\mu_n = \int f\, d\mu',$$

for all $f \in C_b(S)$. Let F be a closed set, and let $\{f_k\}_{k \in \mathbb{N}}$ be as in Problem 7.4, with $f_k = f_{F;\varepsilon}$ corresponding to $\varepsilon = 1/k$. If we set $F_k = \{x \in S : d(x, F) \leq 1/k\}$, then F_k is a closed set (why?) and we have $\mathbf{1}_F \leq f_k \leq \mathbf{1}_{F_k}$. By (7.1), we have

$$\mu(F) \leq \int f_k\, d\mu = \int f_k d\mu' \leq \mu'(F_k),$$

and, similarly, $\mu'(F) \leq \mu(F_k)$, for all $k \in \mathbb{N}$. Since $F_k \searrow F$ (why?), we have $\mu(F_k) \searrow \mu(F)$ and $\mu'(F_k) \searrow \mu'(F)$, and it follows that $\mu(F) = \mu'(F)$.

It remains to note that the family of all closed sets is a π-system which generates the σ-algebra $\mathcal{S}$ to conclude that $\mu = \mu'$. ∎

We have seen in the proof of Proposition 7.5 that an operational characterization of weak convergence is needed. Here is a useful one. We start with a lemma; remember that ∂A denotes the topological boundary $\partial A = \mathrm{Cl}\, A \setminus \mathrm{Int}\, A$ of a set $A \subseteq S$.

Problem 7.6 Let $(F_\gamma)_{\gamma \in \Gamma}$ be a partition of S into (possibly uncountably many) measurable subsets. Show that for any probability measure μ on $\mathcal{S}$, $\mu(F_\gamma) = 0$, for all but countably many $\gamma \in \Gamma$ (*Hint*: For $n \in \mathbb{N}$, define $\Gamma_n = \{\gamma \in \Gamma : \mu(F_\gamma) \geq \frac{1}{n}\}$. Argue that Γ_n has at most n elements.)

Definition 7.7 (μ-continuity sets) *A set $A \in \mathcal{S}$ with the property that $\mu(\partial A) = 0$, is called a μ-continuity set*

Theorem 7.8 (Portmanteau Theorem) *Let μ, $\{\mu_n\}_{n\in\mathbb{N}}$ be probability measures on $\mathcal{S}$. Then, the following are equivalent:*

1. $\mu_n \xrightarrow{w} \mu$,

2. $\int f \, d\mu_n \to \int f \, d\mu$, *for all bounded,* Lipschitz *continuous* $f : S \to \mathbb{R}$,

3. $\limsup_n \mu_n(F) \leq \mu(F)$, *for all closed* $F \subseteq S$,

4. $\liminf_n \mu_n(G) \geq \mu(G)$, *for all open* $G \subseteq S$,

5. $\lim_n \mu_n(A) = \mu(A)$, *for all* μ-*continuity sets* $A \in \mathcal{S}$.

Remark 7.9 Before a proof is given, here is a way to remember whether closed sets go together with the $\liminf$ or the $\limsup$: take a convergent sequence $\{x_n\}_{n\in\mathbb{N}}$ in S, with $x_n \to x$. If μ_n is the Dirac measure concentrated on x_n, and μ the Dirac measure concentrated on x, then clearly $\mu_n \xrightarrow{w} \mu$ (since $\int f \, d\mu_n = f(x_n) \to f(x) = \int f \, d\mu$). Let F be a closed set. It can happen that $x_n \notin F$ for all $n \in \mathbb{N}$, but $x \in F$ (think of x on the boundary of F). Then $\mu_n(F) = 0$, but $\mu(F) = 1$ and so $\limsup \mu_n(F) = 0 < 1 = \mu(F)$.

PROOF (PROOF OF THEOREM 7.8)

$(1) \Rightarrow (2)$: trivial.

$(2) \Rightarrow (3)$: given a closed set F and let $F_k = \{x \in S \, : \, d(x,F) \leq 1/k\}$, $f_k = f_{F;1/k}$, $k \in \mathbb{N}$, be as in the proof of Proposition 7.5. Since $\mathbf{1}_F \leq f_k \leq \mathbf{1}_{F_k}$ and the functions f_k are Lipschitz continuous, we have

$$\limsup_n \mu_n(F) = \limsup_n \int \mathbf{1}_F \, d\mu_n \leq \lim_n \int f_k \, d\mu_n = \int f_k \, d\mu \leq \mu(F_k),$$

for all $k \in \mathbb{N}$. Letting $k \to \infty$ - as the proof of Proposition 7.5 - yields (3).

$(3) \Rightarrow (4)$: follows directly by taking complements.

$(4) \Rightarrow (1)$: Pick $f \in C_b(S)$ and (possibly after applying a linear transformation to it) assume that $0 < f(x)$, for all $x \in S$. Then, by Problem 5.23, we have $\int f \, d\nu = \int_0^1 \nu(f > t) \, dt$, for any probability measure on $\mathcal{B}(\mathbb{R})$. The set $\{f > t\} \subseteq S$ is open, so by (3), $\liminf_n \mu_n(f > t) \geq \mu(f > t)$, for all t. Therefore, by Fatou's lemma,

$$\liminf_n \int f \, d\mu_n = \liminf_n \int_0^1 \mu_n(f > t) \, dt \geq \int_0^1 \liminf_n \mu_n(f > t) \, dt \geq \int_0^1 \mu(f > t) \, dt$$
$$= \int f \, d\mu.$$

We get the other inequality - $\int f \, d\mu \geq \limsup_n \int f \, d\mu_n$, by repeating the procedure with f replaced by $-f$.

$(3), (4) \Rightarrow (5)$: Let A be a μ-continuity set, let $\operatorname{Int} A$ be its interior and $\operatorname{Cl} A$ its closure. Then, since $\operatorname{Int} A$ is open and $\operatorname{Cl} A$ is closed, we have

$$\mu(\operatorname{Int} A) \leq \liminf_n \mu_n(\operatorname{Int} A) \leq \liminf_n \mu_n(A) \leq \limsup_n \mu_n(A)$$
$$\leq \limsup_n \mu_n(\operatorname{Cl} A) \leq \mu(\operatorname{Cl} A).$$

Since $0 = \mu(\partial A) = \mu(\text{Cl } A \setminus \text{Int } A) = \mu(\text{Cl } A) - \mu(\text{Int } A)$, we conclude that all inequalities above are, in fact, equalities so that $\mu(A) = \lim_n \mu_n(A)$

$(5) \Rightarrow (3)$: For $x \in S$, consider the family $\{B_F(r) \ : \ r \geq 0\}$, where

$$B_F(r) = \{x \in S \ : \ d(x, F) \leq r\},$$

of closed sets.

Claim: There exists a countable subset R of $[0, \infty)$ such that $B_F(r)$ is a μ-continuity set for all $r \in [0, \infty) \setminus R$.

PROOF For $r \geq 0$ define $C_F(r) = \{x \in S \ : \ d(x, F) = r\}$, so that $\{C_F(r) \ : \ r \geq 0\}$ forms a measurable partition of S. Therefore, by Problem, 7.6, there exists a countable set $R \subseteq [0, \infty)$ such that $\mu(C_F(r)) = 0$ for $r \in [0, \infty) \setminus R$. It is not hard to see that $\partial B_F(r) \subseteq C_F(r)$ (btw, the inclusion may be strict), for each $r \geq 0$. Therefore, $\mu(\partial B_F(r)) = 0$, for all $r \in [0, \infty) \setminus R$. ∎

The above claim implies that there exists a sequence $r_k \in [0, \infty) \setminus R$ such that $r_k \searrow 0$. By (5) and the Claim above, we have $\mu_n(B_F(r_k)) \to \mu(B_F(r_k))$ for all $k \in \mathbb{N}$. Hence, for $k \in \mathbb{N}$,

$$\mu(B_F(r_k)) = \lim_n \mu_n(B_F(r_k)) \geq \limsup_n \mu_n(F).$$

By continuity of measure we have $\mu(B_F(r_k)) \searrow \mu(F)$, as $k \to \infty$, and so $\mu(F) \geq \limsup_n \mu_n(F)$. ∎

As we will soon see, it is sometimes easy to prove that $\mu_n(A) \to \mu(A)$ for all A in some subset of $\mathcal{B}(\mathbb{R})$. Our next result has something to say about cases when that is enough to establish weak convergence:

Proposition 7.10 (Weak-convergence test families) *Let $\mathcal{I}$ be a collection of open subsets of S such that*

1. *$\mathcal{I}$ is a π-system,*

2. *Each open set in S can be represented as a finite or countable union of elements of $\mathcal{I}$.*

If $\mu_n(I) \to \mu(I)$, for each $I \in \mathcal{I}$, then $\mu_n \overset{w}{\to} \mu$.

PROOF For $I_1, I_2 \in \mathcal{I}$, we have $I_1 \cap I_2 \in \mathcal{I}$, and so

$$\mu(I_1 \cup I_2) = \mu(I_1) + \mu(I_2) - \mu(I_1 \cap I_2) = \lim_n \mu_n(I_1) + \lim_n \mu_n(I_2) - \lim_n \mu_n(I_1 \cap I_2)$$

$$= \lim_n \Big(\mu_n(I_1) + \mu_n(I_2) - \mu_n(I_1 \cap I_2) \Big) = \lim_n \mu_n(I_1 \cup I_2),$$

so that $I_1 \cup I_2 \in \mathcal{I}$, i.e., $\mathcal{I}$ is closed under finite unions, too.

For an open set G, let $G = \cup_k I_k$ be a representation of G as a union of a countable family in $\mathcal{I}$. By continuity of measure, for each $\varepsilon > 0$ there exists $K \in \mathbb{N}$ such that $\mu(G) \leq \mu(\cup_{k=1}^{K} I_k) + \varepsilon$. Since $\cup_{k=1}^{K} I_K \in \mathcal{I}$, we have

$$\mu(G) + \varepsilon = \mu(\cup_{k=1}^{K} I_k) = \lim_n \mu_n(\cup_{k=1}^{K} I_k) \leq \liminf \mu_n(G).$$

Since $\varepsilon > 0$ was arbitrary, we get $\mu(G) \leq \liminf_n \mu_n(G)$. ∎

Remark 7.11 We could have required that $\mathcal{I}$ be closed under finite unions right from the start. Starting from a π-system, however, is much more useful for applications.

Corollary 7.12 (Weak convergence using cdfs) *Suppose that $S = \mathbb{R}$, and let μ_n be a family of probability measures on $\mathcal{B}(\mathbb{R})$. Let $F(x) = \mu((-\infty, x])$ and $F_n(x) = \mu_n((-\infty, x])$, $x \in \mathbb{R}$ be the corresponding cdfs. Then, the following two statements are equivalent:*

1. *$F_n(x) \to F(x)$ for all x such that F is continuous at x, and*

2. *$\mu_n \overset{w}{\to} \mu$.*

PROOF (2) $\Rightarrow$ (1): Let C be the set of all x such that F is continuous at x; eqivalently, $C = \{x \in \mathbb{R} : \mu(\{x\}) = 0\}$. The sets $(-\infty, x]$ are μ-continuity sets for $x \in C$, so the Portmanteau theorem (Theorem 7.8) implies that $F_n(x) = \mu_n((-\infty, x]) \to \mu((-\infty, x]) = F(x)$, for all $x \in C$.

(1) $\Rightarrow$ (2): The set C^c is at most countable (why?) and so the family

$$\mathcal{I} = \{(a, b) : a < b, a, b \in C\},$$

satisfies the the conditions of Proposition 7.10. To show that $\mu_n \overset{w}{\to} \mu$, it will be enough to show that $\mu_n(I) \to \mu(I)$, for all $a, b \in \mathcal{I}$. Since $\mu((a, b)) = F(b-) - F(a)$, where $F(b-) = \lim_{x \nearrow b} F(x)$, it will be enough to show that

$$F_n(x-) \to F(x),$$

for all $x \in C$. Since $F_n(x-) \leq F_n(x)$, we have $\limsup_n F_n(x-) \leq \lim F_n(x) = F(x)$. To prove the other inequality, we pick $\varepsilon > 0$, and, using the continuity of F at x, find $\delta > 0$ such that $x - \delta \in C$ and $F(x - \delta) > F(x) - \varepsilon$. Since $F_n(x - \delta) \to F(x - \delta)$, there exists $n_0 \in \mathbb{N}$ such that $F_n(x - \delta) > F(x) - 2\varepsilon$ for $n \geq n_0$, and, by increase of F_n, $F_n(x-) > F(x) - 2\varepsilon$, for $n \geq n_0$. Consequently $\liminf_n F_n(x-) \geq F(x) - 2\varepsilon$ and the statement follows. ∎

One of the (many) reasons why weak convergence is so important, is the fact that it possesses nice compactness properties. The central result here is the theorem of Prohorov which is, in a sense, an analogue of the Arzelá-Ascoli compactness theorem for families of measures. The statement we give here is not the most general possible, but it will serve all our purposes.

Definition 7.13 (Tightness, weak compactness) *A subset $\mathcal{M}$ of probability measures on $\mathcal{S}$ is said to be*

1. **tight**, *if for each $\varepsilon > 0$ there exists a compact set K such that*

$$\sup_{\mu \in \mathcal{M}} \mu(K^c) \leq \varepsilon.$$

2. **relatively (sequentially) weakly compact** *if any sequence $\{\mu_n\}_{n \in \mathbb{N}}$ in $\mathcal{M}$ admits a weakly-convergent subsequence $\{\mu_{n_k}\}_{k \in \mathbb{N}}$.*

Theorem 7.14 (Prohorov) *Suppose that the metric space (S, d) is complete and separable, and let $\mathcal{M}$ be a set of probability measures on $\mathcal{S}$. Then $\mathcal{M}$ is relatively weakly compact if and only if it is tight.*

PROOF (*Note:* In addition to the fact that the stated version of the theorem is not the most general available, we only give the proof for the so-called *Helly's selection theorem*, i.e., the special case $S = \mathbb{R}$. The general case is technically more involved, but the key ideas are similar.)

(Tight $\Rightarrow$ relatively weakly compact): Suppose that $\mathcal{M}$ is tight, and let $\{\mu_n\}_{n \in \mathbb{N}}$ be a sequence in $\mathcal{M}$. Let Q be a countable and dense subset of $\mathbb{R}$, and let $\{q_k\}_{k \in \mathbb{N}}$ be an enumeration of Q. Since all $\{\mu_n\}_{n \in \mathbb{N}}$ are probability measures, the sequence $\{F_n(q_1)\}_{n \in \mathbb{N}}$, where $F_n(x) = \mu_n((-\infty, x])$ is bounded. Consequently, it admits a convergent subsequence; we denote its indices by $n_{1,k}$, $k \in \mathbb{N}$. The sequence $\{F_{n_{1,k}}(q_2)\}_{k \in \mathbb{N}}$ is also bounded, so we can extract a further subsequence - let's denote it by $n_{2,k}$, $k \in \mathbb{N}$, so that $F_{n_{2,k}}(q_2)$ converges as $k \to \infty$. Repeating this procedure for each element of Q, we arrive to a sequence of increasing sequences of integers $n_{i,k}$, $k \in \mathbb{N}$, $i \in \mathbb{N}$ with the property that $n_{i+1,k}$, $k \in \mathbb{N}$ is a subsequence of $n_{i,k}$, $k \in \mathbb{N}$ and that $F_{n_{i,k}}(q_j)$ converges for each $j \leq i$. Therefore, the diagonal sequence $m_k = n_{k,k}$, is a subsequence of each $n_{i,k}$, $k \in \mathbb{N}$, $i \in \mathbb{N}$, and can define a function $\tilde{F} : Q \to [0, 1]$ by

$$\tilde{F}(q) = \lim_{k \to \infty} F_{m_k}(q).$$

Each F_n is non-decreasing and so if $\tilde{F}$. As a matter of fact the "right-continuous" version

$$F(x) = \inf_{q < x, q \in Q} \tilde{F}(q),$$

is non-decreasing and right-continuous (why?), with values in $[0, 1]$.

Our next task is to show that $F_{m_k}(x) \to F(x)$, for each $x \in C_F$, where C_F is the set of all points where F is continuous. We pick $x \in C_F$, $\varepsilon > 0$ and $q_1, q_2 \in Q$, $y \in \mathbb{R}$ such that $q_1 < q_2 < x < y$ and

$$F(x) - \varepsilon < F(q_1) \leq F(q_2) \leq F(x) \leq F(y) < F(x) + \varepsilon.$$

Since $F_{m_k}(q_2) \to \tilde{F}(q_2) \geq F(q_1)$ and $F_{m_k}(s) \to \tilde{F}(s) \leq F(s)$ (why is $\tilde{F}(s) \leq F(s)$?), we have, for large enough $k \in \mathbb{N}$

$$F(x) - \varepsilon < F_{m_k}(q_2) \leq F_{m_k}(x) \leq F_{m_k}(s) < F(x) + \varepsilon,$$

which implies that $F_{m_k}(x) \to F(x)$.

It remains to show - thanks to Corollary 7.12 - that $F(x) = \mu((-\infty, x])$, for some probability measure μ on $\mathcal{B}(\mathbb{R})$. For that, in turn, it will be enough to show that $F(x) \to 1$, as $x \to \infty$ and $F(x) \to 0$, as $x \to -\infty$. Indeed, in that case, we would have all the conditions needed to use Problem 6.40 to construct a probability space and a random variable X on it so that F is the cdf of X; the required measure μ would be the distribution $\mu = \mu_X$ of X.

To show that $F(x) \to 0, 1$ as $x \to \pm\infty$, we use tightness (note that this is the only place in the proof where it is used). For $\varepsilon > 0$, we pick $M > 0$ such that $\mu_n([-M, M]) \geq 1 - \varepsilon$, for all $n \in \mathbb{N}$. In terms of corresponding cdfs, this implies that

$$F_n(-M) \leq \varepsilon \text{ and } F_n(M) \geq 1 - \varepsilon \text{ for all } n \in \mathbb{N}.$$

We can assume that $-M$ and M are continuity points of F (why?), so that

$$F(-M) = \lim_k F_{m_k}(-M) \leq \varepsilon \text{ and } F(M) = \lim_k F_{m_k}(M) \geq 1 - \varepsilon,$$

so that $\lim_{x\to\infty} F(x) \geq 1 - \varepsilon$ and $\lim_{x\to-\infty} F(x) \leq \varepsilon$. The claim follows from the arbitrariness of $\varepsilon > 0$.

(Relatively weakly compact $\Rightarrow$ tight): Suppose to the contrary, that $\mathcal{M}$ is relatively weakly compact, but not tight. Then, there exists $\varepsilon > 0$ such that for each $n \in \mathbb{N}$ there exists $\mu_n \in \mathcal{M}$ such that $\mu_n([-n, n]) < 1 - \varepsilon$, and, consequently,

(7.2) $$\mu_n([-M, M]) < 1 - \varepsilon \text{ for } n \geq M.$$

The sequence $\{\mu_n\}_{n\in\mathbb{N}}$ admits a weakly-convergent subsequence $\{\mu_{n_k}\}_{k\in\mathbb{N}}$. By (7.2), we have

$$\limsup_k \mu_{n_k}([-M, M]) \leq 1 - \varepsilon, \text{ for each } M > 0,$$

■

so that $\mu([-M, M]) \leq 1 - \varepsilon$ for all $M > 0$. Continuity of probability implies that $\mu(\mathbb{R}) \leq 1 - \varepsilon$ - a contradiction with the fact that μ is a *probability* measure on $\mathcal{B}(\mathbb{R})$.

The following problem cases tightness in more operational terms:

Problem 7.15 Let $\mathcal{M}$ be a non-empty set of probability measures on $\mathbb{R}$. Show that $\mathcal{M}$ is tight if and only if there exists a non-decreasing function $\varphi : [0, \infty) \to [0, \infty)$ such that

1. $\varphi(x) \to \infty$ as $x \to \infty$, and

2. $\sup_{\mu\in\mathcal{M}} \int \varphi(|x|) \, \mu(dx) < \infty$.

Prohorov's theorem goes well with the following problem (it will be used soon):

Problem 7.16 Let μ be a probability measure on $\mathcal{B}(\mathbb{R})$ and let $\{\mu_n\}_{n\in\mathbb{N}}$ be a sequence of probability measures on $\mathcal{B}(\mathbb{R})$ with the property that every subsequence $\{\mu_{n_k}\}_{k\in\mathbb{N}}$ of $\{\mu_n\}_{n\in\mathbb{N}}$ has a (further) subsequence $\{\mu_{n_{k_l}}\}_{l\in\mathbb{N}}$ which converges towards μ. Show that $\{\mu_n\}_{n\in\mathbb{N}}$ is convergent. (*Hint:* If $\mu_n \overset{w}{\not\to} \mu$, then there exists $f \in C_b$ and a subsequence $\{\mu_{n_k}\}_{k\in\mathbb{N}}$ of $\{\mu_n\}_{n\in\mathbb{N}}$ such that $\int f \, d\mu_{n_k}$ converges, but not to $\int f \, d\mu$.)

We conclude with a comparison between convergence in distribution and convergence in probability.

Proposition 7.17 (Relation between $\overset{\mathbb{P}}{\to}$ and $\overset{\mathcal{D}}{\to}$) *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables. Then $X_n \overset{\mathbb{P}}{\to} X$ implies $X_n \overset{\mathcal{D}}{\to} X$, for any random variable X. Conversely, $X_n \overset{\mathcal{D}}{\to} X$ implies $X_n \overset{\mathbb{P}}{\to} X$ if there exists $c \in \mathbb{R}$ such that $\mathbb{P}[X = c] = 1$.*

PROOF Assume that $X_n \overset{\mathbb{P}}{\to} X$. To show that $X_n \overset{\mathcal{D}}{\to} X$, the Portmanteau theorem guarantees that it will be enough to prove that $\limsup_n \mathbb{P}[X_n \in F] \leq \mathbb{P}[X \in F]$, for all closed sets F. For $F \subseteq \mathbb{R}$, we define $F^\varepsilon = \{x \in \mathbb{R} : d(x, F) \leq \varepsilon\}$. Therefore, for a closed set F, we have

$$\mathbb{P}[X_n \in F] = \mathbb{P}[X_n \in F, |X - X_n| > \varepsilon] + \mathbb{P}[X_n \in F, |X - X_n| \leq \varepsilon]$$
$$\leq \mathbb{P}[|X - X_n| > \varepsilon] + \mathbb{P}[X \in F_\varepsilon].$$

because $X \in F_\varepsilon$ if $X_n \in F$ and $|X - X_n| \leq \varepsilon$. Taking a $\limsup$ of both sides yields

$$\limsup \mathbb{P}[X_n \in F] \leq \mathbb{P}[X \in F_\varepsilon] + \limsup_n \mathbb{P}[|X - X_n| > \varepsilon] = \mathbb{P}[X \in F_\varepsilon].$$

Since $\cap_{\varepsilon > 0} F^\varepsilon = F$, the statement follows.

For the second part, without loss of generality, we assume $c = 0$. Given $m \in \mathbb{N}$, let $f_m \in C_b(\mathbb{R})$ be a continuous function with values in $[0, 1]$ such that $f_m(0) = 1$ and $f_m(x) = 0$ for $|x| > 1/m$. Since $f_m(x) \leq \mathbf{1}_{[-1/m, 1/m]}(x)$, we have

$$\mathbb{P}[|X_n| \leq 1/m] \geq \mathbb{E}[f_m(X_n)] \to f_m(0) = 1,$$

for each $m \in \mathbb{N}$. ∎

Remark 7.18 It is not true that $X_n \overset{\mathcal{D}}{\to} X$ implies $X_n \overset{\mathbb{P}}{\to} X$ in general. Here is a simple example: take $\Omega = \{1, 2\}$ with uniform probability, and define $X_n(1) = 1$ and $X_n(2) = 2$, for n odd and $X_n(1) = 2$ and $X_n(2) = 1$, for n even. Then all X_n have the same distribution, so we have $X_n \overset{\mathcal{D}}{\to} X_1$. On the other hand $\mathbb{P}[|X_n - X_1| \geq \frac{1}{2}] = 1$, for n even. In fact, it is not hard to see that $X_n \overset{\mathbb{P}}{\not\to} X$ for any random variable X.

7.2 Characteristic functions

A characteristic function is simply the Fourier transform, in probabilistic language. Since we will be integrating complex-valued functions, we define (both integrals on the right need to exist)

$$\int f \, d\mu = \int \Re f \, d\mu + i \int \Im f \, d\mu,$$

where $\Re f$ and $\Im f$ denote the real and the imaginary part of a function $f : \mathbb{R} \to \mathbb{C}$. The reader will easily figure out which properties of the integral transfer from the real case.

Definition 7.19 (Characteristic functions) *The **characteristic function** of a probability measure μ on $\mathcal{B}(\mathbb{R})$ is the function $\varphi_\mu : \mathbb{R} \to \mathbb{C}$ given by*

$$\varphi_\mu(t) = \int e^{itx} \, \mu(dx)$$

When we speak of the characteristic function φ_X of a random variable X, we have the characteristic function φ_{μ_X} of its distribution μ_X in mind. Note, however, that

$$\varphi_X(t) = \mathbb{E}[e^{itX}].$$

While difficult to visualize, characteristic functions can be used to learn a lot about the random variables they correspond to. We start with some properties which follow directly from the definition:

Proposition 7.20 (First properties of characteristic functions) *Let X, Y and $\{X_n\}_{n\in\mathbb{N}}$ be a random variables.*

1. $\varphi_X(0) = 1$ and $|\varphi_X(t)| \leq 1$, for all t.

2. $\varphi_{-X}(t) = \overline{\varphi_X(t)}$, where bar denotes complex conjugation.

3. φ_X is uniformly continuous.

4. If X and Y are independent, then $\varphi_{X+Y} = \varphi_X \varphi_Y$.

5. For all $t_1 < t_2 < \cdots < t_n$, the matrix $A = (a_{ij})_{1 \leq i,j \leq n}$ given by

$$a_{jk} = \varphi_X(t_j - t_k),$$

is Hermitian and positive semi-definite, i.e., $A^* = A$ and $\xi^T A \overline{\xi} \geq 0$, for any $\xi \in \mathbb{C}^n$,

6. If $X_n \xrightarrow{\mathcal{D}} X$, then $\varphi_{X_n}(t) \to \varphi_X(t)$, for each $t \in \mathbb{R}$.

PROOF

1. Immediate.

2. $\overline{e^{itx}} = e^{-itx}$.

3. We have $|\varphi_X(t) - \varphi_X(s)| = \left| \int (e^{itx} - e^{isx}) \, \mu(dx) \right| \leq h(t-s)$, where $h(u) = \int \left| e^{iux} - 1 \right| \mu(dx)$. Since $\left| e^{iux} - 1 \right| \leq 2$, dominated convergence theorem implies that $\lim_{u\to 0} h(u) = 0$, and, so, φ_X is uniformly continuous.

4. Independence of X and Y implies the independence of $\exp(itX)$ and $\exp(itY)$. Therefore,

$$\varphi_{X+Y}(t) = \mathbb{E}[e^{it(X+Y)}] = \mathbb{E}[e^{itX} e^{itY}] = \mathbb{E}[e^{itX}] \mathbb{E}[e^{itY}] = \varphi_X(t)\varphi_Y(t).$$

5. The matrix A is Hermitian by (2). To see that it is positive semidefinite, note that $a_{jk} = \mathbb{E}[e^{it_j X} e^{-it_k X}]$, and so

$$\sum_{j=1}^n \sum_{k=1}^n \xi_j \overline{\xi_k} a_{jk} = \mathbb{E}\left[\left(\sum_{j=1}^n \xi_j e^{it_j X} \right) \overline{\left(\sum_{k=1}^n \xi_k e^{it_k X} \right)} \right] = \mathbb{E}[| \sum_{j=1}^n \xi_j e^{it_j X}|^2] \geq 0.$$

6. For $f \in C_b(\mathbb{R})$, we have $f(X_n) \to f(X)$, a.s., and so, by the dominated convergence theorem applied to the cases $f(x) = \cos(tx)$ and $f(x) = \sin(tx)$, we have

$$\varphi_X(t) = \mathbb{E}[\exp(itX)] = \mathbb{E}[\lim_n \exp(itX_n)] = \lim_n \mathbb{E}[\exp(itX_n)] = \lim_n \varphi_{X_n}(t). \qquad \blacksquare$$

Remark 7.21 We do not prove (or use) it in these notes, but it can be shown that a function $\varphi : \mathbb{R} \to \mathbb{C}$, continuous at the origin with $\varphi(0) = 1$ is a characteristic function of some probability measure μ on $\mathcal{B}(\mathbb{R})$ if and only if it satisfies (5). This is known as **Bochner's theorem**.

Here is a simple problem you can use to test your understanding of the definitions:

Problem 7.22 Let μ and ν be two probability measures on $\mathcal{B}(\mathbb{R})$, and let φ_μ and φ_ν be their characteristic functions. Show that **Parseval's identity** holds:

$$\int_{\mathbb{R}} e^{-its} \varphi_\mu(t)\, \nu(dt) = \int_{\mathbb{R}} \varphi_\nu(t-s)\, \mu(dt), \text{ for all } s \in \mathbb{R}.$$

Our next result shows μ can be recovered from its characteristic function φ_μ:

Theorem 7.23 (Inversion theorem) *Let μ be a probability measure on $\mathcal{B}(\mathbb{R})$, and let $\varphi = \varphi_\mu$ be its characteristic function. Then, for $a < b \in \mathbb{R}$, we have*

(7.3)
$$\mu((a,b)) + \tfrac{1}{2}\mu(\{a,b\}) = \tfrac{1}{2\pi} \lim_{T\to\infty} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt.$$

PROOF We start by picking $a < b$ and noting that

$$\frac{e^{-ita} - e^{-itb}}{it} = \int_a^b e^{-ity}\, dy,$$

so that, by Fubini's theorem, the integral in (7.3) is well-defined:

$$F(a,b,T) = \int_{[-T,T]\times[a,b]} \exp(-ity)\varphi(t)\, dy\, dt,$$

where

$$F(a,b,T) = \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt.$$

Another use of Fubini's theorem yields:

$$F(a,b,T) = \int_{[-T,T]\times[a,b]\times\mathbb{R}} \exp(-ity)\exp(itx)\, dy\, dt\, \mu(dx)$$

$$= \int_{\mathbb{R}} \left(\int_{[-T,T]\times[a,b]} \exp(-it(y-x))\, dy\, dt \right) \mu(dx)$$

$$= \int_{\mathbb{R}} \left(\int_{[-T,T]} \tfrac{1}{it} \left(e^{-it(a-x)} - e^{-it(b-x)} \right) dt \right) \mu(dx).$$

Set

$$f(a,b,T) = \int_{-T}^{T} \tfrac{1}{it}(e^{-it(a-x)} - e^{-it(b-x)})\, dt \text{ and } K(T,c) = \int_0^T \tfrac{\sin(ct)}{t}\, dt,$$

and note that, since $\cos$ is an even and $\sin$ an odd function, we have

$$f(a,b,T;x) = 2\int_0^T \left(\tfrac{\sin((a-x)t)}{t} - \tfrac{\sin((b-x)t)}{t} \right) dt = 2K(T; a-x) - 2K(T; b-x).$$

(*Note:* The integral $\int_{-T}^{T} \frac{1}{it} \exp(-it(a-x)) \, dt$ is not defined; we *really need* to work with the full $f(a, b, T; x)$ to get the cancellation above.)

Since

$$
(7.4) \qquad K(T; c) = \begin{cases} \int_0^T \frac{\sin(ct)}{ct} d(ct) = \int_0^{cT} \frac{\sin(s)}{s} \, ds = K(cT; 1), & c > 0 \\ 0, & c = 0 \\ -K(|c| T; 1), & c < 0, \end{cases}
$$

Problem 5.24 implies that

$$
\lim_{T \to \infty} K(T; c) = \left\{ \begin{array}{cc} \frac{\pi}{2}, & c > 0 \\ 0, & c = 0 \\ -\frac{\pi}{2}, & c < 0 \end{array} \right\} \text{ and so } \lim_{T \to \infty} f(a, b, T; x) = \begin{cases} 0, & x \in [a, b]^c \\ \pi, & x = a \text{ or } x = b \\ 2\pi, & a < x < b \end{cases}
$$

Observe first that the function $T \mapsto K(T; 1)$ is continuous on $[0, \infty)$ and has a finite limit as $T \to \infty$ so that $\sup_{T \geq 0} |K(T; 1)| < \infty$. Furthermore, (7.4) implies that $|K(T; c)| \leq \sup_{T \geq 0} K(T; 1)$ for any $c \in \mathbb{R}$ and $T \geq 0$ so that

$$
\sup\{|f(a, b, T; x)| \ : \ x \in \mathbb{R}, T \geq 0\} < \infty. \qquad \blacksquare
$$

Therefore, we can use the dominated convergence theorem to conclude that

$$
\lim_{T \to \infty} \frac{1}{2\pi} F(a, b, T; x) = \lim_{T \to \infty} \frac{1}{2\pi} \int f(a, b, T; x) \, \mu(dx) = \frac{1}{2\pi} \int \lim_{T \to \infty} f(a, b, T; x) \, \mu(x)
$$

$$
= \tfrac{1}{2} \mu(\{a\}) + \mu((a, b)) + \tfrac{1}{2} \mu(\{b\}).
$$

Corollary 7.24 ("Characteristic-ness" of characteristic functions) *For probability measures μ_1 and μ_2 on $\mathcal{B}(\mathbb{R})$, the equality $\varphi_{\mu_1} = \varphi_{\mu_2}$ implies that $\mu_1 = \mu_2$.*

PROOF By Theorem 7.23, we have $\mu_1((a, b)) = \mu_2((a, b))$ for all $a, b \in C$ where C is the set of all $x \in \mathbb{R}$ such that $\mu_1(\{x\}) = \mu_2(\{x\}) = 0$. Since C^c is at most countable, it is straightforward to see that the family $\{(a, b) \ : \ a, b \in C\}$ of intervals is a π-system which generates $\mathcal{B}(\mathbb{R})$. $\qquad \blacksquare$

Corollary 7.25 (Inversion for integrably characteristic functions) *Suppose that μ is a probability measure on $\mathcal{B}(\mathbb{R})$ with $\int_{\mathbb{R}} |\varphi_\mu(t)| \, dt < \infty$. Then $\mu \ll \lambda$ and $\frac{d\mu}{d\lambda}$ is a bounded and continuous function given by*

$$
\frac{d\mu}{d\lambda} = f, \text{ where } f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \varphi_\mu(t) \, dt.
$$

PROOF Since φ_μ is integrable and $\left| e^{-itx} \right| = 1$, f is well defined. For $a < b$ we have

$$
\int_a^b f(x) \, dx = \frac{1}{2\pi} \int_a^b \int_{\mathbb{R}} e^{-itx} \varphi_\mu(t) \, dt \, dx = \frac{1}{2\pi} \int_{\mathbb{R}} \varphi_\mu(t) \left(\int_a^b e^{-itx} \, dx \right) dt
$$

$$
(7.5) \qquad = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) \, dt = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) \, dt
$$

$$
= \mu((a, b)) + \tfrac{1}{2} \mu(\{a, b\}),
$$

by Theorem 7.23, where the use of Fubini's theorem above is justified by the fact that the function $(t, x) \mapsto e^{-itx}\varphi_\mu(t)$ is integrable on $[a, b] \times \mathbb{R}$, for all $a < b$. For a, b such that $\mu(\{a\}) = \mu(\{b\}) = 0$, the equation (7.5) implies that $\mu((a, b)) = \int_a^b f(x)\, dx$. The claim now follows by the $\pi - \lambda$-theorem. ∎

Example 7.26 (Common distributions and their characteristic functions) Here is a list of some common distributions and the corresponding characteristic functions:

1. *Continuous distributions.*

	Name	Parameters	Density $f_X(x)$	Ch. function $\varphi_X(t)$
1	Uniform	$a < b$	$\frac{1}{b-a}\,\mathbf{1}_{[a,b]}(x)$	$\frac{e^{-ita}-e^{-itb}}{it(b-a)}$
2	Normal	$\mu \in \mathbb{R},\ \sigma > 0$	$\frac{1}{\sqrt{2\pi\sigma^2}}\,\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\exp(i\mu t - \frac{1}{2}\sigma^2 t^2)$
3	Exponential	$\lambda > 0$	$\lambda\exp(-\lambda x)\mathbf{1}_{[0,\infty)}(x)$	$\frac{1}{1-it}$
4	Double Exponential	$\lambda > 0$	$\frac{1}{2}\,\lambda\exp(-\lambda\,|x|)$	$\frac{1}{1+t^2}$
5	Cauchy	$\mu \in \mathbb{R}, \gamma > 0$	$\frac{\gamma}{\pi(\gamma^2+(x-\mu)^2)}$	$\exp(i\mu t - \gamma\,|t|)$

2. *Discrete distributions.*

	Name	Parameters	$p_n = \mathbb{P}[X = n],\ n \in \mathbb{Z}$	Ch. function $\varphi_X(t)$
6	Dirac	$m \in \mathbb{N}_0$	$\mathbf{1}_{\{m=n\}}$	$\exp(itm)$
7	Coin-toss	$p \in (0,1)$	$p_1 = p, p_{-1} = (1-p)$	$\cos(t)$
8	Geometric	$p \in (0,1)$	$p^n(1-p),\ n \in \mathbb{N}_0$	$\frac{1-p}{1-e^{it}p}$
9	Poisson	$\lambda > 0$	$e^{-\lambda}\frac{\lambda^n}{n!}\,,\ n \in \mathbb{N}_0$	$\exp(\lambda(e^{it} - 1))$

3. *A singular distribution.*

	Name	Ch. function $\varphi_X(t)$
10	Cantor	$e^{\frac{1}{2}it}\prod_{k=1}^{\infty}\cos\left(\frac{t}{3^k}\right)$

7.3 Tail behavior

We continue by describing several methods one can use to extract useful information about the tails of the underlying probability distribution from a characteristic function.

Proposition 7.27 (Existence of moments implies regularity of φ_X at 0) *Let X be a random variable. If $\mathbb{E}[|X|^n] < \infty$, then $\frac{d^n}{(dt)^n}\varphi_X(t)$ exists for all t and*

$$\frac{d^n}{(dt)^n}\varphi_X(t) = \mathbb{E}[e^{itX}(iX)^n].$$

In particular

$$\mathbb{E}[X^n] = (-i)^n \frac{d^n}{(dt)^n}\varphi_X(0).$$

PROOF We give the proof in the case $n = 1$ and leave the general case to the reader:

$$\lim_{h\to 0}\tfrac{\varphi(h)-\varphi(0)}{h} = \lim_{h\to 0}\int_{\mathbb{R}}\tfrac{e^{ihx}-1}{h}\,\mu(dx) = \int_{\mathbb{R}}\lim_{h\to 0}\tfrac{e^{ihx}-1}{h}\,\mu(dx) = \int_{\mathbb{R}} ix\,\mu(dx),$$

where the passage of the limit under the integral sign is justified by the dominated convergence theorem which, in turn, can be used since

$$\left|\tfrac{e^{ihx}-1}{h}\right| \le |x|, \text{ and } \int_{\mathbb{R}}|x|\,\mu(dx) = \mathbb{E}[|X|] < \infty. \qquad \blacksquare$$

Remark 7.28

1. It can be shown that for n *even*, the existence of $\frac{d^n}{(dt)^n}\varphi_X(0)$ (in the appropriate sense) implies the finiteness of the n-th moment $\mathbb{E}[|X|^n]$.

2. When n is odd, it can happen that $\frac{d^n}{(dt)^n}\varphi_X(0)$ exists, but $\mathbb{E}[|X|^n] = \infty$ - see Problem 7.40.

Finer estimates of the tails of a probability distribution can be obtained by finer analysis of the behavior of φ around 0:

Proposition 7.29 (A tail estimate) *Let μ be a probability measure on $\mathcal{B}(\mathbb{R})$ and let $\varphi = \varphi_\mu$ be its characteristic function. Then, for $\varepsilon > 0$ we have*

$$\mu([-\tfrac{2}{\varepsilon}, \tfrac{2}{\varepsilon}]^c) \le \tfrac{1}{\varepsilon}\int_{-\varepsilon}^{\varepsilon}(1 - \varphi(t))\,dt.$$

PROOF Let X be a random variable with distribution μ. We start by using Fubini's theorem to get

$$\tfrac{1}{2\varepsilon}\int_{\varepsilon}^{\varepsilon}(1-\varphi(t))\,dt = \tfrac{1}{2\varepsilon}\mathbb{E}[\int_{-\varepsilon}^{\varepsilon}(1-e^{itX})\,dt] = \tfrac{1}{\varepsilon}\mathbb{E}[\int_0^{\varepsilon}(1-\cos(tX))\,dt] = \mathbb{E}[1 - \tfrac{\sin(\varepsilon X)}{\varepsilon X}].$$

It remains to observe that $1 - \frac{\sin(x)}{x} \ge 0$ and $1 - \frac{\sin(x)}{x} \ge 1 - \frac{1}{|x|}$ for all x. Therefore, if we use the first inequality on $[-2, 2]$ and the second one on $[-2, 2]^c$, we get

$$1 - \tfrac{\sin(x)}{x} \ge \tfrac{1}{2}\mathbf{1}_{\{|x|>2\}} \text{ so that } \tfrac{1}{2\varepsilon}\int_{\varepsilon}^{\varepsilon}(1-\varphi(t))\,dt \ge \tfrac{1}{2}\mathbb{P}[|\varepsilon X| > 2] = \tfrac{1}{2}\mu([-\tfrac{2}{\varepsilon}, \tfrac{2}{\varepsilon}]^c). \qquad \blacksquare$$

Problem 7.30 Use the inequality of Proposition 7.29 to show that if $\varphi(t) = 1 + O(|t|^\alpha)$ for some $\alpha > 0$, then $\int_\mathbb{R} |x|^\beta \, \mu(dx) < \infty$, for all $\beta < \alpha$. Give an example where $\int_\mathbb{R} |x|^\alpha \, \mu(dx) = \infty$.

   (*Note:* "$f(t) = g(t) + O(h(t))$" means that $\sup_{|t| \leq \delta} \frac{|f(t) - g(t)|}{h(t)} < \infty$, for some $\delta > 0$.)

Problem 7.31 (Riemann-Lebesgue theorem) Suppose that $\mu \ll \lambda$. Show that

$$\lim_{t \to \infty} \varphi(t) = \lim_{t \to -\infty} \varphi(t) = 0.$$

(*Hint:* Use (and prove) the fact that $f \in \mathcal{L}^1_+(\mathbb{R})$ can be approximated in $\mathcal{L}^1(\mathbb{R})$ by a function of the form $\sum_{k=1}^n \alpha_k \mathbf{1}_{[a_k, b_k]}$.)

7.4   The continuity theorem

Theorem 7.32 (Continuity theorem) *Let $\{\mu_n\}_{n \in \mathbb{N}}$ be a sequence of probability distributions on $\mathcal{B}(\mathbb{R})$, and let $\{\varphi_n\}_{n \in \mathbb{N}}$ be the sequence of their characteristic functions. Suppose that there exists a function $\varphi : \mathbb{R} \to \mathbb{C}$ such that*

1. *$\varphi_n(t) \to \varphi(t)$, for all $t \in \mathbb{R}$, and*

2. *φ is continuous at $t = 0$.*

Then, φ is the characteristic function of a probability measure μ on $\mathcal{B}(\mathbb{R})$ and $\mu_n \overset{w}{\to} \mu$.

PROOF We start by showing that the continuity of the limit φ implies tightness of $\{\mu_n\}_{n \in \mathbb{N}}$. Given $\varepsilon > 0$ there exists $\delta > 0$ such that $1 - \varphi(t) \leq \varepsilon/2$ for $|t| \leq \delta$. By the dominated convergence theorem we have

$$\limsup_{n \to \infty} \mu_n([-\tfrac{2}{\delta}, \tfrac{2}{\delta}]^c) \leq \limsup_{n \to \infty} \tfrac{1}{\delta} \int_\delta^\delta (1 - \varphi_n(t)) \, dt = \tfrac{1}{\delta} \int_{-\delta}^\delta (1 - \varphi(t)) \, dt \leq \varepsilon.$$

By taking an even smaller $\delta' > 0$, we can guarantee that

$$\sup_{n \in \mathbb{N}} \mu_n([-\tfrac{2}{\delta'}, \tfrac{2}{\delta'}]^c) \leq \varepsilon,$$

which, together with the arbitrariness of $\varepsilon > 0$ implies that $\{\mu_n\}_{n \in \mathbb{N}}$ is tight.

   Let $\{\mu_{n_k}\}_{k \in \mathbb{N}}$ be a convergent subsequence of $\{\mu_n\}_{n \in \mathbb{N}}$, and let μ be its limit. Since $\varphi_{n_k} \to \varphi$, we conclude that φ is the characteristic function of μ. It remains to show that the whole sequence converges to μ weakly. This follows, however, directly from Problem 7.16, since any convergent subsequence $\{\mu_{n_k}\}_{k \in \mathbb{N}}$ has the same limit μ. ∎

Problem 7.33 Let φ be a characteristic function of some probability measure μ on $\mathcal{B}(\mathbb{R})$. Show that $\hat{\varphi}(t) = e^{\varphi(t) - 1}$ is also a characteristic function of some probability measure $\hat{\mu}$ on $\mathcal{B}(\mathbb{R})$.

Problem 7.34 (Scheffé's Theorem) Let $\{X_n\}_{n \in \mathbb{N}}$ be absolutely-continuous random variables with densities f_{X_n}, such that $f_{X_n}(x) \to f(x)$, λ-a.e., where f is the density of the absolutely-continuous random variable X. Show that $X_n \overset{\mathcal{D}}{\to} X$. (*Hint:* Show that $\int_{\mathbb{R}} |f_{X_n} - f| \, d\lambda \to 0$ by writing the integrand in terms of $(f - f_{X_n})^+ \leq f$.)

Problem 7.35 (Convergence of moments) Let $\{X_n\}_{n \in \mathbb{N}}$ and X be random variables with a common uniform bound, i.e., such that

$$\exists\, M > 0, \ \forall\, n \in \mathbb{N}, \ |X_n| \leq M, \ |X| \leq M, \text{ a.s.}$$

Show that the following two statements are equivalent:

1. $X_n \overset{\mathcal{D}}{\to} X$ (where $\overset{\mathcal{D}}{\to}$ denotes convergence in distribution), and

2. $\mathbb{E}[X_n^k] \to \mathbb{E}[X^k]$, as $n \to \infty$, for all $k \in \mathbb{N}$.

(*Hint:* Use the Weierstrass approximation theorem: *Given* $a < b \in \mathbb{R}$, *a continuous function* $f : [a, b] \to \mathbb{R}$ *and* $\varepsilon > 0$ *there exists a polynomial* P *such that* $\sup_{x \in [a,b]} |f(x) - P(x)| \leq \varepsilon$.)

Problem 7.36 (Total-variation convergence) A sequence $\{\mu_n\}_{n \in \mathbb{N}}$ of probability measures on $\mathcal{B}(\mathbb{R})$ is said to converge to the probability measure μ **in (total) variation** if

$$\sup_{A \in \mathcal{B}(\mathbb{R})} |\mu_n(A) - \mu(A)| \to 0 \text{ as } n \to \infty.$$

Compare convergence in variation to weak convergence: if one implies the other, prove it. Give counterexamples, if they are not equivalent.

Problem 7.37 (Convergence of Maxima) Let $\{X_n\}_{n \in \mathbb{N}}$ be an iid sequence of standard normal $(N(0,1))$ random variables. Define the sequence of *up-to-date-maxima* $\{M_n\}_{n \in \mathbb{N}}$ by

$$M_n = \max(X_1, \ldots, X_n).$$

Show that

1. Show that $\lim_{x \to \infty} \frac{\mathbb{P}[X_1 > x]}{x^{-1} \exp(-\frac{1}{2}x^2)} = (2\pi)^{-\frac{1}{2}}$ by establishing the following inequality

   (7.6)
   $$\frac{1}{x} \geq \frac{\mathbb{P}[X_1 > x]}{\phi(x)} \geq \frac{1}{x} - \frac{1}{x^3}, \ x > 0,$$

   where, $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$ is the density of the standard normal. (*Hint:* Use integration by parts).

2. Prove that for any $\theta \in \mathbb{R}$, $\lim_{x \to \infty} \frac{\mathbb{P}[X_1 > x + \frac{\theta}{x}]}{\mathbb{P}[X_1 > x]} = \exp(-\theta)$.

3. Let $\{b_n\}_{n \in \mathbb{N}}$ be a sequence of real numbers with the property that $\mathbb{P}[X_1 > b_n] = 1/n$. Show that $\mathbb{P}[b_n(M_n - b_n) \leq x] \to \exp(-e^{-x})$.

4. Show that $\lim_n \frac{b_n}{\sqrt{2 \log n}} = 1$.

5. Show that $\frac{M_n}{\sqrt{2\log n}} \to 1$ in probability. (*Hint:* Use Problem 7.17.)

Problem 7.38 Check that the expressions for characteristic functions in Example 7.26 are correct. (*Hint:* Not much computing is needed. Use the inversion theorem. For 2., start with the case $\mu = 0$, $\sigma = 1$ and derive a first-order differential equation for φ.)

Problem 7.39 (Atoms from the characteristic function) Let μ be a probability measure on $\mathcal{B}(\mathbb{R})$, and let $\varphi = \varphi_\mu$ be its characteristic function.

1. Show that $\mu(\{a\}) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} e^{-ita}\varphi(t)\,dt$.

2. Show that if $\lim_{t\to\infty} |\varphi(t)| = \lim_{t\to-\infty} |\varphi(t)| = 0$, then μ has no atoms.

3. Show that converse of (2) is false. (*Hint:* Prove that $|\varphi(t_n)| = 1$ along a suitably chosen sequence $t_n \to \infty$, where φ is the characteristic function of the Cantor distribution.)

Problem 7.40 (Existence of $\varphi_X'(0)$ does not imply that $X \in \mathbb{L}^1$) Let X be a random variable which takes values in $\mathbb{Z} \setminus \{-2, -1, 0, 1, 2\}$ with

$$\mathbb{P}[X = k] = \mathbb{P}[X = -k] = \frac{C}{k^2 \log(k)}, \text{ for } k = 3, 4, \ldots,$$

where $C = \frac{1}{2}(\sum_{k\geq 3} \frac{1}{k^2 \log(k)})^{-1} \in (0, \infty)$. Show that $\varphi_X'(0) = 0$, but $X \notin \mathbb{L}^1$.
   (*Hint:* Argue that, in order to establish that $\varphi_X'(0) = 0$, it is enough to show that

$$\lim_{h\to 0} \frac{1}{h} \sum_{k\geq 3} \frac{\cos(hk)-1}{k^2 \log(k)} = 0.$$

Then split the sum at k close to $2/h$ and use (and prove) the inequality $|\cos(x) - 1| \leq \min(x^2/2, x)$. Bounding sums by integrals may help, too.)

Problem 7.41 (Multivariate characteristic functions) Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random vector. The *characteristic function* $\varphi = \varphi_{\boldsymbol{X}} : \mathbb{R}^n \to \mathcal{C}$ is given by

$$\varphi(t_1, t_2, \ldots, t_n) = \mathbb{E}[\exp(i \sum_{k=1}^{n} t_k X_k)].$$

We will also use the shortcut $\boldsymbol{t}$ for $(t_1, \ldots, t_n)$ and $\boldsymbol{t} \cdot \boldsymbol{X}$ for the random variable $\sum_{k=1}^{n} t_k X_k$. Take for granted the following statement (the proof of which is similar to the proof of the 1-dimensional case):

Fact. Suppose that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are random vectors with $\varphi_{\boldsymbol{X}_1}(\boldsymbol{t}) = \varphi_{\boldsymbol{X}_2}(\boldsymbol{t})$ for all $\boldsymbol{t} \in \mathbb{R}^n$. Then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have the same distribution, i.e. $\mu_{\boldsymbol{X}_1} = \mu_{\boldsymbol{X}_2}$.

   Prove the following statements

1. Random variables X and Y are independent if and only if $\varphi_{(X,Y)}(t_1, t_2) = \varphi_X(t_1)\varphi_Y(t_2)$ for all $t_1, t_2 \in \mathbb{R}$.

2. Random vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have the same distribution if and only if random variables $\boldsymbol{t} \cdot \boldsymbol{X}_1$ and $\boldsymbol{t} \cdot \boldsymbol{X}_2$ have the same distribution for all $\boldsymbol{t} \in \mathbb{R}^n$. (This fact is known as *Wald's device.*)

An n-dimensional random vector $\boldsymbol{X}$ is said to be *Gaussian* (or, to have the *multivariate normal distribution*) if there exists a vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and a symmetric positive semi-definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ such that

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(i\, \boldsymbol{t} \cdot \boldsymbol{\mu} - \tfrac{1}{2} \boldsymbol{t}^\tau \boldsymbol{\Sigma} \boldsymbol{t}),$$

where $\boldsymbol{t}$ is interpreted as a column vector, and $()^\tau$ is transposition. This is denoted as $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. $\boldsymbol{X}$ is said to be *non-degenerate* if $\boldsymbol{\Sigma}$ is positive definite.

3. Show that a random vector $\boldsymbol{X}$ is Gaussian, if and only if the random vector $\boldsymbol{t} \cdot \boldsymbol{X}$ is normally distributed (with some mean and variance) for each $\boldsymbol{t} \in \mathbb{R}^n$. (*Note:* Be careful, nothing in the second statement tells you what the mean and variance of $\boldsymbol{t} \cdot \boldsymbol{X}$ are.)

4. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be a Gaussian random vector. Show that X_k and X_l, $k \neq l$, are independent if and only if they are uncorrelated.

5. Construct a random vector (X, Y) such that both X and Y are normally distributed, but that $\boldsymbol{X} = (X, Y)$ is *not* Gaussian.

6. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be a random vector consisting of n independent random variables with $X_i \sim N(0, 1)$. Let $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ be a given positive semi-definite symmetric matrix, and $\boldsymbol{\mu} \in \mathbb{R}^n$ a given vector. Show that there exists an affine transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ such that the random vector $T(\boldsymbol{X})$ is Gaussian with $T(\boldsymbol{X}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

7. Find a necessary and sufficient condition on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ such that the converse of the previous problem holds true: For a Gaussian random vector $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, there exists an affine transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ such that $T(\boldsymbol{X})$ has independent components with the $N(0, 1)$-distribution (i.e. $T(\boldsymbol{X}) \sim N(0, \boldsymbol{y}I)$, where $\boldsymbol{y}I$ is the identity matrix).

Problem 7.42 (Slutsky's Theorem) Let $X, Y, \{X_n\}_{n \in \mathbb{N}}$ and $\{Y_n\}_{n \in \mathbb{N}}$ be random variables defined on the same probability space, such that $X_n \overset{\mathcal{D}}{\to} X$ and $Y_n \overset{\mathcal{D}}{\to} Y$. Show that

1. It is not necessarily true that $X_n + Y_n \overset{\mathcal{D}}{\to} X + Y$. For that matter, we do not necessarily have $(X_n, Y_n) \overset{\mathcal{D}}{\to} (X, Y)$ (where the pairs are considered as random elements in the metric space $\mathbb{R}^2$).

2. If, in addition to (11.13), there exists a constant $c \in \mathbb{R}$ such that $\mathbb{P}[Y = c] = 1$, show that $g(X_n, Y_n) \overset{\mathcal{D}}{\to} g(X, c)$, for any continuous function $g : \mathbb{R}^2 \to \mathbb{R}$. (*Hint:* It is enough to show that $(X_n, Y_n) \overset{\mathcal{D}}{\to} (X_n, c)$. Use Problem 7.41).)

Problem 7.43 (Convergence of a normal sequence)

1. Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of normally-distributed random variables converging weakly towards a random variable X. Show that X must be a normal random variable itself. (*Hint:* Use the following fact: *for a sequence $\{\mu_n\}_{n \in \mathbb{N}}$ of real numbers, the following two statements are equivalent*

$$\mu_n \to \mu \in \mathbb{R}, \ \text{ and } \ \forall\, t \in \mathbb{R}, \ \exp(it\mu_n) \to \exp(it\mu).$$

You don't need to prove it, but feel free to try.)

2. Let X_n be a sequence of normal random variables such that $X_n \overset{a.s.}{\to} X$. Show that $X_n \overset{\mathbb{L}^p}{\to} X$ for all $p \geq 1$.

Chapter 8

Classical Limit Theorems

8.1 The weak law of large numbers

We start with a definitive form of the weak law of large numbers. We need two lemmas, first:

Lemma 8.1 (A simple inequality) *Let $u_1, u_2, \ldots, u_n$ and $w_1, w_2, \ldots, w_n$ be complex numbers, all of modulus at most $M > 0$. Then*

(8.1)
$$\left| \prod_{k=1}^{n} u_k - \prod_{k=1}^{n} w_k \right| \leq M^{n-1} \sum_{k=1}^{n} |u_k - w_k|.$$

PROOF We proceed by induction. For $n = 1$, the claim is trivial. Suppose that (8.1) holds. Then

$$\left| \prod_{k=1}^{n+1} u_k - \prod_{k=1}^{n+1} w_k \right| \leq \left| \prod_{k=1}^{n} u_k \right| |u_{n+1} - w_{n+1}| + |w_{n+1}| \left| \prod_{k=1}^{n} u_k - \prod_{k=1}^{n} w_k \right|$$

$$\leq M^n |u_{n+1} - w_{n+1}| + M \times M^{n-1} \sum_{k=1}^{n} |u_k - w_k|$$

$$= M^{(n+1)-1} \sum_{k=1}^{n+1} |u_k - w_k|.$$

Lemma 8.2 (Convergence to the exponential) *Let $\{z_n\}_{n \in \mathbb{N}}$ be a sequence of complex numbers with $z_n \to z \in \mathbb{C}$. Then $(1 + \frac{z_n}{n})^n \to e^z$.*

PROOF Using Lemma 8.1 with $u_k = 1 + \frac{z_n}{n}$ and $w_k = e^{z_n/n}$ for $k = 1, \ldots, n$, we get

(8.2)
$$\left| (1 + \tfrac{z_n}{n})^n - e^{z_n} \right| \leq n M_n^{n-1} \left| 1 + \tfrac{z_n}{n} - e^{z_n/n} \right|,$$

where $M_n = \max(\left|1 + \frac{z_n}{n}\right|, \left|e^{z_n/n}\right|)$. Let $K = \sup_{n \in \mathbb{N}} |z_n| < \infty$, so that $\left|e^{z_n/n}\right|^n \leq e^{|z_n|} \leq e^K$. Similarly, $\left|1 + \frac{z_n}{n}\right|^n \leq (1 + \frac{K}{n})^n \to e^K$. Therefore

$$(8.3) \qquad L = \sup_{n \in \mathbb{N}} M_n^{n-1} < \infty.$$

To estimate the last term in (8.2), we start with the Taylor expansion $e^b = 1 + b + \sum_{k \geq 2} \frac{b^k}{k!}$, which converges absolutely for all $b \in \mathbb{C}$. Then, we use the fact that $\frac{1}{k!} \leq 2^{-k+1}$, to obtain

$$(8.4) \qquad \left|e^b - 1 - b\right| \leq \sum_{k \geq 2} |b|^k \tfrac{1}{k!} \leq |b^2| \sum_{k \geq 2} 2^{-k+1} = |b|^2 \,, \text{ for } |b| \leq 1.$$

Since $|z_n|/n \leq 1$ for large-enough n, it follows from (8.2), (8.4) and (8.3), that

$$\limsup_n \left|(1 + \tfrac{z_n}{n})^n - e^{z_n}\right| \leq \limsup_n nL \left|\tfrac{z_n}{n}\right|^2 = 0.$$

It remains to remember that $e^{z_n} \to e^z$ to finish the proof. ∎

Theorem 8.3 (Weak law of large numbers) *Let $\{X_n\}_{n \in \mathbb{N}}$ be an iid sequence of random variables with the (common) distribution μ and the characteristic function $\varphi = \varphi_\mu$ such that $\varphi'(0)$ exists. Then, $c = -i\varphi'(0) \in \mathbb{R}$ and*

$$\frac{1}{n} \sum_{k=1}^{n} X_k \to c \text{ in probability.}$$

PROOF Since $\varphi(-s) = \overline{\varphi(s)}$, we have

$$\varphi'(0) = \lim_{s \to 0} \tfrac{\varphi(-s)-1}{-s} = \lim_{s \to 0} \tfrac{\overline{\varphi(s)}-1}{-s} = -\lim_{s \to 0} \overline{\tfrac{\varphi(s)-1}{s}} = -\overline{\varphi'(0)}.$$

Therefore, $c = -i\varphi'(0) \in \mathbb{R}$.

Let $S_n = \sum_{k=1}^n X_k$. According to Proposition 7.17, it will be enough to show that $\frac{1}{n} S_n \xrightarrow{\mathcal{D}} c = -i\varphi'(0) \in \mathbb{R}$. Moreover, by Theorem 7.32, all we need to do is show that $\varphi_{\frac{1}{n} S_n}(t) \to e^{itc} = e^{t\varphi'(0)}$, for all $t \in \mathbb{R}$.

The iid property of $\{X_n\}_{n \in \mathbb{N}}$ and the fact that $\varphi_{\alpha X}(t) = \varphi_X(\alpha t)$ imply that

$$\varphi_{\frac{1}{n} S_n}(t) = (\varphi(\tfrac{t}{n}))^n = (1 + \tfrac{z_n}{n})^n,$$ ∎

where $z_n = n(\varphi(\frac{t}{n}) - 1)$. By the assumption, we have $\lim_{s \to 0} \frac{\varphi(s)-1}{s} = ic$, and so $z_n \to t\varphi'(0)$. Therefore, by Lemma 8.2 above, we have $\varphi_{\frac{1}{n} S_n}(t) \to e^{itc}$.

Remark 8.4

1. It can be shown that the converse of Theorem 8.3 is true in the following sense: if $\frac{1}{n} S_n \xrightarrow{\mathbb{P}} c \in \mathbb{R}$, then $\varphi'(0)$ exists and $\varphi'(0) = ic$. That's why we call the result of Theorem 8.3 definitive.

2. $X_1 \in \mathbb{L}^1$ implies $\varphi'(0) = \mathbb{E}[X_1]$, so that Theorem 8.3 covers the classical case. As we have seen in Problem 7.40, there are cases when $\varphi'(0)$ exists but $\mathbb{E}[|X_1|] = \infty$.

Problem 8.5 Let $\{X_n\}_{n \in \mathbb{N}}$ be iid with the Cauchy distribution (remember, the density of the Cauchy distribution is given by $f_X(x) = \frac{1}{\pi(1+x^2)}$, $x \in \mathbb{R}$). Show that φ_X is not differentiable at 0 and show that there is no constant c such that

$$\tfrac{1}{n} S_n \xrightarrow{\mathbb{P}} c,$$

where $S_n = \sum_{k=1}^{n} X_k$. (*Hint:* What is the distribution of $\frac{1}{n} S_n$?)

8.2   An "iid"-central limit theorem

We continue with a central limit theorem for iid sequences. Unlike in the case of the (weak) law of large numbers, existence of the first moment will not be enough - we will need to assume that the second moment is finite, too. We will see how this assumption can be relaxed when we state and prove the Lindeberg-Feller theorem. We start with an estimate of the "error" term in the Taylor expansion of the exponential function of imaginary argument:

Lemma 8.6 (A tight error estimate for the exponential) *For $\xi \in \mathbb{R}$ we have*

$$\left| e^{i\xi} - \sum_{k=0}^{n} \frac{(i\xi)^k}{k!} \right| \le \min\left(\frac{|\xi|^{n+1}}{(n+1)!}, 2\frac{|\xi|^n}{n!} \right)$$

PROOF If we write the remainder in the Taylor formula in the integral form (derived easily using integration by parts), we get

$$e^{i\xi} - \sum_{k=0}^{n} \frac{(i\xi)^k}{k!} = R_n(\xi), \text{ where } R_n(\xi) = i^{n+1} \int_0^{\xi} e^{iu} \frac{(\xi-u)^n}{n!} \, du.$$

The usual estimate of R_n gives:

$$|R_n(\xi)| \le \tfrac{1}{n!} \int_0^{|\xi|} (|\xi| - u)^n \, du = \frac{|\xi|^{n+1}}{(n+1)!}.$$

We could also transform the expression for R_n by integrating it by parts:

$$R_n(\xi) = \tfrac{i^{n+1}}{n!} \left(\tfrac{1}{i} \xi^n - \tfrac{n}{i} \int_0^{\xi} e^{iu} (\xi-u)^{n-1} \, du \right)$$

$$= \tfrac{i^n}{(n-1)!} \left(\int_0^{\xi} (\xi-u)^{n-1} \, du - \int_0^{\xi} e^{iu} (\xi-u)^{n-1} \, du \right),$$

since $\xi^n = n \int_0^{\xi} (\xi - u)^{n-1} \, du$. Therefore

$$|R_n(\xi)| \le \tfrac{1}{(n-1)!} \int_0^{|\xi|} (|\xi| - u)^{n-1} \left| e^{iu} - 1 \right| du \le \tfrac{2}{n!} \int_0^{|\xi|} n(|\xi| - u)^{n-1} \, du = \tfrac{2|\xi|^n}{n!}. \qquad \blacksquare$$

While the following result can be obtained as a direct consequence of twice-differentiability of the function φ at 0, we use the (otherwise useful) estimate based on Lemma 8.6 above:

Corollary 8.7 (A two-finite-moment regularity estimate) *Let X be a random variable with $\mathbb{E}[X] = \mu$ and $\mathbb{E}[X^2] = \nu < \infty$, and let the function $r : [0, \infty) \to [0, \infty)$ be defined by*

(8.5) $$r(t) = \mathbb{E}[X^2 \min(t |X| , 1)], \ t \geq 0.$$

Then

1. *$\lim_{t \searrow 0} r(t) = 0$ and*

2. *$\left| \varphi_X(t) - 1 - it\mu + \frac{1}{2}\nu t^2 \right| \leq t^2 r(|t|).$*

PROOF The inequality in (2) is a direct consequence of Lemma 8.6 (with the extra factor $\frac{1}{6}$ neglected). As for (1), it follows from the dominated convergence theorem because

$$X^2 \min(1, t |X|) \leq X^2 \in \mathcal{L}^1 \text{ and } \lim_{t \to 0} X^2 \min(1, t |X|) = 0. \qquad \blacksquare$$

Theorem 8.8 (Central Limit Theorem - iid version) *Let $\{X_n\}_{n \in \mathbb{N}}$ be an iid sequence of random variables with $0 < \mathrm{Var}[X_1] < \infty$. Then*

$$\frac{\sum_{k=1}^{n}(X_k - \mu)}{\sqrt{\sigma^2 n}} \xrightarrow{\mathcal{D}} \chi,$$

where $\chi \sim N(0,1)$, $\mu = \mathbb{E}[X_1]$ and $\sigma^2 = \mathrm{Var}[X_1]$.

PROOF By considering the sequence $\{(X_n - \mu)/\sqrt{\sigma^2}\}_{n \in \mathbb{N}}$, instead of $\{X_n\}_{n \in \mathbb{N}}$, we may assume that $\mu = 0$ and $\sigma = 1$. Let φ be the characteristic function of the common distribution of $\{X_n\}_{n \in \mathbb{N}}$ and set $S_n = \sum_{k=1}^{n} X_k$, so that

$$\varphi_{\frac{1}{\sqrt{n}} S_n}(t) = (\varphi(\tfrac{t}{\sqrt{n}}))^n,$$

By Theorem 7.32, the problem reduces to whether the following statement holds:

(8.6) $$\lim_{n \to \infty} (\varphi(\tfrac{t}{\sqrt{n}}))^n = e^{-\frac{1}{2}t^2}, \text{ for each } t \in \mathbb{R}.$$

Corollary 8.7 guarantees that

$$\left| \varphi(t) - 1 + \tfrac{1}{2}t^2 \right| \leq t^2 r(t) \text{ for all } t \in \mathbb{R},$$

where r is given by (8.5), i.e.,

$$\left| \varphi(\tfrac{t}{\sqrt{n}}) - 1 + \tfrac{t^2}{2n} \right| \leq \tfrac{t^2}{n} r(t/\sqrt{n}).$$

109

Lemma 8.1 with $u_1 = \cdots = u_n = \varphi(\frac{t}{\sqrt{n}})$ and $w_1 = \cdots = w_n = (1 - \frac{t^2}{2n})$ yields:

$$\left|(\varphi(\tfrac{t}{\sqrt{n}}))^n - (1 - \tfrac{t^2}{2n})^n\right| \le t^2 r(t/\sqrt{n}),$$

for $n \ge \frac{2}{t^2}$ (so that $\max(|\varphi(\frac{t}{\sqrt{n}})|, |1 - \frac{t^2}{2n}|) \le 1$). Since $\lim_n r(t/\sqrt{n}) = 0$, we have

$$\lim_{n\to\infty} \left|(\varphi(\tfrac{t}{\sqrt{n}}))^n - (1 - \tfrac{t^2}{2n})^n\right| = 0,$$

∎

and (8.6) follows from the fact that $(1 - \frac{t^2}{2n})^n \to e^{-\frac{1}{2}t^2}$, for all t.

8.3 The Lindeberg-Feller Theorem

Unlike Theorem 8.8, the Lindeberg-Feller Theorem does not require summands to be equall distributed - it only asks for no single term to dominated the sum. As usual, we start with a technical lemma:

Lemma 8.9 ((*) Convergence to the exponential for triangular arrays) *Let* $(c_{n,m})$, $n \in \mathbb{N}$, $m = 1, \ldots, n$ *be a (triangular) array of real numbers with*

1. $\sum_{m=1}^{n} c_{n,m} \to c \in \mathbb{R}$, *and* $\sum_{m=1}^{n} |c_{n,m}|$ *is a bounded sequence,*

2. $m_n \to 0$, *as* $n \to \infty$, *where* $m_n = \max_{1 \le m \le n} |c_{n,m}| \to 0$

Then

$$\prod_{m=1}^{n} (1 + c_{n,m}) \to e^c \text{ as } n \to \infty.$$

PROOF Without loss of generality we assume that $m_n < \frac{1}{2}$ for all n, and note that the statement is equivalent to $\sum_{m=1}^{n} \log(1 + c_{n,m}) \to c$, as $n \to \infty$. Since $\sum_{m=1}^{n} c_{n,m} \to c$, this is also equivalent to

$$(8.7) \qquad \sum_{m=1}^{n} (\log(1 + c_{n,m}) - c_{n,m}) \to 0, \text{ as } n \to \infty.$$

Consider the function $f(x) = \log(1 + x) + x^2 - x$, $x > -1$. It is straightforward to check that $f(0) = 0$ and that the derivative $f'(x) = \frac{1}{1+x} + 2x - 1$ satisfies $f'(x) > 0$ for $x > 0$ and $f'(x) < 0$ for $x \in (-1/2, \infty)$. It follows that $f(x) \ge 0$ for $x \in [-1/2, \infty)$ so that (the absolute value can be inserted since $x \ge \log(1 + x)$)

$$|\log(1 + x) - x| \le x^2 \text{ for } x \ge -\tfrac{1}{2}.$$

Since $m_n < \frac{1}{2}$, we have $|\log(1 + c_{n,m}) - c_{n,m}| \le c_{n,m}^2$, and so

$$\left|\sum_{m=1}^{n} (\log(1 + c_{n,m}) - c_{n,m})\right| \le \sum_{m=1}^{n} |\log(1 + c_{n,m}) - c_{n,m}| \le \sum_{m=1}^{n} c_{n,m}^2$$

$$\le m_n \sum_{m=1}^{n} |c_{n,m}| \to 0,$$

because $\sum_{m=1}^{n} |c_{n,m}|$ is bounded and $m_n \to 0$.

∎

Theorem 8.10 (Lindeberg-Feller) *Let $X_{n,m}$, $n \in \mathbb{N}$, $m = 1, \ldots, n$ be a (triangular) array of random variables such that*

1. $\mathbb{E}[X_{n,m}] = 0$, *for all $n \in \mathbb{N}$, $m = 1, \ldots, n$,*

2. $X_{n,1}, \ldots, X_{n,n}$ *are independent,*

3. $\sum_{m=1}^{n} \mathbb{E}[X_{n,m}^2] \to \sigma^2 > 0$, *as $n \to \infty$,*

4. *for each $\varepsilon > 0$, $s_n(\varepsilon) \to 0$, as $n \to \infty$, where $s_n(\varepsilon) = \sum_{m=1}^{n} \mathbb{E}[X_{n,m}^2 \mathbf{1}_{\{|X_{n,m}| \geq \varepsilon\}}]$.*

Then
$$X_{n,1} + \cdots + X_{n,n} \overset{\mathcal{D}}{\to} \sigma \chi, \text{ as } n \to \infty,$$

where $\chi \sim N(0, 1)$.

PROOF **(*)** Set $\varphi_{n,m} = \varphi_{X_{n,m}}$, $\sigma_{n,m}^2 = \mathbb{E}[X_{n,m}^2]$. Just like in the proofs of Theorems 8.3 and 8.8, it will be enough to show that

$$\prod_{m=1}^{n} \varphi_{n,m}(t) \to e^{-\frac{1}{2}\sigma^2 t^2}, \text{ for all } t \in \mathbb{R}.$$

We fix $t \neq 0$ and use Lemma 8.1 with $u_{n,m} = \varphi_{n,m}(t)$ and $w_{n,m} = 1 - \frac{1}{2}\sigma_{n,m}^2 t^2$ to conclude that

$$D_n(t) \leq M_n^{n-1} \sum_{m=1}^{n} \left| \varphi_{n,m}(t) - 1 + \tfrac{1}{2}\sigma_{n,m}^2 t^2 \right|,$$

where

$$D_n(t) = \left| \prod_{m=1}^{n} \varphi_{n,m}(t) - \prod_{m=1}^{n} (1 - \tfrac{1}{2}\sigma_{n,m}^2 t^2) \right|$$

and $M_n = 1 \vee \max_{1 \leq m \leq n} (\left| 1 - \frac{1}{2}t^2\sigma_{n,m}^2 \right|)$. Assumption (4) in the statement implies that

$$\sigma_{n,m}^2 = \mathbb{E}[X_{n,m}^2 \mathbf{1}_{\{|X_{m,n}| \geq \varepsilon\}}] + \mathbb{E}[X_{n,m}^2 \mathbf{1}_{\{|X_{m,n}| < \varepsilon\}}] \leq \varepsilon^2 + \mathbb{E}[X_{n,m}^2 \mathbf{1}_{\{|X_{m,n}| < \varepsilon\}}]$$
$$\leq \varepsilon^2 + s_n(\varepsilon),$$

and so $\sup_{1 \leq m \leq n} \sigma_{n,m}^2 \to 0$, as $n \to \infty$. Therefore, for n large enough, we have $\frac{1}{2}t^2\sigma_{n,m}^2 \leq 2$ and $M_n = 1$.

According to Corollary 8.7 we now have (for large-enough n)

$$D_n(t) \leq t^2 \sum_{m=1}^{n} \mathbb{E}[X_{n,m}^2 \min(t \left| X_{n,m} \right|, 1)]$$

$$\leq t^2 \sum_{m=1}^{n} \left(\mathbb{E}[X_{n,m}^2 \mathbf{1}_{\{|X_{n,m}| \geq \varepsilon\}}] + \mathbb{E}[t \left| X_{n,m} \right|^3 \mathbf{1}_{\{|X_{n,m}| < \varepsilon\}}] \right)$$

$$\leq t^2 s_n(\varepsilon) + t^3 \varepsilon \sum_{m=1}^{n} \mathbb{E}[X_{n,m}^2 \mathbf{1}_{\{|X_{n,m}| < \varepsilon\}}] \leq t^2 s_n(\varepsilon) + 2t^3 \varepsilon \sigma^2.$$

Therefore, $\limsup_n D_n(t) \leq 2t^3 \varepsilon \sigma^2$, and so, since $\varepsilon > 0$ is arbitrary, we have $\lim_n D_n(t) = 0$.

Our last task is to remember that $\max_{1 \leq m \leq n} \sigma_{n,m}^2 \to 0$, note that $\sum_{m=1}^m \sigma_{n,m}^2 \to \sigma^2$ (why?), and use Lemma 8.9 to conclude that

$$\prod_{m=1}^n (1 - \tfrac{1}{2}\sigma_{n,m}^2 t^2) - e^{-\frac{1}{2}\sigma^2 t^2}.$$

∎

Problem 8.11 Show how the iid central limit theorem follows from the Lindeberg-Feller theorem.

Example 8.12 (Cycles in a random permutation) Let $\Pi : \Omega \to S_n$ be a random element taking values in the set S_n of all permutations of the set $\{1, \ldots, n\}$, i.e., the set of all bijections $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$. One usually considers the probability measure on Ω such that Π is uniformly distributed over S_n, i.e. $\mathbb{P}[\Pi = \pi] = \frac{1}{n!}$, for each $\pi \in S_n$. A random element in S_n whose distribution is uniform over S_n is called a **random permutation**.

Remember that each permutation $\pi \in S_n$ be decomposed into cycles; a **cycle** is a collection $(i_1 i_2 \ldots i_k)$ in $\{1, \ldots, n\}$ such that $\pi(i_l) = i_{l+1}$ for $l = 1, \ldots, k - 1$ and $\pi(i_k) = i_1$. For example, the permutation $\pi : \{1, 2, 3, 4\} \to \{1, 2, 3, 4\}$, given by $\pi(1) = 3$, $\pi(2) = 1$, $\pi(3) = 2$, $\pi(4) = 4$ has two cycles: (132) and (4). More precisely, start from $i_1 = 1$ and follow the sequence $i_{k+1} = \pi(i_k)$, until the first time you return to $i_k = 1$. Write these number in order $(i_1 i_2 \ldots i_k)$ and pick $j_1 \in \{1, 2, \ldots, n\} \setminus \{i_1, \ldots, i_k\}$. If no such j_1 exists, π consist of a single cycle. If it does, we repeat the same procedure starting from j_1 to obtain another cycle $(j_1 j_2 \ldots j_l)$, etc. In the end, we arrive at the decomposition

$$(i_1 i_2 \ldots i_k)(j_1 j_2 \ldots j_l) \ldots$$

of π into cycles.

Let us first answer the following, warm-up, question: what is the probability $p(n, m)$ that 1 is a member of a cycle of length m? Equivalently, we can ask for the number $c(n, m)$ of permutations in which 1 is a member of a cycle of length m. The easiest way to solve this is to note that each such permutation corresponds to a choice of $(m - 1)$ district numbers of $\{2, 3, \ldots\}$ - these will serve as the remaining elements of the cycle containing 1. This can be done in $\binom{n-1}{m-1}$ ways. Furthermore, the $m - 1$ elements to be in the same cycle with 1 can be ordered in $(m - 1)!$ ways. Also, the remaining $n - m$ elements give rise to $(n - m)!$ distinct permutations. Therefore,

$$c(n, m) = \binom{n-1}{m-1}(m-1)!(n-m)! = (n-1)!, \text{ and so } p(n, m) = \tfrac{1}{n}.$$

This is a remarkable result - all cycle lengths are equally likely. Note, also, that 1 is not special in any way.

Our next goal is to say something about the number of cycles - a more difficult task. We start by describing a procedure for producing a random permutation by building it from cycles. The reader will easily convince his-/herself that the outcome is uniformly distributed over all permutations. We start with $n - 1$ independent random variables $\xi_2, \ldots, \xi_n$ such that ξ_i is uniformly distributed over the set $\{0, 1, 2, \ldots, n - i + 1\}$. Let the first cycle start from $X_1 = 1$. If $\xi_2 = 0$, then we declare (1) to be a full cycle and start building the next cycle from 2. If $\xi_2 \neq 0$, we pick the ξ_2-th smallest element - let us call it X_2 - from the set of remaining $n - 1$ numbers to be the second element in the first cycle. After that, we close the cycle if $\xi_3 = 0$, or append the ξ_3-th smallest element - let's call it X_3 - in $\{1, 2, \ldots, n\} \setminus \{X_1, X_2\}$ to the cycle. Once the cycle $(X_1 \ X_2 \ \ldots \ X_k)$ is closed, we pick the smallest element in $\{1, 2, \ldots, n\} \setminus \{X_1, X_2, \ldots, X_k\}$ - let's call it X_{k+1} - and repeat the procedure starting from X_{k+1} and using $\xi_{k+1}, \ldots, \xi_n$ as "sources of randomness".

Let us now define the random variables (we stress the dependence on n here) $Y_{n,1}, \ldots, Y_{n,n}$ by $Y_{n,k} = \mathbf{1}_{\{\xi_k=0\}}$. In words, $Y_{n,k}$ is in indicator of the event when a cycle ends right after the position k. It is clear that $Y_{n,1}, \ldots, Y_{n,k}$ are independent (they are functions of independent variables $\xi_1, \ldots, \xi_n$). Also, $p(n,k) = \mathbb{P}[Y_{n,k} = 1] = \frac{1}{n-k+1}$. The number of cycles C_n is the same as the number of closing parenthesis, so $C_n = \sum_{k=1}^n Y_{k,n}$. (Btw, can you derive the identity $p(n,m) = \frac{1}{n}$ by using random variables $Y_{n,1}, \ldots, Y_{n,n}$?)

It is easy to compute

$$\mathbb{E}[C_n] = \sum_{k=1}^n \mathbb{E}[Y_{n,k} = 1] = \sum_{k=1}^n \tfrac{1}{n-k+1} = 1 + \tfrac{1}{2} + \cdots + \tfrac{1}{n} = \log(n) + \gamma + o(1),$$

where $\gamma \approx 0.58$ is the Euler-Mascheroni constant, and $a_n = b_n + o(n)$ means that $|b_n - a_n| \to 0$, as $n \to \infty$.

If we want to know more about the variability of C_n, we can also compute its variance:

$$\mathrm{Var}[C_n] = \sum_{k=1}^n \mathrm{Var}[Y_{n,k}] = \sum_{k=1}^n (\tfrac{1}{n-k+1} - \tfrac{1}{(n-k+1)^2}) = \log(n) + \gamma - \tfrac{\pi^2}{6} + o(1).$$

The Lindeberg-Feller theorem will give us the precise asymptotic behavior of C_n. For $m = 1, \ldots, n$, we define

$$X_{n,m} = \frac{Y_{n,m} - \mathbb{E}[Y_{n,m}]}{\sqrt{\log(n)}},$$

so that $X_{n,m}$, $m = 1, \ldots, n$ are independent and of mean 0. Furthermore, we have

$$\lim_n \sum_{m=1}^n \mathbb{E}[X_{n,m}^2] = \lim_n \frac{\log(n) + \gamma - \frac{\pi^2}{6} + o(1)}{\log(n)} = 1.$$

Finally, for $\varepsilon > 0$ and $\log(n) > 2/\varepsilon$, we have $\mathbb{P}[|X_{n,m}| > \varepsilon] = 0$, so

$$\sum_{m=1}^n \mathbb{E}[X_{n,m}^2 \mathbf{1}_{\{|X_{n,m}| \geq \varepsilon\}}] = 0.$$

Having checked that all the assumption of the Lindeberg-Feller theorem are satisfied, we conclude that

$$\tfrac{C_n - \log(n)}{\sqrt{\log(n)}} \xrightarrow{\mathcal{D}} \chi, \text{ where } \chi \sim N(0,1).$$

It follows that (if we believe that the approximation is good) the number of cycles in a random permutation with $n = 8100$ is at most 18 with probability 99%.

How about variability? Here is histogram of the number of cycles from 1000 simulations for $n = 8100$, together with the appropriately-scaled density of the normal distribution with mean $\log(8100)$ and standard deviation $\sqrt{\log(8100)}$. The quality of approximation leaves something to

be desired, but it seems to already work well in the tails: only 3 of 1000 had more than 17 cycles:
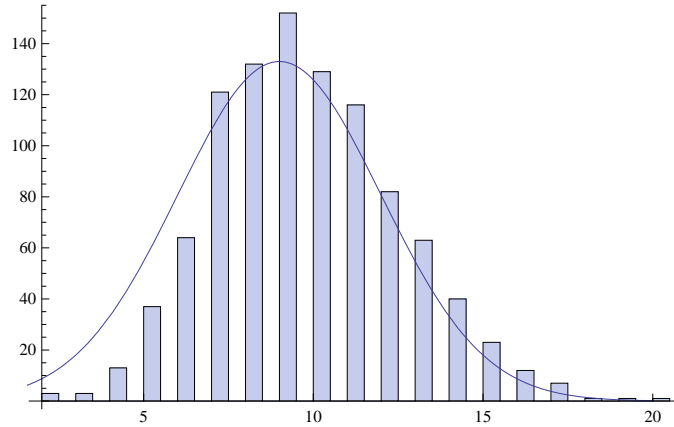
8.4 Additional Problems

Problem 8.13 (Lyapunov's theorem) Let $\{X_n\}_{n\in\mathbb{N}}$ be an independent sequence, let $S_n = \sum_{m=1}^{n} X_m$, and let $\alpha_n = \sqrt{\operatorname{Var}[S_n]}$. Suppose that $\alpha_n > 0$ for all $n \in \mathbb{N}$ and that there exists a constant $\delta > 0$ such that

$$\lim_n \alpha_n^{-(2+\delta)} \sum_{m=1}^{n} \mathbb{E}[|X_m - \mathbb{E}[X_m]|^{2+\delta}] = 0.$$

Show that

$$\frac{S_n - \mathbb{E}[S_n]}{\alpha_n} \xrightarrow{\mathcal{D}} \chi, \text{ where } \chi \sim N(0,1).$$

Problem 8.14 (Self-normalized sums) Let $\{X_n\}_{n\in\mathbb{N}}$ be iid random variables with $\mathbb{E}[X_1] = 0$, $\sigma = \sqrt{\mathbb{E}[X_1^2]} > 0$ and $\mathbb{P}[X_1 = 0] = 0$. Show that the sequence $\{Y_n\}_{n\in\mathbb{N}}$ given by

$$Y_n = \frac{\sum_{k=1}^{n} X_k}{\sqrt{\sum_{k=1}^{n} X_k^2}},$$

converges in distribution, and identify its limit.

(*Hint:* Use Slutsky's theorem of Problem 7.42; you don't have to prove it.)

Chapter 9

Conditional Expectation

9.1 The definition and existence of conditional expectation

For events A, B with $\mathbb{P}[B] > 0$, we recall the familiar object

$$\mathbb{P}[A|B] = \tfrac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

We say that $\mathbb{P}[A|B]$ the **conditional probability of** A**, given** B. It is important to note that the condition $\mathbb{P}[B] > 0$ is crucial. When X and Y are random variables defined on the same probability space, we often want to give a meaning to the expression $\mathbb{P}[X \in A|Y = y]$, even though it is usually the case that $\mathbb{P}[Y = y] = 0$. When the random vector (X, Y) admits a joint density $f_{X,Y}(x, y)$, and $f_Y(y) > 0$, the concept of conditional density $f_{X|Y=y}(x) = f_{X,Y}(x, y)/f_Y(y)$ is introduced and the quantity $\mathbb{P}[X \in A|Y = y]$ is given meaning via $\int_A f_{X|Y=y}(x, y) \, dx$. While this procedure works well in the restrictive case of absolutely continuous random vectors, we will see how it is encompassed by a general concept of a conditional expectation. Since probability is simply an expectation of an indicator, and expectations are linear, it will be easier to work with expectations and no generality will be lost.

Two main conceptual leaps here are: 1) we condition with respect to a σ-algebra, and 2) we view the conditional expectation itself as a random variable. Before we illustrate the concept in discrete time, here is the definition.

Definition 9.1 (Conditional Expectation) *Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$, and let $X \in \mathcal{L}^1$ be a random variable. We say that the random variable ξ is (a version of) the* **conditional expectation of** X **with respect to** $\mathcal{G}$ *- and denote it by $\mathbb{E}[X|\mathcal{G}]$ - if*

1. *$\xi \in \mathcal{L}^1$.*

2. *ξ is $\mathcal{G}$-measurable,*

3. *$\mathbb{E}[\xi \mathbf{1}_A] = \mathbb{E}[X \mathbf{1}_A]$, for all $A \in \mathcal{G}$.*

Example 9.2 Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space where $\Omega = \{a, b, c, d, e, f\}$, $\mathcal{F} = 2^\Omega$ and $\mathbb{P}$ is uniform. Let X, Y and Z be random variables given by (in the obvious notation)

$$X \sim \begin{pmatrix} a & b & c & d & e & f \\ 1 & 3 & 3 & 5 & 5 & 7 \end{pmatrix}, \ Y \sim \begin{pmatrix} a & b & c & d & e & f \\ 2 & 2 & 1 & 1 & 7 & 7 \end{pmatrix} \text{ and } Z \sim \begin{pmatrix} a & b & c & d & e & f \\ 3 & 3 & 3 & 3 & 2 & 2 \end{pmatrix}$$

We would like to think about $\mathbb{E}[X|\mathcal{G}]$ as the average of $X(\omega)$ over all ω which are consistent with the our current information (which is $\mathcal{G}$). For example, if $\mathcal{G} = \sigma(Y)$, then the information contained in $\mathcal{G}$ is exactly the information about the exact value of Y. Knowledge of the fact that $Y = y$ does not necessarily reveal the "true" ω, but certainly rules out all those ω for which $Y(\omega) \neq y$.
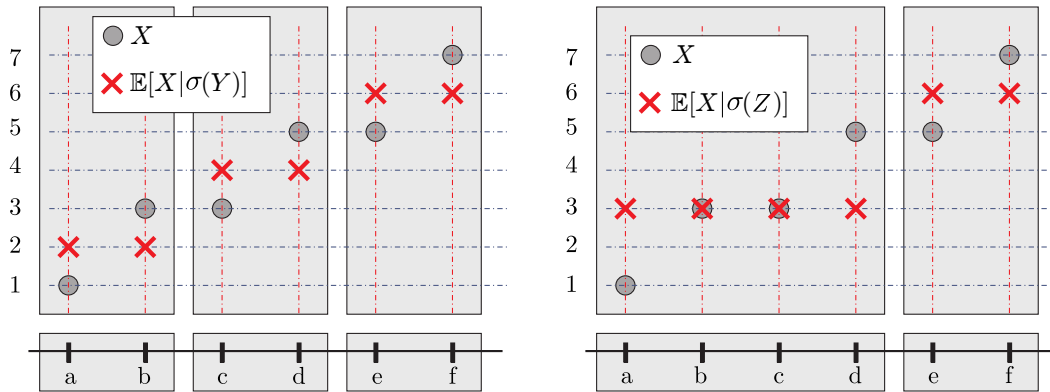
In our specific case, if we know that $Y = 2$, then $\omega = a$ or $\omega = b$, and the expected value of X, given that $Y = 2$, is $\frac{1}{2}X(a) + \frac{1}{2}X(b) = 2$. Similarly, this average equals 4 for $Y = 1$, and 6 for $Y = 7$. Let us show that the random variable ξ defined by this average, i.e.,

$$\xi \sim \begin{pmatrix} a & b & c & d & e & f \\ 2 & 2 & 4 & 4 & 6 & 6 \end{pmatrix},$$

satisfies the definition of $\mathbb{E}[X|\sigma(Y)]$, as given above. The integrability is not an issue (we are on a finite probability space), and it is clear that ξ is measurable with respect to $\sigma(Y)$. Indeed, the atoms of $\sigma(Y)$ are $\{a, b\}$, $\{c, d\}$ and $\{e, f\}$, and ξ is constant over each one of those. Finally, we need to check that

$$\mathbb{E}[\xi \mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A], \text{ for all } A \in \sigma(Y),$$

which for an atom A translates into

$$\xi(\omega) = \tfrac{1}{\mathbb{P}[A]}\mathbb{E}[X\mathbf{1}_A] = \sum_{\omega' \in A} X(\omega')\mathbb{P}[\{\omega'\}|A], \text{ for all } \omega \in A.$$

The moral of the story is that when A is an atom, part (3) of Definition 9.1 translates into a requirement that ξ be constant on A with value equal to the expectation of X over A with respect to the conditional probability $\mathbb{P}[\cdot|A]$. In the general case, when there are no atoms, (3) still makes sense and conveys the same message.

Btw, since the atoms of $\sigma(Z)$ are $\{a, b, c, d\}$ and $\{e, f\}$, it is clear that

$$\mathbb{E}[X|\sigma(Z)](\omega) = \begin{cases} 3, & \omega \in \{a, b, c, d\}, \\ 6, & \omega \in \{e, f\}. \end{cases}$$

Look at the illustrations above and convince yourself that

$$\mathbb{E}[\mathbb{E}[X|\sigma(Y)]|\sigma(Z)] = \mathbb{E}[X|\sigma(Z)].$$

A general result along the same lines - called the *tower property of conditional expectation* - will be stated and proved below.

Our first task is to prove that conditional expectations always exist. When Ω is finite (as explained above) or countable, we can always construct them by averaging over atoms. In the general case, a different argument is needed. In fact, here are two:

Proposition 9.3 (Conditional expectation - existence and a.s.-uniqueness) *Let $\mathcal{G}$ be a sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$. Then*

1. *there exists a conditional expectation $\mathbb{E}[X|\mathcal{G}]$ for any $X \in \mathcal{L}^1$, and*

2. *any two conditional expectations of $X \in \mathcal{L}^1$ are equal $\mathbb{P}$-a.s.*

PROOF (*Uniqueness*): Suppose that ξ and ξ' both satisfy (1),(2) and (3) of Definition 9.1. Then

$$\mathbb{E}[\xi \mathbf{1}_A] = \mathbb{E}[\xi' \mathbf{1}_A], \text{ for all } A \in \mathcal{G}.$$

For $A_n = \{\xi' - \xi \geq \frac{1}{n}\}$, we have $A_n \in \mathcal{G}$ and so

$$\mathbb{E}[\xi \mathbf{1}_{A_n}] = \mathbb{E}[\xi' \mathbf{1}_{A_n}] \geq \mathbb{E}[(\xi + \frac{1}{n})\mathbf{1}_{A_n}] = \mathbb{E}[\xi \mathbf{1}_{A_n}] + \frac{1}{n}\mathbb{P}[A_n].$$

Consequently, $\mathbb{P}[A_n] = 0$, for all $n \in \mathbb{N}$, so that $\mathbb{P}[\xi' > \xi] = 0$. By a symmetric argument, we also have $\mathbb{P}[\xi' < \xi] = 0$.

(*Existence*): By linearity, it will be enough to prove that the conditional expectation exists for $X \in \mathcal{L}^1_+$.

1. *A Radon-Nikodym argument*. Suppose, first, that $X \geq 0$ and $\mathbb{E}[X] = 1$, as the general case follows by additivity and scaling. Then the prescription

$$\mathbb{Q}[A] = \mathbb{E}[X \mathbf{1}_A],$$

defines a probability measure on $(\Omega, \mathcal{F})$, which is absolutely continuous with respect to $\mathbb{P}$. Let $\mathbb{Q}^{\mathcal{G}}$ be the restriction of $\mathbb{Q}$ to $\mathcal{G}$; it is trivially absolutely continuous with respect to the restriction $\mathbb{P}^{\mathcal{G}}$ of $\mathbb{P}$ to $\mathcal{G}$. The Radon-Nikodym theorem - applied to the measure space $(\Omega, \mathcal{G}, \mathbb{P}^{\mathcal{G}})$ and the measure $\mathbb{Q}^{\mathcal{G}} \ll \mathbb{P}^{\mathcal{G}}$ - guarantees the existence of the Radon-Nikodym derivative

$$\xi = \frac{d\mathbb{Q}^{\mathcal{G}}}{d\mathbb{P}^{\mathcal{G}}} \in \mathbb{L}^1_+(\Omega, \mathcal{G}, \mathbb{P}^{\mathcal{G}}).$$

For $A \in \mathcal{G}$, we thus have

$$\mathbb{E}[X \mathbf{1}_A] = \mathbb{Q}[A] = \mathbb{Q}^{\mathcal{G}}[A] = \mathbb{E}^{\mathbb{P}^{\mathcal{G}}}[\xi \mathbf{1}_A] = \mathbb{E}[\xi \mathbf{1}_A].$$

where the last equality follows from the fact that $\xi \mathbf{1}_A$ is $\mathcal{G}$-measurable. Therefore, ξ is (a version of) the conditional expectation $\mathbb{E}[X|\mathcal{G}]$.

1. *An $\mathcal{L}^2$-argument.* Suppose, first, that $X \in \mathcal{L}^2$. Let H be the family of all $\mathcal{G}$-measurable elements in $\mathcal{L}^2$. Let $\bar{H}$ denote the closure of H in the topology induced by $\mathcal{L}^2$-convergence. Being a closed and convex (why?) subset of $\mathcal{L}^2$, $\bar{H}$ satisfies all the conditions of Problem 4.26 so that there exists $\xi \in \bar{H}$ at the minimal $\mathcal{L}^2$-distance from X (when $X \in \bar{H}$, we take $\xi = X$). The same problem states that ξ has the following property:

$$\mathbb{E}[(\eta - \xi)(X - \xi)] \geq 0 \text{ for all } \eta \in \bar{H},$$

and, since $\bar{H}$ is a linear space, we have

$$\mathbb{E}[(\eta - \xi)(X - \xi)] = 0, \text{ for all } \eta \in \bar{H}.$$

It remains to pick η of the form $\eta = \xi + \mathbf{1}_A \in \bar{H}$, $A \in \mathcal{G}$, to conclude that

$$\mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[\xi\mathbf{1}_A], \text{ for all } A \in \mathcal{G}.$$

Our next step is to show that ξ is $\mathcal{G}$-measurable (after a modification on a null set, perhaps). Since $\xi \in \bar{H}$, there exists a sequence $\{\xi_n\}_{n\in\mathbb{N}}$ such that $\xi_n \to \xi$ in $\mathcal{L}^2$. By Corollary 4.20, $\xi_{n_k} \overset{a.s.}{\to} \xi$, for some subsequence $\{\xi_{n_k}\}_{k\in\mathbb{N}}$ of $\{\xi_n\}_{n\in\mathbb{N}}$. Set $\xi' = \liminf_{k\in\mathbb{N}} \xi_{n_k} \in \mathcal{L}^0([-\infty, \infty], \mathcal{G})$ and $\hat{\xi} = \xi'\mathbf{1}_{\{|\xi'|<\infty\}}$, so that $\hat{\xi} = \xi$, a.s., and $\hat{\xi}$ is $\mathcal{G}$-measurable.

We still need to remove the restriction $X \in \mathcal{L}^2_+$. We start with a general $X \in \mathcal{L}^1_+$ and define $X_n = \min(X, n) \in \mathcal{L}^\infty_+ \subseteq \mathcal{L}^2_+$. Let $\xi_n = \mathbb{E}[X_n|\mathcal{G}]$, and note that $\mathbb{E}[\xi_{n+1}\mathbf{1}_A] = \mathbb{E}[X_{n+1}\mathbf{1}_A] \geq \mathbb{E}[X_n\mathbf{1}_A] = \mathbb{E}[\xi_n\mathbf{1}_A]$. It follows (just like in the proof of uniqueness above) that $\xi_n \leq \xi_{n+1}$, a.s. We define $\xi = \sup_n \xi_n$, so that $\xi_n \nearrow \xi$, a.s. Then, for $A \in \mathcal{G}$, the monotone-convergence theorem implies that

$$\mathbb{E}[X\mathbf{1}_A] = \lim_n \mathbb{E}[X_n\mathbf{1}_A] = \lim_n \mathbb{E}[\xi_n\mathbf{1}_A] = \mathbb{E}[\xi\mathbf{1}_A],$$

and it is easy to check that $\xi\mathbf{1}_{\{\xi<\infty\}} \in \mathcal{L}^1(\mathcal{G})$ is a version of $\mathbb{E}[X|\mathcal{G}]$. ∎

Remark 9.4 There is no canonical way to choose "the version" of the conditional expectation. We follow the convention started with Radon-Nikodym derivatives, and interpret a statement such at $\xi \leq \mathbb{E}[X|\mathcal{G}]$, a.s., to mean that $\xi \leq \xi'$, a.s., for any version ξ' of the conditional expectation of X with respect to $\mathcal{G}$.

If we use the symbol $\mathbb{L}^1$ to denote the set of all a.s.-equivalence classes of random variables in $\mathcal{L}^1$, we can write:

$$\mathbb{E}[\cdot|\mathcal{G}] : \mathcal{L}^1(\mathcal{F}) \to \mathbb{L}^1(\mathcal{G}),$$

but $\mathbb{L}^1(\mathcal{G})$ cannot be replaced by $\mathcal{L}^1(\mathcal{G})$ in a natural way. Since $X = X'$, a.s., implies that $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X'|\mathcal{G}]$, a.s. (why?), we consider conditional expectation as a map from $\mathbb{L}^1(\mathcal{F})$ to $\mathbb{L}^1(\mathcal{G})$

$$\mathbb{E}[\cdot|\mathcal{G}] : \mathbb{L}^1(\mathcal{F}) \to \mathbb{L}^1(\mathcal{G}).$$

9.2 Properties

Conditional expectation inherits many of the properties from the "ordinary" expectation. Here are some familiar and some new ones:

Proposition 9.5 (Properties of the conditional expectation) *Let $X, Y, \{X_n\}_{n\in\mathbb{N}}$ be random variables in $\mathcal{L}^1$, and let $\mathcal{G}$ and $\mathcal{H}$ be sub-σ-algebras of $\mathcal{F}$. Then*

1. (linearity) $\mathbb{E}[\alpha X + \beta Y | \mathcal{G}] = \alpha \mathbb{E}[X | \mathcal{G}] + \beta \mathbb{E}[Y | \mathcal{G}]$, *a.s.*

2. (monotonicity) $X \leq Y$, *a.s., implies* $\mathbb{E}[X | \mathcal{G}] \leq \mathbb{E}[Y | \mathcal{G}]$, *a.s.*

3. (identity on $\mathbb{L}^1(\mathcal{G})$) *If X is $\mathcal{G}$-measurable, then* $X = \mathbb{E}[X | \mathcal{G}]$, *a.s. In particular,* $c = \mathbb{E}[c | \mathcal{G}]$, *for any constant $c \in \mathbb{R}$.*

4. (conditional Jensen's inequality) *If $\psi : \mathbb{R} \to \mathbb{R}$ is convex and $\mathbb{E}[|\psi(X)|] < \infty$ then*

$$\mathbb{E}[\psi(X) | \mathcal{G}] \geq \psi(\mathbb{E}[X | \mathcal{G}]), \ \ a.s.$$

5. ($\mathcal{L}^p$-nonexpansivity) *If $X \in \mathcal{L}^p$, for $p \in [1, \infty]$, then $\mathbb{E}[X | \mathcal{G}] \in \mathbb{L}^p$ and*

$$||\mathbb{E}[X | \mathcal{G}]||_{\mathcal{L}^p} \leq ||X||_{\mathcal{L}^p}.$$

   In particular,
$$\mathbb{E}[|X| \ | \mathcal{G}] \geq |\mathbb{E}[X | \mathcal{G}]| \ \ a.s.$$

6. (pulling out what's known) *If Y is $\mathcal{G}$-measurable and $XY \in \mathcal{L}^1$, then*

$$\mathbb{E}[XY | \mathcal{G}] = Y \mathbb{E}[X | \mathcal{G}], \ \ a.s.$$

7. ($\mathbb{L}^2$-projection) *If $X \in \mathcal{L}^2$, then $\xi^* = \mathbb{E}[X | \mathcal{G}]$ minimizes $\mathbb{E}[(X - \xi)^2]$ over all $\mathcal{G}$-measurable random variables $\xi \in \mathcal{L}^2$.*

8. (tower property) *If $\mathcal{H} \subseteq \mathcal{G}$, then*

$$\mathbb{E}[\mathbb{E}[X | \mathcal{G}] | \mathcal{H}] = \mathbb{E}[X | \mathcal{H}], \ \ a.s..$$

9. (irrelevance of independent information) *If $\mathcal{H}$ is independent of $\sigma(\mathcal{G}, \sigma(X))$ then*

$$\mathbb{E}[X | \sigma(\mathcal{G}, \mathcal{H})] = \mathbb{E}[X | \mathcal{G}], \ \ a.s.$$

   In particular, if X is independent of $\mathcal{H}$, then $\mathbb{E}[X | \mathcal{H}] = \mathbb{E}[X]$, a.s.

10. (conditional monotone-convergence theorem) *If $0 \leq X_n \leq X_{n+1}$, a.s., for all $n \in \mathbb{N}$ and $X_n \to X \in \mathcal{L}^1$, a.s., then*
$$\mathbb{E}[X_n | \mathcal{G}] \nearrow \mathbb{E}[X | \mathcal{G}], \ \ a.s.$$

11. (conditional Fatou's lemma) *If $X_n \geq 0$, a.s., for all $n \in \mathbb{N}$, and $\liminf_n X_n \in \mathcal{L}^1$, then*

$$\mathbb{E}[\liminf_n X | \mathcal{G}] \leq \liminf_n \mathbb{E}[X_n | \mathcal{G}], \ \ a.s.$$

12. (conditional dominated-convergence theorem) *If $|X_n| \leq Z$, for all $n \in \mathbb{N}$ and some $Z \in \mathcal{L}^1$, and if $X_n \to X$, a.s., then*

$$\mathbb{E}[X_n | \mathcal{G}] \to \mathbb{E}[X | \mathcal{G}], \ \ a.s. \ and \ in \ \mathcal{L}^1.$$

119

PROOF Only some of the properties are proved in detail. The others are only commented upon, since they are either similar to the other ones or otherwise not hard.

1. (linearity) $\mathbb{E}[(\alpha X + \beta Y)\mathbf{1}_A] = \mathbb{E}[(\alpha\mathbb{E}[X|\mathcal{G}] + \beta\mathbb{E}[Y|\mathcal{G}])\mathbf{1}_A]$, for $A \in \mathcal{G}$.

2. (monotonicity) Use $A = \{\mathbb{E}[X|\mathcal{G}] > \mathbb{E}[Y|\mathcal{G}]\} \in \mathcal{G}$ to obtain a contradiction if $\mathbb{P}[A] > 0$.

3. (identity on $\mathbb{L}^1(\mathcal{G})$) Check the definition.

4. (conditional Jensen's inequality) Use the result of Lemma 4.22 which states that $\psi(x) = \sup_{n\in\mathbb{N}}(a_n + b_n x)$, where $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ are sequences of real numbers.

5. ($\mathbb{L}^p$-nonexpansivity) For $p \in [1, \infty)$, apply conditional Jensen's inequality with $\psi(x) = |x|^p$. The case $p = \infty$ follows directly.

6. (pulling out what's known) For Y $\mathcal{G}$-measurable and $XY \in \mathcal{L}^1$, we need to show that

   $$(9.1) \qquad \mathbb{E}[XY\mathbf{1}_A] = \mathbb{E}[Y\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A], \text{ for all } A \in \mathcal{G}.$$

   Let us prove a seemingly less general statement:

   $$(9.2) \qquad \mathbb{E}[ZX] = \mathbb{E}[Z\mathbb{E}[X|\mathcal{G}]], \text{ for all } \mathcal{G}\text{-measurable } Z \text{ with } ZX \in \mathcal{L}^1.$$

   The statement (9.1) will follow from it by taking $Z = Y\mathbf{1}_A$. For $Z = \sum_{k=1}^{n} \alpha_k \mathbf{1}_{A_k}$, (9.2) is a consequence of the definition of conditional expectation and linearity. Let us assume that both Z and X are nonnegative and $ZX \in \mathcal{L}^1$. In that case we can find a non-decreasing sequence $\{Z_n\}_{n\in\mathbb{N}}$ of non-negative simple random variables with $Z_n \nearrow Z$. Then $Z_n X \in \mathcal{L}^1$ for all $n \in \mathbb{N}$ and the monotone convergence theorem implies that

   $$\mathbb{E}[ZX] = \lim_n \mathbb{E}[Z_n X] = \lim_n \mathbb{E}[Z_n\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[Z\mathbb{E}[X|\mathcal{G}]].$$

   Our next task is to relax the assumption $X \in \mathcal{L}^1_+$ to the original one $X \in \mathcal{L}^1$. In that case, the $\mathcal{L}^p$-nonexpansivity for $p = 1$ implies that

   $$|\mathbb{E}[X|\mathcal{G}]| \leq \mathbb{E}[|X| \, |\mathcal{G}] \text{ a.s., and so } |Z_n\mathbb{E}[X|\mathcal{G}]| \leq Z_n\mathbb{E}[|X| \, |\mathcal{G}] \leq Z\mathbb{E}[|X| \, |\mathcal{G}].$$

   We know from the previous case that

   $$\mathbb{E}[Z\mathbb{E}[|X| \, |\mathcal{G}]] = \mathbb{E}[Z \, |X|], \text{ so that } Z\mathbb{E}[|X| \, |\mathcal{G}] \in \mathcal{L}^1.$$

   We can, therefore, use the dominated convergence theorem to conclude that

   $$\mathbb{E}[Z\mathbb{E}[X|\mathcal{G}]] = \lim_n \mathbb{E}[Z_n\mathbb{E}[X|\mathcal{G}]] = \lim_n \mathbb{E}[Z_n X] = \mathbb{E}[ZX].$$

   Finally, the case of a general Z follows by linearity.

7. ($\mathbb{L}^2$-projection) It is enough to show that $X - \mathbb{E}[X|\mathcal{G}]$ is orthogonal to all $\mathcal{G}$-measurable $\xi \in \mathcal{L}^2$. for that we simply note that for $\xi \in \mathcal{L}^2$, $x \in \mathcal{G}$, we have

   $$\mathbb{E}[(X - \mathbb{E}[X|\mathcal{G}])\xi] = \mathbb{E}[\xi X] - \mathbb{E}[\xi\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[\xi X] - \mathbb{E}[\mathbb{E}[\xi X|\mathcal{G}]] = 0.$$

8. (tower property) Use the definition.

9. (irrelevance of independent information) We assume $X \geq 0$ and show that

(9.3) $$\mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A], \text{ a.s. for all } A \in \sigma(\mathcal{G}, \mathcal{H}).$$

Let $\mathcal{L}$ be the collection of all $A \in \sigma(\mathcal{G}, \mathcal{H})$ such that (9.3) holds. It is straightforward that $\mathcal{L}$ is a λ-system, so it will be enough to establish (9.3) for some π-system that generates $\sigma(\mathcal{G}, \mathcal{H})$. One possibility is $\mathcal{P} = \{G \cap H \: : \: G \in \mathcal{G}, H \in \mathcal{H}\}$, and for $G \cap H \in \mathcal{P}$ we use independence of $\mathbf{1}_H$ and $\mathbb{E}[X|\mathcal{G}]\mathbf{1}_G$, as well as the independence of $\mathbf{1}_H$ and $X\mathbf{1}_G$ to get

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbf{1}_{G \cap H}] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbf{1}_G \mathbf{1}_H] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbf{1}_G]\mathbb{E}[\mathbf{1}_H] = \mathbb{E}[X\mathbf{1}_G]\mathbb{E}[\mathbf{1}_H]$$
$$= \mathbb{E}[X\mathbf{1}_{G \cap H}]$$

10. (conditional monotone-convergence theorem) By monotonicity, $\mathbb{E}[X_n|\mathcal{G}] \nearrow \xi \in \mathcal{L}^0_+(\mathcal{G})$, a.s. The monotone convergence theorem implies that

$$\mathbb{E}[\xi\mathbf{1}_A] = \lim_n \mathbb{E}[\mathbf{1}_A\mathbb{E}[X_n|\mathcal{G}]] = \lim_n \mathbb{E}[\mathbf{1}_A X_n] = \mathbb{E}[\mathbf{1}_A X], \text{ for all } A \in \mathcal{G}.$$

11. (conditional Fatou's lemma) Set $Y_n = \inf_{k \geq n} X_k$, so that $Y_n \nearrow Y = \liminf_k X_k$. By monotonicity,

$$\mathbb{E}[Y_n|\mathcal{G}] \leq \inf_{k \geq n} \mathbb{E}[X_k|\mathcal{G}], \text{ a.s.,}$$

and the conditional monotone-convergence theorem implies that

$$\mathbb{E}[Y|\mathcal{G}] = \lim_{n \in \mathbb{N}} \mathbb{E}[Y_n|\mathcal{G}] \leq \liminf_n \mathbb{E}[X_n|\mathcal{G}], \text{ a.s.}$$

12. (conditional dominated-convergence theorem) By the conditional Fatou's lemma, we have

$$\mathbb{E}[Z + X|\mathcal{G}] \leq \liminf_n \mathbb{E}[Z + X_n|\mathcal{G}], \text{ and } \mathbb{E}[Z - X|\mathcal{G}] \leq \liminf_n \mathbb{E}[Z - X_n|\mathcal{G}], \text{ a.s.,}$$

and the a.s.-statement follows. ∎

Problem 9.6

1. Show that the condition $\mathcal{H} \subseteq \mathcal{G}$ is necessary for the tower property to hold in general. (*Hint:* Take $\Omega = \{a, b, c\}$.)

2. For $X, Y \in \mathcal{L}^2$ and a sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$, show that the following self-adjointness property holds

$$\mathbb{E}[X\,\mathbb{E}[Y|\mathcal{G}]] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\,Y] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\,\mathbb{E}[Y|\mathcal{G}]].$$

3. Let $\mathcal{H}$ and $\mathcal{G}$ be two sub-σ-algebras of $\mathcal{F}$. Is it true that

$$\mathcal{H} = \mathcal{G} \text{ if and only if } \mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X|\mathcal{H}], \text{ a.s., for all } X \in \mathcal{L}^1?$$

4. Construct two random variables X and Y in $\mathcal{L}^1$ such that $\mathbb{E}[X|\sigma(Y)] = \mathbb{E}[X]$, a.s., but X and Y are not independent.

9.3   Regular conditional distributions

Once we have a the notion of conditional expectation defined and analyzed, we can use it to define other, related, conditional quantities. The most important of those is the conditional probability:

Definition 9.7 (Conditional probability) *Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. The **conditional probability** of $A \in \mathcal{F}$, given $\mathcal{G}$ - denoted by $\mathbb{P}[A|\mathcal{G}]$ - is defined by*

$$\mathbb{P}[A|\mathcal{G}] = \mathbb{E}[\mathbf{1}_A|\mathcal{G}].$$

It is clear (from the conditional version of the monotone-convergence theorem) that

(9.4)
$$\mathbb{P}[\cup_{n\in\mathbb{N}} A_n|\mathcal{G}] = \sum_{n\in\mathbb{N}} \mathbb{P}[A_n|\mathcal{G}], \text{ a.s.}$$

We can, therefore, think of the conditional probability as a countably-additive map from events to (equivalence classes of) random variables $A \mapsto \mathbb{P}[A|\mathcal{G}]$. In fact, this map has the structure of a vector measure:

Definition 9.8 (Vector Measures) *Let $(B, ||\cdot||)$ be a Banach space, and let $(S, \mathcal{S})$ be a measurable space. A map $\mu : \mathcal{S} \to B$ is called a **vector measure** if*

1. *$\mu(\emptyset) = 0$, and*

2. *for each pairwise disjoint sequence $\{A_n\}_{n\in\mathbb{N}}$ in $\mathcal{S}$, $\mu(\cup_n A_n) = \sum_{n\in\mathbb{N}} \mu(A_n)$ (where the series in B converges absolutely).*

Proposition 9.9 (Conditional probability as a vector measure) *The conditional probability $A \mapsto \mathbb{P}[A|\mathcal{G}] \in \mathbb{L}^1$ is a vector measure with values in $B = \mathbb{L}^1$.*

PROOF Clearly $\mathbb{P}[0|\mathcal{G}] = 0$, a.s. Let $\{A_n\}_{n\in\mathbb{N}}$ be a pairwise-disjoint sequence in $\mathcal{F}$. Then

$$\left|\left|\mathbb{P}[A_n|\mathcal{G}]\right|\right|_{\mathbb{L}^1} = \mathbb{E}[|\mathbb{E}[\mathbf{1}_{A_n}|\mathcal{G}]|] = \mathbb{E}[\mathbf{1}_{A_n}] = \mathbb{P}[A_n],$$

and so

$$\sum_{n\in\mathbb{N}} \left|\left|\mathbb{P}[A_n|\mathcal{G}]\right|\right|_{\mathbb{L}^1} = \sum_{n\in\mathbb{N}} \mathbb{P}[A_n] = \mathbb{P}[\cup_n A_n] \leq 1 < \infty,$$

which implies that $\sum_{n\in\mathbb{N}} \mathbb{P}[A_n|\mathcal{G}]$ converges absolutely in $\mathbb{L}^1$. Finally, for $A = \cup_{n\in\mathbb{N}} A_n$, we have

$$\left|\left|\mathbb{P}[A|\mathcal{G}] - \sum_{n=1}^{N} \mathbb{P}[A_n|\mathcal{G}]\right|\right|_{\mathbb{L}^1} = \left|\left|\mathbb{E}[\sum_{n=N+1}^{\infty} \mathbf{1}_{A_n}|\mathcal{G}]\right|\right|_{\mathbb{L}^1} = \mathbb{P}[\cup_{n=N+1}^{\infty} A_n] \to 0 \text{ as } N \to \infty. \qquad \blacksquare$$

It is tempting to try to interpret the map $A \mapsto \mathbb{P}[A|\mathcal{G}](\omega)$ as a probability measure for a fixed $\omega \in \Omega$. It will not work in general; the reason is that $\mathbb{P}[A|\mathcal{G}]$ is defined only a.s., and the exceptional sets pile up when uncountable families of events A are considered. Even if we fixed versions

$\mathbb{P}[A|\mathcal{G}] \in \mathcal{L}^0_+$, for each $A \in \mathcal{F}$, the countable additivity relation (9.4) holds only almost surely so there is no guarantee that, for a fixed $\omega \in \Omega$, $\mathbb{P}[\cup_{n \in \mathbb{N}} A_n|\mathcal{G}](\omega) = \sum_{n \in \mathbb{N}} \mathbb{P}[A_n|\mathcal{G}](\omega)$, for all pairwise disjoint sequences $\{A_n\}_{n \in \mathbb{N}}$ in $\mathcal{F}$.

There is a way out of this predicament in certain situations, and we start with a description of an abstract object that corresponds to a well-behaved conditional probability:

Definition 9.10 (Measurable kernels) *Let* $(R, \mathcal{R})$ *and* $(S, \mathcal{S})$ *be measurable spaces. A map* $\nu : R \times \mathcal{S} \to \mathbb{R}$ *is called a* **(measurable) kernel** *if*

1. $x \mapsto \nu(x, B)$ *is* $\mathcal{R}$-*measurable for each* $B \in \mathcal{S}$, *and*

2. $B \mapsto \nu(x, B)$ *is a measure on* $\mathcal{S}$ *for each* $x \in R$.

Definition 9.11 (Regular conditional distributions) *Let* $\mathcal{G}$ *be a sub-*σ-*algebra of* $\mathcal{F}$, *let* $(S, \mathcal{S})$ *be a measurable space, and let* $e : \Omega \to S$ *be a random element in* S. *A kernel* $\mu_{e|\mathcal{G}} : \Omega \times \mathcal{S} \to [0, 1]$ *is called the* **regular conditional distribution of** e, **given** $\mathcal{G}$, *if*

$$\mu_{e|\mathcal{G}}(\omega, B) = \mathbb{P}[e \in B|\mathcal{G}](\omega), \text{ a.s., for all } B \in \mathcal{S}.$$

Remark 9.12

1. When $(S, \mathcal{S}) = (\Omega, \mathcal{F})$, and $e(\omega) = \omega$, the regular conditional distribution of e (if it exists) is called the **regular conditional probability**. Indeed, in this case, $\mu_{e|\mathcal{G}}(\cdot, B) = \mathbb{P}[e \in B|\mathcal{G}] = \mathbb{P}[B|\mathcal{G}]$, a.s.

2. It can be shown that regular conditional distributions not need to exist in general if S is "too large".

When $(S, \mathcal{S})$ is "small enough", however, regular conditional distributions can be constructed. Here is what we mean by "small enough":

Definition 9.13 (Borel spaces) *A measurable space* $(S, \mathcal{S})$ *is said to be a* **Borel space** (*or a* **nice space**) *if it is isomorphic to a Borel subset of* $\mathbb{R}$, *i.e., if there one-to-one map* $\rho : S \to \mathbb{R}$ *such that both* ρ *and* ρ^{-1} *are measurable.*

Problem 9.14 Show that $\mathbb{R}^n$, $n \in \mathbb{N}$ (together with their Borel σ-algebras) are Borel spaces. (*Hint:* Show, first, that there is a measurable bijection $\rho : [0, 1] \to [0, 1] \times [0, 1]$ such that ρ^{-1} is also measurable. Use binary (or decimal, or . . .) expansions.)

Remark 9.15 It can be show that any Borel subset of any complete and separable metric space is a Borel space. In particular, the coin-toss space is a Borel space.

Proposition 9.16 (A criterion for existence of regular conditional distributions) *Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$, and let $(S, \mathcal{S})$ be a Borel space. Any random element $e : \Omega \to S$ admits a regular conditional distribution.*

PROOF **(*)** Let us, first, deal with the case $S = \mathbb{R}$, so that $e = X$ is a random variable. Let Q be a countable dense set in $\mathbb{R}$. For $q \in Q$, consider the random variable P^q, defined as an arbitrary version of

$$P^q = \mathbb{P}[X \leq q | \mathcal{G}].$$

By redefining each P^q on a null set (and aggregating the countably many null sets - one for each $q \in Q$), we may suppose that $P^q(\omega) \leq P^r(\omega)$, for $q \leq r$, $q, r \in Q$, for *all* $\omega \in \Omega$ and that $\lim_{q \to \infty} P^q(\omega) = 1$ and $\lim_{q \to -\infty} P^q(\omega) = 0$, for *all* $\omega \in \Omega$. For $x \in \mathbb{R}$, we set

$$F(\omega, x) = \inf_{q \in Q, q > x} P^q(\omega),$$

so that, for each $\omega \in \Omega$, $F(\omega, \cdot)$ is a right-continuous non-decreasing function from $\mathbb{R}$ to $[0, 1]$, with $\lim_{x \to \infty} F(\omega, x) = 1$ and $\lim_{x \to -\infty} F(\omega, x) = 0$, for all $\omega \in \Omega$. Moreover, as an infimum of countably many random variables, the map $\omega \mapsto F(\omega, x)$ is a random variable for each $x \in \mathbb{R}$.

By (the proof of) Proposition 6.40, for each $\omega \in \Omega$, there exists a unique probability measure $\mu_{e|\mathcal{G}}(\omega, \cdot)$ on $\mathbb{R}$ such that $\mu_{e|\mathcal{G}}(\omega, (-\infty, x]) = F(\omega, x)$, for all $x \in \mathbb{R}$. Let $\mathcal{L}$ denote the set of all $B \in \mathcal{B}$ such that

1. $\omega \mapsto \mu_{e|\mathcal{G}}(\omega, B)$ is a random variable, and

2. $\mu_{e|\mathcal{G}}(\cdot, B)$ is a version of $\mathbb{P}[X \in B | \mathcal{G}]$.

It is not hard to check that $\mathcal{L}$ is a λ-system, so we need to prove that (1) and (2) hold for all B in some π-system which generates $\mathcal{B}(\mathbb{R})$. A convenient π-system to use is $\mathcal{P} = \{(-\infty, x] : x \in \mathbb{R}\}$. For $B = (-\infty, x] \in \mathcal{P}$, we have $\mu_{e|\mathcal{G}}(\omega, B) = F(\omega, x)$, so that (1) holds. To check (2), we need to show that $F(x, \omega) = \mathbb{P}[X \leq x | \mathcal{G}]$, a.s. This follows from the fact that

$$F(\cdot, x) = \inf_{q > x} P^q = \lim_{q \searrow x} P^q = \lim_{q \searrow x} \mathbb{P}[X \leq q | \mathcal{G}] = \mathbb{P}[X \leq x | \mathcal{G}], \text{ a.s.,}$$

by the conditional dominated convergence theorem.

Turning to the case of a general random element e which takes values in a Borel space $(S, \mathcal{S})$, we pick a one-to-one measurable map $f : S \to \mathbb{R}$ whose inverse ρ^{-1} is also measurable. Then $X = \rho(e)$ is a random variable, and so, by the above, there exists a kernel $\mu_{X|\mathcal{G}} : \Omega \times \mathcal{B}(\mathbb{R}) \to [0, 1]$ such that

$$\mu_{X|\mathcal{G}}(\cdot, A) = \mathbb{P}[\rho(e) \in A | \mathcal{G}], \text{ a.s.}$$

We define the kernel $\mu_{e|\mathcal{G}} : \Omega \times \mathcal{S} \to [0, 1]$ by

$$\mu_{e|\mathcal{G}}(\omega, B) = \mu_{X|\mathcal{G}}(\omega, \rho(B)). \qquad \blacksquare$$

Then, $\mu_{e|\mathcal{G}}(\cdot, B)$ is a random variable for each $B \in \mathcal{S}$ and for a pairwise disjoint sequence $\{B_n\}_{n \in \mathbb{N}}$ in $\mathcal{S}$, we have

$$\mu_{e|\mathcal{G}}(\omega, \cup_n B_n) = \mu_{X|\mathcal{G}}(\omega, \rho(\cup_n B_n)) = \mu_{X|\mathcal{G}}(\omega, \cup_n \rho(B_n))$$

$$= \sum_{n \in \mathbb{N}} \mu_{X|\mathcal{G}}(\omega, \rho(B_n)) = \sum_{n \in \mathbb{N}} \mu_{e|\mathcal{G}}(\omega, B_n),$$

which shows that $\mu_{e|\mathcal{G}}$ is a kernel; we used the measurability of ρ^{-1} to conclude that $\rho(B_n) \in \mathcal{B}(\mathbb{R})$ and the injectivity of ρ to ensure that $\{\rho(B_n)\}_{n\in\mathbb{N}}$ is pairwise disjoint. Finally, we need to show that $\mu_{e|\mathcal{G}}(\cdot, B)$ is a version of the conditional probability $\mathbb{P}[e \in B|\mathcal{G}]$. By injectivity of ρ, we have

$$\mathbb{P}[e \in B|\mathcal{G}] = \mathbb{P}[\rho(e) \in \rho(B)|\mathcal{G}] = \mu_{X|\mathcal{G}}(\cdot, \rho(B)) = \mu_{e|\mathcal{G}}(\cdot, B), \text{ a.s.}$$

Remark 9.17 Note that the regular conditional distribution is not unique, in general. Indeed, we can redefine it arbitrarily (as long as it remains a kernel) on a set of the form $N \times \mathcal{S} \subseteq \Omega \times \mathcal{S}$, where $\mathbb{P}[N] = 0$, without changing any of its defining properties. This will, in these notes, never be an issue.

One of the many reasons why regular conditional distributions are useful is that they sometimes allow non-conditional thinking to be transferred to the conditional case:

Proposition 9.18 (Conditional expectation as a parametrized integral) *Let X be an $\mathbb{R}^n$-valued random vector, let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$, and let $g : \mathbb{R}^n \to \mathbb{R}$ be a Borel function with the property $g(X) \in \mathbb{L}^1$. Then $\int_{\mathbb{R}^n} g(x)\mu_{X|\mathcal{G}}(\cdot, dx)$ is a $\mathcal{G}$-measurable random variable and*

$$\mathbb{E}[g(X)|\mathcal{G}] = \int_{\mathbb{R}^n} g(x)\mu_{X|\mathcal{G}}(\cdot, dx), \text{ a.s.}$$

PROOF When $g = \mathbf{1}_B$, for $B \in \mathbb{R}^n$, the statement follows by the very definition of the regular condition distribution. For the general case, we simply use the standard machine. ∎

Just like we sometimes express the distribution of a random variable or a vector in terms of its density, cdf or characteristic function, we can talk about the conditional density, conditional cdf or the conditional characteristic function. All of those will correspond to the case covered in Proposition 9.16 and all conditional distributions will be assumed to be regular. For $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, $y \leq_n x$ means $y_1 \leq x_1, \ldots, y_n \leq x_n$.

Definition 9.19 (Other regular conditional quantities) *Let $X : \Omega \to \mathbb{R}^n$ be a random vector, let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$, and let $\mu_{X|\mathcal{G}} : \Omega \times \mathcal{B}(\mathbb{R}^n) \to [0, 1]$ be the regular conditional distribution of X given $\mathcal{G}$.*

1. The **(regular) conditional cdf of X, given** $\mathcal{G}$ *is the map* $F : \Omega \times \mathbb{R}^n \to [0, 1]$*, given by*

$$F(\omega, x) = \mu_{X|\mathcal{G}}(\omega, \{y \in \mathbb{R}^n \,:\, y \leq_n x\}), \text{ for } x \in \mathbb{R}^n,$$

2. A map $f_{X|\mathcal{G}} : \Omega \times \mathbb{R}^n \to [0, \infty)$ is called the **conditional density of X with respect to** $\mathcal{G}$ *if*

   a) $f_{X|\mathcal{G}}(\omega, \cdot)$ is Borel measurable for all $\omega \in \Omega$,

   b) $f_{X|\mathcal{G}}(\cdot, x)$ is $\mathcal{G}$-measurable for each $x \in \mathbb{R}^n$, and

   c) $\int_B f_{X|\mathcal{G}}(\omega, x)\, dx = \mu_{X|\mathcal{G}}(\omega, B)$, for all $\omega \in \Omega$ and all $B \in \mathcal{B}(\mathbb{R}^n)$,

---

3. *The* **conditional characteristic function of X, given** $\mathcal{G}$ *is the map* $\varphi_{X|\mathcal{G}} : \Omega \times \mathbb{R}^n \to \mathbb{C}$, *given by*

$$\varphi_{X|\mathcal{G}}(\omega, t) = \int_{\mathbb{R}^n} e^{it \cdot x} \, \mu_{X|\mathcal{G}}(\omega, dx), \text{ for } t \in \mathbb{R}^n \text{ and } \omega \in \Omega.$$

To illustrate the utility of the above concepts, here is a versatile result (see Example 9.23 below):

Proposition 9.20 (Regular conditional characteristic funtions and independence) *Let X be a random vector in $\mathbb{R}^n$, and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. The following two statements are equivalent:*

1. *There exists a (deterministic) function $\varphi : \mathbb{R}^n \to \mathbb{C}$ such that for $\mathbb{P}$-almost all $\omega \in \Omega$,*

$$\varphi_{X|\mathcal{G}}(\omega, t) = \varphi(t), \text{ for all } t \in \mathbb{R}^n.$$

2. *$\sigma(X)$ is independent of $\mathcal{G}$.*

Moreover, whenever the two equivalent statements hold, φ is the characteristic function of X.

PROOF $(1) \Rightarrow (2)$. By Proposition 9.18, we have $\varphi_{X|\mathcal{G}}(\cdot, t) = \mathbb{E}[e^{it \cdot X}|\mathcal{G}]$, a.s. If we replace $\varphi_{X|\mathcal{G}}$ by φ, multiplying both sides by a bounded $\mathcal{G}$-measurable random variable Y and take expectations, we get

$$\varphi(t)\mathbb{E}[Y] = \mathbb{E}[Ye^{it \cdot X}].$$

In particular, for $Y = 1$ we get $\varphi(t) = \mathbb{E}[e^{it \cdot X}]$, so that

(9.5) $$\mathbb{E}[Ye^{it \cdot X}] = \mathbb{E}[Y]\mathbb{E}[e^{it \cdot X}],$$

for all $\mathcal{G}$-measurable and bounded Y, and all $t \in \mathbb{R}^n$. For Y of the form $Y = e^{isZ}$, where Z is a $\mathcal{G}$-measurable random variable, relation (9.5) and (a minimal extension of) part (1) of Problem 7.41, we conclude that X and Z are independent. Since Z is arbitrary and $\mathcal{G}$-measurable, X and $\mathcal{G}$ are independent.

$(2) \Rightarrow (1)$. If $\sigma(X)$ is independent of $\mathcal{G}$, so is $e^{it \cdot X}$, and so, the "irrelevance of independent information" property of conditional expectation implies that

$$\varphi(t) = \mathbb{E}[e^{it \cdot X}] = \mathbb{E}[e^{it \cdot X}|\mathcal{G}] = \varphi_{X|\mathcal{G}}(\cdot, t), \text{ a.s.} \qquad \blacksquare$$

One of the most important cases used in practice is when a random vector $(X_1, \ldots, X_n)$ admits a density and we condition on the σ-algebra generated by several of its components. To make the notation more intuitive, we denote the first d components $(X_1, \ldots, X_d)$ by X^o (for *observed*) and the remaining $n - d$ components $(X_{d+1}, \ldots, X_n)$ by X^u (for *unobserved*).

Proposition 9.21 (Conditional densities) *Suppose that the random vector $\boldsymbol{X} = (\boldsymbol{X}^o, \boldsymbol{X}^u) = (X_1, \ldots, X_d, X_{d+1}, \ldots, X_n)$ admits a density $f_{\boldsymbol{X}} : \mathbb{R}^n \to [0, \infty)$ and that the σ-algebra $\mathcal{G} = \sigma(\boldsymbol{X}^o)$ is generated by the random vector $\boldsymbol{X}^o = (X_1, \ldots, X_d)$, for some $d \in \{1, \ldots, n-1\}$. Then, for $\boldsymbol{X}^u = (X_{d+1}, \ldots, X_n)$, there exists a conditional density $f_{\boldsymbol{X}^u|\mathcal{G}} : \Omega \times \mathbb{R}^{n-d} \to [0, \infty)$, of $\boldsymbol{X}^u$ given $\mathcal{G}$, and (a version of it) is given by*

$$f_{\boldsymbol{X}^u|\mathcal{G}}(\omega, \boldsymbol{x}^u) = \begin{cases} \frac{f_{\boldsymbol{X}}(\boldsymbol{X}^o(\omega), \boldsymbol{x}^u)}{\int_{\mathbb{R}^{n-d}} f_{\boldsymbol{X}}(\boldsymbol{X}^o(\omega), \boldsymbol{y}) \, d\boldsymbol{y}}, & \int_{\mathbb{R}^{n-d}} f(\boldsymbol{X}^o, \boldsymbol{y}) \, d\boldsymbol{y} > 0, \\ f_0(\boldsymbol{x}^u), & otherwise, \end{cases}$$

for $x \in \mathbb{R}^{n-d}$ and $\omega \in \Omega$, where $f_0 : \mathbb{R}^{n-d} \to \mathbb{R}$ is an arbitrary density function.

PROOF First, we note that $f_{\boldsymbol{X}^u|\mathcal{G}}$ is constructed from the jointly Borel-measurable function $f_{\boldsymbol{X}}$ and the random vector $\boldsymbol{X}^o$ in an elementary way, and is, thus, jointly measurable in $\mathcal{G} \times \mathcal{B}(\mathbb{R}^{n-d})$. It remains to show that

$$\int_A f_{\boldsymbol{X}^u|\mathcal{G}}(\cdot, \boldsymbol{x}^u) \, d\boldsymbol{x}^u \text{ is a version of } \mathbb{P}[\boldsymbol{X}^u \in A|\mathcal{G}], \text{ for all } A \in \mathcal{B}(\mathbb{R}^{n-d}). \qquad \blacksquare$$

Equivalently, we need to show that

$$\mathbb{E}[\mathbf{1}_{\{\boldsymbol{X}^o \in A^o\}} \int_{A^u} f_{\boldsymbol{X}^u|\mathcal{G}}(\cdot, \boldsymbol{x}^u) \, d\boldsymbol{x}^u] = \mathbb{E}[\mathbf{1}_{\{\boldsymbol{X}^o \in A^o\}} \mathbf{1}_{\{\boldsymbol{X}^u \in A^u\}}],$$

for all $A^o \in \mathcal{B}(\mathbb{R}^d)$ and $A^u \in \mathcal{B}(\mathbb{R}^{n-d})$.

Fubini's theorem, and the fact that $f_{\boldsymbol{X}^o}(\boldsymbol{x}^o) = \int_{\mathbb{R}^{n-d}} f(\boldsymbol{x}^o, \boldsymbol{y}) \, d\boldsymbol{y}$ is the density of $\boldsymbol{X}^o$ yield

$$
\begin{aligned}
\mathbb{E}[\mathbf{1}_{\{\boldsymbol{X}^o \in A^o\}} \int_{A^u} f_{\boldsymbol{X}^u|\mathcal{G}}(\cdot, \boldsymbol{x}^u) \, d\boldsymbol{x}^u] &= \int_{A^u} \mathbb{E}[\mathbf{1}_{\{\boldsymbol{X}^o \in A^o\}} f_{\boldsymbol{X}^u|\mathcal{G}}(\cdot, \boldsymbol{x}^u)] \, d\boldsymbol{x}^o \\
&= \int_{A^u} \int_{A^o} f_{\boldsymbol{X}^u|\mathcal{G}}(\boldsymbol{x}^o, \boldsymbol{x}^u) f_{\boldsymbol{X}^o}(\boldsymbol{x}^o) \, d\boldsymbol{x}^o \, d\boldsymbol{x}^u \\
&= \int_{A^u} \int_{A^o} f_{\boldsymbol{X}}(\boldsymbol{x}^o, \boldsymbol{x}^u) \, d\boldsymbol{x}^o \, d\boldsymbol{x}^u \\
&= \mathbb{P}[\boldsymbol{X}^o \in A^o, \boldsymbol{X}^u \in A^u].
\end{aligned}
$$

The above result expresses a conditional density, given $\mathcal{G} = \sigma(\boldsymbol{X}^o)$, as a (deterministic) function of $\boldsymbol{X}^o$. Such a representation is possible even when there is no joint density. The core of the argument is contained in the following problem:

Problem 9.22 Let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^d$, and let $\mathcal{G} = \sigma(\boldsymbol{X})$ be the σ-algebra generated by $\boldsymbol{X}$. Then, a random variable Z is $\mathcal{G}$-measurable if and only if there exists a Borel function $f : \mathbb{R}^d \to \mathbb{R}$ with the property that $Z = f(\boldsymbol{X})$.

Let $\boldsymbol{X}^o$ be a random vector in $\mathbb{R}^d$. For $X \in \mathcal{L}^1$ the conditional expectation $\mathbb{E}[X|\sigma(\boldsymbol{X}^o)]$ is $\sigma(\boldsymbol{X}^o)$-measurable, so there exists a Borel function $f : \mathbb{R}^d \to R$ such that $\mathbb{E}[X|\sigma(\boldsymbol{X}^o)] = f(\boldsymbol{X}^o)$, a.s. Note that f is uniquely defined only up to $\mu_{\boldsymbol{X}^o}$-null sets. The value $f(\boldsymbol{x}^o)$ at $\boldsymbol{x}^o \in \mathbb{R}^d$ is usually denoted by $\mathbb{E}[X|\boldsymbol{X}^o = \boldsymbol{x}^o]$.

Example 9.23 (Conditioning normals on their components) Let $\boldsymbol{X} = (\boldsymbol{X}^o, \boldsymbol{X}^u) \in \mathbb{R}^d \times \mathbb{R}^{n-d}$ be a multivariate normal random vector with mean $\boldsymbol{\mu} = (\boldsymbol{\mu}^o, \boldsymbol{\mu}^u)$ and the variance-covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\tilde{\boldsymbol{X}} \tilde{\boldsymbol{X}}^T]$, where $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{\mu}$. A block form of the matrix $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{oo} & \boldsymbol{\Sigma}_{ou} \\ \boldsymbol{\Sigma}_{uo} & \boldsymbol{\Sigma}_{uu} \end{pmatrix},$$

Where

$$\boldsymbol{\Sigma}_{oo} = \mathbb{E}[\tilde{\boldsymbol{X}}^o (\tilde{\boldsymbol{X}}^o)^T] \in \mathbb{R}^{d \times d}$$
$$\boldsymbol{\Sigma}_{ou} = \mathbb{E}[\tilde{\boldsymbol{X}}^o (\tilde{\boldsymbol{X}}^u)^T] \in \mathbb{R}^{d \times (n-d)}$$
$$\boldsymbol{\Sigma}_{uo} = \mathbb{E}[\tilde{\boldsymbol{X}}^u (\tilde{\boldsymbol{X}}^o)^T] \in \mathbb{R}^{(n-d) \times d}$$
$$\boldsymbol{\Sigma}_{uu} = \mathbb{E}[\tilde{\boldsymbol{X}}^u (\tilde{\boldsymbol{X}}^u)^T] \in \mathbb{R}^{(n-d) \times (n-d)}.$$

We assume that $\boldsymbol{\Sigma}_{oo}$ is invertible. Otherwise, we can find a subset of components of $\boldsymbol{X}^o$ whose variance-covariance matrix in invertible and which generate the same σ-algebra (why?). The matrix $A = \boldsymbol{\Sigma}_{uo} \boldsymbol{\Sigma}_{oo}^{-1}$ has the property that $\mathbb{E}[(\tilde{\boldsymbol{X}}^u - A\tilde{\boldsymbol{X}}^o)(\tilde{\boldsymbol{X}}^o)^T] = 0$, i.e., that the random vectors $\tilde{\boldsymbol{X}}^o - A\tilde{\boldsymbol{X}}^o$ and $\tilde{\boldsymbol{X}}^o$ are uncorrelated. We know, however, that $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{X}}^o, \tilde{\boldsymbol{X}}^u)$ is a Gaussian random vector, so, by Problem 7.41, part (3), $\tilde{\boldsymbol{X}}^o - A\tilde{\boldsymbol{X}}^o$ is independent of $\tilde{\boldsymbol{X}}^o$. It follows from Proposition 9.20 that the conditional characteristic function of $\tilde{\boldsymbol{X}}^o - A\tilde{\boldsymbol{X}}^o$, given $\mathcal{G} = \sigma(\tilde{\boldsymbol{X}}^o)$ is deterministic and given by

$$\mathbb{E}[e^{i\boldsymbol{t}(\tilde{\boldsymbol{X}}^u - A\tilde{\boldsymbol{X}}^o)} | \mathcal{G}] = \varphi_{\tilde{\boldsymbol{X}}^u - A\tilde{\boldsymbol{X}}^o}(\boldsymbol{t}), \text{ for } \boldsymbol{t} \in \mathbb{R}^{n-d}.$$

Since $A\tilde{\boldsymbol{X}}^o$ is $\mathcal{G}$-measurable, we have

$$\mathbb{E}[e^{i\boldsymbol{t}\boldsymbol{X}^u} | \mathcal{G}] = e^{i\boldsymbol{t}\boldsymbol{\mu}^u} e^{i\boldsymbol{t}A\tilde{\boldsymbol{X}}^o} e^{-\frac{1}{2}\boldsymbol{t}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{t}}, \text{ for } \boldsymbol{t} \in \mathbb{R}^{n-d}.$$

where $\hat{\boldsymbol{\Sigma}} = \mathbb{E}[(\tilde{\boldsymbol{X}}^u - A\tilde{\boldsymbol{X}}^o)(\tilde{\boldsymbol{X}}^u - A\tilde{\boldsymbol{X}}^o)^T]$. A simple calculation yields that, conditionally on $\mathcal{G}$, $\boldsymbol{X}^u$ is multivariate normal with mean $\boldsymbol{\mu}_{\boldsymbol{X}^u|\mathcal{G}}$ and variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}^u|\mathcal{G}}$ given by

$$\boldsymbol{\mu}_{\boldsymbol{X}^u|\mathcal{G}} = \boldsymbol{\mu}^o + A(\boldsymbol{X}^o - \boldsymbol{\mu}^o), \ \boldsymbol{\Sigma}_{\boldsymbol{X}^u|\mathcal{G}} = \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{uo} \boldsymbol{\Sigma}_{oo}^{-1} \boldsymbol{\Sigma}_{ou}.$$

Note how the mean gets corrected by a multiple of the difference between the observed value $\boldsymbol{X}^o$ and its (unconditional) expected value. Similarly, the variance-covariance matrix also gets corrected by $\boldsymbol{\Sigma}_{uo} \boldsymbol{\Sigma}_{oo}^{-1} \boldsymbol{\Sigma}_{ou}$, but this quantity does not depend on the observation $\boldsymbol{X}^o$.

Problem 9.24 Let (X_1, X_2) be a bivariate normal vector with $\text{Var}[X_1] > 0$. Work out the exact form of the conditional distribution of X_2, given X_1 in terms of $\mu_i = \mathbb{E}[X_i]$, $\sigma_i^2 = \text{Var}[X_i]$, $i = 1, 2$ and the correlation coefficient $\rho = \text{corr}(X_1, X_2)$.

9.4   Additional Problems

Problem 9.25 (Conditional expectation for non-negative random variables) A parallel definition of conditional expectation can be given for random variables in $\mathcal{L}_+^0$. For $X \in \mathcal{L}_+^0$, we say that Y is a conditional expectation of X with respect to $\mathcal{G}$ - and denote it by $\mathbb{E}[X|\mathcal{G}]$ - if

(a) Y is $\mathcal{G}$-measurable and $[0, \infty]$-valued, and

(b) $\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A] \in [0, \infty]$, for $A \in \mathcal{G}$.

Show that

1. $\mathbb{E}[X|\mathcal{G}]$ exists for each $X \in \mathcal{L}^0_+$.

2. $\mathbb{E}[X|\mathcal{G}]$ is unique a.s. (*Hint:* The argument in the proof of Proposition 9.3 needs to be modified before it can be used.)

3. $\mathbb{E}[X|\mathcal{G}]$ no longer necessarily exists for all $X \in \mathcal{L}^0_+$ if we insist that $\mathbb{E}[X|\mathcal{G}] < \infty$, a.s., instead of $\mathbb{E}[X|\mathcal{G}] \in [0,\infty]$, a.s.

Problem 9.26 (How to deal with the independent component) Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a bounded Borel-measurable function, and let X and Y be independent random variables. Define the function $g : \mathbb{R} \to \mathbb{R}$ by
$$g(y) = \mathbb{E}[f(X,y)].$$
Show that the function g is Borel-measurable, and that

$$\mathbb{E}[f(X,Y)|Y = y] = g(y), \ \mu_Y - a.s.$$

Problem 9.27 (Some exercises in conditional probability)

1. Let X, Y_1, Y_2 be random variables. Show that the random vectors (X, Y_1) and (X, Y_2) have the same distribution if and only if $\mathbb{P}[Y_1 \in B|\sigma(X)] = \mathbb{P}[Y_2 \in B|\sigma(X)]$, for all $B \in \mathcal{B}(\mathbb{R})$.

2. Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of non-negative integrable random variables, and let $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ be sub-σ-algebras of $\mathcal{F}$. Show that $X_n \xrightarrow{\mathbb{P}} 0$ if $\mathbb{E}[X_n|\mathcal{F}_n] \xrightarrow{\mathbb{P}} 0$. Does the converse hold? (*Hint:* Prove that for $X_n \in \mathcal{L}^0_+$, we have $X_n \xrightarrow{\mathbb{P}} 0$ if and only if $\mathbb{E}[\min(X_n, 1)] \to 0$.)

3. Let $\mathcal{G}$ be a complete sub-σ-algebra of $\mathcal{F}$. Suppose that for $X \in \mathcal{L}^1$, $\mathbb{E}[X|\mathcal{G}]$ and X have the same distribution. Show that X is $\mathcal{G}$-measurable. (*Hint:* Use the conditional Jensen's inequality.)

Problem 9.28 (A characterization of $\mathcal{G}$-measurability) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Show that for a random variable $X \in \mathcal{L}^1$ the following two statements are equivalent:

1. X is $\mathcal{G}$-measurable.

2. For all $\xi \in \mathcal{L}^\infty$, $\mathbb{E}[X\xi] = \mathbb{E}[X\mathbb{E}[\xi|\mathcal{G}]]$.

Problem 9.29 (Conditioning a part with respect to the sum) Let $X_1, X_2, \ldots$ be a sequence of iid r.v.'s with finite first moment, and let $S_n = X_1 + X_2 + \cdots + X_n$. Define $\mathcal{G} = \sigma(S_n)$.

1. Compute $\mathbb{E}[X_1|\mathcal{G}]$.

2. Supposing, additionally, that X_1 is normally distributed, compute $\mathbb{E}[f(X_1)|\mathcal{G}]$, where $f : \mathbb{R} \to \mathbb{R}$ is a Borel function such that $f(X_1) \in \mathbb{L}^1$.

Chapter 10

Discrete Martingales

10.1 Discrete-time filtrations and stochastic processes

One of the uses of σ-algebras is to single out the subsets of Ω to which probability can be assigned. This is the role of $\mathcal{F}$. Another use, as we have seen when discussing conditional expectation, is to encode information. The arrow of time, as we perceive it, points from less information to more information. A useful mathematical formalism is the one of a *filtration*.

Definition 10.1 (Filtered probability spaces) *A **filtration** is a sequence $\{\mathcal{F}_n\}_{n\in\mathbb{N}_0}$, where $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, of sub-$\sigma$-algebras of $\mathcal{F}$ such that $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$, for all $n \in \mathbb{N}_0$. A probability space with a filtration - $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}_{n\in\mathbb{N}_0}, \mathbb{P})$ - is called a **filtered probability space**.*

We think of $n \in \mathbb{N}_0$ as the time-index and of $\mathcal{F}_n$ as the information available at time n.

Definition 10.2 (Discrete-time stochastic process) *A **(discrete-time) stochastic process** is a sequence $\{X_n\}_{n\in\mathbb{N}_0}$ of random variables.*

A stochastic process is a generalization of a random vector; in fact, we can think of a stochastic processes as an infinite-dimensional random vector. More precisely, a stochastic process is a random element in the space $\mathbb{R}^{\mathbb{N}_0}$ of real sequences. In the context of stochastic processes, the sequence $(X_0(\omega), X_1(\omega), \dots)$ is called a **trajectory** of the stochastic process $\{X_n\}_{n\in\mathbb{N}_0}$. This dual view of stochastic processes - as random trajectories (sequences) or as sequences of random variables - can be supplemented by another interpretation: a stochastic process is also a map from the product space $\Omega \times \mathbb{N}_0$ into $\mathbb{R}$.

Definition 10.3 (Adapted processes) *A stochastic process $\{X_n\}_{n\in\mathbb{N}_0}$ is said to be **adapted** with respect to the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}_0}$ if X_n is $\mathcal{F}_n$-measurable for each $n \in \mathbb{N}_0$.*

Intuitively, the process $\{X_n\}_{n\in\mathbb{N}_0}$ is adapted with respect to the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}_0}$ if its value X_n is fully known at time n (assuming that the time-n information is given by $\mathcal{F}_n$).

The most common way of producing filtrations is by generating them from stochastic processes. More precisely, for a stochastic process $\{X_n\}_{n\in\mathbb{N}_0}$, the filtration $\{\mathcal{F}_n^X\}_{n\in\mathbb{N}_0}$, given by

$$\mathcal{F}_n^X = \sigma(X_0, X_1, \ldots, X_n),\ n \in \mathbb{N}_0,$$

is called the **filtration generated by** $\{X_n\}_{n\in\mathbb{N}_0}$. Clearly, X is always adapted to the filtration generated by X.

10.2 Martingales

Definition 10.4 ((Sub-, super-) martingales) *Let $\{\mathcal{F}_n\}_{n\in\mathbb{N}_0}$ be a filtration. A stochastic process $\{X_n\}_{n\in\mathbb{N}_0}$ is called an $\{\mathcal{F}_n\}_{n\in\mathbb{N}_0}$-**supermartingale** if*

1. *$\{X_n\}_{n\in\mathbb{N}_0}$ is $\{\mathcal{F}_n\}_{n\in\mathbb{N}_0}$-adapted,*

2. *$X_n \in \mathcal{L}^1$, for all $n \in \mathbb{N}_0$, and*

3. *$\mathbb{E}[X_{n+1}|\mathcal{F}_n] \leq X_n$, a.s., for all $n \in \mathbb{N}_0$.*

*A process $\{X_n\}_{n\in\mathbb{N}_0}$ is called a **submartingale** if $\{-X_n\}_{n\in\mathbb{N}_0}$ is a supermartingale. A **martingale** is a process which is both a supermartingale and a submartingale at the same time, i.e., for which the equality holds in (3).*

Remark 10.5 Very often, the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}_0}$ is not explicitly mentioned. Then, it is often clear from the context, i.e., the existence of an underlying filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ is assumed throughout. Alternatively, if no filtration is pre-specified, the filtration $\{\mathcal{F}_n^X\}_{n\in\mathbb{N}_0}$, generated by $\{X_n\}_{n\in\mathbb{N}_0}$ is used. It is important to remember, however, that the notion of a (super-, sub-) martingale only makes sense in relation to a filtration.

The fundamental examples of martingales are (additive or multiplicative) random walks:

Example 10.6

1. An additive random walk. Let $\{\xi_n\}_{n\in\mathbb{N}}$ be a sequence of iid random variables with $\xi_n \in \mathcal{L}^1$ and $\mathbb{E}[\xi_n] = 0$, for all $n \in \mathbb{N}$. We define

$$X_0 = 0,\ X_n = \sum_{k=1}^{n} \xi_k,\ \text{for } n \in \mathbb{N}.$$

The process $\{X_n\}_{n\in\mathbb{N}_0}$ is a martingale with respect to the filtration $\{\mathcal{F}_n^X\}_{n\in\mathbb{N}_0}$ generated by it (which is the same as $\sigma(\xi_1, \ldots, \xi_n)$). Indeed, $X_n \in \mathcal{L}^1(\mathcal{F}_n^X)$ for all $n \in \mathbb{N}$ and

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] = \mathbb{E}[\xi_{n+1} + X_n|\mathcal{F}_n] = X_n + \mathbb{E}[\xi_{n+1}|\mathcal{F}_n] = X_n + \mathbb{E}[\xi_{n+1}] = X_n,\ \text{a.s.},$$

where we used the "irrelevance of independent information"-property of conditional expectation (in this case ξ_{n+1} is independent of $\mathcal{F}_n^X = \sigma(X_0, \ldots, X_n) = \sigma(\xi_1, \ldots, \xi_n)$).

It is easy to see that if $\{\xi_n\}_{n\in\mathbb{N}}$ are still iid, but $\mathbb{E}[\xi_n] > 0$, then $\{X_n\}_{n\in\mathbb{N}_0}$ is a submartingale. When $\mathbb{E}[\xi_n] < 0$, we get a supermartingale.

2. A multiplicative random walk. Let $\{\xi_n\}_{n \in \mathbb{N}}$ be an iid sequence in $\mathcal{L}^1$ such that $\mathbb{E}[\xi_n] = 1$. We define

$$X_0 = 1, \ X_n = \prod_{k=1}^{n} \xi_k, \text{ for } n \in \mathbb{N}.$$

The process $\{X_n\}_{n \in \mathbb{N}_0}$ is a martingale with respect to the filtration $\{\mathcal{F}_n^X\}_{n \in \mathbb{N}_0}$ generated it. Indeed, $X_n \in \mathcal{L}^1(\mathcal{F}_n^X)$ for all $n \in \mathbb{N}$ and

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] = \mathbb{E}[\xi_{n+1} X_n|\mathcal{F}_n] = X_n \mathbb{E}[\xi_{n+1}|\mathcal{F}_n] = X_n \mathbb{E}[\xi_{n+1}] = X_n, \text{ a.s.},$$

where, in addition to the "irrelevance of independent information" we also used "pulling our what's known".

Is it true that, if $\mathbb{E}[\xi_n] > 1$, we get a submartingale and that if $\mathbb{E}[\xi_n] < 1$, we get a super-martingale?

3. Wald's martingales Let $\{\xi_n\}_{n \in \mathbb{N}}$ be an independent sequence, and let $\varphi_n(t)$ be the characteristic function of ξ_n. Assuming that $\varphi_n(t) \neq 0$, for $n \in \mathbb{N}$, the previous example implies that the process $\{X_n\}_{n \in \mathbb{N}_0}$, defined by

$$X_0^t = 1, \ X_n^t = \prod_{k=1}^{n} \frac{e^{it\xi_k}}{\varphi_k(t)}, \ n \in \mathbb{N},$$

is a martingale. Actually, it is complex-valued, so it would be better to say that its real and imaginary parts are both martingales. This martingale will be important in the study of hitting times of random walks.

4. Lévy martingales. For $X \in \mathcal{L}^1$, we define

$$X_n = \mathbb{E}[X|\mathcal{F}_n], \text{ for } n \in \mathbb{N}_0.$$

The tower property of conditional expectation implies that $\{X_n\}_{n \in \mathbb{N}_0}$ is a martingale.

5. An urn scheme. An urn contains b black and w white balls on day $n = 0$. On each subsequent day, a ball is chosen (each ball in the urn has the same probability of being picked) and then put back together with another ball of the same color. Therefore, at the end of day n, here are $n + b + w$ balls in the urn. Let B_n denote the number of black balls in the urn at day n, and let define the process $\{X_n\}_{n \in \mathbb{N}_0}$ by

$$X_n = \frac{B_n}{b+w+n}, \ n \in \mathbb{N}_0,$$

to be the proportion of black balls in the urn at time n. Let $\{\mathcal{F}_n\}_{n \in \mathbb{N}_0}$ denote the filtration generated by $\{X_n\}_{n \in \mathbb{N}_0}$. The conditional probability - given $\mathcal{F}_n$ - of picking a black ball at time n is X_n, i.e.,

$$\mathbb{P}[B_{n+1} = B_n + 1|\mathcal{F}_n] = X_n \text{ and } \mathbb{P}[B_{n+1} = B_n|\mathcal{F}_n] = 1 - X_n.$$

Therefore,

$$\begin{aligned}
\mathbb{E}[X_{n+1}|\mathcal{F}_n] &= \mathbb{E}[X_{n+1}\mathbf{1}_{\{B_{n+1}=B_n\}}|\mathcal{F}_n] + \mathbb{E}[X_{n+1}\mathbf{1}_{\{B_{n+1}=B_n+1\}}|\mathcal{F}_n] \\
&= \mathbb{E}[\tfrac{B_n}{b+w+n+1}\mathbf{1}_{\{B_{n+1}=B_n\}}|\mathcal{F}_n] + \mathbb{E}[\tfrac{B_n+1}{b+w+n+1}\mathbf{1}_{\{B_{n+1}=B_n+1\}}|\mathcal{F}_n] \\
&= \tfrac{B_n}{b+w+n+1}(1-X_n) + \tfrac{B_n+1}{b+w+n+1}X_n = \tfrac{B_n(1-X_n)+(B_n+1)X_n}{b+w+n+1} = X_n.
\end{aligned}$$

How does this square with your intuition? Should not a high number of black balls translate into a high probability of picking a black ball? This will, in turn, only increase the number of black balls with high probability. In other words, why is $\{X_n\}_{n\in\mathbb{N}_0}$ not a submartingale (which is not a martingale), at least for large $\frac{b}{b+w}$?

To get some feeling for the definition, here are several simple exercises:

Problem 10.7

1. Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a martingale. Show that $\mathbb{E}[X_n] = \mathbb{E}[X_0]$, for all $n \in \mathbb{N}_0$. Give an example of a process $\{Y_n\}_{n\in\mathbb{N}_0}$ with $Y_n \in \mathcal{L}^1$, for all $n \in \mathbb{N}_0$ which is *not* a martingale, but $\mathbb{E}[Y_n] = \mathbb{E}[Y_0]$, for all $n \in \mathbb{N}_0$.

2. Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a martingale. Show that

$$\mathbb{E}[X_m|\mathcal{F}_n] = X_n, \text{ for all } m > n.$$

3. Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a submartingale, and let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function such that $\varphi(X_n) \in \mathcal{L}^1$, for all $n \in \mathbb{N}_0$. Show that $\{\varphi(X_n)_n\}_{n\in\mathbb{N}_0} n \in \mathbb{N}_0$ is a submartingale, provided that either

   a) φ is nondecreasing, or

   b) $\{X_n\}_{n\in\mathbb{N}_0}$ is a martingale.

   In particular, if $\{X_n\}_{n\in\mathbb{N}}$ is a submartingale, so are $\{X_n^+\}_{n\in\mathbb{N}_0}$ and $\{e^{X_n}\}_{n\in\mathbb{N}_0}$. (*Hint:* Use conditional Jensen's inequality.)

10.3   Predictability and martingale transforms

> **Definition 10.8 (Predictable processes)** *A process $\{H_n\}_{n\in\mathbb{N}}$ is said to be **predictable** with respect to the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}_0}$ if H_n is $\mathcal{F}_{n-1}$-measurable for $n \in \mathbb{N}$.*

A process is predictable if you can predict its tomorrow's value today. We often think of predictable processes as strategies: let $\{\xi_n\}_{n\in\mathbb{N}}$ be a sequence of random variables which we interpret as gambles. At time n we can place a bet of H_n dollars, thus realizing a gain/loss of $H_n\xi_n$. Note that a negative H_n is allowed - the player wins money if $\xi_n < 0$ and loses if $\xi_n > 0$ in that case. If $\{\mathcal{F}_n\}_{n\in\mathbb{N}_0}$ is a filtration generated by the gambles, i.e., $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_n = \sigma\{\xi_1, \ldots, \xi_n\}$, for $n \in \mathbb{N}$, then $H_n \in \mathcal{F}_{n-1}$, so that it does not use any information about ξ_n: we are allowed to adjust our bet according to the outcomes of previous gambles, but we don't know the outcome of ξ_n until after the bet is placed. Therefore, the sequence $\{H_n\}_{n\in\mathbb{N}}$ is a predictable sequence with respect to $\{\mathcal{F}_n\}_{n\in\mathbb{N}_0}$.

Problem 10.9 Characterize predictable submartingales and predictable martingales. (*Note:* To comply with the setting in which the definition of predictability is given (processes defined on $\mathbb{N}$ and not on $\mathbb{N}_0$), simply discard the value at 0.)

Definition 10.10 (Martingale transforms) *Let $\{\mathcal{F}_n\}_{n\in\mathbb{N}_0}$ be a filtration and let $\{X_n\}_{n\in\mathbb{N}_0}$ be a process adapted to $\{\mathcal{F}_n\}_{n\in\mathbb{N}_0}$. The stochastic process $\{(H \cdot X)_n\}_{n\in\mathbb{N}_0}$, defined by*

$$(H \cdot X)_0 = 0, \ (H \cdot X)_n = \sum_{k=1}^{n} H_k(X_k - X_{k-1}), \ \text{for } n \in \mathbb{N},$$

is called the **martingale transform of X by H**.

Remark 10.11

1. The process $\{(H \cdot X)_n\}_{n\in\mathbb{N}_0}$ is called the martingale transform of X, even if neither H nor X is a martingale. It is most often applied to a martingale X, though - hence the name.

2. In terms of the gambling interpretation given above, X plays the role of the cumulative gain (loss) when a \$1-bet is placed each time:

$$X_0 = 0, \ X_n = \sum_{k=1}^{n} \xi_k, \ \text{where } \xi_k = X_n - X_{n-1}, \ \text{for } n \in \mathbb{N}.$$

   If we insist that $\{\xi_n\}_{n\in\mathbb{N}}$ is a sequence of *fair* bets, i.e., that there are no expected gains/losses in the n-th bet, even after we had the opportunity to learn from the previous $n-1$ bets, we arrive to the condition

$$\mathbb{E}[\xi_n | \mathcal{F}_{n-1}] = 0, \ \text{i.e., that } \{X_n\}_{n\in\mathbb{N}_0} \text{ is a martingale.}$$

The following proposition states that no matter how well you choose your bets, you cannot make (or loose) money by betting on a sequence of fair games. A part of result is stated for submartingales; this is for convenience only. The reader should observe that almost any statement about submartingales can be turner into a statement about supermartingales by a simple change of sign.

Proposition 10.12 (Stability of martingales under martingale transforms) *Let $\{X_n\}_{n\in\mathbb{N}_0}$ be adapted, and let $\{H_n\}_{n\in\mathbb{N}}$ be predictable. Then, the martingale transform $H \cdot X$ of X by H is*

1. *a martingale, provided that $\{X_n\}_{n\in\mathbb{N}_0}$ is a martingale and $H_n(X_n - X_{n-1}) \in \mathcal{L}^1$, for all $n \in \mathbb{N}$,*

2. *a submartingale, provided that $\{X_n\}_{n\in\mathbb{N}_0}$ is a submartingale, $H_n \geq 0$, a.s., and $H_n(X_n - X_{n-1}) \in \mathcal{L}^1$, for all $n \in \mathbb{N}$.*

PROOF Just check the definition and use properties of conditional expectation. ∎

Remark 10.13 The martingale transform is the discrete-time analogue of the *stochastic integral*. Note that it is crucial that H be predictable if we want a martingale transform of a martingale to be a martingale. Otherwise, we just take $H_n = \text{sgn}(X_n - X_{n-1}) \in \mathcal{F}_n$ and obtain a process which is not a martingale unless X is constant. This corresponds to a player who knows the outcome of the game before the bet is placed and places the bet of \$ ± 1 which is guaranteed to win.

10.4 Stopping times

> **Definition 10.14 (Random and stopping times)** *A random variable T with values in $\mathbb{N}_0 \cup \{\infty\}$ is called a **random time**. A random time is said to be a **stopping time** with respect to the filtration $\{\mathcal{F}_n\}_{n \in \mathbb{N}_0}$ if*
> $$\{T \leq n\} \in \mathcal{F}_n, \text{ for all } n \in \mathbb{N}.$$

Remark 10.15

1. Stopping times are simply random instances with the property that at every instant you can answer the question "Has T already happened?" using only the currently-available information.

2. The additional element $+\infty$ is used as a placeholder for the case when T "does not happen".

Example 10.16

1. Constant (deterministic) times $T = m$, $m \in \mathbb{N}_0 \cup \{\infty\}$ are obviously stopping time. The set of all stopping times can be thought of as an enlargement of the set of "time-instances". The meaning of "when Red Sox win the World Series again" is clear, but it does not correspond to a deterministic time.

2. Let $\{X_n\}_{n \in \mathbb{N}_0}$ be a stochastic process adapted to the filtration $\{\mathcal{F}_n\}_{n \in \mathbb{N}_0}$. For a subset $B \in \mathcal{B}(\mathbb{R})$, we define the random time T_B by

$$T_B = \min\{n \in \mathbb{N}_0 \,:\, X_n \in B\}.$$

T_b is called the **hitting time** of the set B and is a stopping time. Indeed,

$$\{T_B \leq n\} = \{X_0 \in B\} \cup \{X_1 \in B\} \cup \cdots \cup \{X_n \in B\} \in \mathcal{F}_n.$$

3. Let $\{\xi_n\}_{n \in \mathbb{N}}$ be an iid sequence of coin tosses, i.e. $\mathbb{P}[\xi_i = 1] = \mathbb{P}[\xi_i = -1] = \frac{1}{2}$, and let $X_n = \sum_{k=1}^{n} \xi_k$ be the corresponding random walk. For $N \in \mathbb{N}$, let S be the random time defined by
$$S = \max\{n \leq N \,:\, X_n = 0\}.$$

S is called the **last visit time** to 0 before time $N \in \mathbb{N}$. Intuitively, S is not a stopping time since, in order to know whether S had already happened at time $m < N$, we need to know that $X_k \neq 0$, for $k = m + 1, \ldots, N$, and, for that, we need the information which is not contained in $\mathcal{F}_m$. We leave it to the reader to make this comment rigorous.

Stopping times have good stability properties, as the following proposition shows. All stopping times are with respect to an arbitrary, but fixed filtration $\{\mathcal{F}_n\}_{n \in \mathbb{N}_0}$.

Proposition 10.17 (Properties of stopping times)

1. *A random time T is a stopping time if and only if the process $X_n = \mathbf{1}_{\{n \geq T\}}$ is $\{\mathcal{F}_n\}_{n \in \mathbb{N}_0}$-adapted.*

2. *If S and T are stopping times, then so are $S + T$, $\max(S, T)$, $\min(S, T)$.*

3. *Let $\{T_n\}_{n \in \mathbb{N}}$ be a sequence of stopping times such that $T_1 \leq T_2 \leq \ldots$, a.s. Then $T = \sup_n T_n$ is a stopping time.*

4. *Let $\{T_n\}_{n \in \mathbb{N}}$ be a sequence of stopping times such that $T_1 \geq T_2 \geq \ldots$, a.s. Then $T = \inf_n T_n$ is a stopping time.*

PROOF

1. Immediate.

2. Let us show that $S + T$ is a stopping time and leave the other two to the reader:
$$\{S + T \leq n\} = \cup_{k=0}^{n}(\{S \leq k\} \cap \{T \leq n - k\}) \in \mathcal{F}_n. \qquad \blacksquare$$

3. For $m \in \mathbb{N}_0$, we have $\{T \leq m\} = \cap_{n \in \mathbb{N}}\{T_n \leq m\} \in \mathcal{F}_m$.

4. For $m \in \mathbb{N}_0$, we have $\{T_n \geq m\} = \{T_n < m\}^c = \{T_n \leq m - 1\}^c \in \mathcal{F}_{m-1}$. Therefore,
$$\{T \leq m\} = \{T < m + 1\} = \{T \geq m + 1\}^c = \cup_{n \in \mathbb{N}}\{T_n \geq m + 1\}^c \in \mathcal{F}_m.$$

Stopping times are often used to produce new processes from old ones. The most common construction runs the process X until time T and after that keeps it constant and equal to its value at time T. More precisely:

Definition 10.18 (Stopped processes) *Let $\{X_n\}_{n \in \mathbb{N}_0}$ be a stochastic process, and let T be a stopping time. The process $\{X_n\}_{n \in \mathbb{N}_0}$ **stopped at** T, denoted by $\{X_n^T\}_{n \in \mathbb{N}_0}$ is defined by*

$$X_n^T(\omega) = X_{T(\omega) \wedge n}(\omega) = X_n(\omega)\mathbf{1}_{\{n \leq T(\omega)\}} + X_{T(\omega)}\mathbf{1}_{\{n > T(\omega)\}}.$$

The (sub)martingale property is stable under stopping:

Proposition 10.19 (Stability under stopping) *Let $\{X_n\}_{n \in \mathbb{N}_0}$ be a (sub)martingale, and let T be a stopping time. Then the stopped process $\{X_n^T\}_{n \in \mathbb{N}}$ is also a (sub)martingale.*

PROOF Let $\{X_n\}_{n \in \mathbb{N}_0}$ be a (sub)martingale. We note that the process $K_n = \mathbf{1}_{\{n \leq T\}}$, is predictable, non-negative and bounded, so its martingale transform $(K \cdot X)$ is a (sub)martingale. Moreover,

$$(K \cdot X)_n = X_{T \wedge n} - X_0 = X_n^T - X_0,$$

and so, X^T is a (sub)martingale, as well. $\qquad \blacksquare$

10.5 Convergence of martingales

A judicious use of a predictable processes in a martingale transform yields the following important result:

Theorem 10.20 (Martingale convergence) *Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a martingale such that*

$$\sup_{n\in\mathbb{N}_0} \mathbb{E}[|X_n|] < \infty.$$

Then, there exists a random variable $X \in \mathcal{L}^1(\mathcal{F})$ such that $X_n \to X$, a.s.

PROOF We pick two real numbers $a < b$ and define two sequences of stopping times as follows:

$$T_0 = 0,$$
$$S_1 = \inf\{n \geq 0 : X_n \leq a\}, \quad T_1 = \inf\{n \geq S_1 : X_n \geq b\}$$
$$S_2 = \inf\{n \geq T_1 : X_n \leq a\}, \quad T_2 = \inf\{n \geq S_2 : X_n \geq b\}, \text{ etc.}$$

In words, let S_1 be the first time X falls under a. Then, T_1 is the first time after S_1 when X exceeds b, etc. We leave it to the reader to check that $\{T_n\}_{n\in\mathbb{N}}$ and $\{S_n\}_{n\in\mathbb{N}}$ are stopping times. These two sequences of stopping times allow us to construct a predictable process $\{H_n\}_{n\in\mathbb{N}}$ which takes values in $\{0,1\}$. Simply, we "buy low and sell high":

$$H_n = \sum_{k\in\mathbb{N}} \mathbf{1}_{\{S_k < n \leq T_k\}} = \begin{cases} 1, & S_k < n \leq T_k \text{ for some } k \in \mathbb{N}, \\ 0, & \text{otherwise.} \end{cases}$$

Let $U_n^{a,b}$ be the number of " completed upcrossings by time n", i.e., the process defined by

$$U_n^{a,b} = \inf\{k \in \mathbb{N} : T_k \leq n\}$$

A bit of accounting yields:

(10.1) $$(H \cdot X)_n \geq (b-a)U_n^{a,b} - (X_n - a)^-.$$

Indeed, the total gains from the strategy H can be split into two components. First, every time a passage from below a to above b is completed, we pocket at least $(b-a)$. After that, if X never falls below a again, H remains 0 and our total gains exceeds $(b-a)U_n^{a,b}$, which, in turn, trivially dominates $(b-a)U_n^{a,b} - (X_n - a)^-$. The other possibility is that after the last upcrossing, the process does reach the value below a at a certain point. The last upcrossing already happened, so the process never hits a value above b after that; it may very well happen that we "lose" on this transaction. The loss is overestimated by $(X_n - a)^-$.

Then the inequality (10.1) and fact that the martingale transform (by a bounded process) of a martingale is a martingale yield

$$\mathbb{E}[U_n^{a,b}] \leq \tfrac{1}{b-a}\mathbb{E}[(H \cdot X)_n] + \tfrac{1}{b-a}\mathbb{E}[(X_n - a)^-] \leq \tfrac{\mathbb{E}[|X_n|]+|a|}{b-a}.$$

Let $U_\infty^{a,b}$ be the total number of upcrossings, i.e., $U_\infty^{a,b} = \lim_n U_n^{a,b}$. Using the monotone convergence theorem, we get the so-called **upcrossings inequality**

$$\mathbb{E}[U_\infty^{a,b}] \leq \frac{|a| + \sup_{n \in \mathbb{N}_0} \mathbb{E}[|X_n|]}{b - a}$$

The assumption that $\sup_{n \in \mathbb{N}_0} \mathbb{E}[|X_n|] < \infty$, implies that $\mathbb{E}[U_\infty^{a,b}] < \infty$ and so $\mathbb{P}[U_\infty^{a,b} < \infty] = 1$. In words, the number of upcrossings is almost surely finite (otherwise, we would be able to make money by betting on an unfair game).

It remains to use the fact that $U_\infty^{a,b} < \infty$, a.s., to deduce that $\{X_n\}_{n \in \mathbb{N}_0}$ converges. First of all, by passing to rational numbers and taking countable intersections of probability-one sets, we can assert that

$$\mathbb{P}[U_\infty^{a,b} < \infty, \text{ for all } a < b \text{ rational}] = 1.$$

Then, we assume, contrary to the statement, that $\{X_n\}_{n \in \mathbb{N}_0}$ does not converge, so that

$$\mathbb{P}[\liminf_n X_n < \limsup_n X_n] > 0.$$

This can be strengthened to

$$\mathbb{P}[\liminf_n X_n < a < b < \limsup_n X_n, \text{ for some } a < b \text{ rational}] > 0,$$

which is, however, a contradiction since, on the event $\{\liminf_n X_n < a < b < \limsup_n X_n\}$, the process X completes infinitely many upcrossings.

We conclude that there exists an $[-\infty, \infty]$-valued random variable X_∞ such that $X_n \overset{a.s.}{\to} X_\infty$. In particular, we have $|X_n| \overset{a.s.}{\to} |X_\infty|$, and Fatou's lemma yields

$$\mathbb{E}[|X_\infty|] \leq \liminf_n \mathbb{E}[|X_n|] \leq \sup_n \mathbb{E}[|X_n|] < \infty. \qquad \blacksquare$$

Some, but certainly not all, results about martingales can be transferred to submartingales (supermartingales) using the following proposition:

Proposition 10.21 (Doob-Meyer decomposition) *Let $\{X_n\}_{n \in \mathbb{N}_0}$ be a submartingale. Then, there exists a martingale $\{M_n\}_{n \in \mathbb{N}_0}$ and a predictable process $\{A_n\}_{n \in \mathbb{N}}$ (with $A_0 = 0$ adjoined) such that $A_n \in \mathcal{L}^1$, $A_n \leq A_{n+1}$, a.s., for all $n \in \mathbb{N}_0$ and*

$$X_n = M_n + A_n, \text{ for all } n \in \mathbb{N}_0.$$

PROOF Define

$$A_n = \sum_{k=1}^{n} \mathbb{E}[(X_k - X_{k-1})|\mathcal{F}_{k-1}], \quad n \in \mathbb{N}. \qquad \blacksquare$$

Then $\{A_n\}_{n \in \mathbb{N}}$ is clearly predictable and $A_{n+1} \geq A_n$, a.s., thanks to the submartingale property of X. Finally, set $M_n = X_n - A_n$, so that

$$\begin{aligned}
\mathbb{E}[M_n - M_{n-1}|\mathcal{F}_{n-1}] &= \mathbb{E}[X_n - X_{n-1} - (A_n - A_{n-1})|\mathcal{F}_{n-1}] \\
&= \mathbb{E}[X_n - X_{n-1}|\mathcal{F}_{n-1}] - (A_n - A_{n-1}) = 0.
\end{aligned}$$

Corollary 10.22 (Submartingale convergence) *Let $\{X_n\}_{n \in \mathbb{N}_0}$ be a submartingale such that*

$$\sup_{n \in \mathbb{N}_0} \mathbb{E}[X_n^+] < \infty.$$

Then, there exists a random variable $X \in \mathcal{L}^1(\mathcal{F})$ such that $X_n \to X$, a.s.

PROOF Let $\{M_n\}_{n \in \mathbb{N}_0}$ and $\{A_n\}_{n \in \mathbb{N}}$ be as in Proposition 10.21. Since $M_n = X_n - A_n \leq X_n$, a.s., we have $\mathbb{E}[M_n^+] \leq \mathbb{E}[X_n^+] \leq \sup_n \mathbb{E}[X_n^+] < \infty$. Finally, since $\mathbb{E}[M_n^-] = \mathbb{E}[M_n^+] - \mathbb{E}[M_n] \leq \sup_n \mathbb{E}[X_n^+] - \mathbb{E}[M_0] < \infty$, we have

$$\sup_{n \in \mathbb{N}} \mathbb{E}[|M_n|] < \infty.$$

Therefore, $M_n \overset{a.s.}{\to} M_\infty$, for some $M_\infty \in \mathcal{L}^1$. Since $\{A_n\}_{n \in \mathbb{N}}$ is non-negative and non-decreasing, there exists $A_\infty \in \mathcal{L}^0_+$ such that $A_n \to A_\infty \geq 0$, a.s., and so

$$X_n \overset{a.s.}{\to} X_\infty = M_\infty + A_\infty.$$

It remains to show that $X_\infty \in \mathcal{L}^1$, and for that, it suffices to show that $\mathbb{E}[A_\infty] < \infty$. Since $\mathbb{E}[A_n] = \mathbb{E}[X_n] - \mathbb{E}[M_n] \leq C = \sup_n \mathbb{E}[X_n^+] + \mathbb{E}[|M_0|] < \infty$, monotone convergence theorem yields $\mathbb{E}[A_\infty] \leq C < \infty$. ∎

Remark 10.23 Corollary 10.22 - or the simple observation that $\mathbb{E}[|X_n|] = 2\mathbb{E}[X_n^+] - \mathbb{E}[X_0]$ when $\mathbb{E}[X_n] = \mathbb{E}[X_0]$ - implies that it is enough to assume $\sup_n \mathbb{E}[X_n^+] < \infty$ in the original martingale-convergence theorem (Theorem 10.20).

Corollary 10.24 (Convergence of non-negative supermartingales) *Let $\{X_n\}_{n \in \mathbb{N}_0}$ be a non-negative supermartingale (or a non-positive submartingale). Then there exists a random variable $X \in \mathcal{L}^1(\mathcal{F})$ such that $X_n \to X$, a.s.*

To convince yourself that things can go wrong if the boundedness assumptions are not met, here is a problem:

Problem 10.25 Give an example of a submartingale $\{X_n\}_{n \in \mathbb{N}}$ with the property that $X_n \to -\infty$, a.s. and $\mathbb{E}[X_n] \to \infty$. (*Hint:* Use the Borel-Cantelli lemma.)

10.6 Additional problems

Problem 10.26 (Combining (super)martingales) In 1., 2. and 3. below, $\{X_n\}_{n \in \mathbb{N}_0}$ and $\{Y_n\}_{n \in \mathbb{N}_0}$ are martingales. In 4., they are only supermartingales.

1. Show that the process $\{Z_n\}_{n \in \mathbb{N}_0}$ given by $Z_n = X_n \vee Y_n = \max\{X_n, Y_n\}$ is a submartingale.

2. Give an example of $\{X_n\}_{n \in \mathbb{N}_0}$ and $\{Y_n\}_{n \in \mathbb{N}_0}$ such that $\{Z_n\}_{n \in \mathbb{N}_0}$ (defined above) is *not* a martingale.

3. Does the product $\{X_n Y_n\}_{n \in \mathbb{N}_0}$ have to be a martingale? A submartingale? A supermartingale? (Provide proofs or counterexamples).

4. Let T be an $\{\mathcal{F}_n\}_{n \in \mathbb{N}_0}$-stopping time (and remember that $\{X_n\}_{n \in \mathbb{N}_0}$ and $\{Y_n\}_{n \in \mathbb{N}_0}$ are super-martingales). Show that the process $\{Z_n\}_{n \in \mathbb{N}_0}$, given by

$$
Z_n(\omega) = \begin{cases} X_n(\omega), & n < T(\omega) \\ Y_n(\omega), & n \geq T(\omega), \end{cases}
$$

is a supermartingale, provided that $X_T \geq Y_T$, a.s. (*Note:* This result is sometimes called the *switching principle.* It says that if you switch from one supermartingale to a smaller one at a stopping time, the resulting process is still a supermartingale.)

Problem 10.27 (An urn model) An urn contains $B_0 \in \mathbb{N}$ black and $W_0 \in \mathbb{N}$ white balls at time 0. At each time we draw a ball (each ball in the urn has the same probability of being picked), throw it away, and replace it with $C \in \mathbb{N}$ balls of the same color. Let B_n denote the number of black balls at time n, and let X_n denote the proportion of black balls in the urn.

1. Show that there exists a random variable X such that

$$
X_n \xrightarrow{\mathbb{L}^p} X, \text{ for all } p \in [1, \infty).
$$

2. Find an expression for $\mathbb{P}[B_n = k]$, $k = 1, \dots, n+1$, $n \in \mathbb{N}_0$, when $B_0 = W_0 = 1$, $C = 2$, and use it to determine the distribution of X in that case. (*Hint:* Guess the form of the solution for $n = 1, 2, 3$, and prove that you are correct for all n.)

Problem 10.28 (Stabilization of integer-valued submartingales)

1. Let $\{M_n\}_{n \in \mathbb{N}_0}$ be an integer-valued (i.e., $\mathbb{P}[M_n \in \mathbb{Z}] = 1$, for $n \in \mathbb{N}_0$) submartingale bounded from above. Show that there exists an $\mathbb{N}_0$-valued random variable T with the property that

$$
\forall n \in \mathbb{N}_0, \ M_n = M_T \text{ on } \{T \leq n\}, \text{ a.s.}
$$

2. Can such T always be found in the class of stopping times? (*Note:* This is quite a bit harder than the other two parts.)

3. Let $\{X_n\}_{n \in \mathbb{N}_0}$ be a simple biased random walk, i.e.,

$$
X_0 = 0, \ X_n = \sum_{k=1}^{n} \xi_k, \ n \in \mathbb{N},
$$

where $\{\xi_n\}_{n \in \mathbb{N}}$ are iid with $\mathbb{P}[\xi_1 = 1] = p$ and $\mathbb{P}[\xi_1 = -1] = 1 - p$, for some $p \in (0, 1)$. Under the assumption that $p \geq \frac{1}{2}$, show that X hits any nonnegative level with probability 1, i.e., that for $a \in \mathbb{N}$, we have $\mathbb{P}[\tau_a < \infty] = 1$, where $\tau_a = \inf\{n \in \mathbb{N} : X_n = a\}$.

Problem 10.29 (An application to gambling) Let $\{\varepsilon_n\}_{n \in \mathbb{N}_0}$ be an iid sequence with $\mathbb{P}[\varepsilon_n = 1] = 1 - \mathbb{P}[\varepsilon_n = -1] = p \in (\frac{1}{2}, 1)$. We interpret $\{\varepsilon_n\}_{n \in \mathbb{N}_0}$ as outcomes of a series of gambles. A gambler starts with $Z_0 > 0$ dollars, and in each play wagers a certain portion of her wealth. More precisely, the wealth of the gambler at time $n \in \mathbb{N}$ is given by

$$
Z_n = Z_0 + \sum_{k=1}^{n} C_k \varepsilon_k,
$$

where $\{C_n\}_{n\in\mathbb{N}_0}$ is a predictable process such that $C_k \in [0, Z_{k-1})$, for $k \in \mathbb{N}$. The goal of the gambler is to maximize the "return" on her wealth, i.e., to choose a strategy $\{C_n\}_{n\in\mathbb{N}_0}$ such that the expectation $\frac{1}{T}\mathbb{E}[\log(Z_T/Z_0)]$, where $T \in \mathbb{N}$ is some fixed time horizon, is the maximal possible. (*Note:* It makes sense to call the random variable R such that $Z_T = Z_0 e^{RT}$ the *return*, and, consequently $\mathbb{E}[R] = \frac{1}{T}\mathbb{E}[\log(Z_T/Z_0)]$, the *expected return*. Indeed, if you put Z_0 dollars in a bank and accrue (a compound) interest with rate $R \in (0, \infty)$, you will get $Z_0 e^{RT}$ dollars after T years. In our case, R is not deterministic anymore, but the interpretation still holds.)

1. Define $\alpha = H(\frac{1}{2}) - H(p)$, where $H(p) = -p \log p - (1-p) \log(1-p)$ and show that the process $\{W_n\}_{n\in\mathbb{N}_0}$ given by
$$W_n = \log(Z_n) - \alpha n, \text{ for } n \in \mathbb{N}_0$$
   is a supermartingale. Conclude that
$$\mathbb{E}[\log(Z_T)] \le \log(Z_0) + \alpha T,$$
   for any choice of $\{C_n\}_{n\in\mathbb{N}_0}$.

2. Show that the upper bound above is attained for some strategy $\{C_n\}_{n\in\mathbb{N}_0}$.

(*Note:* The quantity $H(p)$ is called the **entropy** of the distribution of ξ_1. This problem shows how it appears naturally in a gambling-theoretic context: the optimal rate of return equals to "excess" entropy $H(\frac{1}{2}) - H(p)$.)

Problem 10.30 (An application to analysis) Let $\Omega = [0, 1)$, $\mathcal{F} = \mathcal{B}[0, 1)$, and $\mathbb{P} = \lambda$, where λ denotes the Lebesgue measure on $[0, 1)$. For $n \in \mathbb{N}$ and $k \in \{0, 1, \ldots, 2^n - 1\}$, we define
$$I_{k,n} = [k2^{-n}, (k+1)2^{-n}), \ \mathcal{F}_n = \sigma(I_{0,n}, I_{1,n}, \ldots, I_{2^{n-1},n}).$$

In words, $\mathcal{F}_n$ is generated by the n-th dyadic partition of $[0, 1)$. For $x \in [0, 1)$, let $k_n(x)$ be the unique number in $\{0, 1, \ldots, 2^n - 1\}$ such that $x \in I_{k_n(x),n}$. For a function $f : [0, 1) \to \mathbb{R}$ we define the process $\{X_n^f\}_{n\in\mathbb{N}_0}$ by
$$X_n^f(x) = 2^n \left(f\big((k_n(x)+1)2^{-n}\big) - f\big(k_n(x)2^{-n}\big) \right), \ x \in [0, 1).$$

1. Show that $\{X_n^f\}_{n\in\mathbb{N}_0}$ is a martingale.

2. Assume that the function f is Lipschitz, i.e., that there exists $K > 0$ such that $|f(y) - f(x)| \le K|y - x|$, for all $x, y \in [0, 1)$. Show that the limit $X^f = \lim_n X_n^f$ exists a.s.

3. Show that, for f Lipschitz, X^f has the property that
$$f(y) - f(x) = \int_x^y X^f(\xi)\, d\xi, \text{ for all } 0 \le x < y < 1.$$

(*Note:* This problem gives an alternative proof of the fact that Lipschitz functions are absolutely continuous.)

Chapter 11

Uniform Integrability

11.1   Uniform integrability

Uniform integrabilty is a compactness-type concept for families of random variables, not unlike that of tightness.

Definition 11.1 (Uniform integrability) *A non-empty family $\mathcal{X} \subseteq \mathbb{L}^0$ of random variables is said to be **uniformly integrable (UI)** if*

$$\lim_{K \to \infty} \left(\sup_{X \in \mathcal{X}} \mathbb{E}[|X| \, \mathbf{1}_{\{|X| \geq K\}}] \right) = 0.$$

Remark 11.2 It follows from the dominated convergence theorem (prove it!) that

$$\lim_{K \to \infty} \mathbb{E}[|X| \, \mathbf{1}_{\{|X| \geq K\}}] = 0 \text{ if and only if } X \in \mathcal{L}^1,$$

i.e., that for integrable random variables, far tails contribute little to the expectation. Uniformly integrable families are simply those for which the size of this contribution can be controlled uniformly over all elements.

We start with a characterization and a few basic properties of uniform-integrable families:

Proposition 11.3 (A criterion for uniform integrability) *A family $\mathcal{X} \subseteq \mathcal{L}^0$ of random variables is uniformly integrable if and only if*

1. *there exists $C > 0$ such that $\mathbb{E}[|X|] \leq C$, for all $X \in \mathcal{X}$, and*

2. *for each $\varepsilon > 0$ there exists $\delta > 0$ such that for any $A \in \mathcal{F}$, we have*

$$\mathbb{P}[A] \leq \delta \Rightarrow \sup_{X \in \mathcal{X}} \mathbb{E}[|X| \, \mathbf{1}_A] \leq \varepsilon.$$

PROOF UI $\Rightarrow$ (1), (2). Assume $\mathcal{X}$ is UI and choose $K > 0$ such that $\sup_{X \in \mathcal{X}} \mathbb{E}[|X| \, \mathbf{1}_{\{|X| > K\}}] \leq 1$.
Since

$$\mathbb{E}[|X|] = \mathbb{E}[|X| \, \mathbf{1}_{\{|X| \leq K\}}] + \mathbb{E}[|X| \, \mathbf{1}_{\{|X| > K\}}] \leq K + \mathbb{E}[|X| \, \mathbf{1}_{\{|X| > K\}}],$$

for any X, we have

$$\sup_{X \in \mathcal{X}} \mathbb{E}[|X|] \leq K + 1,$$

and (1) follows with $C = K + 1$.

For (2), we take $\varepsilon > 0$ and use the uniform integrability of $\mathcal{X}$ to find a constant $K > 0$ such that $\sup_{X \in \mathcal{X}} \mathbb{E}[|X| \, \mathbf{1}_{\{|X| > K\}}] < \varepsilon/2$. For $\delta = \frac{\varepsilon}{2K}$ and $A \in \mathcal{F}$, the condition $\mathbb{P}[A] \leq \delta$ implies that

$$\mathbb{E}[|X| \, \mathbf{1}_A] = \mathbb{E}[|X| \, \mathbf{1}_A \mathbf{1}_{\{|X| \leq K\}}] + \mathbb{E}[|X| \, \mathbf{1}_A \mathbf{1}_{\{|X| > K\}}] \leq K\mathbb{P}[A] + \mathbb{E}[|X| \, \mathbf{1}_{\{|X| > K\}}] \leq \varepsilon.$$

(1), (2) $\Rightarrow$ UI. Let $C > 0$ be the bound from (1), pick $\varepsilon > 0$ and let $\delta > 0$ be such that (2) holds.
For $K = \frac{C}{\delta}$, Markov's inequality gives

$$\mathbb{P}[|X| \geq K] \leq \tfrac{1}{K}\mathbb{E}[|X|] \leq \delta,$$

for all $X \in \mathcal{X}$. Therefore, by (2), $\mathbb{E}[|X| \, \mathbf{1}_{\{|X| \geq K\}}] \leq \varepsilon$ for all $X \in \mathcal{X}$. ∎

Remark 11.4 Boundedness in $\mathcal{L}^1$ is not enough for uniform integrability. To construct the counterexample, take $(\Omega, \mathcal{F}, \mathbb{P}) = ([0, 1], \mathcal{B}([0, 1]), \lambda)$, and define

$$X_n(\omega) = \begin{cases} n, & \omega \leq \frac{1}{n}, \\ 0, & \text{otherwise.} \end{cases}$$

Then $\mathbb{E}[X_n] = n\frac{1}{n} = 1$, but, for $K > 0$, $\mathbb{E}[|X_n| \, \mathbf{1}_{\{|X_n| \geq K\}}] = 1$, for all $n \geq K$, so $\{X_n\}_{n \in \mathbb{N}}$ is not UI.

Problem 11.5 Let $\mathcal{X}$ and $\mathcal{Y}$ be two uniformly-integrable families (on the same probability space).
Show that the following families are also uniformly integrable:

1. $\{Z \in \mathcal{L}^0 : |Z| \leq |X| \text{ for some } X \in \mathcal{X}\}$.

2. $\{X + Y : X \in \mathcal{X}, Y \in \mathcal{Y}\}$.

Another useful characterization of uniform integrability uses a class of functions which converge to infinity faster than any linear function:

Definition 11.6 (Test functions of uniform integrability) *A Borel function $\varphi : [0, \infty) \to [0, \infty)$ is called a* **test function of uniform integrability** *if*

$$\lim_{x \to \infty} \frac{\varphi(x)}{x} = \infty.$$

Proposition 11.7 (A UI criterion using test functions) *A nonempty family $\mathcal{X} \subseteq \mathcal{L}^0$ is uniformly integrable if and only if there exists a test function of uniform integrability φ such that*

$$(11.1) \qquad\qquad\qquad \sup_{X \in \mathcal{X}} \mathbb{E}[\varphi(|X|)] < \infty.$$

Moreover, if it exists, the function φ can be chosen in the class of non-decreasing convex functions.

PROOF Suppose, first, that (11.1) holds for some test function of uniform integrability and that the value of the supremum is $0 < M < \infty$. For $n > 0$, there exists $C_n \in \mathbb{R}$ such that $\varphi(x) \geq nMx$, for $x \geq C_n$. Therefore,

$$M \geq \mathbb{E}[\varphi(|X|)] \geq \mathbb{E}[\varphi(|X|)\mathbf{1}_{\{|X| \geq C_n\}}] \geq nM\mathbb{E}[|X|\,\mathbf{1}_{\{|X| \geq C_n\}}], \text{ for all } X \in \mathcal{X}.$$

Hence, $\sup_{X \in \mathcal{X}} \mathbb{E}[|X|\,\mathbf{1}_{\{|X| \geq C_n\}}] \leq \frac{1}{n}$, and the uniform integrability of $\mathcal{X}$ follows.

Conversely, suppose that $\mathcal{X}$ is uniformly integrable. By definition, there exists a sequence $\{C_n\}_{n \in \mathbb{N}}$ (which can always be chosen so that $0 < C_n < C_{n+1}$ for $n \in \mathbb{N}$, $C_n \to \infty$) such that

$$\sup_{X \in \mathcal{X}} \mathbb{E}[|X|\,\mathbf{1}_{\{|X| \geq C_n\}}] \leq \frac{1}{n^3}.$$

Let the function $\varphi : [0, \infty) \to [0, \infty)$ be continuous and piecewise affine with $\varphi(x) = 0$ for $x \in [0, C_1]$, and the derivative equal to n on (C_n, C_{n+1}), so that

$$\lim_{x \to \infty} \tfrac{\varphi(x)}{x} = \lim_{x \to \infty} \varphi'(x) = \infty.$$

Then,

$$\mathbb{E}[\varphi(|X|)] = \mathbb{E}\Big[\int_0^{|X|} \varphi'(\xi)\,d\xi\Big] = \int_{C_1}^{\infty} \mathbb{E}[\varphi'(\xi)\mathbf{1}_{\{\xi \leq |X|\}}]\,d\xi = \sum_{n=1}^{\infty} n \int_{C_n}^{C_{n+1}} \mathbb{E}[\mathbf{1}_{\{\xi \leq |X|\}}]\,d\xi$$

$$= \sum_{n=1}^{\infty} n\,(\mathbb{E}[|X| \wedge C_{n+1}] - \mathbb{E}[|X| \wedge C_n])$$

Clearly,

$$\mathbb{E}[|X| \wedge C_{n+1}] - \mathbb{E}[|X| \wedge C_n] = \mathbb{E}[(|X| - C_n)\mathbf{1}_{\{C_n \leq |X| < C_{n+1}\}}] + (C_{n+1} - C_n)\mathbb{E}[\mathbf{1}_{\{|X| \geq C_{n+1}\}}]$$

$$\leq \mathbb{E}[|X|\,\mathbf{1}_{\{|X| \geq C_n\}}] + \mathbb{E}[C_{n+1}\mathbf{1}_{\{|X| \geq C_{n+1}\}}]$$

$$\leq \mathbb{E}[|X|\,\mathbf{1}_{\{|X| \geq C_n\}}] + \mathbb{E}[|X|\,\mathbf{1}_{\{|X| \geq C_{n+1}\}}] \leq \frac{2}{n^3},$$

and so

$$\mathbb{E}[\varphi(|X|)] \leq \sum_{n \in \mathbb{N}} \frac{2}{n^2} < \infty. \qquad\blacksquare$$

Corollary 11.8 ($\mathcal{L}^p$-boundedness, $p > 1$ implies UI) *For $p > 1$, let $\mathcal{X}$ be a nonempty family of random variables bounded in $\mathcal{L}^p$, i.e., such that $\sup_{X \in \mathcal{X}} ||X||_{\mathcal{L}^p} < \infty$. Then $\mathcal{X}$ is uniformly integrable.*

Problem 11.9 Let $\mathcal{X}$ be a nonempty uniformly-integrable family in $\mathcal{L}^0$. Show that $\mathrm{conv}\,\mathcal{X}$ is uniformly-integrable, where $\mathrm{conv}\,\mathcal{X}$ is the smallest convex set in $\mathcal{L}^0$ which contains $\mathcal{X}$, i.e., $\mathrm{conv}\,\mathcal{X}$ is the set of all random variables of the form $X = \alpha_1 X_1 + \cdots + \alpha_n X_n$, for $n \in \mathbb{N}$, $\alpha_k \geq 0$, $k = 1, \ldots, n$, $\sum_{k=1}^n \alpha_k = 1$ and $X_1, \ldots, X_n \in \mathcal{X}$.

Problem 11.10 Let $\mathcal{C}$ be a non-empty family of sub-σ-algebras of $\mathcal{F}$, and let X be a random variable in $\mathcal{L}^1$. The family

$$\mathcal{X} = \{\mathbb{E}[X|\mathcal{F}] \,:\, \mathcal{F} \in \mathcal{C}\},$$

is uniformly integrable. (*Hint:* Argue that it follows directly from Proposition 11.7 that $\mathbb{E}[\varphi(|X|)] < \infty$ for some test function of uniform integrability. Then, show that the same φ can be used to prove that $\mathcal{X}$ is UI.)

11.2 First properties of uniformly-integrable martingales

When it is known that the martingale $\{X_n\}_{n\in\mathbb{N}}$ is uniformly integrable, a lot can be said about its structure. We start with a definitive version of the dominated convergence theorem:

> **Proposition 11.11 (A master dominated-convergence theorem)** *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables in $\mathcal{L}^p$, $p \geq 1$, which converges to $X \in \mathcal{L}^0$ in probability. Then, the following statements are equivalent:*
>
> 1. *the sequence $\{|X|_n^p\}_{n\in\mathbb{N}}$ is uniformly integrable,*
>
> 2. *$X_n \xrightarrow{\mathcal{L}^p} X$, and*
>
> 3. *$||X_n||_{\mathcal{L}^p} \to ||X||_{\mathcal{L}^p} < \infty$.*

PROOF (1) $\Rightarrow$ (2): Since there exists a subsequence $\{X_{n_k}\}_{k\in\mathbb{N}}$ such that $X_{n_k} \xrightarrow{a.s.} X$, Fatou's lemma implies that

$$\mathbb{E}[|X|^p] = \mathbb{E}[\liminf_k |X_{n_k}|^p] \leq \liminf_k \mathbb{E}[|X_{n_k}|^p] \leq \sup_{X\in\mathcal{X}} \mathbb{E}[|X|^p] < \infty,$$

where the last inequality follows from the fact that uniformly-integrable families are bounded in $\mathcal{L}^1$.

Now that we know that $X \in \mathcal{L}^p$, uniform integrability of $\{|X_n|^p\}_{n\in\mathbb{N}}$ implies that the family $\{|X_n - X|^p\}_{n\in\mathbb{N}}$ is UI (use Problem 11.5, (2)). Since $X_n \xrightarrow{\mathbb{P}} X$ if and only if $X_n - X \xrightarrow{\mathbb{P}} 0$, we can assume without loss of generality that $X = 0$ a.s., and, consequently, we need to show that $\mathbb{E}[|X_n|^p] \to 0$. We fix an $\varepsilon > 0$, and start by the following estimate

$$(11.2) \qquad \mathbb{E}[|X_n|^p] = \mathbb{E}[|X_n|^p \mathbf{1}_{\{|X_n|^p \leq \varepsilon/2\}}] + \mathbb{E}[|X_n|^p \mathbf{1}_{\{|X_n|^p > \varepsilon/2\}}] \leq \varepsilon/2 + \mathbb{E}[|X_n|^p \mathbf{1}_{\{|X_n|^p > \varepsilon/2\}}].$$

By uniform integrability there exists $\rho > 0$ such that $\sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|^p \mathbf{1}_A] < \varepsilon/2$, whenever $\mathbb{P}[A] \leq \rho$. Convergence in probability now implies that there exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$, we have $\mathbb{P}[|X_n|^p > \varepsilon/2] \leq \rho$. It follows directly from (11.2) that for $n \geq n_0$, we have $\mathbb{E}[|X_n|^p] \leq \varepsilon$.

(2) $\Rightarrow$ (3): $\left|||X_n||_{\mathcal{L}^p} - ||X||_{\mathcal{L}^p}\right| \leq ||X_n - X||_{\mathcal{L}^p} \to 0$

(3) $\Rightarrow$ (1): For $M \geq 0$, define the function $\psi_M : [0, \infty) \to [0, \infty)$ by

$$\psi_M(x) = \begin{cases} x, & x \in [0, M-1] \\ 0, & x \in [M, \infty) \\ \text{interpolated linearly,} & x \in (M-1, M). \end{cases}$$

For a given $\varepsilon > 0$, dominated convergence theorem guarantees the existence of a constant $M > 0$ (which we fix throughout) such that

$$\mathbb{E}[|X|^p] - \mathbb{E}[\psi_M(|X|^p)] < \frac{\varepsilon}{2}.$$

Convergence in probability, together with continuity of ψ_M, implies that $\psi_M(X_n) \to \psi_M(X)$ in probability, for all M, and it follows from boundedness of ψ_M and the bounded convergence theorem that

(11.3) $$\mathbb{E}[\psi_M(|X_n|^p)] \to \mathbb{E}[\psi_M(|X|^p)].$$

By the assumption (3) and (11.3), there exists $n_0 \in \mathbb{N}$ such that

$$\mathbb{E}[|X_n|^p] - \mathbb{E}[|X|^p] < \varepsilon/4 \text{ and } \mathbb{E}[\psi_M(|X|^p)] - \mathbb{E}[\psi_M(|X_n|^p)] < \varepsilon/4, \text{ for } n \geq n_0.$$

Therefore, for $n \geq n_0$.

$$\mathbb{E}[|X_n|^p \mathbf{1}_{\{|X_n|^p > M\}}] \leq \mathbb{E}[|X_n|^p] - \mathbb{E}[\psi_M(|X_n|^p)] \leq \varepsilon/2 + \mathbb{E}[|X|^p] - \mathbb{E}[\psi_M(|X|^p)] \leq \varepsilon.$$

Finally, to get uniform integrability of the entire sequence, we choose an even larger value of M to get $\mathbb{E}[|X_n|^p \mathbf{1}_{\{|X_n|^p > M\}}] \leq \varepsilon$ for the remaining $n < n_0$. ∎

Problem 11.12 For $Y \in \mathcal{L}^1_+$, show that the family $\{X \in \mathcal{L}^0 : |X| \leq Y, \text{ a.s.}\}$ is uniformly integrable. Deduce the dominated convergence theorem from Proposition 11.11

Since convergence in $\mathcal{L}^p$ implies convergence in probability, we have:

Corollary 11.13 ($\mathcal{L}^p$-convergence, $p \geq 1$, implies UI) *Let* $\{X_n\}_{n \in \mathbb{N}_0}$ *be an $\mathcal{L}^p$-convergent sequence, for some $p \geq 1$. Then family $\{X_n : n \in \mathbb{N}_0\}$ is UI.*

Since UI (sub)martingales are bounded in $\mathcal{L}^1$, they converge by Theorem 10.20. Proposition 11.11 guarantees that, additionally, convergence holds in $\mathcal{L}^1$:

Corollary 11.14 (Convergence of UI (sub)martingales) *Uniformly-integrable (sub)martingales converge a.s., and in $\mathcal{L}^1$.*

For martingales, uniform integrability implies much more:

Proposition 11.15 (Structure of UI martingales) *If $\{X_n\}_{n\in\mathbb{N}_0}$ be a martingale. Then, the following are equivalent:*

1. $\{X_n\}_{n\in\mathbb{N}_0}$ *is a Lévy martingale, i.e., it admits a representation of the form $X_n = \mathbb{E}[X|\mathcal{F}_n]$, a.s., for some $X \in \mathcal{L}^1(\mathcal{F})$,*

2. $\{X_n\}_{n\in\mathbb{N}_0}$ *is uniformly integrable.*

3. $\{X_n\}_{n\in\mathbb{N}_0}$ *converges in $\mathcal{L}^1$,*

In that case, convergence also holds a.s., and the limit is given by $\mathbb{E}[X|\mathcal{F}_\infty]$, where $\mathcal{F}_\infty = \sigma(\cup_{n\in\mathbb{N}_0}\mathcal{F}_n)$.

PROOF $(1) \Rightarrow (2)$. The representation $X_n = \mathbb{E}[X|\mathcal{F}_n]$, a.s., and Problem 11.10 imply that $\{X_n\}_{n\in\mathbb{N}_0}$ is uniformly integrable.

$(2) \Rightarrow (3)$. Corollary 11.14.

$(3) \Rightarrow (2)$. Corollary 11.13.

$(2) \Rightarrow (1)$. Corollary 11.14 implies that there exists a random variable $Y \in \mathcal{L}^1(\mathcal{F})$ such that $X_n \to Y$ a.s., and in $\mathcal{L}^1$. For $m \in \mathbb{N}$ and $A \in \mathcal{F}_m$, we have $|\mathbb{E}[X_n\mathbf{1}_A - Y\mathbf{1}_A]| \leq \mathbb{E}[|X_n - Y|] \to 0$, so $\mathbb{E}[X_n\mathbf{1}_A] \to \mathbb{E}[Y\mathbf{1}_A]$. Since $\mathbb{E}[X_n\mathbf{1}_A] = \mathbb{E}[\mathbb{E}[X|\mathcal{F}_n]\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A]$, for $n \geq m$, we have

$$\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A], \text{ for all } A \in \cup_n\mathcal{F}_n.$$

The family $\cup_n\mathcal{F}_n$ is a π-system which generated the sigma algebra $\mathcal{F}_\infty = \sigma(\cup_n\mathcal{F}_n)$, and the family of all $A \in \mathcal{F}$ such that $\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A]$ is a λ-system. Therefore, by the $\pi - \lambda$ Theorem, we have

$$\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A], \text{ for all } A \in \mathcal{F}_\infty.$$

Therefore, since $Y \in \mathcal{F}_\infty$, we conclude that $Y = \mathbb{E}[X|\mathcal{F}_\infty]$. ■

Example 11.16 There exists a non-negative (and therefore a.s.-convergent) martingale which is not uniformly integrable (and therefore, not $\mathcal{L}^1$-convergent). Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a simple random walk starting from 1, i.e. $X_0 = 1$ and $X_n = 1 + \sum_{k=1}^n \xi_k$, where $\{\xi_n\}_{n\in\mathbb{N}}$ is an iid sequence with $\mathbb{P}[\xi_n = 1] = \mathbb{P}[\xi_n = -1] = \frac{1}{2}$, $n \in \mathbb{N}$. Clearly, $\{X_n\}_{n\in\mathbb{N}_0}$ is a martingale, and so is $\{Y_n\}_{n\in\mathbb{N}_0}$, where $Y_n = X_n^T$ and $T = \inf\{n \in \mathbb{N} : X_n = 0\}$. By convention, $\inf \emptyset = +\infty$. It is well known that a simple symmetric random walk hits any level eventually, with probability 1 (we will prove this rigorously later), so $\mathbb{P}[T < \infty] = 1$, and, since $Y_n = 0$, for $n \geq T$, we have $Y_n \to 0$, a.s., as $n \to \infty$. On the other hand, $\{Y_n\}_{n\in\mathbb{N}_0}$ is a martingale, so $\mathbb{E}[Y_n] = \mathbb{E}[Y_0] = 1$, for $n \in \mathbb{N}$. Therefore, $\mathbb{E}[Y_n] \not\to \mathbb{E}[X]$, which can happen only if $\{Y_n\}_{n\in\mathbb{N}_0}$ is *not* uniformly integrable.

11.3   Backward martingales

If, instead of $\mathbb{N}_0$, we use $-\mathbb{N}_0 = \{\ldots, -2, -1, 0\}$ as the time set, the notion of a filtration is readily extended: it is still a family of sub-σ-algebras of $\mathcal{F}$, parametrized by $-\mathbb{N}_0$, such that $\mathcal{F}_{n-1} \subseteq \mathcal{F}_n$, for $n \in -\mathbb{N}_0$.

Definition 11.17 (Backward submartingales) *A stochastic process* $\{X_n\}_{n \in -\mathbb{N}_0}$, *is said to be a* **backward submartingale** *with respect to the filtration* $\{\mathcal{F}_n\}_{n \in -\mathbb{N}_0}$, *if*

1. $\{X_n\}_{n \in -\mathbb{N}_0}$ *is* $\{\mathcal{F}_n\}_{n \in -\mathbb{N}_0}$-*adapted,*

2. $X_n \in \mathcal{L}^1$, *for all* $n \in \mathbb{N}_0$, *and*

3. $\mathbb{E}[X_n | \mathcal{F}_{n-1}] \geq X_{n-1}$, *for all* $n \in -\mathbb{N}_0$.

If, in addition to (1) and (2), the inequality in (3) is, in fact, an equality, we say that $\{X_n\}_{n \in -\mathbb{N}_0}$ *is a* **backward martingale**.

One of the most important facts about backward submartingales is that they (almost) always converge a.s., and in $\mathcal{L}^1$.

Proposition 11.18 (Backward submartingale convergence) *Suppose that* $\{X_n\}_{n \in -\mathbb{N}_0}$ *is a backward submartingale such that*
$$\lim_{n \to -\infty} \mathbb{E}[X_n] > -\infty.$$

Then $\{X_n\}_{n \in -\mathbb{N}_0}$ *is uniformly integrable and there exists a random variable* $X_{-\infty} \in \mathcal{L}^1(\cap_n \mathcal{F}_n)$ *such that*

(11.4) $$X_n \to X_{-\infty} \text{ a.s. and in } \mathcal{L}^1,$$

and

(11.5) $$X_{-\infty} \leq \mathbb{E}[X_m | \cap_n \mathcal{F}_n], \text{ a.s., for all } m \in -\mathbb{N}_0.$$

PROOF We start by decomposing $\{X_n\}_{n \in -\mathbb{N}_0}$ in the manner of Doob and Meyer. For $n \in -\mathbb{N}_0$, set $\Delta A_n = \mathbb{E}[X_n - X_{n-1} | \mathcal{F}_{n-1}] \geq 0$, a.s., and $A_{-n} = \sum_{k=0}^{n} \Delta A_{-k}$, for $n \in \mathbb{N}_0$. The backward submartingale property of $\{X_n\}_{n \in \mathbb{N}_0}$ implies that $\mathbb{E}[X_n] \geq L = \lim_{n \to -\infty} \mathbb{E}[X_n] > -\infty$, so
$$\mathbb{E}[A_n] = \mathbb{E}[X_0 - X_n] \leq \mathbb{E}[X_0] - L, \text{ for all } n \in \mathbb{N}_0.$$

The monotone convergence theorem implies that $\mathbb{E}[A_{-\infty}] < \infty$, where $A_{-\infty} = \sum_{n=0}^{\infty} A_{-n}$. The process $\{M_n\}_{n \in -\mathbb{N}_0}$ defined by $M_n = X_n - A_n$ is a backward martingale. Indeed,
$$\mathbb{E}[M_n - M_{n-1} | \mathcal{F}_{n-1}] = \mathbb{E}[X_n - X_{n-1} - \Delta A_n | \mathcal{F}_{n-1}] = 0.$$

Since all backward martingales are uniformly integrable (why?) and the sequence $\{A_n\}_{n \in -\mathbb{N}_0}$ is uniformly dominated by $A_{-\infty} \in \mathcal{L}^1$ - and therefore uniformly integrable - we conclude that $\{X_n\}_{n \in -\mathbb{N}_0}$ is also uniformly integrable.

To prove convergence, we start by observing that the uniform integrability of $\{X_n\}_{n \in -\mathbb{N}_0}$ implies that $\sup_{n \in -\mathbb{N}_0} \mathbb{E}[X_n^+] < \infty$. A slight modification of the proof of the martingale convergence

theorem (left to a very diligent reader) implies that $X_n \to X_{-\infty}$, a.s. for some random variable $X_\infty \in \cap_n \mathcal{F}_n$. Uniform integrability also ensures that the convergence holds in $\mathcal{L}^1$ and that $X_{-\infty} \in \mathcal{L}^1$.

In order to show (11.5), it is enough to show that

(11.6) $$\mathbb{E}[X_{-\infty} \mathbf{1}_A] \leq \mathbb{E}[X_m \mathbf{1}_A],$$

for any $A \in \cap_n \mathcal{F}_n$, and any $m \in -\mathbb{N}_0$. We first note that since $X_n \leq \mathbb{E}[X_m | \mathcal{F}_n]$, for $n \leq m \leq 0$, we have

$$\mathbb{E}[X_n \mathbf{1}_A] \leq \mathbb{E}[\mathbb{E}[X_m | \mathcal{F}_n] \mathbf{1}_A] = \mathbb{E}[X_m \mathbf{1}_A],$$

for any $A \in \cap_n \mathcal{F}_n$. It remains to use the fact the $\mathcal{L}^1$-convergence of $\{X_n\}_{n \in -\mathbb{N}_0}$ implies that $\mathbb{E}[X_n \mathbf{1}_A] \to \mathbb{E}[X_{-\infty} \mathbf{1}_A]$, for all $A \in \mathcal{F}$. ∎

Remark 11.19 Even if $\lim \mathbb{E}[X_n] = -\infty$, the convergence $X_n \to X_{-\infty}$ still holds, but not in $\mathcal{L}^1$ and $X_{-\infty}$ may take the value $-\infty$ with positive probability.

Corollary 11.20 (Backward martingale convergence) *If $\{X_n\}_{n \in -\mathbb{N}_0}$ is a backward martingale, then $X_n \to X_{-\infty} = \mathbb{E}[X_0 | \cap_n \mathcal{F}_n]$, a.s., and in $\mathcal{L}^1$.*

11.4 Applications of backward martingales

We can use the results about the convergence of backward martingales to give a non-classical proof of the strong law of large numbers. Before that, we need a useful classical result.

Proposition 11.21 (Kolmogorov's 0-1 law) *Let $\{\xi_n\}_{n \in \mathbb{N}}$ be a sequence of independent random variables, and let the **tail** σ-**algebra** $\mathcal{F}_{-\infty}$ be defined by*

$$\mathcal{F}_{-\infty} = \bigcap_{n \in \mathbb{N}} \mathcal{F}_{-n}, \text{ where } \mathcal{F}_{-n} = \sigma(\xi_n, \xi_{n+1}, \dots).$$

Then $\mathcal{F}_{-\infty}$ is $\mathbb{P}$-trivial, i.e., $\mathbb{P}[A] \in \{0, 1\}$, for all $A \in \mathcal{F}_{-\infty}$.

PROOF Define $\mathcal{F}_n = \sigma(\xi_1, \dots, \xi_n)$, and note that $\mathcal{F}_{n-1}$ and $\mathcal{F}_{-n}$ are independent σ-algebras. Therefore, $\mathcal{F}_{-\infty} \subseteq \mathcal{F}_{-n}$ is also independent of $\mathcal{F}_n$, for each $n \in \mathbb{N}$. This, in turn, implies that $\mathcal{F}_{-\infty}$ is independent of the σ-algebra $\mathcal{F}_\infty = \sigma(\cup_n \mathcal{F}_n)$. On the other hand, $\mathcal{F}_{-\infty} \subseteq \mathcal{F}_\infty$, so $\mathcal{F}_{-\infty}$ is independent of itself. This implies that $\mathbb{P}[A] = \mathbb{P}[A \cap A] = \mathbb{P}[A]\mathbb{P}[A]$, for each $A \in \mathcal{F}_{-\infty}$, i.e., that $\mathbb{P}[A] \in \{0, 1\}$. ∎

Theorem 11.22 (Strong law of large numbers) *Let $\{\xi_n\}_{n \in \mathbb{N}}$ be an iid sequence of random variables in $\mathcal{L}^1$. Then*

$$\tfrac{1}{n}(\xi_1 + \dots + \xi_n) \to \mathbb{E}[\xi_1], \text{ a.s. and in } \mathcal{L}^1.$$

PROOF For notational reasons, backward martingales are indexed by $-\mathbb{N}$ instead of $-\mathbb{N}_0$. For $n \in -\mathbb{N}$, let $S_n = \xi_1 + \cdots + \xi_n$, and let $\mathcal{F}_n$ be the σ-algebra generated by $S_n, S_{n+1}, \ldots$. The process $\{X_n\}_{n \in -\mathbb{N}}$ is given by

$$X_{-n} = \mathbb{E}[\xi_1 | \mathcal{F}_n], \text{ for } n \in \mathbb{N}_0.$$

Since $\sigma(S_n, S_{n+1}, \ldots) = \sigma(\sigma(S_n), \sigma(\xi_{n+1}, \xi_{n+2}, \ldots))$, and $\sigma(\xi_{n+1}, \xi_{n+2}, \ldots)$ is independent of ξ_1, for $n \in \mathbb{N}$, we have

$$X_{-n} = \mathbb{E}[\xi_1 | \mathcal{F}_n] = \mathbb{E}[\xi_1 | \sigma(S_n)] = \tfrac{1}{n} S_n,$$

where the last equality follows from Problem 9.29. Backward martingales converge a.s., and in $\mathcal{L}^1$, so for the random variable $X_{-\infty} = \lim_n \tfrac{1}{n} S_n$ we have

$$\mathbb{E}[X_{-\infty}] = \lim_n \mathbb{E}[\tfrac{1}{n} S_n] = \mathbb{E}[\xi_1].$$

On the other hand, since $\lim_n \tfrac{1}{n} S_k = 0$, for all $k \in \mathbb{N}$, we have $X_{-\infty} = \lim_n \tfrac{1}{n}(\xi_{k+1} + \cdots + \xi_n)$, for any $k \in \mathbb{N}$, and so $X_{-\infty} \in \sigma(\xi_{k+1}k, \xi_{k+2}, \ldots)$. By Proposition 11.21, $X_{-\infty}$ is measurable in a $\mathbb{P}$-trivial σ-algebra, and is, thus, constant a.s. (why?). Since $\mathbb{E}[X_{-\infty}] = \mathbb{E}[\xi_1]$, we must have $X_{-\infty} = \mathbb{E}[\xi_1]$, a.s. ∎

11.5   Exchangeability and de Finetti's theorem (*)

We continue with a useful generalization of the Kolmogorov's 0-1 law, where we extend the ideas about the use of symmetry in the proof of Theorem 11.22 above.

Definition 11.23 (Symmetric functions) *A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be* **symmetric** *if*

$$f(x_1, \ldots, x_n) = f(x_{\pi(1)}, \ldots, x_{\pi(n)}),$$

for each permutation $\pi \in S_n$.

For $k, n \in \mathbb{N}$, $k \leq n$, let S_n^k denote the set of all injections $\beta : \{1, 2, \ldots, k\} \to \{1, 2, \ldots, n\}$. Note that S_n^k has $n(n-1) \ldots (n-k+1) = \frac{n!}{(n-k)!} = k!\binom{n}{k}$ elements and that for $k = n$, $S_n = S_n^n$ is the set of all permutations of the set $\{1, 2, \ldots, n\}$.

Definition 11.24 (Symmetrization) *For a function $f : \mathbb{R}^k \to \mathbb{R}$, and $n \geq k$ the function $f_n^{sim} : \mathbb{R}^n \to \mathbb{R}$, given by*

$$f_n^{sim}(x_1, \ldots, x_n) = \tfrac{1}{n(n-1)\ldots(n-k+1)} \sum_{\beta \in S_n^k} f(x_{\beta(1)}, \ldots, x_{\beta(k)}),$$

*is called the n-**symmetrization** of f.*

Problem 11.25

1. Show that a function $f : \mathbb{R}^n \to \mathbb{R}$ is symmetric if and only if $f = f_n^{sim}$.

2. Show that for each $1 \leq k \leq n$, and each function $f : \mathbb{R}^k \to \mathbb{R}$, the n-symmetrization f_n^{sim} of f is a symmetric function.

Definition 11.26 (Exchangeable σ-algebra) *For $n \in \mathbb{N}$, let $\mathcal{E}_n$ be the σ-algebra generated by X_{n+1}, $X_{n+2}, \ldots$, in addition to all random variables of the form $f(X_1, \ldots, X_n)$, where f is a symmetric Borel function $f : \mathbb{R}^n \to \mathbb{R}$.*

*The **exchangeable** σ-algebra $\mathcal{E}$ is defined by*

$$\mathcal{E} = \bigcap_{n \in \mathbb{N}} \mathcal{E}_n.$$

Remark 11.27 The exchangeable σ-algebra clearly contains the tail σ-algebra, and we can interpret it as the collection of all events whose occurrence is not affected by a permutation of the order of $X_1, X_2, \ldots$.

Example 11.28 Consider the event

$$A = \{\omega \in \Omega \, : \, \limsup_{k \in \mathbb{N}} \sum_{j=1}^{k} X_j(\omega) \geq 0\}.$$

This event is not generally in the tail σ-algebra (why?), but it is always in the exchangeable σ-algebra. Indeed, A can be written as

$$\{\omega \in \Omega \, : \, \limsup_{k} \sum_{j=n+1}^{k} X_j(\omega) \geq -(X_1(\omega) + \cdots + X_n(\omega))\},$$

and as such belongs to $\mathcal{E}_n$, since it can be represented as a combination of a random variable $\limsup_k \sum_{j=n+1}^{k} X_j$ measurable in $\sigma(X_{n+1}, \ldots)$ and a symmetric function $f(x_1, \ldots, x_n) = x_1 + \cdots + x_n$ of $(X_1, \ldots, X_n)$.

For iid sequences, there is no real difference between the exchangeable and the tail σ-algebra: they are both trivial. Before we prove this fact, we need a lemma and a problem.

Lemma 11.29 (Symmetrization as conditional expectation) *Let $\{X_n\}_{n \in \mathbb{N}}$ be an iid sequence and let $f : \mathbb{R}^k \to \mathbb{R}$, $k \in \mathbb{N}$ be a bounded Borel function. Then*

(11.7) $$f_n^{sim}(X_1, \ldots, X_n) = \mathbb{E}[f(X_1, \ldots, X_k)|\mathcal{E}_n], \text{ a.s., for } n \geq k.$$

Moreover,

(11.8) $$f_n^{sim}(X_1, \ldots, X_n) \to \mathbb{E}[f(X_1, \ldots, X_k)|\mathcal{E}], \text{ a.s., and in } \mathcal{L}^1, \text{ as } n \to \infty.$$

PROOF By Problem 11.25, for $n \geq k$, $f_n^{sim}(X_1, \ldots, X_n) \in \mathcal{E}_n$, and so

$$f_n^{sim}(X_1, \ldots, X_n) = \mathbb{E}[f_n^{sim}(X_1, \ldots, X_n)|\mathcal{E}_n]$$
$$= \tfrac{1}{n(n-1)\ldots(n-k+1)} \sum_{\beta \in S_n^k} \mathbb{E}[f(X_{\beta(1)}, \ldots, X_{\beta(k)})|\mathcal{E}_n].$$

By symmetry and definition of $\mathcal{E}_n$, we expect $\mathbb{E}[f(X_{\beta(1)}, \ldots, X_{\beta(k)})|\mathcal{E}_n]$ not to depend on β. To prove this in a rigorous way, we must show that,

$$\mathbb{E}[f(X_{\beta(1)}, \ldots, X_{\beta(k)}) - f(X_1, \ldots, X_k)|\mathcal{E}_n] = 0, \text{ a.s.,}$$

for each $\beta \in S_n^k$. For that, in turn, it will be enough to pick $\beta \in S_n^k$ and show that

$$\mathbb{E}[g(X_1, \ldots, X_n)(f(X_{\beta(1)}, \ldots, X_{\beta(k)}) - f(X_1, \ldots, X_k))] = 0,$$

for any bounded symmetric function $g : \mathbb{R}^n \to \mathbb{R}$. Notice that the iid property implies that for any permutation $\pi \in S_n$, and any bounded Borel function $h : \mathbb{R}^n \to \mathbb{R}$, we have

$$\mathbb{E}[h(X_1, \ldots, X_n)] = \mathbb{E}[h(X_{\pi(1)}, \ldots, X_{\pi(n)})].$$

In particular, for the function $h : \mathbb{R}^n \to \mathbb{R}$ given by

$$h(x_1, \ldots, x_n) = g(x_1, \ldots, x_n) f(x_{\beta(1)}, \ldots, x_{\beta(k)}),$$

and the permutation $\pi \in S_n$ with $\pi(\beta(i)) = i$, for $i = 1, \ldots, k$ (it exists since β is an injection), we have

$$\mathbb{E}[g(X_1, \ldots, X_n) f(X_{\beta(1)}, \ldots, f_{\beta(k)})] = \mathbb{E}[h(X_1, \ldots, X_n)]$$
$$= \mathbb{E}[h(X_{\pi(1)}, \ldots, X_{\pi(n)}]$$
$$= \mathbb{E}[g(X_{\pi(1)}, \ldots, X_{\pi(n)}) f(X_1, \ldots, X_n)]$$
$$= \mathbb{E}[g(X_1, \ldots, X_n) f(X_1, \ldots, X_k)],$$

where the last equality follows from the fact that g is symmetric.

Finally, to prove (11.8), we simply combine (11.7), the definition $\mathcal{E} = \cap_n \mathcal{E}_n$ of the exchangeable σ-algebra and the backward martingale convergence theorem (Proposition 11.18). ∎

Problem 11.30 Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$, and let X be a random variable with $\mathbb{E}[X^2] < \infty$ such that $\mathbb{E}[X|\mathcal{G}]$ is independent of X. Show that $X = \mathbb{E}[X]$, a.s.

(*Hint:* Use the fact that $\mathbb{E}[X|\mathcal{G}]$ is the $\mathcal{L}^2$-projection of X onto $\mathcal{L}^2(\mathcal{G})$, and, consequently, that $\mathbb{E}[(X - \mathbb{E}[X|\mathcal{G}])\mathbb{E}[X|\mathcal{G}]] = 0$.)

Proposition 11.31 (Hewitt-Savage 0-1 Law) *The exchangeable σ-algebra of an iid sequence is trivial, i.e., $\mathbb{P}[A] \in \{0, 1\}$, for $A \in \mathcal{E}$.*

PROOF We pick a Borel function $f : \mathbb{R}^k \to \mathbb{R}$ such that $|f(x)| \leq C$, for $x \in \mathbb{R}^k$. The idea of the proof is to improve the conclusion (11.8) of Lemma 11.29 to $f_n^{sim} \to \mathbb{E}[f(X_1, \ldots, X_k)]$. We start by observing that for $n > k$, out of $\frac{n!}{(n-k)!}$ terms in f_n^{sim}, exactly $\frac{n!}{(n-k)!} - \frac{(n-1)!}{(n-1-k)!} = \frac{k}{n}\frac{n!}{(n-k)!}$

involve X_1, so that the sum of all those terms does not exceed $\frac{k}{n}C$. It follows that all occurrences of X_1 in $f_n^{sim}(X_1, \ldots, X_n)$ disappear in the limit, i.e., more precisely, that the limit of the sequence consisting of only those terms in $f_n^{sim}(X_1, \ldots, X_n)$ which do not contain X_1 coincides with the limit of (all terms in) $f_n^{sim}(X_1, \ldots, X_n)$. Therefore, the limit $\mathbb{E}[f(X_1, \ldots, X_k)|\mathcal{E}]$ can be obtained as a limit of a sequence of combinations of the random variables $X_2, X_3, \ldots$ and is, hence, measurable with respect to $\sigma(X_2, X_3, \ldots)$.

We can repeat the above argument with X_1 replaced by X_2, X_3,..., X_k, to conclude that $\mathbb{E}[f(X_1, \ldots, X_k)|\mathcal{E}]$ is measurable with respect to the σ-algebra $\sigma(X_{k+1}, X_{k+1}, \ldots)$, which is independent of $f(X_1, \ldots, X_k)$. Therefore, by Problem 11.30, we have

$$\mathbb{E}[f(X_1, \ldots, X_k)|\mathcal{E}] = \mathbb{E}[f(X_1, \ldots, X_k)], \text{ a.s.}$$

In particular, for $A \in \mathcal{E}$ we have $\mathbb{E}[\mathbf{1}_A f(X_1, \ldots, X_k)] = \mathbb{P}[A]\mathbb{E}[f(X_1, \ldots, X_k)]$, for any bounded Borel function $f : \mathbb{R}^k \to \mathbb{R}$. It follows that the σ-algebras $\sigma(X_1, \ldots, X_k)$ and $\mathcal{E}$ are independent, for all $k \in \mathbb{N}$. The π-λ theorem implies that $\mathcal{E}$ is then also independent of $\mathcal{F}_\infty = \sigma(X_1, X_2, \ldots) = \sigma(\cup_k \sigma(X_1, \ldots, X_k))$. On the other hand $\mathcal{E} \subseteq \mathcal{F}_\infty$, and so $\mathcal{E}$ is independent of itself. It follows that $\mathbb{P}[A] = \mathbb{P}[A \cap A] = \mathbb{P}[A]^2$, for all $A \in \mathcal{E}$, and, so, $\mathbb{P}[A] \in \{0, 1\}$. ∎

The following generalization of the strong law of large numbers follows directly from (11.8) and Proposition 11.31.

Corollary 11.32 (A strong law for symmetric functions) *Let $\{X_n\}_{n \in \mathbb{N}}$ be an iid sequence, and let $f : \mathbb{R}^k \to \mathbb{R}$, $k \in \mathbb{N}$, be a Borel function with $f(X_1, \ldots, X_k) \in \mathcal{L}^1$. Then*

$$f_n^{sim}(X_1, \ldots, X_n) \to \mathbb{E}[f(X_1, \ldots, X_k)].$$

Remark 11.33 The random variables of the form $f_n^{sim}(X_1, \ldots, X_n)$, for some Borel function $f : \mathbb{R}^k \to \mathbb{R}$ are sometimes called **U-statistics**, and are used as estimators in statistics. Corollary 11.32 can be interpreted as *consistency* statement for U-statistics.

Example 11.34 For $f(x_1, x_2) = (x_1 - x_2)^2$, we have

$$f_n^{sim}(x_1, \ldots, x_n) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} (x_i - x_j)^2,$$

and so, by Corollary 11.32, we have that for an iid sequence $\{X_n\}_{n \in \mathbb{N}}$ with $\sigma^2 = \text{Var}[X_1] < \infty$, we have

$$\frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} (X_i - X_j)^2 \to \mathbb{E}[(X_1 - X_2)^2] = 2\sigma^2, \text{ a.s. and in } \mathcal{L}^1.$$

As a last application of the backward martingale convergence theorem, we prove de Finetti's theorem.

Definition 11.35 (Exchangeable sequences) *A sequence $\{X_n\}_{n\in\mathbb{N}}$ of random variables is said to be* **exchangeable** *If*

$$(X_1, \ldots, X_n) \overset{(d)}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)}),$$

for all $n \in \mathbb{N}$ and all permutations $\pi \in S_n$

Example 11.36 Iid sequences are clearly exchangeable, but the inclusion is strict. Here is a typical example of an exchangeable sequence which is not iid. Let ξ, $\{Y_n\}_{n\in\mathbb{N}}$ be independent uniformly distributed random variable in $(0, 1)$. We define

$$X_n = \mathbf{1}_{\{Y_n \leq \xi\}}.$$

Think of a situation where the parameter $\xi \in (0, 1)$ is chosen randomly, and then an unfair coin with the probability of obtaining heads ξ is minted and tossed repeatedly. The distribution of the coin-toss is the same for every toss, but the values are not independent. Intuitively, if we are told the value of X_1, we have a better idea about the value of ξ, which, in turn, affects our distribution of Y_2. To show that that $\{X_n\}_{n\in\mathbb{N}}$ is an exchangeable sequence which is not iid we compute

$$\mathbb{E}[X_1 X_2] = \mathbb{E}[\mathbb{P}[Y_1, Y_2 \leq \xi]|\sigma(\xi)] = \int_0^1 \mathbb{P}[Y_1, Y_2 \leq x]\,dx = \int_0^1 x^2\,dx = \tfrac{1}{3},$$

as, well as

$$\mathbb{E}[X_1]\mathbb{E}[X_2] = \mathbb{P}[Y_1 \leq \xi]\mathbb{P}[Y_2 \leq \xi] = (\int_0^1 x\,dx)^2 = \tfrac{1}{4}.$$

To show that $\{X_n\}_{n\in\mathbb{N}}$ is exchangeable, we need to compare the distribution of $(X_1, \ldots, X_n)$ and $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ for $\pi \in S_n$, $n \in \mathbb{N}$. For a choice $(b_1, \ldots, b_n) \in \{0, 1\}^n$, we have

$$
\begin{aligned}
\mathbb{P}[(X_{\pi(1)}, &\ldots, X_{\pi(n)}) = (b_1, \ldots, b_n)] \\
&= \mathbb{E}[\mathbb{E}[\mathbf{1}_{\{(X_{\pi(1)},\ldots,X_{\pi(n)})=(b_1,\ldots,b_n)\}}|\sigma(\xi)]] \\
&= \int_0^1 \mathbb{P}[\mathbf{1}_{\{Y_{\pi(1)}\leq x\}} = b_1, \ldots, \mathbf{1}_{\{Y_{\pi(n)}\leq x\}} = b_n]\,dx \\
&= \int_0^1 \prod_{i=1}^n \mathbb{P}[\mathbf{1}_{\{Y_{\pi(i)}\leq x\}} = b_i]\,dx \\
&= \int_0^1 \prod_{i=1}^n \mathbb{P}[\mathbf{1}_{\{Y_1\leq x\}} = b_i]\,dx
\end{aligned}
$$

where the last equality follows from the fact that $\{Y_n\}_{n\in\mathbb{N}}$ are iid. Since the final expression above does not depend on π, the sequence $\{X_n\}_{n\in\mathbb{N}}$ is indeed exchangeable.

Problem 11.37 (A consequence of exchangeability) Let $\{X_n\}_{n\in\mathbb{N}}$ be an exchangeable sequence with $\mathbb{E}[X_1^2] < \infty$. Show that
$$\mathbb{E}[X_1 X_2] \geq 0.$$

(*Hint:* Expand the inequality $\mathbb{E}[(X_1 - \mathbb{E}[X_1] + \cdots + X_n - \mathbb{E}[X_n])^2] \geq 0$ and use exchangeability.)

Definition 11.38 (Conditional iid) *A sequence $\{X_n\}_{n\in\mathbb{N}}$ is said to be* **conditionally iid** *with respect to a sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$, if*

$$\mathbb{P}[X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n | \mathcal{G}] = \mathbb{P}[X_1 \in A_1 | \mathcal{G}] \cdots \mathbb{P}[X_1 \in A_n | \mathcal{G}], \text{ a.s.,}$$

for all $n \in \mathbb{N}$, and all $A_1, \ldots, A_n \in \mathcal{B}(\mathbb{R})$.

Problem 11.39 Let the sequence $\{X_n\}_{n\in\mathbb{N}}$ be as in Example 11.36. Show that

1. $\{X_n\}_{n\in\mathbb{N}}$ is conditionally iid with respect to $\sigma(\xi)$, and

2. $\sigma(\xi)$ is the exchangeable σ-algebra for $\{X_n\}_{n\in\mathbb{N}}$. (*Hint:* Consider the limit $\lim_n \frac{1}{n} \sum_{k=1}^n X_k$.)

The result of Problem 11.39 is not a coincidence. In a sense, all exchangeable sequences have the structure similar to that of $\{X_n\}_{n\in\mathbb{N}}$ above:

Theorem 11.40 (de Finetti) *Let $\{X_n\}_{n\in\mathbb{N}}$ be an exchangeable sequence, and let $\mathcal{E}$ be the corresponding exchangeable σ-algebra. Then, conditionally on $\mathcal{E}$, $\{X_n\}_{n\in\mathbb{N}}$ are iid.*

PROOF Let f be of the form $f = gh$, where $g : \mathbb{R}^{k-1} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ are Borel functions with $|g(x)| \leq C_g$, for all $x \in \mathbb{R}^{k-1}$ and $|h(x)| \leq C_h$, for all $x \in \mathbb{R}$. The product $P_n = n(n-1)\ldots(n-k+2)g_n^{sim}(X_1, \ldots, X_n) \, nh_n^{sim}(X_1, \ldots, X_n)$ can be expanded into

$$P_n = \sum_{\beta \in S_n^{k-1}} g(X_{\beta(1)}, \ldots, X_{\beta(k-1)}) \sum_{i \in \{1, \ldots, n\}} h(X_i)$$

$$= \sum_{\beta \in S_n^k} g(X_{\beta(1)}, \ldots, X_{\beta(k-1)}) h(X_{\beta(k)}) + \sum_{\beta \in S_n^{k-1}} g(X_{\beta(1)}, \ldots, X_{\beta(k-1)}) \sum_{j=1}^{k-1} h(X_{\beta(j)}).$$

A bit of simple algebra, where $f^j(x_1, \ldots, x_{k-1}) = g(x_1, \ldots, x_{k-1})h(x_j)$, for $j = 1, \ldots, k-1$ yields

$$f_n^{sim} = \frac{n}{n-k+1} g_n^{sim} h_n^{sim} - \frac{1}{n-k+1} \sum_{j=1}^{k-1} f_n^{j,sim}.$$

The sum $\sum_{j=1}^{k-1} f_n^{j,sim}$ is bounded by $(k-1)C_g C_h$, and so, upon letting $n \to \infty$, we get

$$\lim_n \left| f_n^{sim} - g_n^{sim} h_n^{sim} \right| = 0.$$

Therefore, the relation (11.8) of Lemma 11.29 applied to f^{sim}, g^{sim} and h^{sim} implies that

$$\mathbb{E}[f(X_1, \ldots, X_k)|\mathcal{E}] = \mathbb{E}[g(X_1, \ldots, X_{k-1})|\mathcal{E}] \, \mathbb{E}[h(X_k)|\mathcal{E}].$$

We can repeat this procedure with $g = g'h'$, for some bounded Borel functions $g' : \mathbb{R}^{k-1} \to \mathbb{R}$ and $h' : \mathbb{R} \to \mathbb{R}$ to split the conditional expectation $\mathbb{E}[g(X_1, \ldots, X_{k-1})|\mathcal{E}]$ into the product $\mathbb{E}[g'(X_1, \ldots, X_{k-2})|\mathcal{E}]\mathbb{E}[h'(X_{k-1})|\mathcal{E}]$. After $k - 1$ such steps, we get

$$\mathbb{E}[h_1(X_1)h_2(X_2) \ldots, h_k(X_k)|\mathcal{E}] = \mathbb{E}[h_1(X_1)|\mathcal{E}] \cdots \mathbb{E}[h_k(X_k)|\mathcal{E}],$$

for any selection of bounded Borel functions $h_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, k$. We pick $h_i = \mathbf{1}_{A_i}$, for $A_i \in \mathcal{B}(\mathbb{R})$, to conclude that

$$\mathbb{P}[X_1 \in A_1, \ldots, X_k \in A_k|\mathcal{E}] = \mathbb{P}[X_1 \in A_1|\mathcal{E}] \cdots \mathbb{P}[X_k \in A_k|\mathcal{E}],$$

for all $A_i \in \mathcal{B}(\mathbb{R})$, $i = 1, \ldots, k$. The full conditional iid property follows from exchangeability. ∎

11.6 Additional Problems

Problem 11.41 (A UI martingale not in H^1) Set $\Omega = \mathbb{N}$, $\mathcal{F} = 2^\mathbb{N}$, and $\mathbb{P}$ is the probability measure on $\mathcal{F}$ characterized by $\mathbb{P}[\{k\}] = 2^{-k}$, for each $k \in \mathbb{N}$. Define the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ by

$$\mathcal{F}_n = \sigma\Big(\{1\}, \{2\}, \ldots, \{n-1\}, \{n, n+1, \ldots\}\Big), \text{ for } n \in \mathbb{N}.$$

Let $Y : \Omega \to [1, \infty)$ be a random variable such that $\mathbb{E}[Y] < \infty$ and $\mathbb{E}[YK] = \infty$, where $K(k) = k$, for $k \in \mathbb{N}$.

1. (3pts) Find an explicit example of a random variable Y with the above properties.

2. (5pts) Find an expression for $X_n = \mathbb{E}[Y|\mathcal{F}_n]$ in terms of the values $Y(k)$, $k \in \mathbb{N}$.

3. (12pts) Using the fact that $X_\infty^*(k) := \sup_{n\in\mathbb{N}} |X_n(k)| \geq X_k(k)$ for $k \in \mathbb{N}$, show that $\{X_n\}_{n\in\mathbb{N}}$ is a uniformly integrable martingale which is not in H^1. (*Note:* A martingale $\{X_n\}_{n\in\mathbb{N}}$ is said to be in H^1 if $X_\infty^* \in \mathbb{L}^1$.)

Problem 11.42 (Scheffé's lemma) Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a sequence of random variables in $\mathcal{L}_+^1$ such that $X_n \to X$, a.s., for some $X \in \mathcal{L}_+^1$. Show that $\mathbb{E}[X_n] \to \mathbb{E}[X]$ if and only if the sequence $\{X_n\}_{n\in\mathbb{N}_0}$ is UI.

Problem 11.43 (Hunt's lemma) Let $\{\mathcal{F}_n\}_{n\in\mathbb{N}_0}$ be a filtration, and let $\{X_n\}_{n\in\mathbb{N}_0}$ be a sequence in $\mathcal{L}^0$ such that $X_n \to X$, for some $X \in \mathcal{L}^0$, both in $\mathcal{L}^1$ and a.s.

1. *(Hunt's lemma).* Assume that $|X_n| \leq Y$, a.s., for all $n \in \mathbb{N}$ and some $Y \in \mathcal{L}_+^1$ Prove that

   (11.9) $$\mathbb{E}[X_n|\mathcal{F}_n] \to \mathbb{E}[X|\sigma(\cup_n \mathcal{F}_n)], \text{ a.s.}$$

   (*Hint:* Define $Z_n = \sup_{m\geq n} |X_m - X|$, and show that $Z_n \to 0$, a.s., and in $\mathcal{L}^1$.)

2. Find an example of a sequence $\{X_n\}_{n\in\mathbb{N}}$ in $\mathcal{L}^1$ such that $X_n \to 0$, a.s., and in $\mathcal{L}^1$, but $\mathbb{E}[X_n|\mathcal{G}]$ does not converge to 0, a.s., for some $\mathcal{G} \subseteq \mathcal{F}$. (*Hint:* Look for X_n of the form $X_n = \xi_n \frac{\mathbf{1}_{A_n}}{\mathbb{P}[A_n]}$ and $\mathcal{G} = \sigma(\xi_n; n \in \mathbb{N})$.)

   (*Note:* The existence of such a sequence proves that (11.9) is not true without an additional assumption, such as the one of uniform domination in (1). It provides an example of a property which does not generalize from the unconditional to the conditional case.)

Problem 11.44 (Krickeberg's decomposition) Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a martingale. Show that the following two statements are equivalent:

1. There exists martingales $\{X_n^+\}_{n\in\mathbb{N}_0}$ and $\{X_n^-\}_{n\in\mathbb{N}_0}$ such that $X_n^+ \geq 0$, $X_n^- \geq 0$, a.s., for all $n \in \mathbb{N}_0$ and $X_n = X_n^+ - X_n^-$, $n \in \mathbb{N}_0$.

2. $\sup_{n\in\mathbb{N}_0} \mathbb{E}[|X_n|] < \infty$.

(*Hint:* Consider $\lim_n \mathbb{E}[X_{m+n}^+|\mathcal{F}_m]$, for $m \in \mathbb{N}_0$.)

Problem 11.45 (Branching processes) Let ν be a probability measure on $\mathcal{B}(\mathbb{R})$ with $\nu(\mathbb{N}_0) = 1$, which we call the **offspring distribution**. A population starting from one individual ($Z_0 = 1$) evolves as follows. The initial member leaves a random number Z_1 of children and dies. After that, each of the Z_1 children of the initial member, produces a random number of children and dies. The total number of all children of the Z_1 members of the generation 1 is denoted by Z_2. Each of the Z_2 members of the generation 2 produces a random number of children, etc. Whenever an individual procreates, the number of children has the distribution ν, and is independent of the sizes of all the previous generations including the present one, as well as of the numbers of children of other members of the present generation.

1. Suppose that a probability space and iid sequence $\{\eta_n\}_{n\in\mathbb{N}}$ of random variables with the distribution μ is given. Show how you would construct a sequence $\{Z_n\}_{n\in\mathbb{N}_0}$ with the above properties. (*Hint:* Z_{n+1} is a sum of iid random variables with the number of summands equal to Z_n.)

2. For a distribution ρ on $\mathbb{N}_0$, we define the the **generating function** $P_\rho : [0,1] \to [0,1]$ of ρ by

$$P_\rho(x) = \sum_{k\in\mathbb{N}_0} \rho(\{k\})x^k.$$

   Show that each P_ρ is continuous, non-decreasing and convex on $[0,1]$ and continuously differentiable on $(0,1)$.

3. Let $P = P_\nu$ be the generating function of the offspring distribution ν, and for $n \in \mathbb{N}_0$, we define $P_n(x)$ as the generating function of the distribution of Z_n, i.e., $P_n(x) = \sum_{k\in\mathbb{N}_0} \mathbb{P}[Z_n = k]x^k$. Show that $P_n(x) = P(P(\ldots P(x)\ldots))$ (there are n Ps). (*Hint:* Note that $P(x) = \mathbb{E}[x^{Z_n}]$ for $x > 0$ and use the result of Problem 9.26)

4. Define the **extinction probability** p_e by $p_e = \mathbb{P}[Z_n = 0, \text{ for some } n \in \mathbb{N}]$. Prove that p_e is a fixed point of the map P, i.e., that $P(p_e) = p_e$. (*Hint:* Show that $p_e = \lim_n P^{(n)}(0)$, where $P^{(n)}$ is the n-fold composition of P with itself.)

5. Let $\mu = \mathbb{E}[Z_1]$, be the expected number of offspring. Show that when $\mu \leq 1$ and $\nu(\{1\}) < 1$, we have $p_e = 1$, i.e., the population dies out with certainty if the expected number of offspring does not exceed 1. (*Hint:* Draw a picture of the functions x and $P(x)$ and use (and prove) the fact that, as a consequence of the assumption $\mu \leq 1$, we have $P'(x) < 1$ for all $x < 1$.)

6. Assuming that $0 < \mu < \infty$, show that the process $\{X_n\}_{n\in\mathbb{N}_0}$, given by $X_n = Z_n/\mu^n$, is a martingale (with respect to the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}_0}$, where $\mathcal{F}_n = \sigma(Z_0, Z_1, \ldots, Z_n)$).

7. Identify all probability measures ν with $\nu(\mathbb{N}_0) = 1$, and $\sum_{k\in\mathbb{N}_0} k\nu(\{k\}) = 1$ such that the branching process $\{Z_n\}_{n\in\mathbb{N}_0}$ with the offspring distribution ν is uniformly integrable.

Index